Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Data & Knowledge Engineering 45 (2003) 131–153

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Bounding the cardinality of aggregate views through domain-derived constraints ☆

Paolo Ciaccia, Matteo Golfarelli, Stefano Rizzi *

DEIS, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

Abstract

Accurately estimating the cardinality of aggregate views is crucial for logical and physical design of data warehouses. This paper proposes an approach based on cardinality constraints, derived a-priori from the application domain, which may bound either the cardinality of a view or the ratio between the cardinalities of two views. We face the problem by first computing satisfactory bounds for the cardinality, then by capitalizing on these bounds to determine a good probabilistic estimate for it. In particular, we propose a bounding strategy which achieves an effective trade-off between the tightness of the bounds produced and the computational complexity.

Keywords: Data warehousing; Logical design; View materialization

1. Introduction and motivation

The multidimensional model is the foundation for data representation and querying in multidimensional databases and data warehouses. It represents facts of interest for the decision process into *cubes* in which each cell contains numerical *measures* which quantify the fact from different points of view, while each axis represents an interesting *dimension* for analysis. For instance, within a four-dimensional cube modeling the phone calls supported by a telecommunication company, the dimensions might be the calling number, the number called, the date, and the time segment in which the call is placed; each cube cell could be associated to a measure of the total duration of the calls made from a given number to another number on a given time segment and date.

☆ This work has been partially supported by the D2I MURST project.
* Corresponding author. Tel.: +39-051-2093542; fax: +39-051-2093540.
E-mail addresses: pciaccia@deis.unibo.it (P. Ciaccia), mgolfarelli@deis.unibo.it (M. Golfarelli), srizzi@deis.unibo.it (S. Rizzi).

0169-023X/03/$ - see front matter
doi:10.1016/S0169-023X(02)00177-5

The basic mechanism to extract significant information from the huge quantity of data stored in cubes is aggregation according to hierarchies of attributes rooted in dimensions [11]. In most application cases, cubes are significantly sparse (for instance, most couples of telephone numbers are never connected by a call in a given date), and so are the aggregate views.

Accurately estimating the actual cardinality of each view is crucial for logical and physical design as well as for query processing and optimization [19]. As a relevant case, consider the view materialization problem, where the aggregate views which are the most useful in answering the workload queries have to be selected for materialization (see [18] for a survey). Since the number of possible views which can be derived by aggregating a cube is exponential in the number of attributes, most approaches assume that a constraint on the total disk space occupied by materialization is posed, and attempt to find the subset of views which contemporarily satisfies this constraint and minimizes the workload cost [7,9,12]. Another case where estimation of view cardinalities is relevant is index selection [10].

If the data warehouse has already been loaded, view cardinalities can be quite accurately estimated by using statistical techniques based, say, on histograms [15] or sampling [13]. However, such techniques cannot be applied at all if the data warehouse is still under development, and the estimation of view cardinalities is needed for design purposes. To obviate this, current approaches are based on estimation models that only exploit the cardinality of the base cube and that of the single attribute domains [16,17], which however leads to significant overestimation.

In this paper we propose a novel approach to estimate the cardinality of views based on a-priori information derived from the application domain. Similarly to what is done when estimating the cardinality of projections in relational databases [6], we face the problem by first computing satisfactory bounds for the cardinality, then by capitalizing on these bounds to determine a good probabilistic estimate for it. Besides the functional dependencies (FD's) expressed by the multidimensional scheme, the bounds we determine also take into account additional domain-derived information expressed in the form of *cardinality constraints*, namely, bounds of the cardinality of some views and bounds (called *k-dependencies*) on the ratio between the cardinalities of two views. The main contribution of the paper is a *bounding strategy* comprising (1) a bounding function which computes effective upper bounds of cardinalities and (2) a set of formal results aimed at reducing the complexity of computation.

The paper is organized as follows. In Section 2, we provide some basic definitions, describe the basic principles of our approach and provide a motivating example. Section 3 introduces the basic properties of cardinality constraints, and Section 4 introduces the cover-based bounding strategy. Section 5 presents all major formal results on the computation of bounds. Section 6 proposes a simple probabilistic model to show how the bounds derived may be used to improve the cardinality estimates. Section 7 discusses some interesting open issues. Finally, the most complex proofs are included in the Appendix.

2. Outline of the approach

Before introducing the framework for our approach, we need to provide some basic definitions on views and on their associated lattice.

Definition 1 (*Dimensional scheme*). We call dimensional scheme $\mathscr{D}$ a couple $(\mathscr{U}, \mathscr{F})$ where $\mathscr{U}$ is a set of attributes and $\mathscr{F} = \{A_i \rightarrow A_j : A_i, A_j \in \mathscr{U}\}$ is a set of FD's which relate the attributes of $\mathscr{U}$ into a set of pairwise disjoint directed trees. We call dimensions the attributes $A_k \in \mathscr{U}$ in which the trees are rooted, i.e., such that $\forall A_i \in \mathscr{U}\ (A_i \rightarrow A_k) \notin \mathscr{F}$; let $\dim(\mathscr{D}) \subseteq \mathscr{U}$ denote the set of dimensions of $\mathscr{D}$.

Definition 2 (*View*). Let $\mathscr{D} = (\mathscr{U}, \mathscr{F})$ be a dimensional scheme. We call view on $\mathscr{D}$ any subset of attributes $V \subseteq \mathscr{U}$ such that $\forall A_i, A_j \in V\ (A_i \rightarrow A_j) \notin \mathscr{F}^+$ where $\mathscr{F}^+$ denotes the set of all FD's logically implied by $\mathscr{F}$. [1]

Example 1. Consider an enterprise with branches in different cities. A simple dimensional scheme Transfers modeling the transfers of employees between offices might include:

$$\mathscr{U} = \{\mathsf{date}, \mathsf{month}, \mathsf{year}, \mathsf{fromOffice}, \mathsf{fromDept}, \mathsf{fromCity}, \mathsf{toOffice}, \mathsf{toDept}, \mathsf{toCity}, \mathsf{employee}\}$$
$$\mathscr{F} = \{\mathsf{date} \rightarrow \mathsf{month}, \mathsf{month} \rightarrow \mathsf{year}, \mathsf{fromOffice} \rightarrow \mathsf{fromDept},$$
$$\mathsf{fromOffice} \rightarrow \mathsf{fromCity}, \mathsf{toOffice} \rightarrow \mathsf{toDept}, \mathsf{toOffice} \rightarrow \mathsf{toCity}\}$$

thus the base cube is characterized by

$$\dim(\mathscr{D}) = \{\mathsf{date}, \mathsf{fromOffice}, \mathsf{toOffice}, \mathsf{employee}\}$$

Examples of views on the Transfers scheme are:

$$V = \{\mathsf{month}, \mathsf{fromOffice}, \mathsf{toCity}, \mathsf{employee}\}$$
$$W = \{\mathsf{month}, \mathsf{fromCity}, \mathsf{fromDept}, \mathsf{employee}\}$$
$$Z = \{\mathsf{year}, \mathsf{fromOffice}, \mathsf{toCity}\}$$

In this work we face the problem of accurately estimating the cardinality of a view when the source data cannot be directly queried, which is the case during off-line logical design of multidimensional databases. We assume that a set $\mathscr{I}$ of *cardinality constraints* is available instead, and we look for effective ways to exploit them for estimation. Without loss of generality, suppose that estimates are needed for the purpose of a view materialization algorithm. As sketched in Fig. 1, whenever the materialization algorithm requires information about a candidate view V, our approach works in two steps. First, the *bounder* uses the set $\mathscr{I}$ of cardinality constraints supplied by the user to determine an effective upper bound for the cardinality of V; then, the *estimator* uses this bound to derive a probabilistic estimate for the cardinality of V. Note that this two-steps approach generalizes well-known parametric models for the estimation of the cardinality of relational queries [14] and in particular those for projection size estimation [6], for which bounds are typically given as input parameters.

[1] We are using the term *view* to denote the set of grouping attributes used for aggregation, while the actual views will also include one or more measures. This slight abuse in terminology is allowable in the context of this work since the cardinality of a view only depends on its grouping attributes.

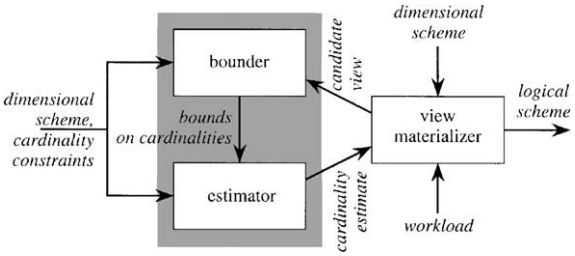

Fig. 1. Overall architecture for logical design.

We consider two different forms of cardinality constraints:

(1) the upper bound w^+ of the cardinality w of a view W;
(2) *k-dependencies*, expressing an upper bound of the ratio between the cardinalities of two views W and Z (see Section 3.2).

We will assume that at least the upper bounds of the cardinalities of all the single attributes in the dimensional scheme are known. This assumption, which is perfectly reasonable in all application domains, is necessary in order to guarantee that at least one upper bound can be determined for each view.

The set $\mathscr{I}$, together with the dimensional scheme $\mathscr{D}$, univocally determines the *least upper bound* v^+ of the cardinality of V, meaning that:

(1) in each instance of $\mathscr{D}$ that does not violate any constraint in $\mathscr{I}$, the cardinality v of V is such that $v \leqslant v^+$; and
(2) there exists an instance compatible with $\mathscr{I}$ where $v = v^+$.

We say a constraint $c \in \mathscr{I}$ is *redundant* iff all the bounds determined by $\mathscr{I}$ are equal to those determined by $\mathscr{I} - \{c\}$.

Definition 3 (*Sound and minimal input*). Let $\mathscr{I}$ be a set of cardinality constraints on dimensional scheme $\mathscr{D}$. We say $\mathscr{I}$ is sound iff there exists at least one non-empty instance of $\mathscr{D}$ which satisfies all the constraints in $\mathscr{I}$. We say $\mathscr{I}$ is minimal iff no constraint in $\mathscr{I}$ is redundant.

In this paper we will assume that the input $\mathscr{I}$ is sound and minimal. It is straightforward to derive that, in this case, all the bounds in $\mathscr{I}$ are least upper bounds (whereas the opposite is not necessarily true).

