Axiomatization of Frequent Itemsets

T. Calders [1] J. Paredaens

Departement Wiskunde-Informatica, Universiteit Antwerpen, Universiteitsplein 1,B-2610 Wilrijk, Belgium

Abstract

Mining *association rules* is very popular in the data mining community. Most algorithms designed for finding association rules start with searching for *frequent itemsets*. Typically, in these algorithms, *counting phases* and *pruning phases* are interleaved. In the counting phase, partial information about the frequencies of selected itemsets is gathered. In the pruning phase as much as possible of the search space is pruned, based on the counting information. We introduce *frequent set expressions* to represent (possible partial) information acquired in the counting phase. A frequent set expression is a pair containing an itemset and a fraction that is a lower bound on the actual frequency of the itemset. A *system of frequent sets* is a collection of such pairs. We give an axiomatization for those systems that are *complete* in the sense that they explicitly contain all information they logically imply. Every system of frequent sets has a unique completion that actually represents all knowledge that can be derived. We also study *sparse systems*, in which not for every frequent set an expression is given. Furthermore, we explore the links with probabilistic logics.

Key words: Data Mining, Association Rules, Frequent sets, Probabilistic Logic.

1 Introduction

1.1 Association Rules

Association rules are one of the most studied topics in data mining. Since their introduction [1], many algorithms have been proposed to find association

Email addresses: calders@uia.ua.ac.be (T. Calders), pareda@uia.ua.ac.be (J. Paredaens).

[1] Research Assistant of the Fund for Scientific Research - Flanders.

rules [1,2,11].

We start with a formal definition of the association rule mining problem as stated in [1]: Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ be a set of symbols, called *items*. Let $\mathcal{D}$ be a set of *transactions*, where each transaction T is a set of items, $T \subseteq \mathcal{I}$, and a unique transaction ID. We say that a transaction T *contains* X, a set of some items in $\mathcal{I}$, if $X \subseteq T$. The fraction of transactions containing X is called the *frequency* of X. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$, and $X \cap Y = \phi$. The rule holds in the transaction set $\mathcal{D}$ with *confidence* c if the fraction of the transactions containing X, that also contain Y is at least c. The rule $X \Rightarrow Y$ has *support* s in the transaction set $\mathcal{D}$ if the fraction of the transactions in $\mathcal{D}$ that contain $X \cup Y$ is at least s.

Most algorithms start with searching for itemsets that are contained in at least a fraction s of the transactions. To optimize the search for these frequent itemsets, the algorithms use the following monotonicity principle:

> If $X \subseteq Y$, then the frequency of X will never be smaller than the frequency of Y.

This information is then used to *prune* parts of the search space *a-priori*. To exploit this monotonicity as much as possible, the "A-priori"-algorithm [2] starts by counting the single itemsets. In the second step, we count only itemsets $\{i_1, i_2\}$ where $\{i_1\}$ and $\{i_2\}$ are both frequent. All other 2-itemsets are *pruned*. In the third step, the algorithm proceeds with the 3-itemsets that contain only frequent 2-itemsets as subsets. This iteration continues until no more itemsets can be generated. The search of frequent itemsets is thus basically an interleaving of a *counting phase* and *a meta-phase*. In the counting phase, the frequencies of some predetermined itemsets, the so-called *candidates* are counted. In the meta-phase the results of the counting phase are evaluated. Based on the monotonicity principle, some itemsets are a-priori excluded.

Although the monotonicity of frequency is commonly used, there is to our knowledge no previous work that discusses whether in the general case this rule is *complete*, in the sense that it tells us everything we can derive from a given set of frequencies.

1.2 Frequent Set Expressions

In this paper we consider the notion of *a system of frequent sets*. A system of frequent sets contains (possibly incomplete) information about the frequency of every itemset. For example, $S = \{\phi{::}0.5, A{::}0.6, B{::}0.6, AB{::}0.1\}$ is a system of frequent sets. This system of frequent sets represents partial information (e.g. obtained in counting phases.) In this system, $A{::}0.6$ expresses the

knowledge that itemset A has a frequency of at least 0.6. The system S can be improved. Indeed, we can conclude that AB::0.2 holds, since A::0.6 and B::0.6 and there must be an overlap of at least 0.2 between the transactions containing A and the transactions containing B. We can also improve ϕ::0.5, because ϕ::1 always holds. Therefore, S is called *incomplete*. The *completion* of a system represents the maximal information that can be assumed in the meta-phase. The completion of S, denoted $C(S)$, is $\{\phi::1, A::0.6, B::0.6, AB::0.2\}$. $C(S)$ explicitly contains all information *logically implied* by S.

In this paper, we give three rules $\boldsymbol{\mathcal{F}_1}$, $\boldsymbol{\mathcal{F}_2}$, and $\boldsymbol{\mathcal{F}_3}$ that characterize complete systems of frequent sets; e.g. a system is complete iff it satisfies $\boldsymbol{\mathcal{F}_1}$, $\boldsymbol{\mathcal{F}_2}$, and $\boldsymbol{\mathcal{F}_3}$. We show that, after a small modification to $\boldsymbol{\mathcal{F}_3}$, this axiomatization is finite and every logical implication can be inferred/proved using these axioms.

We also address *sparse systems*. These are systems that do not contain a frequent set expression for every itemset. We show that the axiomatization can be adapted to handle such systems efficiently.

