# Structure and Complexity of Bag Consistency

Albert Atserias
Universitat Politècnica de Catalunya
Barcelona, Catalonia, Spain
atserias@cs.upc.edu

Phokion G. Kolaitis
University of California Santa Cruz and IBM Research
Santa Cruz, CA, USA
kolaitis@ucsc.edu

## ABSTRACT

Since the early days of relational databases, it was realized that acyclic hypergraphs give rise to database schemas with desirable structural and algorithmic properties. In a by-now classical paper, Beeri, Fagin, Maier, and Yannakakis established several different equivalent characterizations of acyclicity; in particular, they showed that the sets of attributes of a schema form an acyclic hypergraph if and only if the local-to-global consistency property for relations over that schema holds, which means that every collection of pairwise consistent relations over the schema is globally consistent. Even though real-life databases consist of bags (multisets), there has not been a study of the interplay between local consistency and global consistency for bags. We embark on such a study here and we first show that the sets of attributes of a schema form an acyclic hypergraph if and only if the local-to-global consistency property for bags over that schema holds. After this, we explore algorithmic aspects of global consistency for bags by analyzing the computational complexity of the global consistency problem for bags: given a collection of bags, are these bags globally consistent? We show that this problem is in NP, even when the schema is part of the input. We then establish the following dichotomy theorem for fixed schemas: if the schema is acyclic, then the global consistency problem for bags is solvable in polynomial time, while if the schema is cyclic, then the global consistency problem for bags is NP-complete. The latter result contrasts sharply with the state of affairs for relations, where, for each fixed schema, the global consistency problem for relations is solvable in polynomial time.

## CCS CONCEPTS

• **Theory of computation → Database theory**.

## KEYWORDS

acyclic schemas; acyclic hypergraphs; local consistency; pairwise consistency; global consistency; bag semantics; universal relation

**ACM Reference Format:**
Albert Atserias and Phokion G. Kolaitis. 2021. Structure and Complexity of Bag Consistency. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '21), June 20–25, 2021, Virtual Event, China.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3452021.3458329



*PODS '21, June 20–25, 2021, Virtual Event, China.*

ACM ISBN 978-1-4503-8381-3/21/06.
https://doi.org/10.1145/3452021.3458329

## 1 INTRODUCTION

Early investigations in database theory led to the discovery that many fundamental algorithmic problems about relational databases are intractable. In particular, the relational join evaluation problem is NP-complete: given relations $R_1, \ldots, R_m$ and a tuple $t$, does $t$ belong to the join $R_1 \bowtie \ldots \bowtie R_m$ of the given relations? This motivated the pursuit of tractable cases of the relational join evaluation problem. In an influential paper [28], Yannakakis showed that the relational join evaluation problem is solvable in polynomial time if the schema of the given relations is acyclic, i.e., if the sets of the attributes of the given relations are the hyperedges of an acyclic hypergraph. The notion of hypergraph acyclicity turned out to have several other desirable properties in relational databases that were explored in depth by Beeri, Fagin, Maier, and Yannakakis [8]. Arguably the most prominent such property has to do with the universal relation problem, also known as the global consistency problem [6, 25]. This problem asks: given relations $R_1, \ldots, R_m$, is there a relation $R$ such that, for every $i \leq m$, the projection of $R$ on the attributes of $R_i$ is equal to $R_i$? If the answer is positive, then the relations $R_1, \ldots, R_m$ are said to be globally consistent relations and $R$ is said to be a universal relation for them. Honeyman, Ladner, and Yannakakis [15] showed that the universal relation problem is NP-complete, even when all input relations are binary. It is easy to see that if the relations $R_1, \ldots, R_m$ are globally consistent, then they are pairwise consistent, i.e., every two of them are globally consistent; the converse, however, does not hold, in general. Beeri et al. [8] showed that a schema is acyclic if and only if the local-to-global consistency property for relations over that schema holds, which means that every collection of pairwise consistent relations over the schema is globally consistent. Thus, for acyclic schemas, pairwise consistency is both a necessary and sufficient condition for global consistency; therefore, the universal relation problem is solvable in polynomial time.

In all aforementioned results, relations are assumed to be sets. In 1993, Chaudhuri and Vardi [13] pointed out that there is a gap between database theory and database practice because "real" databases use bags (multisets). They then called for a re-examination of the foundations of databases where the fundamental concepts and algorithmic problems are investigated under bag semantics, instead of set semantics. In particular, Chaudhuri and Vardi [13] raised the question of the decidability of the conjunctive query containment problem under bags semantics (the same problem under set semantics is known to be NP-complete [12]). Various efforts in the past and some recent progress notwithstanding [17, 18], this question remains unanswered at present.

It is perhaps surprising that a study of consistency notions under bag semantics has not been carried out to date. Our main goal in this paper is to embark on such a study and to explore both structural and algorithmic aspects of pairwise consistency and of

global consistency under bag semantics. In this study, the notions of consistency of bags are, of course, defined using bag semantics in the computation of projections.

In general, properties of relations do not automatically carry over to similar properties of bags. This phenomenon manifests itself in the context of consistency properties. Indeed, it is well known that if a collection of relations is globally consistent, then their relational join is a witness to their global consistency (see, e.g., [15]); in other words, their relational join is a universal relation for them and, in fact, it is the biggest universal relation. In contrast, as we point out in Section 3, this property fails for bags, i.e., there is a collection of bags that is globally consistent but the bag-join of the bags in the collection is not a witness to their global consistency. In fact, this holds even for two consistent bags and, furthermore, there may be no biggest witness to the consistency of these bags. Our first result establishes that two bags are consistent if and only if they have the same projection on their common attributes. While the analogous fact for relations is rather trivial, here we need to bring in tools from the theory of linear programming and maximum flow problems. As a corollary, we obtain a polynomial-time algorithm for checking whether two given bags are consistent and returning a witness to their consistency, if they are consistent. After this, we establish our main result concerning the structure of bag consistency. Specifically, we show that the sets of attributes of a schema form an acyclic hypergraph if and only if the local-to-global consistency for bags over that schema holds. This shows that the main finding by Beeri et al.[8] about acyclicity and consistency extends to bags. As we explain in Section 4, however, the architecture of the proof is different from that in [8]. In particular, if a schema is cyclic, we give an explicit construction of a collection of bags that are pairwise consistent, but not globally consistent; the inspiration for our construction comes from an earlier construction of hard-to-prove tautologies in propositional logic by Tseitin [24].

We then explore algorithmic aspects of global consistency for bags by analyzing the computational complexity of the global consistency problem for bags: given a collection of bags, are these bags globally consistent? Using a sparse-model property of integer programming that is reminiscent of Carathéodory's Theorem for conic hulls [14], we first show that this problem is in NP, even when the schema is part of the input. After this, we establish the following dichotomy theorem for fixed schemas: if the schema is acyclic, then the global consistency problem for bags is solvable in polynomial time, while if the schema is cyclic, then the global consistency problem for bags is NP-complete. The latter result contrasts sharply with the state of affairs for relations, where, for each fixed schema, the global consistency problem for relations is solvable in polynomial time. Our NP-hardness results build on an earlier NP-hardness result about three-dimensional statistical data tables by Irving and Jerrum [16], which was later on refined by De Loera and Onn [19]. Translated into our context, this result asserts the NP-hardness of the global consistency problem for bags over the triangle hypergraph, i.e., the hypergraph with hyperedges of the form $\{A_1, A_2\}$, $\{A_2, A_3\}$, $\{A_3, A_1\}$. Finally, we give a polynomial-time algorithm for the following problem: given an acyclic schema and a collection of pairwise consistent bags over that schema, construct a (small) witness to their global consistency. For this, we use Carathéodory's classical theorem for conic hulls and the existence of strongly polynomial algorithms for maximum flow problems (for the latter, see, e.g., [20]).

*Related Work.* The interplay between local consistency and global consistency arises naturally in several different settings. Already in 1962, Vorob'ev [26] studied this interplay in the setting of probability distributions and characterized the local-to-global consistency property for probability distributions in terms of a structural property of hypergraphs that turned out to be equivalent to hypergraph acyclicity. It appears that Beeri et al.[8] were unaware of Vorob'ev work, but later on Vorob'ev's work was cited in a survey of database theory by Yannakakis [29]. In recent years, the interplay between local consistency and global consistency has been explored at great depth in the setting of quantum mechanics by Abramsky and his collaborators (see, e.g., [3–5]). In that setting, the interest is in contextuality phenomena, which are situations where collections of measurements are locally consistent but globally inconsistent - the celebrated Bell's Theorem [9] is an instance of this. The similarities between these different settings (probability distributions, relational databases, and quantum mechanics) were pointed out explicitly by Abramsky [1, 2]. This also raised the question of developing a unifying framework in which, among other things, the results by Vorob'ev and the results by Beeri et al. are special cases of a single result. Using a relaxed notion of consistency, we recently established such a result for $K$-relations, where $K$ is a positive semiring [7][1]. By definition, a $K$-relation is a relation such that each of its tuples has an associated element from the semiring $K$ as value. In particular, if $\mathbb{Z}^{\geq 0}$ is the semiring of non-negative integers (also known as the bag semiring), then the $\mathbb{Z}^{\geq 0}$-relations are precisely the bags. For $\mathbb{Z}^{\geq 0}$-relations, however, the relaxed notion of consistency that we studied in [7] is essentially equivalent to the consistency of probability distributions with rational values. This left open the question of exploring the interplay between (the standard notions of) local consistency and global consistency for bags, which is what we set to do in the present paper. Furthermore, as described earlier, here we also explore algorithmic aspects of global consistency, which were not addressed at all in [7].

## 2 PRELIMINARIES

An *attribute* $A$ is a symbol with an associated set $\mathrm{Dom}(A)$ called its *domain*. If $X$ is a finite set of attributes, then we write $\mathrm{Tup}(X)$ for the set of $X$-*tuples*; this means that $\mathrm{Tup}(X)$ is the set of functions that take each attribute $A \in X$ to an element of its domain $\mathrm{Dom}(A)$. Note that $\mathrm{Tup}(\emptyset)$ is non-empty as it contains the *empty tuple*, i.e., the unique function with empty domain. If $Y \subseteq X$ is a subset of attributes and $t$ is an $X$-tuple, then the *projection of* $t$ *on* $Y$, denoted by $t[Y]$, is the unique $Y$-tuple that agrees with $t$ on $Y$. In particular, $t[\emptyset]$ is the empty tuple.

Let $X$ be a set of attributes. We will view relations and bags over $X$ as functions from the set $\mathrm{Tup}(X)$ to, respectively, the Boolean semiring and the semiring of non-negative integers. The *Boolean semiring* $\mathbb{B} = (\{0, 1\}, \vee, \wedge, 0, 1)$ has disjunction $\vee$ and conjunction $\wedge$ as operations, and 0 (false) and 1 (true) as the identity elements of $\vee$ and $\wedge$. The semiring $\mathbb{Z}^{\geq 0} = (\{0, 1, 2, \ldots\}, +, \times, 0, 1)$ of non-negative

[1]This paper will appear in a forthcoming volume in honor of Samson Abramsky's contributions to logic.

integers has the standard arithmetic operations of addition + and multiplication ×, and 0 and 1 as the identity elements of + and ×.