Computing the bounds implied by $\mathscr{I}$ turns out to be a challenging combinatorial problem, even for simple forms of cardinality constraints. For instance, it is known that the problem is NP-hard for arbitrary patterns of FD's [5]. Furthermore, the actual computational effort needed to compute these bounds might limit applicability in real-world cases. For this reason, the bounder is built around the concept of *bounding strategy*. A bounding strategy s is characterized by a bounding function that, given $\mathscr{I}$, $\mathscr{D}$, and V, computes an upper bound v^+_{s} such that $v^+ \leqslant v^+_{\mathsf{s}}$ holds.

In other terms, a bounding strategy never computes bounds which are more restrictive than the ones logically implied by the input constraints, trading-off accuracy for speed of evaluation.

Turning to the estimator, our framework supports different probabilistic models. A probabilistic model is a function that, given $\mathscr{I}$, $\mathscr{D}$, V as well as bounds computed by the bounder, provides an estimate, $\bar{v}$, for the cardinality of V. In general, this step can use further information from the application domain that is not suitable to derive bounds. Typically this is the case with information concerning average values (e.g., the number of transfers of each employee on each year is 1.5 on the average).

Example 2. Let 10^4 be the number of employees who have been transferred at least once, and let the enterprise consist of 10^3 offices distributed over 10 cities and belonging to one of 10 departments; let 10^3 days be the observation period. Let $V = \{\mathsf{date}, \mathsf{fromOffice}, \mathsf{toOffice}\}$. Since each office is involved in transfers at most with every other office on each date, the first trivial upper bound of v is $10^3 \times 10^3 \times 10^3 = 10^9$. If a constraint states that the maximum number of transfers for an employee during one year is 2, and since we consider three years, it is derived that the cardinality of the base cube is at most six times the number of transferred employees, i.e., 6×10^4. Thus, the upper bound of v can be improved to 6×10^4 as well (the cardinality of a view cannot exceed that of its base cube). Now, by using the probabilistic model in Section 6, the cardinality of V is estimated as $\bar{v} = 3.8 \times 10^4$.

The following compact notation is used for some examples throughout the rest of the paper. Uppercase letters from the beginning of the alphabet $(A, B, \ldots)$ denote dimensions. Attributes which are functionally determined by another attribute, i.e. attributes other than dimensions, are denoted by the corresponding primed letters (e.g., $A \rightarrow A'$, $A' \rightarrow A''$). The sets of attributes which define views are represented by omitting braces, thus writing ABC for $\{A, B, C\}$. Lowercase letters are used for the cardinalities of views and attributes (e.g., a is the cardinality of attribute A, ab is the cardinality of the view with attributes AB, and so on).

3. Basics on cardinality constraints

The possibility of exploiting cardinality constraints to bound the size of a view relies on the partial order induced on views by the FD's in the multidimensional scheme:

Definition 4 (*Roll-up*). Given the set $\mathscr{V}_{\mathscr{D}}$ of all possible views on $\mathscr{D}$, we define on $\mathscr{V}_{\mathscr{D}}$ the roll-up partial order $\preceq$ as follows: $V \preceq W$ iff $W \rightarrow V$, i.e., iff $\forall A_i \in V \exists A_j \in W : (A_j \rightarrow A_i) \in \mathscr{F}^+$.

It is straightforward to verify that, given two views $V \in \mathscr{V}_{\mathscr{D}}$ and $W \in \mathscr{V}_{\mathscr{D}}$, both their *sup* and *inf* views always exist; we will denote them with $V \oplus W$ and $V \otimes W$, respectively. [2] Thus, the roll-up partial order determines a lattice, which we will call *multidimensional lattice* for $\mathscr{D}$, whose top and bottom elements are $\dim(\mathscr{D})$ and the empty view $\emptyset$, respectively. The multidimensional lattice is isomorphic to the lattice of the order ideals of a partially ordered set [1]; thus, it is distributive.

[2] From a relational point of view, $V \oplus W$ is obtained by dropping from the natural join $V \bowtie W$ the functionally dependent attributes.

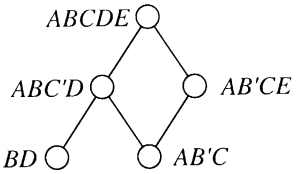

Fig. 2. Roll-up relationships of views in Example 3.

Given two views V and W, we will denote with $V \ominus W$ the least view Z such that $Z \oplus (V \otimes W) = V$. A nice property of this operator, which we will use in the proof of Lemma 7, is the following:

$$(V \ominus W) \oplus W = V \oplus W \tag{1}$$

Proof. By definition of $\ominus$, if we let $Z = V \ominus W$, it is $Z \oplus (V \times W) = V$ which implies $Z \oplus (V \otimes W) \oplus W = V \oplus W$. On the other hand, for the absorption property of distributive lattices it is $Z \oplus (V \otimes W) \oplus W = Z \oplus W$, from which $Z \oplus W = V \oplus W$. □

Example 3. It is $ABC'D \oplus AB'CE = ABCDE$, $ABC'D \otimes AB'CE = AB'C'$, $ABC'D \ominus AB'CE = BD$ (see Fig. 2).

3.1. Upper bounds

The basic observation to determine bounds for view cardinalities using bounds of the cardinalities of other views is that the multidimensional lattice induces an isomorphic structure over such cardinalities; this fact is expressed by the following two lemmas, where we use the notation $[\![E]\!]$ to denote the cardinality of a view that is the result of an expression E.

Lemma 1. *If $W \preceq Z$, then $w^+ \leqslant z^+$.*

Proof. From Definition 4 it follows that $W \preceq Z$ implies $w \leqslant z$ in each instance of $\mathscr{D}$, since $Z \to W$ holds. Now, assume that $w^+ > z^+$. Then, there is an instance of $\mathscr{D}$ in which $z \leqslant z^+ < w \geqslant w^+$, thus $z < w$, which is a contradiction. □

Lemma 2. *Let S be a set of views and let $S' \subseteq S$ consist of all the views $W \in S$ such that for no $Z \in S$ it is $W \preceq Z$. Then, it is $[\![\oplus (S)]\!]^+ = [\![\oplus (S')]\!]^+ \leqslant \prod_{W \in S'} w^+$.*

Proof. When $S' = S$ the result immediately follows, since the least upper bound of a set of views corresponds to their natural join, whose size can never exceed that of the Cartesian product of the views. When $S' \subset S$, the result follows from the observation that $\oplus(S) = \oplus(S')$, since each view in $S - S'$ is functionally determined by some view in S'. □

Example 4. Let $S = \{AB, B'C, A'\}$; it is $[\![\oplus (S)]\!]^+ = [\![\oplus (\{AB, B'C\})]\!]^+ = abc^+ \leqslant ab^+ \cdot b'c^+$.

3.2. The k-dependencies

A k-dependency (kD) is a relevant case of cardinality constraint which naturally generalizes a FD. In the authors' experience, kD's are particularly useful to characterize the knowledge of the business domain held by the experts in the field. For instance, in the transfer domain, we might have some information concerning the number of destination cities for an employee, or on the number of distinct departments moved to from each department. If such information is in the form of bounds, it can be effectively used to improve the bounds of view cardinality.

Definition 5 (*k-dependency*). Let X and Y be two views on $\mathscr{D}$. We say that a kD holds between X and Y, and denote it with $X \xrightarrow{k} Y$, when $k (k \geqslant 1)$ is an upper bound of the number of distinct tuples of Y which correspond to each distinct tuple of X within view $X \oplus Y$.

Example 5. In the Transfers scheme, assume the domain expert provides the following information: The maximum number of inter-department transfers of an employee during one year is 2. This constraint can be formalized by the following $k\mathrm{D} : X \xrightarrow{2} Y$, where $X = \{\mathsf{year}, \mathsf{employee}\}$, $Y = \{\mathsf{toDept}\}$. Intuitively, from this we can derive that the cardinality of the view $\{\mathsf{year}, \mathsf{employee}, \mathsf{toDept}\}$ cannot exceed twice the cardinality of X.

The kD's have been studied in the context of relational database theory, where they are also known as *numerical dependencies*. Grant and Minker [8] have proven that kD's are not finitely axiomatizable, thus no fixed set of inference rules can be used to determine whether or not a given kD is logically implied by a set of kD's. Nonetheless, a basic set of rules, which naturally extend those for FD's, was proposed in [8]. The rules we use, here generalized to work with the multi-dimensional lattice, are:

R1: $X \xrightarrow{k} Y \vdash X \oplus Z \xrightarrow{k} Y \oplus Z$

R2: $X \xrightarrow{k} Y \wedge Y \xrightarrow{l} Z \vdash X \xrightarrow{k \cdot l} Y \oplus Z$

R3: $X \xrightarrow{k} Y \oplus Z \vdash X \xrightarrow{k} Y$

R4: $X \xrightarrow{k} Y \wedge X \xrightarrow{l} Z \vdash X \xrightarrow{k \cdot l} Y \oplus Z$

R5: $X \oplus W \xrightarrow{k} Y \vdash X \xrightarrow{k \cdot [\![W \oplus Z]\!]^+} Y \oplus Z$

Note that the "union" rule R4 can be easily derived from R1 ("extension"), R2 ("transitivity"), and R3 ("decomposition"). As to R5, it may be proved by considering that, for each W and Z, we can write a dummy $k\mathrm{D}$ $X \xrightarrow{[\![W \oplus Z]\!]^+} W \oplus Z$, from which by applying R1 we obtain $X \xrightarrow{[\![W \oplus Z]\!]^+} W \oplus Z \oplus X$. On the other hand, by applying R1 to $X \oplus W \xrightarrow{k} Y$ we obtain $X \oplus W \oplus Z \xrightarrow{k} Y \oplus Z$. Applying R2 to $X \xrightarrow{[\![W \oplus Z]\!]^+} W \oplus Z \oplus X$ and $X \oplus W \oplus Z \xrightarrow{k} Y \oplus Z$ we have $X \xrightarrow{k \cdot [\![W \oplus Z]\!]^+} Y \oplus Z$, which proves R5.

The influence of kD's on the determination of bounds is summarized by the following lemma.

Lemma 3. *If* $X \xrightarrow{k} Y$, *then* $[\![X \oplus Y]\!]^+ \leqslant k \cdot x^+$.