1.3 Outline

The structure of the paper is as follows. In Section 2 we sketch the framework. In Section 3 we introduce frequent set expressions and systems of frequent sets. In Section 4, an axiomatization for complete systems of frequent sets is given, after introducing *rare set expressions* as an intermediate stage in the proofs. In Section 5 we discuss how these axioms can be used as rules of inference, and an algorithm for computing completion is given. Section 6 describes some results about sparse systems and gives a revised set of axioms suited for sparse systems. Section 7 discusses related work, in particular links with probabilistic logics. Section 8 concludes the paper.

Parts of the results presented in this paper were already included in [17,4,5].

2 Framework

As discussed in the introduction, most algorithms for finding the frequent itemsets fit more or less in the following framework:

1. C_1 =initial set of candidate frequent itemsets
2. $k = 1$
3. **while** $C_k \neq \{\}$ **loop**
4. *Info*=Count(C_k)

5. C_{k+1} =Generate($Info$)
6. $k = k + 1$
7. **end-loop**

In the a-priori-algorithm, C_1 is initialized to the set of single-itemsets (1). The information obtained in the k^{th} count-step is the frequency of all itemsets in C_k (4). In the generate-step, we use the information obtained in the count-steps to select exactly those $k+1$-itemsets that have no infrequent subsets (5). The other itemsets are *pruned*.

A-priori relies heavily on the monotonicity principle, stating that the frequency of a set is always smaller than or equal to the frequencies of its subsets. In this paper we concentrate on the derivation of information. In every algorithm, an itemset K can be pruned if from the information gathered in the counting phases can be inferred that the frequency of K must be lower than the threshold. This inference can be seen as completing the information obtained by counts. In this paper, we do not commit ourselves to a specific algorithm. Instead, we concentrate on the following question:

Given a set of expressions, what information can be derived from it?

3 Systems of Frequent Sets

We formally define a system of frequent sets. Frequent set expressions describe positive information about the frequencies of the itemsets.

To represent a databases with transactions, we use a matrix. The columns of the matrix represent the items, and the rows represent the transactions. The matrix contains a 1 in the (i, j)-entry if transaction i contains item j; else this entry is 0. When R is a matrix whose columns represent the items in I, we say that R is a matrix over I. In our running example we regularly refer to the items with capital letters. With this notation, we get the following definition:

Definition 1 *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, and R be a matrix over I. The* frequency *of an itemset $K \subseteq I$ in R, denoted $freq(K, R)$ is the fraction of rows in R that have a 1 in every column of K.*

Example 2 *In Fig. 1, a matrix is given, together with some frequencies. The frequency of DEF is 0.2, because 2 rows out of 10 have a one in every column of DEF.*[2] *Note that, unlike a relation, a matrix can have identical rows.*

[2] $\overline{DEF}$ denotes the set $\{D, E, F\}$.

Matrix R

A	B	C	D	E	F
1	0	1	0	1	1
1	0	1	0	1	1
0	1	0	1	1	0
1	1	1	0	0	1
1	0	0	1	0	1
0	1	0	1	1	1
1	1	0	1	1	1
0	0	1	0	0	1
1	1	1	0	1	0
1	0	0	1	0	1

$freq(A, R) = 0.7$,
$freq(B, R) = 0.5$,
$freq(AB, R) = 0.3$,
$freq(DEF, R) = 0.2$.

R satisfies A::0.5, AB::0.3, and DEF::0.1.
R does not satisfy A::0.8, ABC::0.4, or DEF::0.3.

Fig. 1. A matrix together with some frequent set expressions

3.1 Complete Sytems

We introduce a system of frequent sets as a collection of frequent set expressions. Logical implication and completeness of systems are defined.

Definition 3 *Let $I = \{I_1, \ldots, I_n\}$ be a set of items.*

- *A* frequent set expression over I *is an expression K::p_K with $K \subseteq I$ and p_K a rational number with $0 \leq p_K \leq 1$.*
- *A matrix R over I* satisfies *K::p_K, denoted $R \models K$::p_K iff $freq(K, R) \geq p_K$. Hence itemset K has frequency at least p_K.*
- *A* system of frequent sets *over I is a collection*

  $$\Big\{_{K \subseteq I}\ K::p_K$$

  of frequent set expressions, with one expression for each $K \subseteq I$.
- *A matrix R over I* satisfies *the system $S = \Big\{_{K \subseteq I}\ K::p_K$, denoted $R \models S$, iff R satisfies all K::p_K.*
- *A system of frequent sets S* logically implies *K::p_K, denoted $S \models K::p_K$, iff every matrix that satisfies S, also satisfies K::p_K. System S_1* logically implies *system S_2, denoted $S_1 \models S_2$, iff every K::p in S_2 is logically implied by S_1.*
- *A system of frequent sets $S = \Big\{_{K \subseteq I}\ K::p_K$ is* complete *iff for each K::p logically implied by S, $p \leq p_K$ holds.*

Example 4 *In Fig. 1, the matrix R satisfies A::0.6. R does not satisfy B::0.7.*

3.2 Proof-matrices

Very important in the completeness proof of the axiomatization are the so-called proof-matrices.