A *relation* over $X$ is a function $R : \mathrm{Tup}(X) \to \{0, 1\}$, while a *bag* over $X$ is a function $R : \mathrm{Tup}(X) \to \{0, 1, 2, \ldots\}$. We write $R(X)$ to emphasize the fact that $R$ is a relation or a bag over *schema* $X$. If $R$ is a relation or a bag, then the *support* of $R$, denoted by $\mathrm{Supp}(R)$, is the set of $X$-tuples $t$ that are assigned non-zero value, i.e.,

$$\mathrm{Supp}(R) := \{t \in \mathrm{Tup}(X) : R(t) \neq 0\}. \tag{1}$$

Whenever no confusion arises, we write $R'$ to denote $\mathrm{Supp}(R)$. We say that $R$ is *finite* if its support $R'$ is a finite set. In what follows, we will make the blanket assumption that all relations and bags considered are finite, so we will omit the term "finite". Every relation $R$ can be identified with its support $R'$, thus every relation $R$ can be viewed as a finite set of $X$-tuples. If $R$ is a bag and $t$ is an $X$-tuple, then the non-negative integer $R(t)$ is called the *multiplicity* of $t$ in $R$; we will often write $t : R(t)$ to denote that the multiplicity of $t$ in $R$ is equal to $R(t)$. Therefore, relations are bags in which the multiplicity of each tuple is 0 or 1. Every bag $R$ can be viewed as a finite set of elements of the form $t : R(t)$, where $t \in R'$. Thus, if $X = \{A, B\}$, then $R(A, B) = \{(a_1, b_1) : 2, (a_2, b_2) : 1, (a_3, b_3) : 5\}$ represents the bag $R$ over $X$ such that $R(a_1, b_1) = 2$, $R(a_2, b_2) = 1$, $R(a_3, b_3) = 5$, and $R(a, b) = 0$, for all other pairs $(a, b)$. This bag can also be represented in tabular form as follows:

| $A$ | $B$ | # |
|---|---|---|
| $a_1$ | $b_1$ : | 2 |
| $a_2$ | $b_2$ : | 1 |
| $a_3$ | $b_3$ : | 5 |

If $R$ and $S$ are two bags over the schema $X$, then $R$ is *bag-contained* in $S$, denoted by $R \subseteq_b S$, if $R(t) \leq S(t)$ for every $X$-tuple $t$.

Let $R$ be a relation over $X$ and assume that $Z \subseteq X$. The *projection of $R$ on $Z$*, denoted by $R[Z]$, is the relation over $Z$ consisting of all projections $t[Y]$ as $t$ ranges over $R$.

Let $R$ be a bag over $X$ and assume that $Z \subseteq X$. If $t$ is a $Z$-tuple, then the *marginal of $R$ over $t$* is defined by

$$R(t) := \sum_{\substack{r \in R': \\ r[Z]=t}} R(r). \tag{2}$$

Thus, every bag $R$ over $X$ induces a bag over $Z$, which is called the *marginal of $R$ on $Z$* and is denoted by $R[Z]$. Note that the preceding equation defines also the projection of a relation, provided the sum is interpreted as the disjunction $\vee$ over the Boolean semiring. It is easy to verify that the following facts hold for every bag $R$ over $X$.

- For all $Z \subseteq X$, we have $R'[Z] = R[Z]'$.
- For all $W \subseteq Z \subseteq X$, we have $R[Z][W] = R[W]$.

If $X$ and $Y$ are sets of attributes, then we write $XY$ as shorthand for the union $X \cup Y$. Accordingly, if $x$ is an $X$-tuple and $y$ is a $Y$-tuple with the property that $x[X \cap Y] = y[X \cap Y]$, then we write $xy$ to denote the $XY$-tuple that agrees with $x$ on $X$ and on $y$ on $Y$. We say that *$x$ joins with $y$*, and that *$y$ joins with $x$*, to *produce* the tuple $xy$.

If $R$ is a relation over $X$ and $S$ is a relation over $Y$, then their *join* $R \bowtie S$ is the relation over $XY$ consisting of all tuples $XY$-tuples $t$ such that $t[X]$ is in $R$ and $t[Y]$ is in $S$, i.e., all tuples of the form $xy$ such that $x \in R'$, $y \in R'$, and $x$ joins with $y$. If $R$ is a bag over $X$ and $S$ is a bag over $Y$, then their *bag join* $R \bowtie_b S$ is the bag over $XY$ with support $R' \bowtie S'$ and such that every $XY$-tuple $t \in R' \bowtie S'$ has multiplicity $(R \bowtie_b S)(t) = R(t[X]) \times S(t[Y])$.

## 3 CONSISTENCY OF TWO BAGS

We say that two relations $R(X)$ and $S(Y)$ are *consistent* if there exists a relation $T(XY)$ with $T[X] = R$ and $T[Y] = S$. Similarly, we say that two bags $R(X)$ and $S(Y)$ are *consistent* if there exists a bag $T(XY)$ with $T[X] = R$ and $T[Y] = S$, where now the projections are computed according to Equation (2). In such a case, we say that $T$ *witnesses* the consistency of $R$ and $S$. A simple calculation shows that if $R(X)$ and $S(Y)$ are consistent bags and $T$ is a bag that witnesses their consistency, then the support $T'$ of $T$ is a subset of the join $R' \bowtie S'$ of the supports.

Lemma 1. *If $R(X)$ and $S(Y)$ are consistent bags and $T(XY)$ is a bag that witnesses their consistency, then $T' \subseteq R' \bowtie S'$.*

Proof. If $t \in T'$, then $T(t) \geq 1$, so $R(t[X]) \geq 1$ by $R = T[X]$, and $S(t[Y]) \geq 1$ by $S = T[Y]$. Hence $t[X] \in R'$ and $t[Y] \in S'$, so $t \in R' \bowtie S'$. □

If two relations $R(X)$ and $S(Y)$ are consistent, then their join $R \bowtie S$ witnesses their consistency; in fact, $R \bowtie S$ is the largest relation that has this property. In contrast, there are consistent bags $R(X)$ and $S(Y)$ such that the support $T'$ of every bag $T$ witnessing their consistency is a proper subset of $R' \bowtie S'$. An example of this is provided by the bags $R_1(AB) = \{(1, 2) : 1,\ (2, 2) : 1\}$ and $S_1(BC) = \{(2, 1) : 1,\ (2, 2) : 1\}$; their consistency (as bags) is witnessed by the bags $T_1(ABC) = \{(1, 2, 2) : 1,\ (2, 2, 1) : 1\}$ and $T_2(ABC) = \{(1, 2, 1) : 1,\ (2, 2, 2) : 1\}$, but, as one can easily verify, no other bag. This example can be extended as follows. For $n \geq 2$, let $R_{n-1}(A, B)$ and $S_{n-1}(B, C)$ be the bags

$$\{(1, 2) : 1,\ (2, 2) : 1,\ (1, 3) : 1,\ (3, 3) : 1,\ \ldots,\ (1, n) : 1, (n, n) : 1\}$$
$$\{(2, 1) : 1,\ (2, 2) : 1,\ (3, 1) : 1,\ (3, 3) : 1,\ \ldots,\ (n, 1) : 1, (n, n) : 1\},$$

respectively. For every $n \geq 2$, the bags $R_{n-1}$ and $S_{n-1}$ are consistent and there are exactly $2^{n-1}$ bags witnessing their consistency. Furthermore, these witnesses are pairwise incomparable in the bag-containment sense and their supports are properly contained in the support $(R_{n-1} \bowtie_b S_{n-1})'$ of the bag join $R_{n-1} \bowtie_b S_{n-1}$. Note that the bags $R_{n-1}$ and $S_{n-1}$ are actually relations and that their join $R_{n-1} \bowtie S_{n-1}$ witnesses their consistency as relations, but not as bags (where Equation (2) is used to compute the marginals).

With each pair of bags $R(X)$ and $S(Y)$, we associate the following linear program $P(R, S)$. Let $J = R' \bowtie S'$ be the join of the supports of $R$ and $S$. For each $t \in J$, there is a variable $x_t$. For each $t \in J$ and $r \in R'$, define $a_{r,t} = 1$ if $t[X] = r$ and $a_{r,t} = 0$ if $t[X] \neq r$. Similarly, for each $t \in J$ and $s \in S'$, define $a_{s,t} = 1$ if $t[Y] = s$ and $a_{s,t} = 0$ if $t[Y] \neq s$. The constraints of $P(R, S)$ are:

$$\begin{array}{ll} \sum_{t \in J} a_{r,t} x_t = R(r) & \text{for } r \in R', \\ \sum_{t \in J} a_{s,t} x_t = S(s) & \text{for } s \in S', \\ x_t \geq 0 & \text{for } t \in J. \end{array} \tag{3}$$

If we write the equations of $P(R, S)$ in matrix form as $Ax = b$, then the matrix $A$ has special structure: its set of rows is partitioned into two sets in such a way that every column has at most one 1 entry in each part, and the rest of entries of the column are 0. This means that $A$ is the vertex-edge incidence matrix of a bipartite

graph, so by Example 1 in Section 19.3 of Schrijver's book [22], the matrix $A$ is totally unimodular. By the Hoffman-Kruskal Theorem (Corollary 19.2a in [22]), the polytope defined by $P(R,S)$ is either empty or has integral vertices. Consequently, $P(R,S)$ is feasible over the rationals if and only if $P(R,S)$ is feasible over the integers. As we will soon see, a different proof of this fact can be obtained using the integrality theorem for max flow; for this, we will view $P(R,S)$ as the set of *flow constraints* of an instance of the max-flow problem, as we discuss next.

A network $N = (V, E, c, s, t)$ is a directed graph $G = (V, E)$ with a non-negative weight $c(u,v)$, called the *capacity*, assigned to each edge $(u,v) \in E$, and two distinguished vertices $s, t \in V$, called the *source* and the *sink*. A *flow* for the network is an assignment of non-negative weights $f(u,v)$ on the edges $(u,v) \in E$ in such a way that both the capacity constraints and the flow constraints are respected, i.e., $f(u,v) \leq c(u,v)$ for each $(u,v) \in E$, and $\sum_{v \in N^-(u)} f(v,u) = \sum_{w \in N^+(u)} f(u,w)$ for each $u \in V \setminus \{s,t\}$, where $N^-(u)$ and $N^+(u)$ denote the sets of in-neighbors and out-neighbors of $u$ in $G$. The *value* of such a flow is the quantity $\sum_{w \in N^+(s)} f(s,w) = \sum_{v \in N^-(t)} f(v,t)$, where the equality follows from the flow constraints. In the *max-flow problem*, the goal is to find a max flow, that is, a flow of maximum value. We say that a flow is *saturated* if $f(s,w) = c(s,w)$ for every $w \in N^+(s)$ and $f(v,t) = c(v,t)$ for every $v \in N^-(t)$. It is obvious that if a saturated flow exists, then every max flow is saturated.

With each pair $R(X)$ and $S(Y)$ of bags, we associate the following network $N(R,S)$. The network has $1 + |R'| + |S'| + 1$ vertices: one source vertex $s^*$, one vertex for each tuple $r$ in the support $R'$ of $R$, one vertex for each tuple $s$ in the support $S'$ of $S$, and one target vertex $t^*$. There is an arc of capacity $R(r)$ from $s^*$ to $r$ for each $r \in R'$, an arc of capacity $S(s)$ from $s$ to $t^*$ for each $s \in S'$, and an arc of unbounded (i.e., very large) capacity from $t[X]$ to $t[Y]$ for each $t \in R' \bowtie S'$.

The next result yields several different characterization of the consistency of two bags.