Proof. From Definition 5 it follows immediately that, if $X \xrightarrow{k} Y$, the cardinality of $X \oplus Y$ is related to the cardinality of X by inequality $[\![X \oplus Y]\!] \leqslant k \cdot x$. The inequalities on bounds follow immediately. □

The following inequalities summarize Lemmas 2 and 3, thus relating upper bounds of view cardinalities to kD's:

$$X \xrightarrow{k} Y \Rightarrow [\![X \oplus Y]\!]^+ \leqslant k \cdot x^+ \leqslant x^+ \cdot y^+ \tag{2}$$

where $k \leqslant y^+$ follows from the assumption of sound and minimal input.

4. The cover-based bounding strategy

In this section we introduce a bounding strategy, called *cover-based*, which relies on the concept of cover of a view to compute upper bounds.

Definition 6 (*Candidate set*). We call candidate set a couple $\mathscr{C} = (S, K)$, where S is a set of views and K is a set of kD's. We denote with $\mathrm{lhs}(K)$ and $\mathrm{rhs}(K)$, respectively, the sets of views which are left- and right-hand sides for the kD's in K; besides, let $N_{\mathscr{C}} = S \cup \mathrm{lhs}(K) \cup \mathrm{rhs}(K)$.

Definition 7 (*Cover*). Let $V \in \mathscr{V}_D$ be a view on $\mathscr{D}$ and $\mathscr{C} = (S, K)$ be a candidate set. $\mathscr{C}$ is called a V-cover iff $V \preceq \oplus(N_{\mathscr{C}})$.

Thanks to Lemmas 1 and 2, V-covers can be used to bound from above the cardinality of V, since $v^+ \leqslant [\![\oplus (N_{\mathscr{C}})]\!]^+$. More precisely, the cover-based bounding strategy cb computes v^+_{cb} as:

$$v^+_{\mathsf{cb}} = \begin{cases} v^+ & \text{if } v^+ \in \mathscr{I} \\ \min\{u_{\mathsf{cb}}(\mathscr{C}) : \mathscr{C} \text{ is a } V-\text{cover}\} & \text{if } v^+ \notin \mathscr{I} \end{cases} \tag{3}$$

where $u_{\mathsf{cb}}(\mathscr{C})$ is the bound yielded by cover $\mathscr{C}$. In general, since $u_{\mathsf{cb}}(\mathscr{C})$ depends in turn on the bounds of the views in $\mathscr{C}$, evaluating the cover-based bound may lead to a recursive computational flow; note however that the "case-0" of recursion, $v^+_{\mathsf{cb}} = v^+$, is correctly defined since we assumed the input $\mathscr{I}$ to be minimal.

To start simple, assume that the V-cover $\mathscr{C}$ does not include any kD, thus $N_{\mathscr{C}} = S$. In this case the bound provided by $\mathscr{C}$ is directly derived from Lemma 2:

$$u_{\mathsf{cb}}(\mathscr{C}) = u_{\mathsf{cb}}((S, \emptyset)) = \prod_{W_j \in S'} w^+_{j,\mathsf{cb}} \tag{4}$$

where S' is defined as in Lemma 2.

Let us now turn to the more general (and complex) case when also kD's are present: due to Eq. (2), it is clear that their presence can be exploited to strenghten bounds. Intuitively, if $V \preceq X \oplus Y$ and only x^+ and y^+ are known, the best we can do is to infer that $v \leqslant x^+ \cdot y^+$. On the other hand, if $X \xrightarrow{k} Y$ also holds, then the bound can be improved to $v \leqslant k \cdot x^+$, which can be much better than $x^+ \cdot y^+$. In the following we precisely characterize how sets of views and kD's can be combined together in a common framework.

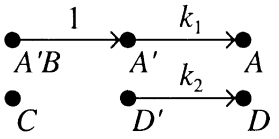

Fig. 3. The $\mathscr{C}$-graph associated to the candidate set in Example 6 (arrows denote kD's).

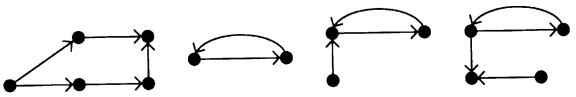

Fig. 4. Four examples of $\mathscr{C}$-graphs (from left to right): acyclic and reachable, cyclic and unreachable without roots, cyclic and reachable, cyclic and unreachable with root.

Definition 8 (*$\mathscr{C}$-graph*). The $\mathscr{C}$-graph $\mathscr{G}_{\mathscr{C}}$ associated with $\mathscr{C} = (S, K)$ is a labeled directed graph [3] whose nodes correspond to the views in $N_{\mathscr{C}}$ and whose arcs are defined as follows:

(1) For each kD $W \xrightarrow{k} Z \in K$ there is an arc, labeled k, from node W to node Z;
(2) For each pair of nodes W and Z such that $Z \preceq W$ there is an arc, labeled 1, from node W to node Z.

The nodes of $\mathscr{G}_{\mathscr{C}}$ with no incoming arcs are called the roots of $\mathscr{G}_{\mathscr{C}}$, and their set is denoted as root($\mathscr{G}_{\mathscr{C}}$). The $\mathscr{C}$-graph $\mathscr{G}_{\mathscr{C}}$ is called *reachable* iff for each node W there exists at least one directed path from a root of $\mathscr{G}_{\mathscr{C}}$ to W.

Example 6. The reachable $\mathscr{C}$-graph associated to $\mathscr{C} = (\{A'B, C\}, \{A' \xrightarrow{k_1} A, D' \xrightarrow{k_2} D\})$ is depicted in Fig. 3. It is root$(\mathscr{G}_{\mathscr{C}}) = \{A'B, C, D'\}$.

Note that an acyclic $\mathscr{C}$-graph is always reachable, whereas the converse is not necessarily true; on the other hand, $\mathscr{G}_{\mathscr{C}}$ is reachable only if each connected component of $\mathscr{G}_{\mathscr{C}}$ has at least one root (see Fig. 4).

While each candidate set $\mathscr{C}$ is univocally associated to one $\mathscr{C}$-graph, different candidate sets may be associated to the same $\mathscr{C}$-graph. We say that $\mathscr{C}$ is *reduced* iff there exists no $\mathscr{C}' \subset \mathscr{C}$ such that $\mathscr{G}_{\mathscr{C}'} = \mathscr{G}_{\mathscr{C}}$ (for instance, $\mathscr{C}' = (\{C\}, \{A \xrightarrow{k} B\})$ is the reduced form of $\mathscr{C} = (\{A, B, C\}, \{A \xrightarrow{k} B\})$).

Lemma 4. *A candidate set $\mathscr{C} = (S, K)$ is reduced iff $S \cap (\mathrm{lhs}(K) \cup \mathrm{rhs}(K)) = \emptyset$.*

Proof (*Only if*). Trivial, since if $W \in S \cap (\mathrm{lhs}(K) \cup \mathrm{rhs}(K))$ then $\mathscr{C} = (S, K)$ and $\mathscr{C}' = (S - \{W\}, K)$ have the same $\mathscr{C}$-graph.

(If.) Assume that $\mathscr{C}$ is not reduced. Then we can reduce either S or K and obtain the same $\mathscr{C}$-graph. Clearly, we cannot reduce K by dropping a kD, since the corresponding graph will have

[3] Technically, $\mathscr{G}_{\mathscr{C}}$ is a *multi-graph*, since two arcs may share the same couple of nodes. This, however, does not influence the following arguments.

one arc less than $\mathscr{G}_{\mathscr{C}}$. Then, we can only remove a view W in S. This will leave unaltered the set of nodes only if W appears as the left or right side of some kD in K, as claimed. □

The following theorem, whose proof is reported in the Appendix, precisely characterizes how the bound obtained from a candidate set is related to its $\mathscr{C}$-graph. For brevity, from now on we will write $\prod_E k_i$ to denote the product of all the labels of the arcs in set E.

Theorem 1. *Let V be a view, $\mathscr{C} = (S, K)$ be a V-cover, and $\mathscr{G}_{\mathscr{C}}$ be the $\mathscr{C}$-graph associated with $\mathscr{C}$. If $\mathscr{G}_{\mathscr{C}}$ is reachable it is*:

$$u_{\mathsf{cb}}(\mathscr{C}) = u_{\mathsf{cb}}((S,K)) = \prod_{E_{\mathscr{C}}} k_i \cdot \prod_{W_j \in \mathrm{root}(\mathscr{G}_{\mathscr{C}})} w^{+}_{j,\mathsf{cb}} \tag{5}$$

Example 7. Let $V = ABC$. Below we consider some examples of V-covers and, for each of them, show how the bound of v provided by Theorem 1 can be justified considering the lemmas proved so far and the inference rules for kD's. In order to help the reader, Fig. 5 depicts the roll-up relationships between the views involved.

- $\mathscr{C}_1 = (\{ABCD\}, \emptyset)$ is a V-cover since $V \preceq \oplus(N_{\mathscr{C}_1}) = ABCD$. From Lemma 1 it is derived $abc \leqslant abcd^{+}$.
- $\mathscr{C}_2 = (\{AB, BC\}, \emptyset)$ is a V-cover since $V \preceq \oplus(N_{\mathscr{C}_2}) = ABC$. Since the natural join between two views is a subset of their Cartesian product, it is $abc \leqslant ab^{+} \cdot bc^{+}$.
- $\mathscr{C}_3 = (\emptyset, \{AB \overset{k}{\rightarrow} C\})$. From Lemma 3 it immediately follows $abc \leqslant ab^{+} \cdot k$.
- $\mathscr{C}_4 = (\emptyset, \{A \overset{k_1}{\rightarrow} B, B \overset{k_2}{\rightarrow} C\})$. By applying rule R2, we derive $A \overset{k_1 k_2}{\rightarrow} BC$, thus $abc \leqslant a^{+} \cdot k_1 \cdot k_2$.
- $\mathscr{C}_5 = (\emptyset, \{A \overset{k_1}{\rightarrow} B, A \overset{k_2}{\rightarrow} C\})$. Rule R4 is now used to derive $A \overset{k_1 k_2}{\rightarrow} BC$, thus $abc \leqslant a^{+} \cdot k_1 \cdot k_2$.
- $\mathscr{C}_6 = (\{A'B, C\}, \{A' \overset{k}{\rightarrow} A\})$. According to rule R1 it is $A'B \overset{k}{\rightarrow} AB$, and from Lemma 3 $ab^{+} \leqslant k \cdot a'b^{+}$. On the other hand, $abc \leqslant ab^{+} \cdot c^{+}$, thus $abc \leqslant k \cdot a'b^{+} \cdot c^{+}$.