Definition 5 *Let I be a set of items, $S = \left\{_{K \subseteq I} K::p_K$ a system of frequent sets, and $L \subseteq I$. A matrix M over I is a* proof-matrix *of L in S iff $M \models S$ and $freq(L, M) = p_L$.*

In order to show that a certain system $S = \left\{_{K \subseteq I} K::p_K$ is complete, for every $K \subseteq I$, we need to construct a proof-matrix M_K for K in S. Suppose $S \models K::p$. Then $freq(K, M_K) \geq p$, since M_K satisfies S. Hence, $p \leq p_K$. Thus, a proof-matrix for K is S shows that the frequency p_K given in the system S cannot be improved. [3]

Example 6 *Let $I = \{A, B, C, D, E, F\}$. Consider the following system: $S = \left\{_{K \subseteq I} K::p_K$, where $p_A = 0.7$, $p_B = 0.5$, $p_{AB} = 0.3$, $p_{DEF} = 0.2$, and $p_K = 0$ for all other itemsets K. The matrix in Fig. 1 satisfies S. S is not complete, because in every matrix satisfying DEF::0.2, the frequency of DE must be at least 0.2, and S contains DE::0. Furthermore, S* does not *logically imply EF::0.5, since R satisfies S, and R does not satisfy EF::0.5.*

Consider the following system over $I = \{A, B, C\}$:

$$\{\phi::1, A::0.6, B::0.8, C::0.8, AB::0.6, AC::0.4, BC::0.6, ABC::0.4\} \quad .$$

This system is complete. In Fig. 2, a possible set of proof-matrices is given.

Notice that when a system is complete, it is not necessary that there exists one matrix that is a proof-matrix for all itemsets at once. Consider for example the following system:

$$\{\phi::1, A::0.5, B::0.5, C::0.1, AB::0, AC::0, BC::0, ABC::0\} \quad .$$

This system is complete. However, we will never find a matrix in which the following six conditions are simultaneously true: $freq(A) = 0.5$, $freq(B) = 0.5$, $freq(C) = 0.1$, $freq(AB) = 0$, $freq(AC) = 0$, and $freq(BC) = 0$, because due to $freq(A) = 0.5$, $freq(B) = 0.5$, and $freq(AB) = 0$, every row has a 1 in A or

[3] Observe the similarities with Armstrong relations in dependency theory [7].

Proof-matrix for B, C, BC, ABC

A	B	C
1	1	0
1	1	1
1	0	1
1	1	1
0	1	1

Proof-matrix for A, AB, AC

A	B	C
1	1	0
1	1	1
1	1	1
0	1	1
0	1	1

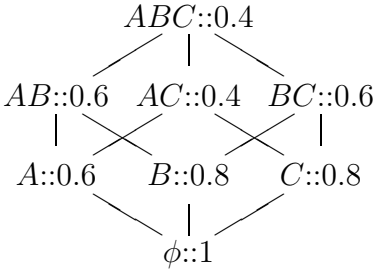

Fig. 2. Proof-matrices for a system of frequent sets

in B. So, every row having a 1 in C has also a 1 in A or a 1 in B, and thus violates either $freq(AC) = 0$, or $freq(BC) = 0$.

3.3 Completion

When a system S is not complete, we can "improve" this system. Suppose $S = \left\{_{K \subseteq I} K::p_K \right.$ is not complete. Then there exists a frequent set expression $K::p'_k$ with $p'_K > p_K$ that is logically implied by S. We can improve S by replacing $K::p_K$ by $K::p'_K$. The next theorem states that for every system S, there exists a unique complete system $C(S)$ logically implied by S.

Theorem 7 *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, and $S = \left\{_{K \subseteq I} K::p_K \right.$ be a system of frequent sets. There exists a unique system $C(S)$, the* completion of S, *such that $S \models C(S)$, and $C(S)$ is a complete system.*

PROOF. Let $L_K = \{p_K \mid S \models K::p_K\}$. L_K always contains its own supremum: suppose a matrix M satisfies S. Let $p := freq(K, M)$. M satisfies S, hence for all $p_K \in L_K$, $p \geq p_K$ holds, and therefore $p \geq sup(L_K)$ holds. Hence, every matrix satisfying S, also satisfies $K::sup(L_K)$, and thus $S \models K::sup(L_K)$. It is now straightforward that the system $\left\{_{K \subseteq I} K::sup(L_K) \right.$ is the unique completion of S. □

Example 8 *$I = \{A, B, C\}$. The system*

$$S_1 = \{\boldsymbol{\phi}\textbf{::1}, A::0.6, B::0.8, C::0.8, AB::0.6, AC::0.4, \boldsymbol{BC}\textbf{::0.6}, ABC::0.4\}$$

is the unique completion of the system

$$S_2 = \{\boldsymbol{\phi}\textbf{::0.8}, A{::}0.6, B{::}0.8, C{::}0.8, AB{::}0.6, AC{::}0.4, \boldsymbol{BC}\textbf{::0.4}, ABC{::}0.4\}\,.$$

$BC{::}0.6$ *is implied by* S_2*, since there is an overlap of at least 0.6 between the rows having a 1 in* B *and the rows having a 1 in* C*.*

4 Axiomatization

We give an axiomatization for frequent sets. An axiomatization in this context is a set of rules that are satisfied by the system if and only if it is complete. In order to simplify the notation we first introduce rare sets. In Section 5 we will show how we can build finite proofs for all logical implications using the axioms.