Lemma 2. *Let $R(X)$ and $S(Y)$ be two bags. The following statements are equivalent:*

(1) *$R(X)$ and $S(Y)$ are consistent.*
(2) *$R[X \cap Y] = S[X \cap Y]$.*
(3) *$P(R,S)$ is feasible over the rationals.*
(4) *$P(R,S)$ is feasible over the integers.*
(5) *$N(R,S)$ admits a saturated flow.*

Proof. Let $Z = X \cap Y$. For (1) implies (2), assume that $T$ witnesses the consistency of $R$ and $S$. Then $T[X] = R$ and $T[Y] = S$ and hence $R[Z] = T[X][Z] = T[Z] = T[Y][Z] = S[Z]$. For (2) implies (3), assume that $R[Z] = S[Z]$. We show that $P(R,S)$ is feasible over the rationals. Let $J = R' \bowtie S'$ and for each $t \in J$ set $x_t := R(t[X])S(t[Y])/R(t[Z]) = R(t[X])S(t[Y])/S(t[Z])$, where the equality follows from the assumption that $R[Z] = S[Z]$. For each fixed $r \in R'$, let $u = r[Z]$ and note that $\sum_{t \in J} a_{r,t} x_t = (R(r)/S(u)) \sum_{t \in J : t[X] = r} S(t[Y]) = (R(r)/S(u)) \sum_{s \in S' : s[Z] = u} S(s) = R(r)$. For each fixed $s \in R'$, let $u = s[Z]$ and note that $\sum_{t \in J} a_{s,t} x_t = (S(s)/R(u)) \sum_{t \in J : t[Y] = s} R(t[X]) = (S(s)/R(u)) \sum_{r \in R' : r[Z] = u} R(s) = S(s)$. Therefore, since $x_t \geq 0$, we have shown that $P(R,S)$ is feasible over the rationals. For (3) implies (5), let $x^* = (x^*_t)_{t \in J}$ be a rational solution for $P(R,S)$ and let $f$ be the following assignment for $N(R,S)$:

$$\begin{array}{ll} f(s^*, r) := c(s^*, r) = R(r) & \text{for each } r \in R'; \\ f(t[X], t[Y]) := x^*_t & \text{for each } t \in J; \\ f(s, t^*) := c(s, t^*) = S(s) & \text{for each } s \in S'. \end{array}$$

This assignment is a flow since the equations of $P(R,S)$ say that the flow-constraints are satisfied; furthermore, it is a saturated flow by construction. For (5) implies (1), let $g$ be a saturated flow for $N(R,S)$; in particular, this is a max flow for $N(R,S)$. Since all capacities in $N(R,S)$ are integers, the integrality theorem for the max-flow problem asserts that there is a max flow $f$ consisting of integers (see, e.g., [27]), which, of course, is also a saturated flow. Let $T(XY)$ be the bag defined by setting $T(t) := f(t[X], t[Y])$ for each $t \in R' \bowtie S'$. Since $f$ is saturated, we have that $f(s^*, r) = c(s^*, r) = R(r)$ for each $r \in R'$ and $f(s, t^*) = c(s, t^*) = S(s)$ for each $s \in S'$. This means that the flow-constraints imply that $T$ witnesses the consistency of $R$ and $S$. Thus, we have established that statements (1), (2), (3), and (5) are equivalent. The equivalence of statements (1) and (4) is immediate from the definitions. □

The equivalence of statements (1) and (2) in Lemma 2 yields a simple polynomial-time test to determine the consistency of two bags, namely, given two bags $R(X)$ and $S(y)$, check whether or not $R[X \cap Y] = S[X \cap Y]$. Furthermore, the equivalence of statements (1) and (5) implies that there is a polynomial-time algorithm for constructing a witness to the consistency of two consistent bags. This is so, because it is well known that there are polynomial-time algorithms for the max-flow problem. As a matter of fact, there are strongly polynomial algorithms for this problem, such as Orlin's algorithm [20], which finds a maximum flow in time $O(|V||E|)$. Thus, we have the following result.

Corollary 1. *There is a strongly polynomial-time algorithm that, given two bags, determines whether the bags are consistent and, if they are, constructs a bag witnessing their consistency.*

We note that it is not known whether a strongly polynomial algorithm for linear programming exists. However, any algorithm for solving linear programming in time polynomial in the bit-complexity of its data could be used to find a witness to the consistency of two consistent bags. Simultaneously, the algorithm could be asked to minimize any given linear function of the multiplicities of the witnessing bag. Furthermore, it would accomplish these tasks in time polynomial in the bit-complexity representation of the input bags and the objective function. This follows from Lemma 2 combined with the fact that, by the Hoffman-Kruskal Theorem, all vertices of the polytope defined by $P(R,S)$ are integral.

## 4 CONSISTENCY OF THREE OR MORE BAGS

Let $R_1(X_1), \ldots, R_m(X_m)$ be bags over the schemas $X_1, \ldots, X_m$. We say that the collection $R_1, \ldots, R_m$ is *globally consistent* if there is a bag $T$ over $X_1 \cup \cdots \cup X_m$ such that $R_i = T[X_i]$ for all $i \in [m]$. We say that such a bag *witnesses* the global consistency of $R_1, \ldots, R_m$. We also say that the bags $R_1, \ldots, R_m$ are *pairwise consistent* if for every $i, j \in [m]$ we have that $R_i(X_i)$ and $R_j(X_j)$ are consistent.

Corresponding notions of global consistency and pairwise consistency can be defined for relations, the only difference being that

the $T[X_i]$'s and the $R_i[X_i]$'s are projections of relations, instead of marginals of bags. The following facts are well known (see, e.g., [15]):

- If $T$ is a relation witnessing the global consistency of the relations $R_1, \ldots, R_m$, then $T \subseteq R_1 \bowtie \cdots \bowtie R_m$.
- A collection $R_1, \ldots, R_m$ of relations is globally consistent if and only if $(R_1 \bowtie \cdots \bowtie R_m)[X_i] = R_i$ for all $i = 1, \ldots, m$.

Consequently, if the collection $R_1, \ldots, R_m$ is globally consistent, then the join $R_1 \bowtie \cdots \bowtie R_m$ is the *largest* relation witnessing their consistency. As seen in Section 2, there are bags that are consistent, but their consistency is not witnessed by their bag-join.

From the definitions, it follows that if $R_1, \ldots, R_m$ are globally consistent bags, then they are also pairwise consistent. The converse, however, need not be true, in general. In fact, the converse fails even for relations. For example, the relations $R(AB) = \{00, 11\}$, $S(BC) = \{01, 10\}$, $T(AC) = \{00, 11\}$ are pairwise consistent but not globally consistent. The interplay between pairwise consistency and global consistency of relations has been extensively studied in database theory. We summarize some of the main findings next.

Pairwise consistency is a necessary, but not sufficient, condition for global consistency of relations. Beeri, Fagin, Maier, and Yannakakis [8] characterized the set of schemas for which pairwise consistency is a necessary and sufficient condition for global consistency of relations. Their characterization involves notions from hypergraph theory that we now review.

*Acyclic Hypergraphs.* A *hypergraph* is a pair $H = (V, E)$, where $V$ is a set of *vertices* and $E$ is a set of *hyperedges*, each of which is a nonempty subset of $V$. Every collection $X_1, \ldots, X_m$ of sets of attributes can be identified with a hypergraph $H = (V, E)$, where $V = X_1 \cup \cdots \cup X_m$ and $E = \{X_1, \ldots, X_m\}$. Conversely, every hypergraph $H = (V, E)$ gives rise to a collection $X_1, \ldots, X_m$ of sets of attributes, where $X_1, \ldots, X_m$ are the hyperedges of $H$. Thus, we can move seamlessly from collections of sets of attributes to hypergraphs, and vice versa. The notion of an *acyclic* hypergraph generalizes the notion of an acyclic graph. Since we will not work directly with the definition of an acyclic hypergraph, we refer the reader to [8] for the precise definition. Instead, we focus on other notions that are equivalent to hypergraph acyclicity and will be of interest to us in the sequel.

*Conformal and Chordal Hypergraphs.* The *primal* graph of a hypergraph $H = (V, E)$ is the undirected graph that has $V$ as its set of vertices and has an edge between any two distinct vertices that appear together in at least one hyperedge of $H$. A hypergraph $H$ is *conformal* if the set of vertices of every clique (i.e., complete subgraph) of the primal graph of $H$ is contained in some hyperedge of $H$. A hypergraph $H$ is *chordal* if its primal graph is chordal, that is, if every cycle of length at least four of the primal graph of $H$ has a chord. To illustrate these concepts, let $V_n = \{A_1, \ldots, A_n\}$ be a set of $n$ vertices and consider the hypergraphs

$$P_n = (V_n, \{A_1, A_2\}, \ldots, \{A_{n-1}, A_n\}) \quad (4)$$

$$C_n = (V_n, \{A_1, A_2\}, \ldots, \{A_{n-1}, A_n\}, \{A_n, A_1\}) \quad (5)$$

$$H_n = (V_n, \{V_n \setminus \{A_i\} : 1 \leq i \leq n\}) \quad (6)$$

If $n \geq 2$, then the hypergraph $P_n$ is both conformal and chordal. The hypergraph $C_3 = H_3$ is chordal, but not conformal. For every $n \geq 4$, the hypergraph $C_n$ is conformal, but not chordal, while the hypergraph $H_n$ is chordal, but not conformal.

*Running Intersection Property.* We say that a hypergraph $H$ has the *running intersection property* if there is a listing $X_1, \ldots, X_m$ of all hyperedges of $H$ such that for every $i \in [m]$ with $i \geq 2$, there exists a $j < i$ such that $X_i \cap (X_1 \cup \cdots \cup X_{i-1}) \subseteq X_j$.

*Join Tree.* A *join tree* for a hypergraph $H$ is an undirected tree $T$ with the set $E$ of the hyperedges of $H$ as its vertices and such that for every vertex $v$ of $H$, the set of vertices of $T$ containing $v$ forms a subtree of $T$, i.e., if $v$ belongs to two vertices $X_i$ and $X_j$ of $T$, then $v$ belongs to every vertex of $T$ in the unique simple path from $X_i$ to $X_j$ in $T$.

*Local-to-Global Consistency Property for Relations.* Let $H$ be a hypergraph and let $X_1, \ldots, X_m$ be a listing of all hyperedges of $H$. We say that $H$ has the *local-to-global consistency property for relations* if every pairwise consistent collection $R_1(X_1), \ldots, R_m(X_m)$ of relations of schema $X_1, \ldots, X_m$ is globally consistent.

We are now ready to state the main result in Beeri et al. [8].

Theorem 1 (Theorem 3.4 in [8]). *Let $H$ be a hypergraph. The following statements are equivalent:*

(a) *$H$ is an acyclic hypergraph.*
(b) *$H$ is a conformal and chordal hypergraph.*
(c) *$H$ has the running intersection property.*
(d) *$H$ has a join tree.*
(e) *$H$ has the local-to-global consistency property for relations.*

As an illustration, if $n \geq 2$, the hypergraph $P_n$ is acyclic, hence it has the local-to-global consistency property for relations. In contrast, if $n \geq 3$, the hypergraphs $C_n$ and $H_n$ are cyclic, hence they do not have the local-to-global consistency property for relations.

In what follows, we will show that the preceding Theorem 1 also holds for bags. We need the following definition. Let $H$ be a hypergraph and let $X_1, \ldots, X_m$ be a listing of all hyperedges of $H$. We say that $H$ has the *local-to-global consistency property for bags* if every pairwise consistent collection $R_1(X_1), \ldots, R_m(X_m)$ of bags of schema $X_1, \ldots, X_m$ is globally consistent.

Theorem 2. *Let $H$ be a hypergraph. The following statements are equivalent:*

(a) *$H$ is an acyclic hypergraph.*
(b) *$H$ is a conformal and chordal hypergraph.*
(c) *$H$ has the running intersection property.*
(d) *$H$ has a join tree.*
(e) *$H$ has the local-to-global consistency property for bags.*

As an immediate consequence of Theorem 1 and Theorem 2, we obtain the following result.

Corollary 2. *Let $H$ be a hypergraph. The following statements are equivalent:*

(a) *$H$ has the local-to-global consistency property for relations.*
(b) *$H$ has the local-to-global consistency property for bags.*

Before embarking on the proof of Theorem 2, we need some additional notions about hypergraphs and two technical lemmas. We begin with the definitions of the notions needed.

Let $H = (V, E)$ be a hypergraph. The *reduction* of $H$ is the hypergraph $R(H)$ whose set of vertices is $V$ and whose hyperedges are those hyperedges $X \in E$ that are not included in any other hyperedge of $H$. A hypergraph $H$ is *reduced* if $H = R(H)$. If $W \subseteq V$, then the *hypergraph induced by* $W$ *on* $H$ is the hypergraph $H[W]$ whose set of vertices is $W$ and whose hyperedges are the non-empty subsets of the form $X \cap W$, where $X \in E$ is a hyperedge of $H$; in symbols, $H[W] = (W, \{X \cap W : X \in E\} \setminus \{\emptyset\})$.

Let $H = (V, E)$ be a hypergraph. For a vertex $u \in V$, we write $H \setminus u$ for the hypergraph induced by $V \setminus \{u\}$ on $H$. For an edge $e \in E$, we write $H \setminus e$ for the hypergraph with $V$ as the set of its vertices and with $E \setminus \{e\}$ as the set of its edges. Let $H' = (V', E')$ be another hypergraph. We say that $H'$ is obtained from $H$ by a *vertex-deletion* if $H' = H \setminus u$ for some $u \in V$. We say that $H'$ is obtained from $H$ by a *covered-edge-deletion* if $H' = H \setminus e$ for some $e \in E$ such that $e \subseteq f$ for some $f \in E \setminus \{e\}$. In either case, we say that $H'$ is obtained from $H$ by a *safe-deletion operation*. We say that a sequence of safe-deletion operations *transforms* $H$ *to* $H'$ if $H'$ can be obtained from $H$ by starting with $H$ and applying the operations in order.