Note that Theorem 1 correctly captures also the case when no kD's are present, thus generalizing Eq. (4); in fact, in this case $\mathrm{root}(\mathscr{G}_{\mathscr{C}}) = S'$, where S' is as in Lemma 2. The reason why

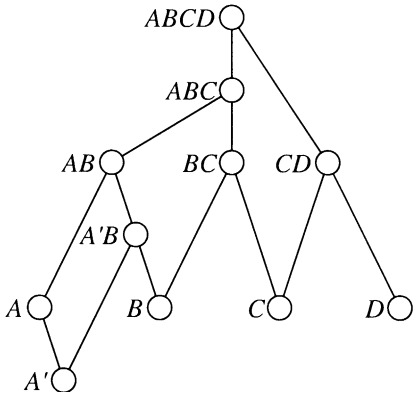

Fig. 5. Roll-up relationships of views in Example 7.

Theorem 1 applies only to reachable $\mathscr{C}$-graphs can be understood by means of a simple example. Let $\mathscr{C} = (\emptyset, \{A \xrightarrow{k_1} B, B \xrightarrow{k_2} A\})$, thus $\mathscr{G}_{\mathscr{C}}$ has no roots and is not reachable (see Fig. 4). If we try to apply Eq. (5) to such graph we obtain $u(\mathscr{C}) = k_1 \cdot k_2$, which is clearly wrong. On the other hand, a non-reachable $\mathscr{C}$-graph can be easily transformed into a reachable one by dropping one or more kD's (for instance, $B \xrightarrow{k_2} A$ in the example above): for this reason, all the $\mathscr{C}$-graphs considered from now on are implicitly assumed to be reachable. Besides, since according to Theorem 1 the bound yielded by a candidate set only depends on its associated $\mathscr{C}$-graph, in the following we will always consider reduced candidate sets.

In order to ensure that the cover-based strategy is consistent, we need to verify that the properties expressed by Lemmas 1–3 are valid also for the cover-based upper bounds.

Lemma 5. *If $W \preceq Z$, then $w^+_{\mathsf{cb}} \leqslant z^+_{\mathsf{cb}}$. If S is a set of views, then $[\![\oplus(S)]\!]^+_{\mathsf{cb}} \leqslant \prod_{W_i \in S'} w^+_{i,\mathsf{cb}}$. If $X \xrightarrow{k} Y$, then $[\![X \oplus Y]\!]^+_{\mathsf{cb}} \leqslant k \cdot x^+_{\mathsf{cb}}$.*

Proof. The first property is obvious, since the set of the W-covers includes Z and all the Z-covers. The second one derives from the fact that $\mathscr{C} = (S, \emptyset)$ is a cover for $\oplus(S) = \oplus(S')$, and $u_{\mathsf{cb}}(\mathscr{C}) = \prod_{W_i \in S'} w^+_{i,\mathsf{cb}}$. The third one is true since $\mathscr{C} = (\emptyset, \{X \xrightarrow{k} Y\})$ is a cover for $X \oplus Y$, and $u_{\mathsf{cb}}(\mathscr{C}) = k \cdot x^+_{\mathsf{cb}}$. □

5. Domination between candidate sets

In this section we complete the presentation of the cover-based bounding strategy by presenting some formal results useful to limit the complexity of algorithms aiming to compute upper bounds of view cardinalities. The problem is inherently difficult, since the space of V-covers to be analyzed in order to determine v^+_{cb} has exponential size. On the other hand, under some circumstances, a V-cover $\mathscr{C}$ can be discarded from the search space without even computing its associated bound $u_{\mathsf{cb}}(\mathscr{C})$.

Definition 9 (*Domination*). Let $\mathscr{C}$ and $\mathscr{C}'$ be two candidate sets. We say that $\mathscr{C}'$ *dominates* $\mathscr{C}$, written $\mathscr{C}' \sqsubseteq \mathscr{C}$, iff $u_{\mathsf{cb}}(\mathscr{C}') \leqslant u_{\mathsf{cb}}(\mathscr{C})$ for every possible input $\mathscr{I}$.

Again, let us consider the basic case where no kD's are present first. Note that, when $\mathscr{C}_1 = (S_1, \emptyset)$ and $\mathscr{C}_2 = (S_2, \emptyset)$, $\mathscr{C}_1$ can dominate $\mathscr{C}_2$ only if $\oplus(S_1) \preceq \oplus(S_2)$. In fact, if $\oplus(S_1) \not\preceq \oplus(S_2)$, there exists at least one attribute $A \in \oplus(S_1)$ which does not belong to $\oplus(S_2)$: intuitively, this means that the bound provided by S_2 is independent of the cardinality of A, whereas the bound yielded by S_1 is not, which is enough to show that $\mathscr{C}_1$ cannot dominate $\mathscr{C}_2$.

The following theorem, whose proof is reported in the Appendix, provides a necessary and sufficient condition for checking dominance between two candidate sets without kD's.

Theorem 2 (Domination between sets of views). *Let $\mathscr{C}_1 = (S_1, \emptyset)$ and $\mathscr{C}_2 = (S_2, \emptyset)$, with $S_1 = \{W_{1,1}, \ldots, W_{1,i}, \ldots, W_{1,m}\}$, and let S'_1 as in Lemma 2, thus $S'_1 = \{W_{1,1}, \ldots, W_{1,i}, \ldots, W_{1,m'}\}$ with $m' \leqslant m$. $\mathscr{C}_1$ dominates $\mathscr{C}_2$ iff S_2 can be partitioned into m' subsets $S_{2,1}, \ldots, S_{2,m'}$ such that $W_{1,i} \preceq \oplus(S_{2,i})$ $\forall i = 1, \ldots, m'$.*

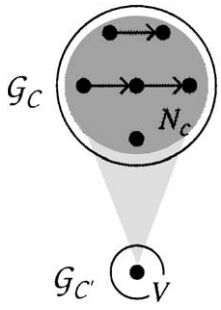

Fig. 6. Two $\mathscr{C}$-graphs in Lemma 6.

Example 8. It is $\{A'B, C\} \sqsubseteq \{AB, CD, E\}$, since $A'B \preceq AB$ and $C \preceq \oplus(\{CD, E\}) = CDE$.

Domination between candidate sets including kD's is much more complex to analyze; in the following we present a set of partial results, beginning with a simple corollary of Theorem 1.

Corollary 1. *Let $\mathscr{C}$ be a candidate set associated to a non-forest* [4] *$\mathscr{C}$-graph $\mathscr{G}_{\mathscr{C}}$, and let $\mathscr{C}'$ be a candidate set associated to a forest $\mathscr{C}$-graph $\mathscr{G}_{\mathscr{C}'}$ such that $N_{\mathscr{C}'} = N_{\mathscr{C}}$ and $E'_{\mathscr{C}} \subset E_{\mathscr{C}}$. Then $\mathscr{C}' \sqsubseteq \mathscr{C}$.*

As discussed in the proof of Theorem 1, a forest $\mathscr{C}$-graph satisfying the conditions above exists when the reason why $\mathscr{G}_{\mathscr{C}}$ is not a forest is that two or more arcs corresponding to kD's converge in the same node. Since Corollary 1 states that this kind of non-forest $\mathscr{C}$-graphs are always dominated, in the following they will not be considered. On the other hand, the other reason why $\mathscr{G}_{\mathscr{C}}$ is not a forest may be that at least two arcs labeled 1 converge in the same node: in this case, since removing such arcs would violate the very definition of $\mathscr{C}$-graph, Corollary 1 cannot be applied.

Next lemma expresses a sufficient condition for domination when one of the two candidate sets includes no kD's.

Lemma 6. *Let $\mathscr{C}' = (\{V\}, \emptyset)$ and $\mathscr{C} = (S, K)$ be two candidate sets. If $V \preceq \oplus(N_{\mathscr{C}})$ then $\mathscr{C}' \sqsubseteq \mathscr{C}$.*

Proof. $\mathscr{C}$ is a V-cover since $V \preceq \oplus(N_{\mathscr{C}})$, hence, from Theorem 1 it is $v^+_{\mathsf{cb}} \leqslant u_{\mathsf{cb}}(\mathscr{C})$. Since $u_{\mathsf{cb}}(\mathscr{C}') = v^+_{\mathsf{cb}}$, the result follows immediately (see Fig. 6). □

The following definition is preliminary to Lemma 7:

Definition 10 (*$\mathscr{C}$-subgraph*). Given $\mathscr{C} = (S, K)$ with associated $\mathscr{C}$-graph $\mathscr{G}_{\mathscr{C}} = (N_{\mathscr{C}}, E_{\mathscr{C}})$, let $N_{\mathscr{C}_1} \subset N_{\mathscr{C}}$. The $\mathscr{C}$-subgraph induced by $N_{\mathscr{C}_1}$ is defined as $\mathscr{G}_{\mathscr{C}_1} = (N_{\mathscr{C}_1}, E_{\mathscr{C}_1})$, where $E_{\mathscr{C}_1} = \{(X, Y) \in E_{\mathscr{C}} : X \in N_{\mathscr{C}_1} \wedge Y \in N_{\mathscr{C}_1}\}$. The $\mathscr{C}$-subgraph is called *proper* iff $\forall (X, Y) \in E_{\mathscr{C}}$ if $Y \in N_{\mathscr{C}_1}$ then $X \in N_{\mathscr{C}_1}$, i.e., iff there is no arc entering $\mathscr{G}_{\mathscr{C}_1}$. The (possibly empty) set

[4] A forest is a set of disjoint directed trees.

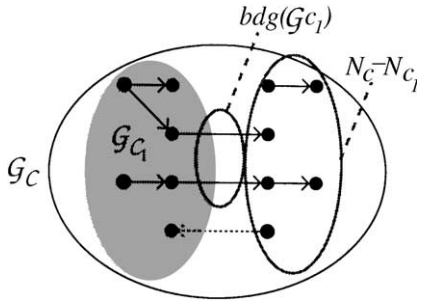

Fig. 7. A proper $\mathscr{C}$-subgraph.

$\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) = \{(X,Y) \in E_{\mathscr{C}} : X \in N_{\mathscr{C}_1} \wedge Y \in N_{\mathscr{C}} - N_{\mathscr{C}_1}\}$, i.e., the set of the arcs exiting $\mathscr{G}_{\mathscr{C}_1}$, is called the *bridge* induced by $N_{\mathscr{C}_1}$.

Example 9. Consider the $\mathscr{C}$-graph shown in Fig. 7. The $\mathscr{C}$-subgraph induced by $N_{\mathscr{C}_1}$ is proper; if also the arc shown in dotted line were included in $\mathscr{C}$, $\mathscr{G}_{\mathscr{C}_1}$ would not be proper.