4.1 Systems of Rare Sets

Definition 9 *Let* $I = \{I_1, \ldots, I_n\}$ *be a set of items.*

- *Let* R *be a matrix over* I*. The* rareness *of an itemset* $K \subseteq I$ *in* R*, denoted* $rare(K, R)$*, is the fraction of rows in* R *that have a 0 in at least one column of* K*.*
- *A* rare set expression *over* I *is an expression* $K{:}p_K$ *with* $K \subseteq I$ *and* p_K *a rational number with* $0 \leq p_K \leq 1$*.*
- *A matrix* R *over* I satisfies $K{:}p_K$*, denoted* $R \models K{:}p_K$*, iff* $rare(K, R) \leq p_K$*. Hence itemset* K *has rareness at most* p_K*.*
- *A* system of rare sets *over* I *is a collection* $\left\{_{K \subseteq I}\ K{:}P_K$ *of rare set expressions, with one expression for each* $K \subseteq I$*.*
- *A matrix* R *over* I satisfies *the system* $S = \left\{_{K \subseteq I}\ K{:}p_K$*, denoted* $R \models S$*, iff* R *satisfies all* $K{:}p_K$*.*
- *A system of rare sets* S logically implies $K{:}p$*, denoted* $S \models K{:}p$ *iff every matrix that satisfies* S *also satisfies* $K{:}p$*. System* S_1 *logically implies system* S_2*, denoted* $S_1 \models S_2$*, iff every* $K{:}p$ *in* S_2 *is logically implied by* S_1*.*
- *A system of rare sets* $S = \left\{_{K \in I}\ K{:}p_K$ *is* complete *iff for each* $K{:}p$ *logically implied by* S*,* $p_K \leq p$ *holds.*

Example 10 *In Fig. 1, the matrix* R *satisfies* $A{:}0.4$*, because fewer than* 0.4 *of the rows have 0 in* A*.* R *does not satisfy* $B{:}0.3$*. Let* $I = \{A, B\}$*. The system* $\{\phi{:}0.4, A{:}0.3, B{:}0.4, \boldsymbol{AB}\textbf{:0.8}\}$ *is not complete. The unique completion of this system is* $\{\phi{:}0, A{:}0.3, B{:}0.4, \boldsymbol{AB}\textbf{:0.7}\}$*.*

The next proposition connects rare sets with frequent sets. The connection between the two is straightforward. Indeed: the rows that have a zero in at least one column on K are exact the complement of the rows having only ones in these columns.

Proposition 11 *Let $I = \{I_1 \dots I_n\}$ be a set of items. For every matrix R over I and every subset K of I holds that*

- *freq*$(K, R) + $ *rare*$(K, R) = 1$.
- *R satisfies $K{:}p_K$ iff R satisfies $K{::}1 - p_K$.*

Notice that a proof-matrix M for an itemset K in a system of frequent sets $\left\{_{K \subseteq I} K{::}p_K$ is also a proof-matrix for K in the system of rare sets $\left\{_{K \subseteq I} K{:}1 - p_K$.

In the following subsection we prove an axiomatization for complete systems of rare sets. From this axiomatization, we can easily derive an axiomatization for frequent sets, using Proposition 11.

4.2 Axiomatization of Rare Sets

We first define bags.

Definition 12 *Let S be a finite set, and $s, s_1, \dots, s_k \in S$.*

(a) A bag over *S is a total function from S into $\mathbb{N}$. Intuitively, a bag is a set in which elements can appear more then once.*
(b) $\mathcal{M} = \langle s_1, \dots, s_k \rangle$ denotes the bag over S where for all $s \in S$, $\mathcal{M}(s)$ is the number of occurrences of s in the list $\langle s_1, \dots, s_k \rangle$. As a shorthand, we denote c occurrences of s by $c \cdot s$.

Let $\mathcal{M}, \mathcal{N}$ be bags over S.

(c) $|\mathcal{M}| := \sum_{s \in S} \mathcal{M}(s)$ is the cardinality of *$\mathcal{M}$.*
(d) s appears *n* times in *$\mathcal{M}$ iff $\mathcal{M}(s) = n$. $s \in \mathcal{M}$ iff $\mathcal{M}(s) \geq 1$.*
(e) The bag-union *$\mathcal{M} \bigcup \mathcal{N}$ is defined as follows: for all $t \in S$, $(\mathcal{M} \bigcup \mathcal{N})(t) = \mathcal{M}(t) + \mathcal{N}(t)$.*
(f) Associate with each element $s \in S$ a real number n_s. $\sum_{s \in \mathcal{M}} n_s$ is shorthand for $\sum_{s \in S} \mathcal{M}(s) n_s$.
(g) Let $\phi(m)$ be a condition on m. $\langle m \in \mathcal{M} \mid \phi(m) \rangle$ denotes the bag $\mathcal{K}$ with for each $s \in S$, $\mathcal{K}(s) = \mathcal{M}(s)$ if $\phi(s)$ holds; else $\mathcal{K}(s) = 0$.

Let $\mathcal{K}$ be a bag over the subsets of S; i.e., the elements of $\mathcal{K}$ are subsets of S.