Note that if $H = (V, E)$ is a hypergraph and $W$ is a subset of $V$, then the hypergraph $R(H[W])$ is obtained from $H$ by a sequence of safe-deletion operations. Indeed, we can first obtain the hypergraph $H[W]$ from $H$ by a sequence of vertex-deletions in which the vertices of the set of $V \setminus W$ are removed one-by-one; after this, we can obtain the hypergraph $R(H[W])$ from $H[W]$ by a sequence of covered-edge deletions.

Lemma 3. *For every hypergraph $H = (V, E)$ the following statements hold:*

(1) *$H$ is not chordal if and only if there exists $W \subseteq V$ with $|W| \geq 4$ and such that $R(H[W]) \cong C_n$, where $n = |W|$.*
(2) *$H$ is not conformal if and only if there exists $W \subseteq V$ with $|W| \geq 3$ and such that $R(H[W]) \cong H_n$, where $n = |W|$.*

*Moreover, there exist a polynomial-time algorithm that, given a hypergraph $H$ that is not chordal or not conformal, finds both a set $W$ as stated in (1) or (2) and a sequence of safe-deletion operations that transforms $H$ to $R(H[W])$.*

Proof. The proof of (1) is straightforward. For the proof of (2) see [11]. Since there exist polynomial-time algorithms that test whether a graph is chordal (see, e.g., [21]), an algorithm to find a $W$ as stated in (1), when $H$ is not chordal, is to iteratively delete vertices whose removal leaves a hypergraph with a non-chordal primal graph until no more vertices can be removed. Also, since there exist polynomial-time algorithms that test whether a hypergraph is conformal (see, e.g., Gilmore's Theorem in page 31 of [10]), an algorithm to find a $W$ stated in (2), when $H$ is not conformal, is to iteratively delete vertices whose removal leaves a non-conformal hypergraph until no more vertices can be removed. In both cases, once the set $W$ is found, a sequence of safe-deletion operations that transforms $H$ to $R(H[W])$ if obtained by first deleting all vertices in $V \setminus W$, and then deleting all covered edges. □

Let $A_1, \ldots, A_n$ be attributes and let $H = (V, E)$ be a hypergraph with vertices $V = \{A_1, \ldots, A_n\}$ and edges $E = \{X_1, \ldots, X_m\}$. A *collection of bags over* $H$ is a collection $D$ of bags $R_1(X_1), \ldots, R_m(X_m)$ for some $m \geq 1$, i.e., each $R_i$ is a bag over the schema $X_i$. For an integer $k \in [m]$, we say that a collection $D$ of bags over $H$ is *$k$-wise consistent* if for every $I \subseteq [m]$ with $|I| \leq k$, the collection $\{R_i : i \in I\}$ is globally consistent. Observe that $D$ is pairwise consistent if and only if it is 2-wise consistent. Furthermore, it is easy to see that $D$ is globally consistent if and only if it is $m$-wise consistent; this uses the fact that for every bag $T$ over a set $X$ and for all $W \subseteq Z \subseteq X$, we have that $T[Z][W] = T[W]$.

Lemma 4. *Let $H_0$ and $H_1$ be hypergraphs such that $H_0$ is obtained from $H_1$ by a sequence of safe-deletion operations, and let $m$ be the number of hyperedges of $H_0$. For every collection $D_0$ of bags over $H_0$, there exists a collection $D_1$ of bags over $H_1$ such that, for every integer $k \in [m]$, it holds that $D_0$ is $k$-wise consistent if and only if $D_1$ is $k$-wise consistent. Moreover, there is a polynomial-time algorithm that, given $D_0$ and a sequence of safe-deletion operations that transforms $H_1$ to $H_0$, computes $D_1$.*

Proof. We first define the collection $D_1$ in the case in which $H_0$ is obtained from $H_1$ by a single safe-deletion operation. In the case of a sequence of safe-deletion operations, the collection $D_1$ in the statement of the lemma will be the result of iterating the construction in the first case $t$ many times, where $t$ is the number of operations that transforms $H_1$ to $H_0$. After the construction is spelled out, we analyse the run-time of the underlying algorithm and then prove its main property. In what follows, suppose that $H_1 = (V_1, E_1)$, where $V_1 = \{A_1, \ldots, A_n\}$ and $E_1 = \{X_1, \ldots, X_n\}$.

Assume first that $H_0 = H_1 \setminus X$ where $X \in E_1$ is such that $X \subseteq X_j$ for some $j \in [m]$ with $X \neq X_j$; i.e., $H_0$ is obtained from $H_1$ by deleting a covered edge. In particular, $V_0 = V_1$ and $E_0 = E_1 \setminus \{X\}$. If the bags of $D_0$ are $S_i(X_i)$ for $i \in [m]$ with $X_i \neq X$, then $D_1$ is defined as the collection with bags $R_i(X_i)$ for $i \in [m]$ defined as follows: For each $i \in [m]$, if $X_i \neq X$, then $R_i := S_i$; else let $R_i := S_j[X]$.

Assume next that $H_0 = H_1 \setminus A$ where $A \in V_1$; i.e., $H_0$ is obtained from $H_1$ by deleting a vertex. In particular, $V_0 = V_1 \setminus \{A\}$ and $E_0 = \{Y_1, \ldots, Y_m\}$ where $Y_i = X_i \setminus \{A\}$ for $i = 1, \ldots, m$. Fix a default value $u_0$ in the domain $\mathrm{Dom}(A)$ of the attribute $A$. If the bags of $D_0$ are $S_i(Y_i)$ for $i \in [m]$, then $D_1$ is defined as the collection with bags $R_i(X_i)$ for $i \in [m]$ defined as follows: For each $i \in [m]$, if $A \notin X_i$, let $R_i := S_i$; else let $R_i$ be the bag of schema $X_i = Y_i \cup \{A\}$ defined for every $X_i$-tuple $t$ by $R_i(t) := 0$ if $t(A) \neq u_0$ and $R_i(t) := S_i(t[Y_i])$ if $t(A) = u_0$. We note that in case $X_i = \{A\}$, the bag $R_i$ has empty schema $Y_i = \emptyset$ and consists of the empty tuple with multiplicity $S_i(u_0)$.

It follows from the definitions that, in both cases, each bag $R$ of $D_1$ has its multiset cardinality bounded by $S(\emptyset)$ for some bag $S$ of $D_0$. In the case $H_0 = H_1 \setminus X$, this follows from the fact that each bag of $D_1$ is either a bag of $D_0$ or the marginal of a bag of $D_0$. In the case $H_0 = H_1 \setminus A$, this follows from the fact that each bag of $D_1$ is either a bag of $D_0$ or a bag with the same multiset cardinality as a bag of $D_0$. It follows by induction that if $H_0$ is obtained from $H_1$ by a sequence of $t$ many safe-deletion operations, then the collection $D_1$ of bags that results by applying the construction $t$ many times starting at $D_0$ has each bag $R$ of multiset cardinality bounded by $S(\emptyset)$ for some bag $S$ of $D_0$. Thus, $D_1$ has size at most $t$ times the size of $D_0$ and can be constructed in time polynomial in the size of $D_0$ and the length $t$ of the sequence.

We prove the main property by cases. Fix an integer $k \geq 1$.

Claim 1. *Assume $H_0 = H_1 \setminus A$ for some vertex $A \in V_1$. Then, the bags $S_i(Y_i)$ of $D_0$ are $k$-wise consistent if and only if the bags $R_i(X_i)$ of $D_1$ are $k$-wise consistent.*

Proof. Fix $I \subseteq [m]$ with $|I| \leq k$, let $X = \bigcup_{i \in I} X_i$ and $Y = \bigcup_{i \in I} Y_i$. Observe that $Y = X \setminus \{A\}$. In particular $Y = X$ if $A$ is not in $X$.

(If): Let $R$ be a bag over $X$ that witnesses the consistency of $\{R_i : i \in I\}$, and let $S := R[Y]$. We claim that $S$ witnesses the consistency of $\{S_i : i \in I\}$. Indeed, $S[Y_i] = R[Y][Y_i] = R[Y_i] = R_i[Y_i] = S_i$, where the first equality follows from the choice of $S$, the second equality follows from $Y_i \subseteq Y$, the third equality follows from the facts that $R[X_i] = R_i$ and $Y_i \subseteq X_i$, and the fourth equality follows from the definition of $R_i$.

(Only if): Consider the two cases: $A \notin X$ or $A \in X$. If $A \notin X$, then $R_i = S_i$ for every $i \in I$ and therefore the bags $\{R_i : i \in I\}$ are consistent because the bags $\{S_i : i \in I\}$ are consistent. If $A \in X$, then let $S$ be a bag over $Y$ that witnesses the consistency of the bags $\{S_i : i \in I\}$, and let $R$ be the bag over $X$ defined for every $X$-tuple $t$ by $R(t) := 0$ if $t(A) \neq u_0$ and by $R(t) := S(t[Y])$ if $t(A) = u_0$. We claim that $R$ witnesses the consistency of the bags $R_i$ for $i \in I$. We show that $R_i = R[X_i]$ for $i \in I$. Towards this, first we argue that $S[Y_i] = R[Y_i]$. Indeed, for every $Y_i$-tuple $r$ we have

$$S(r) = \sum_{\substack{s \in S': \\ s[Y_i] = r}} S(s) = \sum_{\substack{t \in \mathrm{Tup}(X): \\ t[Y_i] = r, \\ t(A) = u_0}} S(t[Y]) = \sum_{\substack{t \in S': \\ t[Y_i] = r}} R(t) = R(r), \quad (7)$$

where the first equality follows from (2), the second equality follows from the fact that the map $t \mapsto t[Y]$ is a bijection between the set of $X$-tuples $t$ such that $t[Y_i] = r$ and $t(A) = u_0$ and the set of $Y$-tuples $s$ such that $s[Y_i] = r$, the third equality follows from the definition of $R$, and the fourth equality follows from (2).

In case $A \notin X_i$, we have that $Y_i = X_i$, hence Equation (7) already shows that $R_i = S_i = S[Y_i] = R[Y_i] = R[X_i]$. In case $A \in X_i$, we use the fact that $S_i = S[Y_i]$ to show that $R_i = R[X_i]$. For every $X_i$-tuple $r$ with $r(A) \neq u_0$, we have $R_i(r) = 0$ and also $R(r) = \sum_{t : t[X_i] = r} R(t) = 0$ since $t[X_i] = r$ and $A \in X_i$ implies $t(A) = r(A) \neq u_0$. Thus, $R_i(r) = 0 = R(r)$ in this case. For every $X_i$-tuple $r$ with $r(A) = u_0$, we have

$$R_i(r) = S_i(r[Y_i]) = S(r[Y_i]) = R(r[Y_i]), \quad (8)$$

where the first equality follows from the definition of $R_i$ and the assumption that $r(A) = u_0$, the second equality follows from $S_i = S[Y_i]$, and the third equality follows from (7). Continuing from the right-hand side of (8), we have

$$R(r[Y_i]) = \sum_{\substack{t \in R': \\ t[Y_i] = r[Y_i]}} R(t) = \sum_{\substack{t \in R': \\ t[X_i] = r}} R(t) = R(r), \quad (9)$$

where the first equality follows from (2), the second equality follows from the assumption that $A \in X_i$ and $r(A) = u_0$ together with $R(t) = 0$ in case $t(A) \neq u_0$, and the third equality follows from (2). Combining (8) with (9), we get $R_i(r) = R(r)$ also in this case. This proves that $R_i = R[X_i]$. □

Claim 2. *Assume $H_0 = H_1 \setminus X$ for some edge $X \in E_1$ that is covered in $H_1$. Then, the bags $S_i(X_i)$ of $D_0$ are $k$-wise consistent if and only if the bags $R_i(Y_i)$ of $D_1$ are $k$-wise consistent.*

Proof. Let $l \in [m]$ be such that $X = X_l \subseteq X_j$ for some $j \in [m] \setminus \{l\}$, so $E_0 = \{X_i : i \in [m] \setminus \{l\}\}$.