Lemma 7. *Let $\mathscr{C}' = (\emptyset, \{V_1 \xrightarrow{k'} V_2\})$ and $\mathscr{C} = (S,K)$ be two candidate sets. If $N_{\mathscr{C}}$ can be partitioned into two subsets $N_{\mathscr{C}_1}$ and $N_{\mathscr{C}_2}$ such that:*

(1) *$\mathscr{G}_{\mathscr{C}_1}$ is proper; and*
(2) *$V_1 \preceq \oplus(N_{\mathscr{C}_1})$; and*
(3) *$V_2 \preceq \oplus(N_{\mathscr{C}_2})$; and*
(4) *$\oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}))) \ominus V_1 \preceq \oplus(N_{\mathscr{C}_u} \cup \mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u})))$, where $N_{\mathscr{C}_u} \subseteq N_{\mathscr{C}_2}$ is the subset of nodes of $\mathscr{G}_{\mathscr{C}_2}$ that are not reachable from $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1})$*
then $\mathscr{C}' \sqsubseteq \mathscr{C}$.

Example 10. Given $\mathscr{C}' = (\emptyset, \{AB \xrightarrow{k_1} C\})$ and $\mathscr{C} = (\{A\}, \{BE \xrightarrow{k_2} C, F \xrightarrow{k_3} D, D \xrightarrow{k_4} E\})$ it is $\mathscr{C}' \sqsubseteq \mathscr{C}$. In fact, if $N_{\mathscr{C}_1} = \{A, BE\}$ and $N_{\mathscr{C}_2} = \{C, D, E, F\}$ (see Fig. 8), it is:

- $AB = V_1 \preceq \oplus(N_{\mathscr{C}_1}) = ABE$;
- $C = V_2 \preceq \oplus(N_{\mathscr{C}_2}) = CDEF$;
- $\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1})) = BE$, $\oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}))) \ominus V_1 = E$, $\oplus(N_{\mathscr{C}_u}) = DF$, $\oplus(\mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u}))) = E$, and finally $E \preceq DEF$.

Since both $\mathscr{C}'$ and $\mathscr{C}$ are covers for ABC, we derive that $abc^+ \leqslant ab^+ \cdot k_1 \leqslant a^+ \cdot be^+ \cdot f^+ \cdot k_2 \cdot k_3 \cdot k_4$.

From this lemma we may finally derive the following theorem, stating a sufficient condition which may be used to recursively prove domination in the general case. Both Lemma 7 and Theorem 3 are proved in the Appendix.

Theorem 3. *Let $\mathscr{C}' = (S',K')$ and $\mathscr{C} = (S,K)$ be two candidate sets. Let $V \in N_{\mathscr{C}'}$ be a node of $\mathscr{G}_{\mathscr{C}'}$ with no outgoing arcs, and let $\mathscr{G}_{\mathscr{C}'_1}$ be the $\mathscr{C}$-subgraph induced by $N_{\mathscr{C}'} - \{V\}$. If $N_{\mathscr{C}}$ can be partitioned into two subsets $N_{\mathscr{C}_1}$ and $N_{\mathscr{C}_2}$ such that:*

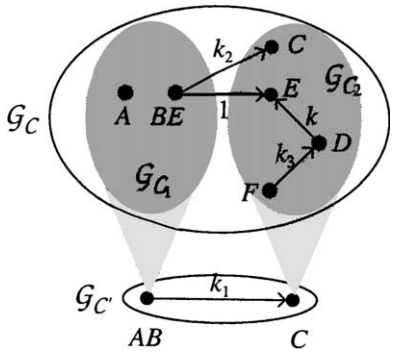

Fig. 8. The two $\mathscr{C}$-graphs in Example 10.

(1) *The* $\mathscr{C}$*-subgraph* $\mathscr{G}_{\mathscr{C}_1}$ *induced by* $N_{\mathscr{C}_1}$ *is proper*; *and*
(2) $\mathscr{C}'_1 \sqsubseteq \mathscr{C}_1$; *and*
(3) $V \preceq \oplus(N_{\mathscr{C}_2})$; *and*
(4) *If* $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \neq \emptyset$ *then* $\exists (W, V) \in E_{\mathscr{C}'} : \oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}))) \ominus W \preceq \oplus(N_{\mathscr{C}_u} \cup \mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u})))$, *where* $N_{\mathscr{C}_u} \subseteq N_{\mathscr{C}_2}$ *is the subset of nodes of* $\mathscr{G}_{\mathscr{C}_2}$ *that are not reachable from* $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1})$ *then* $\mathscr{C}' \sqsubseteq \mathscr{C}$.

Example 11. Given $\mathscr{C}' = (\{AB, D\}, \{A' \xrightarrow{k_1} C\})$ and $\mathscr{C} = (\{A, BE\}, \{A'B \xrightarrow{k_2} E, C \xrightarrow{k_3} BF, D' \xrightarrow{k_4} D\})$ it is $\mathscr{C}' \sqsubseteq \mathscr{C}$. In fact, as shown in Fig. 9, it is possible to apply recursively Theorem 3 to progressively smaller $\mathscr{C}$-graphs:

- *Step 1*. Let $V = D$ and $N_{\mathscr{C}_2} = \{D, D'\}$. Conditions (1), (3) and (4) are immediately verified, we have to verify that $(\{AB\}, \{A' \xrightarrow{k_1} C\}) \sqsubseteq (\{A, BE\}, \{A'B \xrightarrow{k_2} E, C \xrightarrow{k_3} BF\})$.
- *Step 2*. Let $V = C$ and $N_{\mathscr{C}_2} = \{C, E, BF\}$. Conditions (1) and (3) are immediately verified, as to condition (4) it is $W = A'$, $\oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}))) \ominus W = BC \preceq \oplus(N_{\mathscr{C}_u} \cup \mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u}))) = BCF$. It is left to verify that $(\{AB, A'\}, \emptyset) \sqsubseteq (\{A, BE, A'B\}, \emptyset)$, which is immediate due to Theorem 2.

Since both $\mathscr{C}'$ and $\mathscr{C}$ are covers for $ABCD$, we derive that $abcd^+ \leqslant ab^+ \cdot d^+ \cdot k_1 \leqslant a^+ \cdot be^+ \cdot a'b^+ \cdot c^+ \cdot d'^+ \cdot k_2 \cdot k_3 \cdot k_4$.

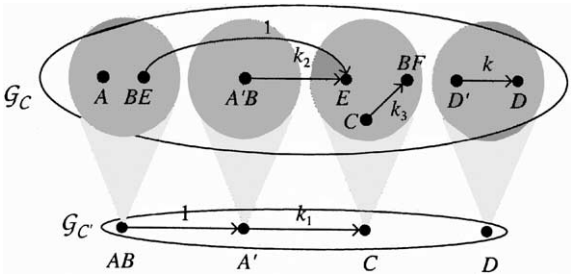

Fig. 9. The two $\mathscr{C}$-graphs in Example 11.

6. The estimator

Assuming that effective bounds have been derived, cardinality estimation must be based on a probabilistic model to derive an estimate, $\bar{v}$, of the cardinality of view V. The model we adopt here is based on the Cardenas' formula [2], which states that, when throwing N distinct objects into B buckets, the expected number of buckets in which at least one object will fall can be estimated as:

$$\Phi(B,N) = B\left(1 - \left(1 - \frac{1}{B}\right)^N\right) \leqslant \min\{B,N\} \tag{6}$$

Within the approach proposed in [17], (6) is used to estimate v by relying on the maximum cardinality of V, defined as the Cartesian product of the cardinalities of the attributes in V, $v_{\max} = \prod_{A_i \in V} a_i$, and on the cardinality of the base cube, $d = [\![\dim(\mathscr{D})]\!]$, that is:

$$\bar{v}_{\mathsf{sdnr}} = \Phi(v_{\max}, d) \leqslant \min\{v_{\max}, d\} \tag{7}$$

This formula turns out to significantly overestimate the cardinalities and can easily lead to violate the constraint $\bar{v}_{\mathsf{sdnr}} \leqslant v^+$.

In our approach, denoted se ("safe-estimate"), the above estimate is improved in two ways: by replacing $v_{\max}$ with the upper bound computed for v, for instance v^+_{cb}, as a measure of the maximum cardinality of V, and by replacing the cardinality of the base cube d with an estimate, $\bar{w}_{\mathsf{se}}$, of the cardinality of a view W such that $V \preceq W$. This leads to:

$$\bar{v}_{\mathsf{se}} = \Phi(v^+_{\mathsf{cb}}, \bar{w}_{\mathsf{se}}) \leqslant \min\{v^+_{\mathsf{cb}}, \bar{w}_{\mathsf{se}}\} \tag{8}$$

Since both v^+_{cb} and $\bar{w}_{\mathsf{se}}$ can be considerably lower than $v_{\max}$ and d, respectively, it is usually the case that $\bar{v}_{\mathsf{se}} \ll \bar{v}_{\mathsf{sdnr}}$. The rationale for (8) is that we can view the problem of estimating v as the one of distributing the tuples of view W, which are estimated to be $\bar{w}_{\mathsf{se}}$, over a number of v^+_{cb} "buckets".

Due to the need to know $\bar{w}_{\mathsf{se}}$, it is obvious that our estimation process must move downward from the top of the lattice (whose cardinality d is typically known) following a path leading to V. Clearly, this represents a simplification of the correct estimation procedure, which would require to determine $\bar{v}$ by following *all* the paths from $\dim(\mathscr{D})$ to V. On the other hand, this would lead to combinatorial explosion and necessitate of highly complex probabilistic models that are well beyond the current state-of-the-art knowledge.

From a more practical (numerical) point of view, it should be noted that moving from upper bounds to estimates leads to significant differences under specific conditions only. Two relevant cases should be considered, which arise from the limit behavior of Cardenas' formula:

(1) When $\bar{w}_{\mathsf{se}} \leqslant 0.1 v^+_{\mathsf{cb}}$ it is $\bar{v}_{\mathsf{se}} \approx \bar{w}_{\mathsf{se}}$
(2) When $\bar{w}_{\mathsf{se}} \geqslant 3 v^+_{\mathsf{cb}}$ it is $\bar{v}_{\mathsf{se}} \approx v^+_{\mathsf{cb}}$

The values 0.1 and 3 can thus be used to predict whether the estimator will deliver results which substantially differ from those directly obtainable from the bounder.