(h) *$\bigcup\mathcal{K}$ is the following bag over S: $\forall s \in S$, $\bigcup\mathcal{K}(s)$ is the number of occurrences of sets in $\mathcal{K}$ that contain s.*

(i) *The* degree of s in $\mathcal{K}$, *denoted $deg(s,\mathcal{K})$ is $(\bigcup\mathcal{K})(s)$. The* minimal degree of $\mathcal{K}$, *denoted $mdeg(\mathcal{K})$, is $min_{s\in K}(deg(s,\mathcal{K}))$.*

Example 13 *$\mathcal{K} = \langle \{a,b\}, 2\cdot\{b,c\}, 2\cdot\{b,d\} \rangle$ is a bag over the subsets of $\{a,b,c,d\}$. $|\mathcal{K}| = 5$, $\bigcup\mathcal{K} = \langle a, 5\cdot b, 2\cdot c, 2\cdot d\rangle$, $deg(b,\mathcal{K}) = 5$, and $mdeg(\mathcal{K}) = 1$.*

Theorem 14 *Let $S = \left\{_{K\subseteq I}\ K{::}p_K$ be a system of rare sets over I. The following two statements are equivalent:*

- *S is a complete system.*
- *S satisfies*
  $\mathcal{R}_1$ *$p_\phi = 0$*
  $\mathcal{R}_2$ *If $K_2 \subseteq K_1$, then $p_{K_2} \le p_{K_1}$*
  $\mathcal{R}_3$ *Let $K \subseteq I$, $\mathcal{M}$ a bag of subsets of K. Then*

$$p_K \le \frac{\sum_{M\in\mathcal{M}} p_M}{k},$$

with $k = mdeg(\mathcal{M})$. [4]

PROOF. Soundness The soundness of $\mathcal{R}_1$ and $\mathcal{R}_2$ is straightforward. For $\mathcal{R}_3$, let $S = \left\{_{K\in I}\ K{:}p_K$ be a complete system, and let $\mathcal{M}$ be a bag over the subsets of $K \subseteq I$. We prove that $\frac{\sum_{M\in\mathcal{M}} p_M}{k} \ge p_K$, with $k = mdeg(\mathcal{M})$. Let R be a matrix over I such that $R \models S$. Let for all $Z \subseteq I$, D_Z denote the bag $\langle t \text{ row in } R \mid (\exists z \in Z) t[z] = 0\rangle$. Every row $t \in D_K$ appears in at least k of the following bags: $\langle D_M \mid M \in \mathcal{M}\rangle$. Therefore, $k\,|D_K| \le \sum_{M\in\mathcal{M}} |D_M| \le n\sum_{M\in\mathcal{M}} p_M$. Since S is complete, and R was arbitrary, we can conclude $p_K \le \frac{\sum_{M\in\mathcal{M}} p_M}{k}$.
Completeness is proved in Appendix A. □

Example 15 *Consider the following systems:*

$S_1 = \{$ **ϕ:0.2**, A:0.8, B:0.4, C:0.4, AB:0.4, AC:0.4, BC:0.8, ABC:1 $\}$,

$S_2 = \{$ ϕ:0, **A:0.8**, B:0.4, C:0.4, **AB:0.4**, AC:0.4, BC:0.8, ABC:1 $\}$,

$S_3 = \{$ ϕ:0, A:0.4, B:0.4, C:0.4, **AB:0.4**, **AC:0.4**, BC:0.8, **ABC:1** $\}$,

$S_4 = \{$ ϕ:0, A:0.4, B:0.4, C:0.4, AB:0.4, AC:0.4, BC:0.8, ABC:0.8 $\}$.

S_1 is not complete, since ϕ:0.2 violates $\mathcal{R}_1$. S_2 is not complete, since AB:0.4 and A:0.8 violate $\mathcal{R}_2$. The system S_3 is not complete, since AB:0.4, AC:0.4,

[4] If $k = 0$, the trivial condition $p_K \le 1$ is assumed.

and $ABC{:}1$ violate $\boldsymbol{\mathcal{R}_3}$. The system S_4 is complete, since it satisfies $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$. S_4 is the unique completion of S_1, S_2, and S_3.

4.3 Why bags are necessary

In the previous section we proved that $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$ are sound and complete for complete systems of rare set expressions. In rule $\boldsymbol{\mathcal{R}_3}$, we state a condition that has to be tested for all bags over the subsets of all itemsets K. Later on we will show that it is not necessary to test all bags. We will describe a finite class of bags that is sufficient to test. Here we prove that in rule $\boldsymbol{\mathcal{R}_3}$, we cannot change the condition "$\mathcal{M}$ is a bag of subsets of K" into "$\mathcal{M}$ is a set of subsets of K". Therefore we will prove that $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and

$\overline{\boldsymbol{\mathcal{R}_3}}$ = Let $K \subseteq I$, $\mathcal{M}$ a **subset** of 2^K. Then

$$p_K \leq \frac{\sum_{M \in \mathcal{M}} p_M}{k},$$

with $k = mdeg(\mathcal{M})$

are not complete.