(If): Fix $I \subseteq [m] \setminus \{l\}$ with $|I| \leq k$ and let $X = \bigcup_{i \in I} X_i$. Let $R$ be a bag over $X$ that witnesses the consistency of $\{R_i : i \in I\}$ and let $S = R$. Since $S_i = R_i$ for every $i \in [m] \setminus \{l\}$, it is obvious that $S$ witnesses the consistency of $\{S_i : i \in I\}$.

(Only if): Fix $I \subseteq [m]$ with $|I| \leq k$ and let $X = \bigcup_{i \in I} X_i$. Let $S$ be a bag over $X$ that witnesses the consistency of $\{S_i : i \in I \setminus \{l\}\}$ and let $R = S$. We have $R_l = S_j[X_l] = S[X_j][X_l] = R[X_j][X_l] = R[X_l]$ where the first equality follows from the definition of $R_l$, the second equality follows from the fact that $S_j = S[X_j]$, the third equality follows from the choice of $R$, and the fourth equality follows from $X_l \subseteq X_j$. □

The proof of Lemma 4 is now complete. □

Lemma 4 implies that the local-to-global consistency property for bags is preserved under induced hypergraphs and under reductions.

Corollary 3. *If a hypergraph $H$ has the local-to-global consistency property for bags, then for every subset $W$ of the set of vertices of $H$, the hypergraph $R(H[W])$ also has the local-to-global consistency property for bags.*

Proof. As discussed earlier, the hypergraph $R(H[W])$ is obtained from the hypergraph $H$ by a sequence of safe-deletion operators. We will apply Lemma 4 with $H_0 = R(H[W])$ and $H_1 = H$. Let $m$ be the number of hyperedges of $R(H[W])$ and let $m'$ be the number of hyperedges of $H$; clearly, we have that $m \leq m'$. Let $R_1, \ldots, R_m$ be a collection of bags over $R(H[W])$ that are pairwise consistent. We have to show that this collection is globally consistent. By Lemma 4, there is a collection of bags $S_1, \ldots, S_{m'}$ over $H$ that are pairwise consistent. Since $H$ has the local-to-global consistency property for bags, it follows that the collection $S_1, \ldots, S_{m'}$ is globally consistent, i.e., it $m'$-wise consistent. Since $m \leq m'$, we have that the collection of bags $S_1, \ldots, S_{m'}$ is also $m$-wise consistent. By Lemma 4 (but in the reverse direction this time), we have that the collection of bags $R_1, \ldots, R_m$ is $m$-wise consistent, which means that it is globally consistent, as it was to be shown. □

We are now ready to give the proof of Theorem 2.

Proof of Theorem 2. Let $H$ be a hypergraph. By Theorem 1, statements (a), (b), (c), and (d) are equivalent, because these statements express "structural" properties of hypergraphs, i.e., their definitions involve only the vertices and the hyperedges of the hypergraph at hand. So, we only have to show that statement (e), which involves "semantics" notions about bags, is equivalent to (one of) the other three statements. This will be achieved in two steps. First, we will show that if $H$ has the running intersection property, then $H$ has the local-to-global consistency property for bags. Second, we will show that if $H$ is not conformal or $H$ is not chordal, then $H$ does not have the local-to-global consistency property for bags.

*Step 1.* Assume that the hypergraph $H$ has the running intersection property. Hence, there is a listing $X_1, \ldots, X_m$ of its hyperedges such that for every $i \in [m]$ with $i \geq 2$, there is a $j \in [i-1]$ such that $X_i \cap (X_1 \cup \cdots \cup X_{i-1}) \subseteq X_j$. Let $R_1(X_1), \ldots, R_m(X_m)$ be a collection of pairwise consistent bags over the schemas $X_1, \ldots, X_m$.

By induction on $i = 1, \ldots, m$, we show that there is a bag $T_i$ over $X_1 \cup \cdots \cup X_i$ that witnesses the global consistency of the bags $R_1, \ldots, R_i$. For $i = 1$ the claim is obvious since $T_1 = R_1$. Assume then that $i \geq 2$ and that the claim is true for all smaller indices. Let $X := X_1 \cup \cdots \cup X_{i-1}$ and, by the running intersection property, let $j \in [i-1]$ be such that $X_i \cap X \subseteq X_j$. By induction hypothesis, there is a bag $T_{i-1}$ over $X$ that witnesses the global consistency of $R_1, \ldots, R_{i-1}$. First, we show that $T_{i-1}$ and $R_i$ are consistent. By Lemma 2, it suffices to show that $T_{i-1}[X \cap X_i] = R_i[X \cap X_i]$. Let $Z = X \cap X_i$, so $Z \subseteq X_j$ by the choice of $j$, and indeed $Z = X_j \cap X_i$. Since $j \leq i-1$, we have $R_j = T_{i-1}[X_j]$. Since $Z \subseteq X_j$, we have $R_j[Z] = T_{i-1}[X_j][Z] = T_{i-1}[Z]$. By assumption, also $R_j$ and $R_i$ are consistent, and $Z = X_j \cap X_i$, which by Lemma 2 implies $R_j[Z] = R_i[Z]$. By transitivity, we get $T_{i-1}[Z] = R_i[Z]$, hence, by Lemma 2, the bags $T_{i-1}$ and $R_i$ are consistent. Let $T_i$ be a bag that witnesses the consistency of the bags $T_{i-1}$ and $R_i$. We show that $T_i$ witnesses the global consistency of $R_1, \ldots, R_i$. Since $T_{i-1}$ and $R_i$ are consistent, first note that $T_{i-1} = T_i[X]$ and $R_i = T_i[X_i]$ by Lemma 2. Now fix $k \leq i-1$ and note that

$$R_k = T_{i-1}[X_k] = T_i[X][X_k] = T_i[X_k], \tag{10}$$

where the first equality follows from the fact that $T_{i-1}$ witnesses the consistency of $R_1, \ldots, R_{i-1}$ and $k \leq i-1$, and the other two equalities follow from $T_{i-1} = T_i[X]$ and the fact that $X_k \subseteq X$. Thus, $T_i$ witnesses the consistency of $R_1, \ldots, R_i$, which was to be shown.

*Step 2.* Assume that the hypergraph $H$ is not conformal or it is not chordal. By Lemma 3, there is a subset $W$ of $V$ such that $|W| \geq 3$ and $R(H[W]) = (W, \{W \setminus \{A\} : A \in W\})$ or there is a subset $W$ of $V$ such that $|W| \geq 4$ and $R(H[W]) = (W, \{\{A_i, A_{i+1}\} : i \in [n]\})$, where $A_1, \ldots, A_n$ is an enumeration of $W$ and $A_{n+1} := A_1$. By Corollary 3, if $H$ has the local-to-global consistency property for bags, then for every subset $W$ of $V$, the hypergraph $R(H[W])$ also has the local-to-global consistency property for bags. It follows that, to show that $H$ does not have the local-to-global consistency property for bags, it suffices to show that no hypergraph of the form $(W, \{W \setminus \{A\} : A \in W\})$ with $|W| \geq 3$ has the local-to-global consistency property for bags, and no hypergraph of the form $(W, \{\{A_i, A_{i+1}\} : i \in [n]\})$, where $|W| \geq 4$, $A_1, \ldots, A_n$ is an enumeration of $W$, and $A_{n+1} := A_1$ has the local-to-global consistency property for bags.

The preceding "minimal" non-conformal and non-chordal hypergraphs share the following properties: 1) all their hyperedges have the same number of vertices, and 2) all their vertices appear in the same number of hyperedges. For hypergraphs $H^*$ that have these properties, we construct a collection $C(H^*)$ of bags that are indexed by the hyperedges of $H^*$

Let $H^* = (V^*, E^*)$ be a hypergraph and let $d$ and $k$ be positive integers. The hypergraph $H^*$ is called *$k$-uniform* if every hyperedge of $H^*$ has exactly $k$ vertices. It is called *$d$-regular* if any vertex of $H^*$ appears in exactly $d$ hyperedges of $H$. Thus, the "minimal" non-conformal hypergraph in Lemma 3 is $k$-uniform and $d$-regular for $k := d := |W| - 1$. Likewise, the "minimal" non-chordal hypergraph in the same lemma is $k$-uniform and $d$-regular for $k := d := 2$. For each $k$-uniform and $d$-regular hypergraph $H^*$ with $d \geq 2$ and with hyperedges $E^* = \{X_1, \ldots, X_m\}$, we construct a collection $C(H^*) := \{R_1(X_1), \ldots, R_m(X_m)\}$ of bags, where $R_i$ is a bag with $X_i$ as its set of attributes. The collection $C(H^*)$ of these bags will turn out to be pairwise consistent but not globally consistent.

For each $i \in [m]$ with $i \neq m$, let $R_i$ be the unique bag over $X_i$ defined as follows: (a) the support $R_i'$ of $R_i$ consists of all tuples $t : X_i \to \{0, \ldots, d-1\}$ whose total sum $\sum_{C \in X_i} t(C)$ is congruent to 0 mod $d$; (b) $R_i(t) := 1$ for each such $X_i$-tuple, and $R_i(t) := 0$ for every other $X_i$-tuple. For $i = m$, let $R_m$ be the unique bag over $X_m$ defined as follows: (a) the support $R_m'$ of $R_m$ consists of all tuples $t : X_m \to \{0, \ldots, d-1\}$ whose total sum $\sum_{C \in X_m} t(C)$ is congruent to 1 mod $d$; (b) $R_m(t) := 1$ for each such $X_m$-tuple, and $R_m(t) := 0$ for every other $X_m$-tuple.

By Lemma 2, to show that the bags $R_1, \ldots, R_m$ are pairwise consistent, it suffices to show that for every two distinct $i, j \in [m]$, we have $R_i[Z] \equiv R_j[Z]$, where $Z := X_i \cap X_j$. In turn, this follows from the claim that for every $i \in [m]$ and every $Z$-tuple $t : Z \to \{0, \ldots, d-1\}$, we have $R_i(t) = d^{k-|Z|-1}$. Indeed, since by $k$-uniformity every hyperedge of $H$ has exactly $k$ vertices, for every $u \in \{0, \ldots, d-1\}$, there are exactly $d^{k-|Z|-1}$ many $X_i$-tuples $t_{i,u,1}, \ldots, t_{i,u,d^{k-|Z|-1}}$ that extend $t$ and have total sum congruent to $u$ mod $d$. It follows then that $R_i[Z] = R_j[Z]$ for every two distinct $i, j \in [m]$, regardless of whether $m \in \{i, j\}$ or $m \notin \{i, j\}$, and hence any two $R_i$ and $R_j$ are consistent by Lemma 2. To argue that the relations $R_1, \ldots, R_m$ are not globally consistent, we proceed by contradiction. If $R$ were a bag that witnesses their consistency, then it would be non-empty and its support would contain a tuple $t$ such that the projections $t[X_i]$ belong to the supports $R_i'$ of the $R_i$, for each $i \in [m]$. In turn this means that

$$\sum_{C \in X_i} t(C) \equiv 0 \bmod d, \qquad \text{for } i \neq m \tag{11}$$

$$\sum_{C \in X_i} t(C) \equiv 1 \bmod d, \qquad \text{for } i = m. \tag{12}$$

Since by $d$-regularity each $C \in V$ belongs to exactly $d$ many sets $X_i$, adding up all the equations in (11) and (12) gives

$$\sum_{C \in V} dt(C) \equiv 1 \bmod d, \tag{13}$$

which is absurd since the left-hand side is congruent to 0 mod $d$, the right-hand side is congruent to 1 mod $d$, and $d \geq 2$ by assumption. This completes the proof of Theorem 2. □

Note that Beeri et al. [8] showed that hypergraph acyclicity is equivalent to several other "structural" properties of hypergraphs, such as Graham's algorithm succeeding on $H$. We chose not to mention these other "structural" properties here because we made no use of them in the proof of Theorem 2; these properties, of course, can be added to the list of equivalent statements in Theorem 2. However, Beeri et al. [8] showed that hypergraph acyclicity is also equivalent to several "semantic" properties of relations other than the local-to-global consistency property for relations, including the existence of a full reducer for relations. As we shall discuss in Section 6, it remains an open problem to formulate a suitable concept of a full reducer for bags and show that the existence of such a full reducer for bags is equivalent to hypergraph acyclicity and, hence to the local-to-global consistency property property for bags. The main technical obstacle is that the bag-join of a globally consistent collection of bags need not witness their global consistency.