Example 12. In the Transfers scheme, we consider three input situations:

$$\mathscr{I}_1 = \big\{\{\mathsf{date}\}^+ = 10^3, \{\mathsf{year}\}^+ = 3, \{\mathsf{employee}\}^+ = 10^4,$$
$$\{\mathsf{fromOffice}\}^+ = \{\mathsf{toOffice}\}^+ = 10^3, \{\mathsf{fromCity}\}^+ = \{\mathsf{toCity}\}^+ = 10,$$
$$\{\mathsf{fromDept}\}^+ = \{\mathsf{toDept}\}^+ = 10\big\}$$
$$\mathscr{I}_2 = \mathscr{I}_1 \cup \Big\{\{\mathsf{employee}, \mathsf{year}\} \xrightarrow{4} \{\mathsf{fromOffice}, \mathsf{toOffice}, \mathsf{date}\}\Big\}$$
$$\mathscr{I}_3 = \mathscr{I}_2 \cup \Big\{\{\mathsf{fromCity}, \mathsf{fromDept}\}^+ = 40, \{\mathsf{toCity}, \mathsf{toDept}\}^+ = 40,$$
$$\{\mathsf{fromCity}, \mathsf{fromDept}\} \xrightarrow{2} \{\mathsf{toCity}, \mathsf{toDept}\}, \{\mathsf{fromCity}, \mathsf{fromDept}\} \xrightarrow{30} \{\mathsf{fromOffice}\},$$
$$\{\mathsf{toCity}, \mathsf{toDept}\} \xrightarrow{30} \{\mathsf{toOffice}\}\Big\}$$

Let $W = \dim(\mathscr{D}) = \{\mathsf{date}, \mathsf{employee}, \mathsf{fromOffice}, \mathsf{toOffice}\}$ be the base cube and $V = \{\mathsf{fromOffice}, \mathsf{toOffice}\}$ be the view whose cardinality is to be estimated.

- When $\mathscr{I} = \mathscr{I}_1$, the best W- and V-cover are, respectively,

$$\mathscr{C}_1 = (\{\{\mathsf{date}\}, \{\mathsf{employee}\}, \{\mathsf{fromOffice}\}, \{\mathsf{toOffice}\}\}, \emptyset)$$
$$\mathscr{C}_2 = (\{\{\mathsf{fromOffice}\}, \{\mathsf{toOffice}\}\}, \emptyset)$$

  Clearly, $\mathscr{C}_2 \sqsubseteq \mathscr{C}_1$.
- When $\mathscr{I} = \mathscr{I}_2$, the best cover for both W and V is

$$\mathscr{C}_3 = (\emptyset, \{\{\mathsf{employee}, \mathsf{year}\} \xrightarrow{4} \{\mathsf{fromOffice}, \mathsf{toOffice}, \mathsf{date}\}\})$$

  No domination relationship exists between $\mathscr{C}_1$ and $\mathscr{C}_3$ and between $\mathscr{C}_2$ and $\mathscr{C}_3$.
- When $\mathscr{I} = \mathscr{I}_3$, the best W-cover is still $\mathscr{C}_3$, while the best V-cover is

$$\mathscr{C}_4 = (\emptyset, \{\{\mathsf{fromCity}, \mathsf{fromDept}\} \xrightarrow{2} \{\mathsf{toCity}, \mathsf{toDept}\}, \{\mathsf{fromCity}, \mathsf{fromDept}\} \xrightarrow{30} \{\mathsf{fromOffice}\},$$
$$\{\mathsf{toCity}, \mathsf{toDept}\} \xrightarrow{30} \{\mathsf{toOffice}\}\})$$

  Still, no domination relationship exists with the other candidate sets.

Table 1 shows how the upper bound w^+_{cb} of W, the upper bound v^+_{cb} of V and the estimate $\bar{v}_{\mathsf{se}}$ improve as new cardinality constraints are progressively supplied. The estimate $\bar{v}_{\mathsf{se}}$ is based on the estimate of w, $\bar{w}_{\mathsf{se}}$, which is assumed to be equal to its upper bound w^+_{cb}. Note how, in the absence of cardinality constraints (i.e. if only the upper bounds of the cardinalities of all the single attributes in the dimensional scheme are known) the only reasonable estimate that can be inferred is $\bar{v}_{\mathsf{sdnr}} = 10^6$ obtained by formula (6) using $B = 10^6$ and $N = 10^{13}$. Even assuming that the cardinality of the base cube is known does not improve the estimate significatively; for example, assuming $w = 1.2 \times 10^5$, the estimate obtained using directly the Cardenas formula as in [17] is $\bar{v}_{\mathsf{sdnr}} = 1.2 \times 10^5$, that is much worse than those obtained exploiting the constraints in $\mathscr{I}_2$ and $\mathscr{I}_3$.

Table 1
Improving upper bounds and estimates for increasing domain-derived information

Input	w^+_{cb}	v^+_{cb}	$\bar{w}_{\mathsf{se}}$	$\bar{v}_{\mathsf{se}}$
$\mathscr{I}_1$	10^{13}	10^6	10^{13}	10^6
$\mathscr{I}_2$	1.2×10^5	1.2×10^5	1.2×10^5	7.6×10^4
$\mathscr{I}_3$	1.2×10^5	7.2×10^4	1.2×10^5	5.8×10^4

7. Conclusions and open issues

In this paper we have shown how cardinality constraints derived from the application domain may be employed to determine effective bounds on the cardinality of aggregate views and how, in turn, such bounds can be used to estimate the cardinality of the views. In order to improve the approach effectiveness, some issues still need to be investigated. In the following we briefly discuss those we believe to be crucial:

- *Computation*. The utility of an approach to cardinality estimation depends on the efficiency of its computation. When no kD's are included among the input constraints $\mathscr{I}$, the search for non-redundant upper bounds can be restricted to the set of *minimal* V-covers, where a minimal V-cover is one which is not dominated by any other V-cover and whose views are all constrained within $\mathscr{I}$ [3]; in the branch-and-bound algorithm sketched in [4], a careful enumeration of minimal covers allows to significatively reduce the (otherwise exponential) search space. On the other hand, considering minimality issues when kD's are involved is much more complex for two main reasons: firstly, a kD may be useful to compute an optimal cover even if the view on its left-hand side is not constrained in $\mathscr{I}$; secondly, the utility of a kD in determining a cover can be evaluated only by considering the other kD's included in the cover itself. Currently we are inclined to pursue a branch-and-bound approach which enumerates forest V-covers by assembling "useful" trees out of the constraints in $\mathscr{I}$. Once a tree covering a view $W \preceq V$ through a set of constraints $\mathscr{I}' \subset \mathscr{I}$ has been built, the problem is reduced to covering view $V \ominus W$ through the constraints in $\mathscr{I} - \mathscr{I}'$, and so on recursively.
- *Cardinality constraints*. The input knowledge may be further extended by considering other forms of cardinality constraints which are typically known to the experts of the application domain. For instance, while in this paper we have defined kD's to express *bounds* on the ratio between the cardinalities of two views, they may also be used to denote the *average* of such ratio; while this kind of knowledge cannot be used by the bounder, it allows the cardinality estimations to be improved. For instance, knowing that the average number of transfers for each employee on each year is 2, would allow the cardinality of the base cube to be estimated as twice the cardinality of view $\{\mathsf{employee}, \mathsf{year}\}$.
- *Probabilistic estimates*. Estimates based on Cardenas' formula can be improved in several ways. In particular, information on *lower bounds* of cardinalities could be considered by exploiting the results in [6], as well as information concerning the distribution of attribute values over their domains. Obviously, this requires to develop a bounding strategy for computing lower bounds; [3] presents some results in this direction.

Appendix A

Proof of Theorem 1. We will first prove that Eq. (5) holds when $\mathcal{G}_{\mathscr{C}}$ is a *forest*, i.e., a set of pairwise disjoint rooted trees. Then we will generalize to arbitrary $\mathscr{C}$-graphs.

Assume that $\mathcal{G}_{\mathscr{C}}$ is a forest. The set root($\mathcal{G}_{\mathscr{C}}$) can be partitioned into two subsets: alone($\mathcal{G}_{\mathscr{C}}$), which contains the "stand-alone" nodes, i.e. those for which the tree consists only of a root node, and $\alpha = \text{root}(\mathcal{G}_{\mathscr{C}}) - \text{alone}(\mathcal{G}_{\mathscr{C}})$, which contains the other roots (see Fig. 10). Let (N_j, E_j) be the tree rooted in $W_j \in \alpha$. Now, since $N_{\mathscr{C}} = \text{alone}(\mathcal{G}_{\mathscr{C}}) \cup (\bigcup_{j:W_j \in \alpha} N_j)$, it is

$$[\![\oplus (N_{\mathscr{C}})]\!] = [\![\oplus (\text{alone}(\mathcal{G}_{\mathscr{C}})) \oplus (\oplus_{j:W_j \in \alpha}(\oplus(N_j)))]\!] \leqslant \prod_{W_j \in \text{alone}(\mathcal{G}_{\mathscr{C}})} w_j^+ \cdot \prod_{j:W_j \in \alpha} [\![\oplus (N_j)]\!]$$

By repeatedly applying rules R2 and R4 to (N_j, E_j) it is derived $W_j \xrightarrow{\prod_{E_j} k_i} \oplus (N_j - \{W_j\})$, hence for Lemma 3

$$[\![\oplus (N_j)]\!] \leqslant w_j^+ \cdot \prod_{E_j} k_i$$

which leads to

$$[\![\oplus (N_{\mathscr{C}})]\!] \leqslant \prod_{W_j \in \text{alone}(\mathcal{G}_{\mathscr{C}})} w_j^+ \cdot \prod_{j:W_j \in \alpha} \left(w_j^+ \cdot \prod_{E_j} k_i \right) = \prod_{W_j \in \text{root}(\mathcal{G}_{\mathscr{C}})} w_j^+ \cdot \prod_{E_{\mathscr{C}}} k_i$$

Since by Definition 7 it is $V \preceq \oplus(N_{\mathscr{C}})$, we obtain $v \leqslant [\![\oplus (N_{\mathscr{C}})]\!]$; thus, $u_{\text{cb}}(\mathscr{C})$ can be written as claimed.