Consider the following system of rare sets:

$$S = \left\{ \begin{array}{llll} \phi : 0, & AB : 0.4, & CD : 0.8, & \boldsymbol{ABCD : 1} \\ A : 0.4, & AC : 0.4, & ABC : 0.8, & \\ B : 0.4, & AD : 0.4, & ABD : 0.8, & \\ C : 0.4, & BC : 0.8, & ACD : 0.8, & \\ D : 0.4, & BD : 0.8, & BCD : 0.8, & \end{array} \right\} \quad (1)$$

This system is not complete as can be seen by $\boldsymbol{\mathcal{R}_3}$ with $K = ABCD$ and

$$\mathcal{M} = \langle AB, AC, AD, 2 \cdot BCD \rangle.$$

Application of $\boldsymbol{\mathcal{R}_3}$ gives:

$$p_{ABCD} \leq \frac{p_{AB} + p_{AC} + p_{AD} + 2p_{BCD}}{3} = \frac{14}{15}.$$

However, we will show next that S satisfies $\overline{\boldsymbol{\mathcal{R}_3}}$.

Lemma 16 *Let I be a finite set of items and for each $K \subseteq I$, $p_K \in [0, 1]$. Let*

$S_1, S_2 \subseteq 2^I$, and $S_1 \cap S_2 = \phi$. If $mdeg(S_1) + mdeg(S_2) = mdeg(S_1 \cup S_2)$, then it holds that

$$\frac{\sum_{M \in (S_1 \cup S_2)} p_M}{mdeg(S_1 \cup S_2)} \geq min\left(\frac{\sum_{M \in S_1} p_M}{mdeg(S_1)}, \frac{\sum_{M \in S_2} p_M}{mdeg(S_2)}\right).$$

PROOF. Let $md_1 = mdeg(S_1)$, $md_2 = mdeg(S_2)$, $md_\cup = mdeg(S_1 \cup S_2)$. Without loss of generality, we can assume that

$$\frac{\sum_{M \in S_1} p_M}{md_1} \leq \frac{\sum_{M \in S_2} p_M}{md_2}.$$

$$\begin{aligned}
\frac{\sum_{M \in (S_1 \cup S_2)} p_M}{md_\cup} &= \frac{\sum_{M \in S_1} p_M}{md_\cup} + \frac{\sum_{M \in S_2} p_M}{md_\cup} \\
&= \frac{\sum_{M \in S_1} p_M}{md_1}\frac{md_1}{md_\cup} + \frac{\sum_{M \in S_2} p_M}{md_2}\frac{md_2}{md_\cup} \\
&\geq \frac{\sum_{M \in S_1} p_M}{md_1}\frac{md_1}{md_\cup} + \frac{\sum_{M \in S_1} p_M}{md_1}\frac{md_2}{md_\cup} \\
&= \frac{\sum_{M \in S_1} p_M}{md_1}.
\end{aligned}$$

□

Proposition 17 *The system of rare sets S given in (1) satisfies $\overline{\boldsymbol{\mathcal{R}_3}}$.*

PROOF. Consider the following three matrices.

M_1

A	B	C	D
1	1	1	1
1	1	0	1
1	0	1	1
1	0	1	0
1	1	0	0

M_2

A	B	C	D
1	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
0	1	1	1

M_3

A	B	C	D
1	1	1	1
1	0	0	1
1	0	0	1
1	1	1	0
1	1	1	0

S

ϕ:0	BC:0.8
A:0.4	BD:0.8
B:0.4	CD:0.8
C:0.4	ABC:0.8
D:0.4	ABD:0.8
AB:0.4	ACD:0.8
AC:0.4	BCD:0.8
AD:0.4	$ABCD$:1

M_1 is a proof-matrix for $B, C, D, AB, AC, AD, BC, ABC$, and BCD in S, M_2 is a proof-matrix for A in S, and M_3 is a proof-matrix for ABD, ACD, BD, and CD is S.

These proof-matrices show for all rare set expressions except for $ABCD$:1 that the system S cannot be improved. Since $\boldsymbol{\mathcal{R}_1}, \boldsymbol{\mathcal{R}_2}, \overline{\boldsymbol{\mathcal{R}_3}}$ are sound, the only way in which S can violate $\boldsymbol{\mathcal{R}_1}, \boldsymbol{\mathcal{R}_2}, \overline{\boldsymbol{\mathcal{R}_3}}$ is in $ABCD$ with rule $\overline{\boldsymbol{\mathcal{R}_3}}$. Therefore, to prove the proposition, we need to show that for every set L of subsets of $\{A, B, C, D\}$, the sum $\frac{\sum_{K \in L} p_K}{mdeg(L)}$ is at least 1, and thus p_{ABCD} cannot be improved with rule $\overline{\boldsymbol{\mathcal{R}_3}}$.

Consider the system S' that we get by replacing AB:0.4 by AB:0.8 in S. S' is complete. M_1 is a proof-matrix for B, C, D, AC, AD, BC, ABC, and BCD in S', M_2 is a proof-matrix for A in S', and M_3 is a proof-matrix for ABD, ACD, BD, and CD is S'. Two proof-matrices M_4, and M_5 for respectively AB and $ABCD$ in S' are given next.