It should also be pointed out that the proof of Theorem 1 in [8] has a different architecture than the proof of our Theorem 2. In particular, in proving the equivalence between the local-to-global consistency property for relations and acyclicity, they make use of Graham's algorithm.

## 5 COMPLEXITY OF BAG CONSISTENCY

In this section, we explore the algorithmic aspects of global consistency. We first discuss known results about global consistency for relations.

### 5.1 The Set Case

The *global consistency problem for relations* asks: given a hypergraph $H = (V, \{X_1, \ldots, X_m\})$ and relations $R_1, \ldots, R_m$ over $H$, are the relations $R_1, \ldots, R_m$ globally consistent? This problem is also known as the *universal relation problem* since a relation $W$ witnessing the global consistency of $R_1, \ldots, R_m$ is called a *universal* relation for $R_1, \ldots, R_m$. Honeyman, Ladner, and Yannakakis [15] showed that the global consistency problem for relations is NP-complete. The proof of NP-hardness is a reduction from 3-Colorability in which each relation is binary and consists of just six pairs. The proof of membership in NP uses the observation that if a collection $R_1, \ldots, R_m$ of relations is globally consistent, then a witness $W$ of this fact can be obtained as follows: for each $i \leq m$ and each tuple $t \in R_i$, pick a tuple in the join $R_1 \bowtie \cdots \bowtie R_m$ that extends $t$ and insert it in $W$. In particular, the cardinality $|W|$ of $W$ is bounded by the sum $\sum_{i=1}^{m} |R_i| \leq m \max\{|R_i| : i \in [m]\}$, and thus the size of $W$ is bounded by a polynomial in the size of the input hypergraph $H$ and the input relations $R_1, \ldots, R_m$.

The main result in Beeri et al. [8] (stated here as Theorem 1) implies that the global consistency problem for relations is solvable in polynomial time when restricted to acyclic hypergraphs, since, in this case, global consistency of relations is equivalent to pairwise consistency of relations. Furthermore, for every fixed hypergraph $H = (V, \{X_1, \ldots, X_m\})$ (be it cyclic or acyclic), the global consistency problem for relations restricted to relations $R_1, \ldots, R_m$ of schemas $X_1, \ldots, X_m$ is also solvable in polynomial time, since one can first compute the join $J = R_1 \bowtie \cdots \bowtie R_m$ in polynomial time and then check whether $J[X_i] = R_i$ holds, for $i = 1, \ldots, m$. While the cardinality $|J|$ of this witness $J$ can only be bounded by $\prod_{i=1}^{m} |R_i| \leq \max\{|R_i| : i \in [m]\}^m$, this cardinality is still polynomial in the size of the input because, in this case, the exponent $m$ is fixed and not part of the input.

### 5.2 Decision Problem for Bags

We now consider the *global consistency problem for bags*, which asks: given a hypergraph $H = (V, \{X_1, \ldots, X_m\})$ and bags $R_1, \ldots, R_m$ over $H$, are the bags $R_1, \ldots, R_m$ globally consistent? We also consider a family of decision problems arising from fixed hypergraphs. Specifically, with every fixed hypergraph $H = (V, \{X_1, \ldots, X_m\})$, we associate the decision problem GCPB($H$), which asks: given bags $R_1, \ldots, R_m$ over $H$, are the bags $R_1, \ldots, R_m$ globally consistent?

The first result we obtain about the global consistency problem for bags is that it is in NP, even if the multiplicities of the tuples in the bags are represented in binary. To prove this, we will show that if $R_1, \ldots, R_m$ are globally consistent bags, then there exists a bag $W$ that witnesses their global consistency and has size polynomial in the size of $R_1, \ldots, R_m$. More precisely, we will establish that the support $W'$ of the bag $W$ has cardinality at most $\sum_{i=1}^{m} \sum_{r \in R_i'} \log(R_i(r) + 1)$, and each tuple $t \in W'$ has multiplicity $W(t)$ bounded by $\max\{R_i(r) : i \in [m], r \in R_i'\}$. In order to establish this, we need an integral version of Carathéodory's Theorem due to Eisenbrand and Shmonin [14]. For a finite set $X \subseteq \mathbb{R}^d$ of real vectors, let intcone($X$) denote the *integer conic hull* of $X$, that is, the set of all vectors of the form $c_1 x_1 + \cdots + c_t x_t$, where $c_1, \ldots, c_t$ are non-negative integers and $x_1, \ldots, x_t$ are vectors in $X$.

Lemma 5 (Lemma 3 in [14]). *Let $X \subseteq \mathbb{Z}^d_{\geq 0}$ be a finite set of non-negative integer vectors and let $b = (b_1, \ldots, b_d)$ be a vector in its integer conic hull* intcone($X$). *If $|X| > \sum_{i=1}^{d} \log(b_i + 1)$, then there exists a proper subset $X_0 \subseteq X$ such that $b$ is in the integer conic hull* intcone($X_0$) *of $X_0$.*

The plan is to apply Lemma 5 on the set $X$ of column vectors of the constraint-matrix $A$ of an integer linear program along the lines of that in (3), but generalized to any number of bags. Precisely, with each collection $R_1(X_1), \ldots, R_m(X_m)$ of bags, we associate a linear program, denoted by $P(R_1, \ldots, R_m)$, that is a direct generalization of the linear program in (3). Let $J = R_1' \bowtie \cdots \bowtie R_m'$ be the join of the supports of $R_1, \ldots, R_m$. For each $t \in J$, the linear program $P(R_1, \ldots, R_m)$ has a variable $x_t$. For each $t \in J$, each $i \in [m]$, and each $r \in R_i'$, define $a_{r,t} = 1$ if $t[X_i] = r$ and $a_{r,t} = 0$ if $t[X_i] \neq r$. Then, the constraints of $P(R_1, \ldots, R_m)$ are

$$\begin{array}{ll} \sum_{t \in J} a_{r,t} x_t = R_i(r) & \text{for } i \in [m],\ r \in R_i', \\ x_t \geq 0 & \text{for } t \in J. \end{array} \tag{14}$$

Writing the equations of $P(R_1, \ldots, R_m)$ in matrix form as $Ax = b$, it is important to note that, unless $m = 2$, the matrix $A$ is no longer the vertex-edge incidence matrix of a bipartite graph as it was when $m = 2$. This means that the matrix $A$ is no longer necessarily totally unimodular. This point notwithstanding, the fact that the integral solutions of $P(R_1, \ldots, R_m)$ are still in 1-to-1 correspondence with the bags that witness the global consistency of $R_1, \ldots, R_m$ is all we need. We elaborate on this in the next result. Before stating the result, we need the following additional concepts.

Let $R_1, \ldots, R_m$ be globally consistent bags. If $W$ is a bag that witnesses the global consistency of $R_1, \ldots, R_m$, then we say that $W$ is a *minimal witness* if there is no other bag $U$ that witnesses the global consistency of $R_1, \ldots, R_m$ and is such that the support $U'$ of $U$ is strictly contained in the support $W'$ of $W$. For a bag $R$, define

- its *support size* by $\|R\|_{\text{supp}} := |R'|$;
- its *multiplicity bound* by $\|R\|_{\text{mu}} := \max\{R(r) : r \in R'\}$;
- its *multiplicity size* by $\|R\|_{\text{mb}} := \max\{\log(R(r)+1) : r \in R'\}$;
- its *unary size* by $\|R\|_{\text{u}} := \sum_{r \in R'} R(r)$;
- its *binary size* by $\|R\|_{\text{b}} := \sum_{r \in R'} \log(R(r) + 1)$.

Clearly, for every bag $R$, the inequalities $\|R\|_{\text{u}} \leq \|R\|_{\text{supp}} \|R\|_{\text{mu}}$ and $\|R\|_{\text{b}} \leq \|R\|_{\text{supp}} \|R\|_{\text{mb}}$ hold.

Theorem 3. *Let $R_1, \ldots, R_m$ be globally consistent bags and let $W$ be a bag that witnesses their global consistency. Then the following statements are true.*

(1) $\|W\|_{\text{mu}} \leq \max\{\|R_i\|_{\text{mu}} : i \in [m]\}$.
(2) $\|W\|_{\text{supp}} \leq \sum_{i=1}^{m} \|R_i\|_{\text{u}}$.
(3) *If $W$ is a minimal witness, then* $\|W\|_{\text{supp}} \leq \sum_{i=1}^{m} \|R_i\|_{\text{b}}$.

Proof. Let $X_1, \ldots, X_m$ be the schemas of $R_1, \ldots, R_m$. The first two statements follow from the fact that if $W$ is a witness of the global consistency of $R_1, \ldots, R_m$, then the equality

$$W(r) = \sum_{\substack{t \in W': \\ t[X_i]=r}} W(t) = R_i(r) \tag{15}$$

holds for each $i \in [m]$ and each $X_i$-tuple $r$, and the quantities $R_i(r)$ and $W(t)$ with $t \in W'$ are non-negative integers.

For the third statement, assume that $W$ is a minimal witness to the global consistency of $R_1, \ldots, R_m$. Setting $J := R'_1 \bowtie \cdots \bowtie R'_m$, by Lemma 1, we have $W' \subseteq J$. For each $t \in J$, define $x_t := W(t)$ and let $x = (x_t : t \in J)$. It follows from the definitions that the vector $x$ is an integer feasible solution for $P(R_1, \ldots, R_m)$. Write the equations of $P(R_1, \ldots, R_m)$ in matrix form as $Ax = b$, where $A$ is a $d \times |J|$ matrix of zeros and ones where $d := \sum_{i=1}^m |R'_i|$, and $b \in \mathbb{Z}^d_{\geq 0}$ is a $d$-dimensional column vector with non-negative integer entries $(R_i(r) : i \in [m], r \in R'_i)$. Let $X = \{c_t : t \in W'\}$ be the subset of the $d$-dimensional column vectors of $A$ that correspond to the non-zero components of $x$. From the definition of $P(R_1, \ldots, R_m)$ it follows that for every two distinct $t, t' \in W'$, we have $c_t \neq c_{t'}$. Hence $|X| = |W'|$.

The fact that $Ax = b$ means that $\sum_{t \in W'} c_t x_t = b$ and therefore the vector $b$ belongs to the integer conic hull intcone$(X)$ of $X$. Likewise, for every subset $Q \subseteq W'$ such that $b$ is in the integer conic hull of $X_0 := \{c_t : t \in Q\}$, there exists a bag $W_0$ with support $Q$ that witnesses the global consistency of $R_1, \ldots, R_m$. Therefore, since $W$ is a minimal witness, it follows from Lemma 5 that $|X| \leq \sum_{i=1}^m \sum_{r \in R'_i} \log(R_i(r) + 1) = \sum_{i=1}^m \|R_i\|_{\mathrm{b}}$. Since $|W'| = |X|$, the third statement has been proved. □

It should be noted that, assuming that all the numbers that are fed into an algorithm are represented with the same number of bits by adding leading zeros when necessary, the size of the representation of a bag $R$ when it is fed into an algorithm is $\|R\|_{\mathrm{supp}}\|R\|_{\mathrm{mu}}$ when the multiplicities are represented in unary, and $\|R\|_{\mathrm{supp}}\|R\|_{\mathrm{mb}}$ when the multiplicities are represented in binary. Therefore, since every globally consistent collection of bags has a minimal witness of their global consistency, Theorem 3 readily implies the following result.

Corollary 4. *The global consistency problem for bags is in NP.*

It is worth noting that the first two statements of Theorem 3 alone already imply the same if the multiplicities of the given bags are bounded, or if they are represented in unary. Nonetheless, as the following example shows, the third statement of Theorem 3 is unavoidable if the multiplicities are represented in binary, even if the schemas form acyclic hypergraphs.

*Example 1.* Consider bags $R_1(A_1A_2), R_2(A_2A_3), \ldots, R_{n-1}(A_{n-1}A_n)$ with supports $\{0,1\}^2$ and multiplicity $2^n$ for each tuples in their support. Let $J$ be the bag of schema $A_1 \cdots A_n$, support $\{0,1\}^n$, and multiplicity 4 for each tuple in its support. Then we have $J[A_iA_{i+1}] = R_i$ for all $i = 1, \ldots, n-1$, and $|J'|$ has cardinality $2^n$, which is exponentially bigger than the size $4(n-1)(n+1)$ of the input $R_1, \ldots, R_{n-1}$, when the multiplicities are written in binary. ⊣

Theorem 2 and Lemma 2 imply that the global consistency problem for bags is solvable in polynomial time when restricted to acyclic hypergraphs, since, in this case, global consistency of bags is equivalent to pairwise consistency of bags, and the latter is checkable in polynomial time.