When $\mathcal{G}_{\mathscr{C}} = (N_{\mathscr{C}}, E_{\mathscr{C}})$ is reachable but is not a forest, it is always possible to find $E'_{\mathscr{C}} \subset E_{\mathscr{C}}$ such that the graph $\mathcal{G}_{\mathscr{C}'} = (N_{\mathscr{C}}, E'_{\mathscr{C}})$ is a forest with the same roots as $\mathcal{G}_{\mathscr{C}}$. If $\mathcal{G}_{\mathscr{C}'}$ satisfies the definition of $\mathscr{C}$-graph, meaning that it is obtained by dropping no arcs labeled 1, it is obviously $v \leqslant u_{\text{cb}}(\mathscr{C}') \leqslant u_{\text{cb}}(\mathscr{C}')$ (the first inequality follows from the first part of the proof, since $\mathcal{G}_{\mathscr{C}'}$ is a forest and $V \preceq \oplus(N_{\mathscr{C}'})$; the second inequality follows from root($\mathcal{G}_{\mathscr{C}'}$) = root($\mathcal{G}_{\mathscr{C}}$) and from the observation that $E'_{\mathscr{C}} \subset E_{\mathscr{C}}$ implies $\prod_{E'_{\mathscr{C}}} k_i < \prod_{E_{\mathscr{C}}} k_i$).

On the other hand, if a forest $\mathcal{G}_{\mathscr{C}'}$ can be obtained from $\mathcal{G}_{\mathscr{C}}$ only by dropping one or more arcs labeled 1, Eq. (5) is still valid since arcs labeled 1 do not contribute to the bound expressed by $u_{\text{cb}}(\mathscr{C})$. □

Proof of Theorem 2. (If.) If $W_{1,i} \preceq \oplus(S_{2,i})\ \forall i = 1, \ldots, m'$ the result directly follows from Lemma 2.

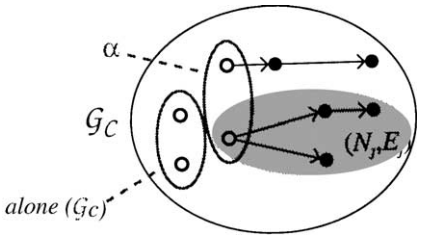

Fig. 10. A forest $\mathscr{C}$-graph; white circles denote views in root($\mathcal{G}_{\mathscr{C}}$).

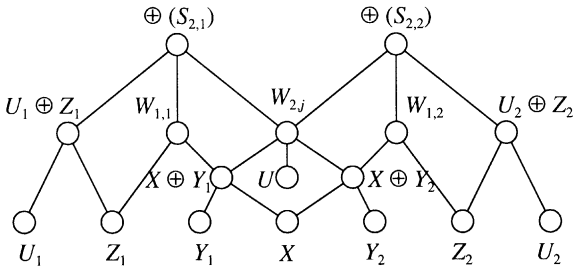

Fig. 11. Roll-up relationships of views in the Proof of Theorem 2.

(Only if.) We only provide a sketch of the complete proof. Without loss of generality assume $m = m' = 2$ and $\oplus(S_1) \preceq \oplus(S_2)$. If S_2 is not partitionable as required by the theorem there exists a view $W_{2,j}$ that is used to cover both $W_{1,1}$ and $W_{1,2}$. We can write $W_{1,1} = X \oplus Y_1 \oplus Z_1$, $W_{1,2} = X \oplus Y_2 \oplus Z_2$, $W_{2,j} = U \oplus X \oplus Y_1 \oplus Y_2$, $\oplus(S_{2,1}) = W_{2,j} \oplus (Z_1 \oplus U_1)$, and $\oplus(S_{2,2}) = W_{2,j} \oplus (Z_2 \oplus U_2)$, which guarantees that $W_{1,1} \preceq \oplus(S_{2,1})$ and $W_{1,2} \preceq \oplus(S_{2,2})$ due to the definition of $\oplus$ (see Fig. 11). In the complete proof it is shown that there exist legal instances such that $w^+_{1,1} \approx [\![X \oplus Y_1]\!]^+ \cdot z^+_1$, $w^+_{1,2} \approx [\![X \oplus Y_2]\!]^+ \cdot z^+_2$, $w^+_{2,j} \approx [\![X \oplus Y_1 \oplus Y_2]\!]^+$, $[\![Z_1 \oplus U_1]\!]^+ \approx z^+_1$, and $[\![Z_2 \oplus U_2]\!]^+ \approx z^+_2$. Under these conditions, the inequality that should hold for domination is: $w^+_{1,1} \cdot w^+_{1,2} = ([\![X \oplus Y_1]\!]^+ \cdot z^+_1) \cdot ([\![X \oplus Y_2]\!]^+ \cdot z^+_2) \leqslant [\![X \oplus Y_1 \oplus Y_2]\!]^+ \cdot z^+_1 \cdot z^+_2 = [\![\oplus (S_2)]\!]^+$, that is $[\![X \oplus Y_1]\!]^+ \cdot [\![X \oplus Y_2]\!]^+ \leqslant [\![X \oplus Y_1 \oplus Y_2]\!]^+$, which can be easily invalidated. □

Proof of Lemma 7. We have to prove that

$$u_{\mathsf{cb}}(\mathscr{C}') = k' \cdot v^+_{1,\mathsf{cb}} \leqslant \prod_{E_{\mathscr{C}}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}})} w^+_{i,\mathsf{cb}} = u_{\mathsf{cb}}(\mathscr{C})$$

Let $N_{\mathscr{C}_r} = N_{\mathscr{C}_2} - N_{\mathscr{C}_u}$ be the set of nodes of $\mathscr{G}_{\mathscr{C}_2}$ reachable from $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1})$; since $\mathscr{G}_{\mathscr{C}_1}$ is proper it is $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}2}) = \emptyset$, thus we may write (see Fig. 12):

$$E_{\mathscr{C}} - E_{\mathscr{C}_1} = \mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_2} = \mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_r} \cup \mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u}) \cup E_{\mathscr{C}_u}$$

where $E_{\mathscr{C}_r}$ and $E_{\mathscr{C}_u}$ are, respectively, the arcs in $\mathscr{G}_{\mathscr{C}_r}$ and $\mathscr{G}_{\mathscr{C}_u}$. In fact, it is necessarily $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_r}) = \emptyset$ since, otherwise, the nodes in $N_{\mathscr{C}_u}$ would become reachable. We can now rewrite $u_{\mathsf{cb}}(\mathscr{C})$ as:

$$\begin{aligned} u_{\mathsf{cb}}(\mathscr{C}) &= \prod_{E_{\mathscr{C}}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}})} w^+_{i,\mathsf{cb}} = \left(\prod_{E_{\mathscr{C}_1}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}_1})} w^+_{i,\mathsf{cb}} \right) \cdot \prod_{E_{\mathscr{C}} - E_{\mathscr{C}1}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}_u})} w^+_{i,\mathsf{cb}} \\ &= u_{\mathsf{cb}}(\mathscr{C}_1) \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_r}} k_i \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u}) \cup E_{\mathscr{C}_u}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}_u})} w^+_{i,\mathsf{cb}} \end{aligned}$$

Since $\mathscr{C}_1$ is a V_1-cover, it is $v^+_{1,\mathsf{cb}} \leqslant u_{\mathsf{cb}}(\mathscr{C}_1)$. Thus, letting $h = \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_r}} k_i$, it is sufficient to prove that

$$k' \leqslant h \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u}) \cup E_{\mathscr{C}_u}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}_u})} w^+_{i,\mathsf{cb}}$$

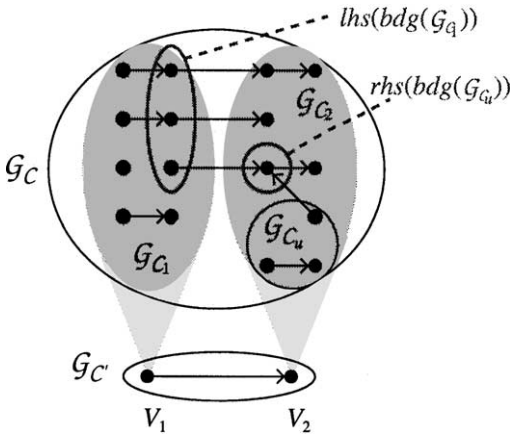

Fig. 12. Two $\mathscr{C}$-graphs in Lemma 7.

From the arguments used in the proof of Theorem 1 it can be easily derived, by a repeated application of rules R2 and R4, that

$$\oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}))) \xrightarrow{h} \oplus(N_{\mathscr{C}_r})$$

thus, by letting $W = \oplus(\mathrm{lhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1})))$, $R = \oplus(N_{\mathscr{C}_r})$, and applying rules R1 and R3, we obtain:

$$V_1 \oplus W \xrightarrow{h} R$$

Due to property (1), we may write $V_1 \oplus (W \ominus V_1) \xrightarrow{h} R$. From here, rule R5 leads to

$$V_1 \xrightarrow{h\cdot[\![(W\ominus V_1)\oplus(V_2\ominus R)]\!]^+} R \oplus (V_2 \ominus R)$$

Again for property (1), this is equivalent to

$$V_1 \xrightarrow{h\cdot[\![(W\ominus V_1)\oplus(V_2\ominus R)]\!]^+} R \oplus V_2$$

$$(\mathrm{R3}) \Rightarrow V_1 \xrightarrow{h\cdot[\![(W\ominus V_1)\oplus(V_2\ominus R)]\!]^+} V_2$$

Since $V_1 \xrightarrow{k'} V_2$ and the input is sound and minimal by hypothesis, it is necessarily

$$k' \leqslant h \cdot [\![(W \ominus V_1) \oplus (V_2 \ominus R)]\!]^+$$

Now, letting $U = \oplus(N_{\mathscr{C}_u} \cup \mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}u})))$ and $Y = V_2 \ominus R$ (see Fig. 13), we will prove that $(W \ominus V_1) \oplus Y \preceq U$. By definition of the $\ominus$ operator it is $Y \oplus (V_2 \otimes R) = V_2$ and for each Y' such that $Y' \oplus (V_2 \otimes R) = V_2$ it is $Y \preceq Y'$. Let $Y' = V_2 \otimes U$; using the distributive and the absorption properties, and considering that $V_2 \preceq \oplus(N_{\mathscr{C}_2}) = \oplus(N_{\mathscr{C}_r} \cup N_{\mathscr{C}_u}) = R \oplus U$ by hypothesis, we get

$$\begin{aligned} Y' \oplus (V_2 \otimes R) &= (V_2 \otimes U) \oplus (V_2 \otimes R) = ((V_2 \otimes U) \oplus V_2) \otimes ((V_2 \otimes U) \oplus R) \\ &= V_2 \otimes (R \oplus V_2) \otimes (R \oplus U) = V_2 \end{aligned}$$

Hence, $Y \preceq V_2 \otimes U \preceq U$. On the other hand, $W \ominus V_1 \preceq U$ by hypothesis. Thus:

$$k' \leqslant h \cdot [\![(W \ominus V_1) \oplus Y]\!]^+ \leqslant h \cdot [\![U]\!]^+$$

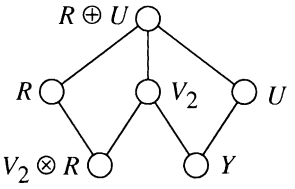

Fig. 13. Roll-up relationships of views in the Proof of Lemma 7.