M_4

A	B	C	D
1	0	1	1
1	0	1	1
1	1	1	1
0	1	1	1
0	1	1	1

M_5

A	B	C	D
1	0	1	1
1	0	1	1
1	1	0	1
1	1	1	0
0	1	1	1

S'

ϕ:0	BC:0.8
A:0.4	BD:0.8
B:0.4	CD:0.8
C:0.4	ABC:0.8
D:0.4	ABD:0.8
AB:0.8	ACD:0.8
AC:0.4	BCD:0.8
AD:0.4	$ABCD$:1

This completeness of system S' shows that for every set L over the subsets of $ABCD$ that does not contain AB, the sum $\frac{\sum_{K \in L} p_K}{mdeg(L)}$ will be bigger than or equal to 1, because $\overline{\boldsymbol{\mathcal{R}_3}}$ is sound, and S' agrees with S on the frequency of every itemset except for AB, and thus, every expression $ABCD{:}p_{ABCD}$, derived from S without using AB, is also implied by S'.

Since every permutation of B, C, D leaves S unchanged, the same result can be proven for AC and AD.

Consider also the system S'' that we get by replacing BCD:0.8 by BCD:1 in the system S. Again we can show that the resulting system S'' is complete,

with the following proof-matrix M_6 for BCD and $ABCD$ in S''.

M_6

A	B	C	D
1	0	1	1
1	0	1	1
1	1	0	1
1	1	0	1
1	1	1	0

S''

ϕ:0	BC:0.8
A:0.4	BD:0.8
B:0.4	CD:0.8
C:0.4	ABC:0.8
D:0.4	ABD:0.8
AB:0.4	ACD:0.8
AC:0.4	BCD:1
AD:0.4	$ABCD$:1

Therefore, for every set L of subsets of $\{A, B, C, D\}$ that is not a superset of $\{AB, AC, AD, BCD\}$, the sum $\frac{\sum_{K \in L} p_K}{mdeg(L)}$ is at least 1. We will now use Lemma 16 to argue that every superset L of $\{AB, AC, AD, BCD\}$ will also give a sum of at least 1. For every possible superset L of $\{AB, AC, AD, BCD\}$ we will identify a subset L' such that L' has the same degree in A, B, C, or D. Then we can split L into L' and $L'' = L - L'$, and $mdeg(L) = mdeg(L') + mdeg(L'')$. According to Lemma 16, the sum over L will be bigger than the minimum of the sum over L' and the sum over L''. Since in all cases neither L' nor L'' will be supersets of $\{AB, AC, AD, BCD\}$, both sums will be at least 1.

$L = \{AB, AC, AD, BCD\}$ $\qquad$ $\frac{p_{AB}+p_{AC}+p_{AD}+p_{BCD}}{2} = 1$

$A \in L$	$L' = \{A, BCD\}$	$B \in L$	$L' = \{B, AC, AD, BCD\}$
$C \in L$	$L' = \{C, AB, AD, BCD\}$	$D \in L$	$L' = \{D, AB, AC, BCD\}$
$BC \in L$	$L' = \{AB, AC, AD, BC\}$	$ABC \in L$	$L' = \{AD, ABC, BCD\}$
$ABD \in L$	$L' = \{AC, ABD, BCD\}$	$ACD \in L$	$L' = \{AB, ACD, BCD\}$
$ABCD \in L$	$L' = \{ABCD\}$		

□

From Proposition 11, we can easily derive the following axiomatization for frequent sets.

Theorem 18 *Let* $S = \left\{_{K \subseteq I} K::p_K \right.$ *be a system of frequent sets over* I*.* S *is a complete system iff* S *satisfies*

$\boldsymbol{\mathcal{F}_1}$ $p_\phi = 1$,
$\boldsymbol{\mathcal{F}_2}$ *If* $K_2 \subseteq K_1$*, then* $p_{K_2} \geq p_{K_1}$,
$\boldsymbol{\mathcal{F}_3}$ *Let* $K \subseteq I$*,* $\mathcal{M}$ *a bag of subsets of* K*. Then*

$$p_K \geq 1 - \frac{|\mathcal{M}| - \sum_{M \in \mathcal{M}} p_M}{k},$$

with $k = mdeg(\mathcal{M})$.

5 Computability

In the rest of the text we continue working with rare sets. The results obtained for rare sets can, just like the axiomatization, easily be carried over to frequent sets.

In the previous section we introduced and proved an axiomatization for complete systems of rare and frequent sets. There is however still one problem with this axiomatization. $\boldsymbol{\mathcal{R}_3}$ states a property that has to be checked for all bags over the subsets of K. This number of bags is infinite. In this section we show that it suffices to check only a finite number of bags: the minimal multi-covers. We show that the number of minimal multi-covers over a set is finite, and that they can be computed.

We also look at the following problem: when an incomplete system is given, can we compute its completion using the axioms? We show that this computation is indeed possible. We use $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$ as inference rules to adjust rareness values in the system; whenever we detect an inconsistency with one of the rules, we improve the system. When the rules are applied in a systematic way, this method leads to a complete system within a finite number of steps.

Actually, the completion of a system of frequent sets can be computed in an obvious way by using linear programming [10]. For all sets K, we can minimize p_K with respect to a system of inequalities expressing that the frequencies obey the system of rare sets. Since the system of inequalities has polynomial size in the number of frequent itemsets, this algorithm is polynomial in the size of

the system. However, as stated in [8], an axiomatization has as advantage that it provides human-readable proofs, and that, when the inference is stopped before termination, still a partial inference of the frequencies is provided.