We now turn to fixed hypergraphs, and state and prove the main result of this section.

Theorem 4. *Let $H = (V, E)$ be a hypergraph. Then the following statements are true.*

(1) *If $H$ is acyclic, then* GCPB$(H)$ *is solvable in polynomial time.*
(2) *If $H$ is cyclic, then* GCPB$(H)$ *is NP-complete.*

Proof. The first part of the theorem follows from Theorem 2 and Lemma 2. To prove the second part of the theorem, first note that membership in NP is a special case of Corollary 4. To prove NP-hardness, we will show that if $H$ is a minimal non-chordal hypergraph or a minimal non-conformal hypergraph, then GCPB$(H)$ is NP-complete. More precisely, we will show in Lemmas 6 and 7 that both problems GCPB$(C_n)$ and GCPB$(H_n)$ are NP-complete for any $n \geq 3$. The desired NP-hardness will then follow from Lemmas 3 and 4.

Lemma 6. *For every $n \geq 3$, the problem* GCPB$(C_n)$ *is NP-complete.*

Proof. The problem GCPB$(C_3)$ generalizes the problem of consistency of 3-dimensional contingency tables (3DCT) from [16]: given a positive integer $n$ and, for each $i, j, k \in [n]$, non-negative integer values $R(i,k)$, $C(j,k)$, $F(i,j)$, is there an $n \times n \times n$ table of non-negative integers $X(i,j,k)$ such that $\sum_{q=1}^n X(i,q,k) = R(i,k)$, $\sum_{q=1}^n X(q,j,k) = C(j,k)$, $\sum_{q=1}^n X(i,j,q) = F(i,j)$ for all indices $i, j, k \in [n]$? To see this, let $X, Y, Z$ be three attributes with domain $[n]$, and let $R(XZ)$, $C(YZ)$, $F(XY)$ be the three bags given by the three tables $R(i,k)$, $C(j,k)$, $F(i,j)$. Therefore, GCPB$(C_3)$ is NP-complete. For $n \geq 4$, we show that there is a polynomial time reduction from GCPB$(C_{n-1})$ to GCPB$(C_n)$. The claim that GCPB$(C_n)$ is NP-complete for every $n \geq 3$ will follow by induction.

Let $R_1(A_1A_2), R_2(A_2A_3), \ldots, R_{n-1}(A_{n-1}A_1)$ be an instance of GCPB$(C_{n-1})$. Let $A_n$ be a new attribute with the same domain as $A_1$. The reduction replaces the bag $R_{n-1}(A_{n-1}A_1)$ by an identical copy $R_{n-1}(A_{n-1}A_n)$ of schema $A_{n-1}A_n$, and adds one more bag $R_n(A_nA_1)$ with support $R'_n = \{(a,a) : a \in \mathrm{Dom}(A_1)\}$, and multiplicities defined by $R_n(a,a) = R_{n-1}(a)$ for every $(a,a) \in R'_n$, where $R_{n-1}(a)$ denotes the multiplicity of $a$ in the $R_{n-1}[A_1]$. If $R$ is a bag that witnesses the global consistency of $R_1, \ldots, R_{n-1}$, then the bag $S(A_1 \cdots A_n)$ defined, for each $A_1 \cdots A_n$-tuple $t$ by $S(t) = R(t[A_1 \cdots A_{n-1}])$ whenever $t[A_n] = t[A_{n-1}]$ and $S(t) = 0$ otherwise, witnesses the global consistency of $R_1, \ldots, R_n$. Conversely, if $S$ is a bag that witnesses the global consistency of $R_1, \ldots, R_n$, then the bag $R(A_1 \cdots A_{n-1})$ defined, for each $A_1 \cdots A_{n-1}$-tuple $t$ by $R(t) = R(t, t[A_{n-1}])$, witnesses the global consistency of $R_1, \ldots, R_{n-1}$. □

Lemma 7. *For every $n \geq 3$, the problem* GCPB$(H_n)$ *is NP-complete.*

Proof. Since $H_3 = C_3$, the problem GCPB$(H_3)$ is NP-complete by the first part of the lemma. For $n \geq 4$, we show that there is a polynomial time reduction from GCPB$(H_{n-1})$ to GCPB$(H_n)$. The claim that GCPB$(H_n)$ is NP-complete for every $n \geq 3$ will follow by induction.

Let $R_1(X_1), \ldots, R_{n-1}(X_{n-1})$ be bags, where $X_i = \{A_1, \ldots, A_{n-1}\} \setminus \{A_i\}$ for $i \in [n-1]$. Let $A_n$ be a new attribute with domain $\{1, 2\}$ and

define new bags $S_1(Y_1), \ldots, S_n(Y_n)$ with $Y_i = \{A_1, \ldots, A_n\} \setminus \{A_i\}$ for $i \in [n]$ as follows. For $i \in [n-1]$, let $D_i$ be the size of the active domain of the attribute $A_i$ in the supports $R'_1, \ldots, R'_{n-1}$ of $R_1, \ldots, R_{n-1}$, and let $M$ be the maximum of all multiplicities in $R_1, \ldots, R_{n-1}$. For $i \in [n-1]$, define $S_i(t,1) = R_i(t)$ and $S_i(t,2) = MD_i - R_i(t)$ for any $X_i$-tuple $t$. For $i = n$, define $S_i(t) = M$ for any $Y_i$-tuple $t$. We claim that this reduction works. Indeed, given a witness $R$ for the global consistency of $R_1, \ldots, R_{n-1}$, we can produce a witness $S$ for the global consistency of $S_1, \ldots, S_n$ by setting $S(t,1) = R(t)$ and $S(t,2) = M - R(t)$ for any $A_1 \cdots A_{n-1}$-tuple $t$. Conversely, given a witness $S$ for the global consistency of $S_1, \ldots, S_n$, we can produce a witness $R$ for the global consistency of $R_1, \ldots, R_{n-1}$ by setting $R(t) = S(t,1)$ for any $A_1, \ldots, A_{n-1}$-tuple $t$. □

The proof of Theorem 4 is now complete. □

## 5.3 Finding the Witness

In this section, we address the question of producing a small witness to global consistency, when a witness to global consistency exists. Theorem 3 ensures that if there is any witness at all, then a small one exists, but it does not tell us how to construct a small witness.

We start by noting that, for any fixed cyclic hypergraph $H$, one cannot hope to find small witnesses to the global consistency of given bags over $H$ in time polynomial in the size of the input, unless P = NP. Indeed, just deciding if a witness exists is already NP-hard by Theorem 4. Since checking if a witness is valid is a problem that can be solved in polynomial time, the problem of finding a witness can only be harder. For acyclic hypergraphs, however, we will see that the structural results of Section 4 provide a way to construct a witness. For this, we will need a strengthening of Corollary 1 to the effect that not only a witness to the consistency of two bags can be found, but even a minimal witness can be found in strongly polynomial time. We will also need a strengthening of Theorem 3 in the special case of two bags.

To describe the algorithm that finds minimal witnesses, we need to introduce some terminology. Let $R(X)$ and $S(Y)$ be two bags and consider the network $N(R,S)$. In what follows, an edge $(u,v)$ of $N(R,S)$ of the form $(t[X], t[Y])$ with $t \in R' \bowtie S'$ is called a *middle edge*. The proof of Lemma 2 established that if $R$ and $S$ are consistent and $f(u,v)$ is a saturated flow of the network $N(R,S)$, then the bag $T(XY)$ defined by setting $T(t) := f(t[X], t[Y])$ for each middle edge $(t[X], t[Y])$ is a witness to the consistency of $R$ and $S$. In particular, the support $T'$ of the witness $T$ is the set of middle edges of $N(R,S)$ that are used by the flow $f$.

In order to find a minimal witness to the consistency of $R$ and $S$, we proceed by self-reducibility, deleting middle edges from $N(R,S)$ one by one. We loop through the middle edges $(u,v)$ of the current network and, for each one, ask: is the middle edge $(u,v)$ used by all saturated flows of the current network? If the answer is no, then it is safe to delete the edge and continue with the new network. If the answer is yes, then we keep the edge and proceed to the next middle edge. To tell whether a middle edge $(u,v)$ is used by all saturated flows of the current network, we can temporarily remove it, compute a maximum flow of the resulting network, and check whether it is saturated. Since the number of middle edges of the initial network $N(R,S)$ is $|R' \bowtie S'|$, a saturated flow along a minimal subset of middle edges will be found after at most $|R' \bowtie S'|$ many such tests. This gives a minimal witness for the consistency of $R$ and $S$.

Before we state the strengthening of Corollary 1, we also need to strengthen the bound on the support-size of minimal witnesses given by Theorem 3 in the case $m = 2$. For this special case, the standard form of Carathéodory's Theorem will suffice. The *conic hull* of a set $X \subseteq R^d$, where $d \geq 1$, is the set of all vectors in $R^d$ that can be written as a linear combination of vectors from $X$ with non-negative coefficients. Carathéodory's Theorem asserts that if $X \subseteq R^d$ for some $d \geq 1$ and if a vector $\mathbf{x}$ belongs to the conic hull of $X$, then there is a subset $X_0$ of $X$ of cardinality at most $d$ such that $x$ belongs to the conic hull of $X_0$ (in Schrijver's book on linear and integer programming, this is stated as Corollary 7.1i and it follows from more general results about linear programming).

Theorem 5. *Let $R$ and $S$ be consistent bags and let $W$ be a bag that witnesses their consistency. If $W$ is a minimal witness to the consistency of $R$ and $S$, then $\|W\|_{\text{supp}} \leq \|R\|_{\text{supp}} + \|S\|_{\text{supp}}$.*

Proof. Let $J := R' \bowtie S'$, so $W' \subseteq J$ by Lemma 1. By setting $x_t := W(t)$ for each $t \in J$, we get a feasible solution for the linear program $P(R,S)$. If we write the constraint matrix of $P(R,S)$ in matrix form as $Ax = b$, this means that the vector $b$ is in the conic hull of the set of columns of $A$ indexed by tuples $t$ in $W'$. By Carathéodory's Theorem, $b$ is also in the conic hull of a subset of at most $d$ many of the columns of $A$ indexed by tuples $t$ in $W'$, where $d := \|R\|_{\text{supp}} + \|S\|_{\text{supp}}$ is the dimension of the vector $b$. This means that there exists $J_0 \subseteq W'$ with $|J_0| \leq d$ and a non-negative vector $y = (y_t : t \in J)$ with $y_t = 0$ for each $t \in J \setminus J_0$ such that $Ay = b$. Setting $f(t[X], t[Y]) = y_t$ for each $t \in J_0$, we get a saturated flow of the subnetwork $N_0$ of $N(R,S)$ in which all middle edges of the form $(t[X], t[Y])$ with $t \in J \setminus J_0$ have been suppressed. Since all capacities of $N(R,S)$ are integers, as in the proof of Lemma 2, the integrality theorem for the max-flow problem gives a max flow $f_0(u,v)$ of $N_0$ with integers. This flow of $N_0$ is also saturated, which means that by setting $W_0(t) := f_0(t[X], t[Y])$ for each $t \in J_0$ we get a witness of the consistency of $R$ and $S$ with support $W'_0$ included in $J_0 \subseteq W'$. Since $W$ is minimal we have $J_0 = W'$, from which it follows that $|W'| = |J_0| \leq d$. That is, $\|W\|_{\text{supp}} \leq \|R\|_{\text{supp}} + \|S\|_{\text{supp}}$. □

Corollary 5. *There is a strongly polynomial-time algorithm that, given two bags $R$ and $S$, determines whether they are consistent and, if they are, constructs a bag $T$ that is a minimal witness of their consistency. In particular, $\|T\|_{\text{supp}} \leq \|R\|_{\text{supp}} + \|S\|_{\text{supp}}$.*

We now put everything together to show that, over acyclic schemas, a witness to global consistency can be found in polynomial time.