Finally, because of Theorem 1 it is

$$[\![U]\!]^+ = [\![\oplus (N_{\mathscr{C}_u} \cup \mathrm{rhs}(\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_u})))]\!]^+ \leqslant \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}u}) \cup E_{\mathscr{C}u}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}_u})} w^+_{i,\mathsf{cb}}$$

which proves the lemma. □

Proof of Theorem 3. We will first examine the case $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}1}) = \emptyset$. In this case it is:

$$\begin{aligned} u_{\mathsf{cb}}(\mathscr{C}) &= u_{\mathsf{cb}}(\mathscr{C}_1) \cdot u_{\mathsf{cb}}(\mathscr{C}_2) \\ u_{\mathsf{cb}}(\mathscr{C}'_1) &\leqslant u_{\mathsf{cb}}(\mathscr{C}_1) \text{ (since } \mathscr{C}'_1 \sqsubseteq \mathscr{C}_1) \\ v^+_{\mathsf{cb}} &\leqslant u_{\mathsf{cb}}(\mathscr{C}_2) \text{ (for Lemma 6)} \end{aligned}$$

If in $\mathscr{G}_{\mathscr{C}'}$ there is an arc, labeled k', entering V (with $1 \leqslant k' \leqslant v^+$ since the input is sound) it is

$$u_{\mathsf{cb}}(\mathscr{C}') = u_{\mathsf{cb}}(\mathscr{C}'_1) \cdot k' \leqslant u_{\mathsf{cb}}(\mathscr{C}_1) \cdot v^+_{\mathsf{cb}} \leqslant u_{\mathsf{cb}}(\mathscr{C}_1) \cdot u_{\mathsf{cb}}(\mathscr{C}_2) = u_{\mathsf{cb}}(\mathscr{C})$$

Otherwise, it is $u_{\mathsf{cb}}(\mathscr{C}') = u_{\mathsf{cb}}(\mathscr{C}_1) \cdot v^+_{\mathsf{cb}} \leqslant u_{\mathsf{cb}}(\mathscr{C})$. In both cases, we have $\mathscr{C}' \sqsubseteq \mathscr{C}$.

Now, let $\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \neq \emptyset$; we have

$$\begin{aligned} u_{\mathsf{cb}}(\mathscr{C}') &= u_{\mathsf{cb}}(\mathscr{C}'_1) \cdot k' \\ u_{\mathsf{cb}}(\mathscr{C}) &= \prod_{E_{\mathscr{C}}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}})} w^+_{i,\mathsf{cb}} \\ &= u_{\mathsf{cb}}(\mathscr{C}_1) \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_r}} k_i \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}u}) \cup E_{\mathscr{C}u}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}u})} w^+_{i,\mathsf{cb}} \end{aligned}$$

where the last equality follows from the proof of Lemma 7. Since $\mathscr{C}'_1 \sqsubseteq \mathscr{C}_1$ it is $u_{\mathsf{cb}}(\mathscr{C}'_1) \leqslant u_{\mathsf{cb}}(\mathscr{C}_1)$; for Lemma 7 it also is $k' \leqslant \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}_1}) \cup E_{\mathscr{C}_r}} k_i \cdot \prod_{\mathrm{bdg}(\mathscr{G}_{\mathscr{C}u}) \cup E_{\mathscr{C}u}} k_i \cdot \prod_{W_i \in \mathrm{root}(\mathscr{G}_{\mathscr{C}u})} w^+_{i,\mathsf{cb}}$. Thus, the theorem is proved. □

References

[1] G. Birkhoff, Lattice Theory, American Mathematical Society, 1995.
[2] A.F. Cardenas, Analysis and performance of inverted database structures, Comm. of the ACM 18 (5) (1975) 253–263.
[3] P. Ciaccia, M. Golfarelli, S. Rizzi, On Estimating the Cardinality of Aggregate Views, in: Proceedings of the DMDW'01, Interlaken, Switzerland, 2001, pp. 12.1–12.10.
[4] P. Ciaccia, M. Golfarelli, S. Rizzi, Using Domain-Derived Constraints to Bound the Cardinality of Aggregate Views, in: Proceedings of the 9th SEBD, Venice, Italy, 2001, pp. 249–256.

[5] P. Ciaccia, D. Maio, On the complexity of finding bounds for projection cardinalities in relational databases, Information Systems 17 (6) (1992) 511–515.
[6] P. Ciaccia, D. Maio, Domains and active domains: What this distinction implies for the estimation of projection sizes in relational databases, IEEE Trans. on Knowledge and Data Engineering 7 (4) (1995) 641–655.
[7] M. Golfarelli, S. Rizzi, View Materialization for Nested GPSJ Queries, in: Proceedings of the DMDW'00, Stockholm, Sweden, 2000.
[8] J. Grant, J. Minker, Numerical Dependencies, in: H. Gallaire, J. Minker, J.-M. Nicolas (Eds.), Advances in Database Theory, vol. II, Plenum Publ. Co., 1983.
[9] H. Gupta, Selection of Views to Materialize in a Data Warehouse, in: Proceedings of the ICDT'97, Delphi, Greece, 1997, pp. 98–112.
[10] H. Gupta, V. Harinarayan, A. Rajaraman, J. Ullman, Index Selection for OLAP, in: Proceedings of the ICDE'97, Birmingham, UK, 1997, pp. 208–219.
[11] M. Gyssens, L.V.S. Lakshmanan, A Foundation for Multi-Dimensional Databases, in: Proceedings of the 23rd VLDB, Athens, Greece, 1997, pp. 106–115.
[12] V. Harinarayan, A. Rajaraman, J. Ullman, Implementing Data Cubes Efficiently, in: Proceedings of the ACM Sigmod Conference, Montreal, Canada, 1996, pp. 205–216.
[13] W. Hou, G. Özsoyoglu, Statistical estimators for aggregate relational algebra queries, ACM Trans. on Database Systems 16 (4) (1991) 600–654.
[14] M.V. Mannino, P. Chu, T. Sager, Statistical profile estimation in database systems, ACM Computing Surveys 20 (3) (1988) 191–221.
[15] M. Muralikrishna, D.J. DeWitt, Equi-depth Histograms for Estimating Selectivity Factors for Multi-Dimensional Queries, in: Proceedings of the ACM Sigmod Conference, Chicago, IL, 1988, pp. 28–36.
[16] K. Ross, D. Srivastava, Fast Computation of Sparse Datacubes, in: Proceedings of the 23rd VLDB, Athens, Greece, 1997, pp. 116–125.
[17] A. Shukla, P. Deshpande, J. Naughton, K. Ramasamy, Storage Estimation for Multidimensional Aggregates in the Presence of Hierarchies, in: Proceedings of the 22nd VLDB, Mumbai, India, 1996, pp. 522–531.
[18] D. Theodoratos, M. Bouzeghoub, A General Framework for the View Selection Problem for Data Warehouse Design and Evolution, in Proceedings of the DOLAP 2000, Washington, DC, 2000, pp. 1–8.
[19] P. Vassiliadis, Gulliver in the Land of Data Warehousing: Practical Experiences and Observations of a Researcher, in: Proceedings of the DMDW'00, Stockholm, Sweden, 2000, pp. 12/1–12/16.

Paolo Ciaccia has a "Laurea" degree in Electronic Engineering (1985) and a Ph.D. in Electronic and Computer Engineering (1992), both from the University of Bologna, Italy, where he is Full Professor of Information Systems. His research interests include similarity and preference-based query processing, data warehousing and data mining. He is one of the designers of the M-tree, an index for metric data used by many multimedia and data mining research groups in the world. He participated several international and national research projects, among which ESPRIT IV LTR HERMES on Multimedia Information Management Systems. Since 2001 he has been the responsible for DEIS within the IST/FET PANDA Network (Patterns for Next-Generation Database Systems). He has published more than 60 refereed papers in the areas of database systems, neural networks, software engineering, and autonomous systems, in major international journals (including IEEE TKDE, IEEE TSE, ACM TODS, ACM TOIS, IS, and Biological Cybernetics) and international conferences (including VLDB, ACM-PODS, EDBT, and ICDE). In 1999 he was programme co-chair of the 1st International Workshop on Similarity Search, and in 2002 he was PC chair of the SEBD italian conference on advanced data bases. He is an Associate Editor of IEEE TKDE.

Matteo Golfarelli received his degree in Computer Science in 1995 and his Ph.D. for his work on autonomous agents in 1998 from the University of Bologna, Italy. In 2000 he joined the Computer Science Department as associate researcher, teaching Information Systems and Computer Architectures. He has published over 30 papers in refereed journals and international conferences in the fields of autonomous agents, pattern recognition and databases. He is currently involved in the EU PANDA thematic network concerning pattern-base management systems. His current research interests include all the aspects related to data warehouse design, in particular multidimensional modeling, view materialization and fragmentation, physical design.

Stefano Rizzi received his degree in Electronic Engineering in 1988 and his Ph.D. for his work on autonomous agents in 1996 from the University of Bologna, Italy. In 1995 he joined the Computer Science Department as a researcher. Since 1998 he is Associate Professor at the University of Bologna, where is the head of the Data Warehousing Laboratory and teaches advanced information systems and software engineering. He has published over 50 papers in refereed journals and international conferences in the fields of data warehousing, pattern recognition, mobile robotics, multi-agent systems, and visual query languages. He joined several national research projects on the above areas and is currently involved in the EU PANDA thematic network concerning pattern-base management systems. He served in the PC of several international conferences and as a reviewer in journals. His current research interests include all the aspects related to data warehouse design, in particular multidimensional modeling, view materialization and fragmentation, physical design.