Theorem 6. *There is a polynomial time algorithm that, given an acyclic hypergraph $H$ and a collection of bags over $H$, determines whether the collection is globally consistent and, if it is, constructs a bag that is a witness to the global consistency of the collection. Furthermore, the bag that the algorithm returns has its support-size bounded by the sum of the support-sizes of the input bags.*

Proof. Let $H = (V, \{X_1, \ldots, X_m\})$ be an acyclic hypergraph and let $R_1(X_1), \ldots, R_m(X_m)$ be a collection of bags over $H$. First, we test

for pairwise consistency. If there are two bags in the collection that are not consistent, then the collection cannot be globally consistent, and we stop. Otherwise, we proceed as in the proof of Theorem 2 to construct a witness of their global consistency as follows.

By first computing a rooted join-tree in polynomial time (see [23]) and then by sorting its vertices in topological order, we may assume that the listing $X_1, \ldots, X_m$ satisfies the running intersection property: for every $i \in [m]$ with $i \geq 2$, there is a $j \in [i-1]$ such that $X_i \cap (X_1 \cup \cdots \cup X_{i-1}) \subseteq X_j$. By induction on $i = 1, \ldots, m$, we construct a bag $T_i$ over $X_1 \cup \cdots \cup X_i$ that is a witness to the global consistency of the bags $R_1, \ldots, R_i$ and satisfies $\|T_i\|_{\text{supp}} \leq \sum_{j=1}^{i} \|R_i\|_{\text{supp}}$. For $i = 1$, we take $T_i = R_i$. For $i \geq 2$, we apply the algorithm given by Corollary 5 on the bags $T_{i-1}$ and $R_i$ to obtain $T_i$. In Step 1 of the proof of Theorem 2, we showed that any bag that witnesses the consistency of $T_{i-1}$ and $R_i$, such as $T_i$, also witnesses the global consistency of $R_1, \ldots, R_i$. By Corollary 5, we also have that $\|T_i\|_{\text{supp}} \leq \|T_{i-1}\|_{\text{supp}} + \|R_i\|_{\text{supp}}$, from which the desired bound $\|T_i\|_{\text{supp}} \leq \sum_{j=1}^{i} \|R_j\|_{\text{supp}}$ follows by the induction hypothesis.

Let $M$ be the maximum multiplicity in the input bags $R_1, \ldots, R_m$ and let $B = \log(M+1)$ be the number of bits it takes to represent them. By Theorem 3 we have that all the multiplicities of every $T_i$ are bounded by $M$. Therefore, the size of each $T_i$ is bounded by $B\|T_i\|_{\text{supp}} \leq B \sum_{j=1}^{i} \|R_j\|_{\text{supp}}$. The runtime of the algorithm is then bounded by $m$ times the runtime of the algorithm in Corollary 5 on inputs of these sizes, and is thus bounded by a polynomial in $B \sum_{j=1}^{m} \|R_j\|_{\text{supp}}$, i.e., the size of the input. □

## 6 CONCLUDING REMARKS

In this paper, we investigated the interplay between local consistency and global consistency for bags. At the structural level, we showed that bags behave like relations as regards the local-to-global consistency property, namely, the local-to-global consistency property for bags holds over a schema if and only if the sets of attributes of that schema form an acyclic hypergraph. At the algorithmic level, however, bags behave different than relations as regards testing for global consistency. Specifically, for every fixed schema, testing relations for global consistency is solvable in polynomial time, while for bags this happens precisely when the schema is acyclic - otherwise, testing bags for global consistency is NP-complete.

We conclude by describing certain open problems that are motivated by the work reported here.

Beeri et al. [8] showed that hypergraph acyclicity is also equivalent to certain semantic conditions other than the local-to-global consistency property for relations, including the existence of a *full reducer* and the existence of a *monotone sequential join expression*. Do analogous results hold bags? One of the difficulties in answering this question is that the bag-join of two consistent relations need not witness their consistency, thus it is not at all clear how to define a suitable semi-join operation for bags or how to find a suitable substitute for a monotone sequential join expression.

As mentioned in the Introduction, we have recently studied a relaxed notion of consistency for $K$-relations, where $K$ is a positive semiring [7]. The goal of that investigation was to find a common generalization of the results by Vorob'ev [26] and by Beeri et al. [8]. The stricter notion of consistency for bags studied here makes perfectly good sense for $K$-relations as well. It is an open problem whether or not the results presented here extend to $K$-relations under the stricter notion of consistency, where $K$ is a positive semiring or some other type of semiring for which there is a good theory for solving systems of linear equations or other combinatorial problems formulated over that semiring.

## ACKNOWLEDGMENTS

The research of Albert Atserias was partially supported by MICIN project PID2019-109137GB-C22 (PROOFS). The research of Phokion Kolaitis was partially supported by NSF Award No. 1814152.

## REFERENCES

[1] Samson Abramsky. 2013. Relational Databases and Bell's Theorem. In *In Search of Elegance in the Theory and Practice of Computation - Essays Dedicated to Peter Buneman (Lecture Notes in Computer Science, Vol. 8000)*, Val Tannen, Limsoon Wong, Leonid Libkin, Wenfei Fan, Wang-Chiew Tan, and Michael P. Fourman (Eds.). Springer, 13–35. https://doi.org/10.1007/978-3-642-41660-6_2

[2] Samson Abramsky. 2014. Contextual Semantics: From Quantum Mechanics to Logic, Databases, Constraints, and Complexity. *Bull. EATCS* 113 (2014). http://eatcs.org/beatcs/index.php/beatcs/article/view/286

[3] Samson Abramsky, Rui Soares Barbosa, Kohei Kishida, Raymond Lal, and Shane Mansfield. 2015. Contextuality, Cohomology and Paradox. In *24th EACSL Annual Conference on Computer Science Logic, CSL 2015, September 7-10, 2015, Berlin, Germany (LIPIcs, Vol. 41)*, Stephan Kreutzer (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 211–228. https://doi.org/10.4230/LIPIcs.CSL.2015.211

[4] Samson Abramsky and Adam Brandenburger. 2011. A Unified Sheaf-Theoretic Account Of Non-Locality and Contextuality. *CoRR* abs/1102.0264 (2011). arXiv:1102.0264 http://arxiv.org/abs/1102.0264

[5] Samson Abramsky, Shane Mansfield, and Rui Soares Barbosa. 2011. The Cohomology of Non-Locality and Contextuality. In *Proceedings 8th International Workshop on Quantum Physics and Logic, QPL 2011, Nijmegen, Netherlands, October 27-29, 2011 (EPTCS, Vol. 95)*, Bart Jacobs, Peter Selinger, and Bas Spitters (Eds.). 1–14. https://doi.org/10.4204/EPTCS.95.1

[6] Alfred V. Aho, Catriel Beeri, and Jeffrey D. Ullman. 1979. The Theory of Joins in Relational Databases. *ACM Trans. Database Syst.* 4, 3 (1979), 297–314. https://doi.org/10.1145/320083.320091

[7] Albert Atserias and Phokion G. Kolaitis. 2020. Consistency, Acyclicity, and Positive Semirings. *CoRR* abs/2009.09488 (2020). arXiv:2009.09488 https://arxiv.org/abs/2009.09488

[8] Catriel Beeri, Ronald Fagin, David Maier, and Mihalis Yannakakis. 1983. On the Desirability of Acyclic Database Schemes. *J. ACM* 30, 3 (July 1983), 479–513. https://doi.org/10.1145/2402.322389

[9] John S Bell. 1964. On the Einstein-Podolsky-Rosen paradox. *Physics Physique Fizika* 1, 3 (1964), 195.

[10] Claude Berge. 1989. *Hypergraphs - combinatorics of finite sets*. North-Holland mathematical library, Vol. 45. North-Holland.

[11] Johann Brault-Baron. 2016. Hypergraph Acyclicity Revisited. *ACM Comput. Surv.* 49, 3 (2016), 54:1–54:26. https://doi.org/10.1145/2983573

[12] Ashok K. Chandra and Philip M. Merlin. 1977. Optimal Implementation of Conjunctive Queries in Relational Data Bases. In *Proceedings of the 9th Annual ACM Symposium on Theory of Computing, May 4-6, 1977, Boulder, Colorado, USA*, John E. Hopcroft, Emily P. Friedman, and Michael A. Harrison (Eds.). ACM, 77–90. https://doi.org/10.1145/800105.803397

[13] Surajit Chaudhuri and Moshe Y. Vardi. 1993. Optimization of *Real* Conjunctive Queries. In *Proceedings of the Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 25-28, 1993, Washington, DC, USA*, Catriel Beeri (Ed.). ACM Press, 59–70. https://doi.org/10.1145/153850.153856

[14] Friedrich Eisenbrand and Gennady Shmonin. 2006. Carathéodory bounds for integer cones. *Operations Research Letters* 34, 5 (2006), 564 – 568. https://doi.org/10.1016/j.orl.2005.09.008

[15] Peter Honeyman, Richard E. Ladner, and Mihalis Yannakakis. 1980. Testing the Universal Instance Assumption. *Inf. Process. Lett.* 10, 1 (1980), 14–19. https://doi.org/10.1016/0020-0190(80)90114-3

[16] Robert W. Irving and Mark Jerrum. 1994. Three-Dimensional Statistical Data Security Problems. *SIAM J. Comput.* 23, 1 (1994), 170–184. https://doi.org/10.1137/S0097539790191010

[17] Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. 2020. Bag Query Containment and Information Theory. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, Dan Suciu, Yufei Tao, and Zhewei Wei (Eds.). ACM, 95–112. https://doi.org/10.1145/3375395.3387645

[18] George Konstantinidis and Fabio Mogavero. 2019. Attacking Diophantus: Solving a Special Case of Bag Containment. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, Dan Suciu, Sebastian Skritek, and Christoph Koch (Eds.). ACM, 399–413. https://doi.org/10.1145/3294052.3319689

[19] Jesús A. De Loera and Shmuel Onn. 2004. The Complexity of Three-Way Statistical Tables. *SIAM J. Comput.* 33, 4 (2004), 819–836. https://doi.org/10.1137/S0097539702403803

[20] James B. Orlin. 2013. Max flows in O(nm) time, or better. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, Dan Boneh, Tim Roughgarden, and Joan Feigenbaum (Eds.). ACM, 765–774. https://doi.org/10.1145/2488608.2488705

[21] Donald J. Rose, Robert Endre Tarjan, and George S. Lueker. 1976. Algorithmic Aspects of Vertex Elimination on Graphs. *SIAM J. Comput.* 5, 2 (1976), 266–283. https://doi.org/10.1137/0205021

[22] Alexander Schrijver. 1986. *Theory of Linear and Integer Programming.* John Wiley & Sons, Inc., USA.

[23] Robert E. Tarjan and Mihalis Yannakakis. 1984. Simple Linear-Time Algorithms to Test Chordality of Graphs, Test Acyclicity of Hypergraphs, and Selectively Reduce Acyclic Hypergraphs. *SIAM J. Comput.* 13, 3 (July 1984), 566–579. https://doi.org/10.1137/0213035

[24] G. S. Tseitin. 1968. On the complexity of derivation in propositional calculus. *Structures in Constructive Mathematics and Mathematical Logic* (1968), 115–125. https://ci.nii.ac.jp/naid/10030021172/en/

[25] Jeffrey D. Ullman. 1982. The U. R. Strikes Back. In *Proceedings of the ACM Symposium on Principles of Database Systems, March 29-31, 1982, Los Angeles, California, USA*, Jeffrey D. Ullman and Alfred V. Aho (Eds.). ACM, 10–22. https://doi.org/10.1145/588111.588114

[26] Nikolai Nikolaevich Vorob'ev. 1962. Consistent families of measures and their extensions. *Theory of Probability & Its Applications* 7, 2 (1962), 147–163.

[27] Herbert S. Wilf. 2002. *Algorithms and complexity (2. ed.).* A K Peters.

[28] Mihalis Yannakakis. 1981. Algorithms for Acyclic Database Schemes. In *Very Large Data Bases, 7th International Conference, September 9-11, 1981, Cannes, France, Proceedings.* IEEE Computer Society, 82–94.

[29] Mihalis Yannakakis. 1996. Perspectives on database theory. *SIGACT News* 27, 3 (1996), 25–49. https://doi.org/10.1145/235666.235670