5.1 Minimal Multi-covers

In the axiomatization for complete systems of rare sets, $\boldsymbol{\mathcal{R}_3}$ expresses a condition that has to be checked for every bag over the subsets of every itemset. Since the number of bags is infinite, rule $\boldsymbol{\mathcal{R}_3}$ cannot be used in a practical implementation. Therefore, we will show that it is not necessary to check *every bag*, but it suffices to check all *minimal bags*, which are finite in number.

Definition 19

- *A* k-cover *of a set* S *is a bag* $\mathcal{K}$ *over the subsets of* S *such that for all* $s \in S$, $deg(s, \mathcal{K}) = k$.
- *A bag* $\mathcal{K}$ *over the subsets of a set* S *is a* multi-cover *of* S *if there exists an integer* k *such that* $\mathcal{K}$ *is a* k*-cover of* S.
- *A* k*-cover* $\mathcal{K}$ *of* S *is* minimal *if it cannot be decomposed as* $\mathcal{K} = \mathcal{K}_1 \bigcup \mathcal{K}_2$, *with* $\mathcal{K}_1$ *and* $\mathcal{K}_2$ *respectively* k_1*- and* k_2*-covers of* S, $\mathcal{K}_1$ *and* $\mathcal{K}_2$ *not empty.*

Example 20 *Let* $K = \{A, B, C, D\}$. $\langle AB, BC, CD, AD, ABCD\rangle$ *is a* 3*-cover of* K. *It is not minimal, because it can be decomposed into the following two minimal multi-covers of* K: $\langle AB, BC, CD, AD\rangle$ *and* $\langle ABCD\rangle$.

The new rule that replaces $\boldsymbol{\mathcal{R}_3}$ states that it is not necessary to check all bags; we only need to check the minimal multi-covers. This adaptation gives the following $\boldsymbol{\mathcal{R}_3}'$:

$\boldsymbol{\mathcal{R}_3}'$ Let $K \subseteq I$, $\mathcal{M}$ a minimal k-cover of K. Then

$$p_K \leq \frac{\sum_{M \in \mathcal{M}} p_M}{k}.$$

Theorem 21 *Let* S *be a system of rare sets over* I. *The following statements are equivalent:*

(1) S *satisfies* $\boldsymbol{\mathcal{R}_1}, \boldsymbol{\mathcal{R}_2}$, *and* $\boldsymbol{\mathcal{R}_3}$.
(2) S *satisfies* $\boldsymbol{\mathcal{R}_1}, \boldsymbol{\mathcal{R}_2}$, *and* $\boldsymbol{\mathcal{R}_3}'$.

Theorem 22 *Let* K *be a finite set. The minimal multi-covers of* K *are finite in number and computable.*

The proof of these two theorems can be found in Appendix B.

We prove that by applying $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$ as rules, we can compute the completion of any given system.

Applying for example rule $\boldsymbol{\mathcal{R}_2}$ means that whenever we see a situation $K_1 \subseteq K_2$, and the system states $K_1{:}p_{K_1}$ and $K_2{:}p_{K_2}$, and $p_{K_2} < p_{K_1}$, we improve the system by replacing $K_1{:}p_{K_1}$ by $K_1{:}p_{K_2}$. $\boldsymbol{\mathcal{R}_1}$ can only be applied once; $\boldsymbol{\mathcal{R}_2}$ and $\boldsymbol{\mathcal{R}_3}$ never create situations in which $\boldsymbol{\mathcal{R}_1}$ can be applied again.

$\boldsymbol{\mathcal{R}_2}$ is a *top-down operation*, in the sense that the rareness values of smaller sets is adjusted using values of bigger sets. So, for a given system S we can easily reach a fixpoint for rule $\boldsymbol{\mathcal{R}_2}$, by going top-down; we first try to improve the frequencies of the biggest itemsets, before continuing with the smaller ones.

$\boldsymbol{\mathcal{R}_3}$ is a *bottom-up operation*; values of smaller sets are used to adjust the values of bigger sets. So, again, for a given system S, we can reach a fixpoint for rule $\boldsymbol{\mathcal{R}_3}$, by applying the rule bottom-up.

A trivial algorithm to compute the completion of a system is the following: apply $\boldsymbol{\mathcal{R}_1}$, and then keep applying $\boldsymbol{\mathcal{R}_2}$ and $\boldsymbol{\mathcal{R}_3}$ randomly until a fixpoint is reached. The *limit* of this approach yields a complete system, but it is not true that always a fixpoint will be reached within a finite number of steps. In Fig. 3 an infinite run is illustrated. The completion of the system is all rareness values equal to 0, because for every matrix satisfying the system, none of the rows have a 0 in AB, and none have a 0 in BC, so there are no 0's at all in the matrix. When we keep applying the rules as in Fig. 3, we never reach this fixpoint, since in step $2n$, the value for ABC is $\left(\frac{1}{2}\right)^n$. We now will show that when we apply the rules $\boldsymbol{\mathcal{R}_2}$ and $\boldsymbol{\mathcal{R}_3}$ in a systematic way, we always reach a fixpoint within a finite number of steps. This systematic approach is illustrated in Fig. 4. First, we apply $\boldsymbol{\mathcal{R}_2}$ top-down until we reach a fixpoint for $\boldsymbol{\mathcal{R}_2}$, and next, we apply $\boldsymbol{\mathcal{R}_3}$ bottom-up until we reach a fixpoint for $\boldsymbol{\mathcal{R}_3}$. The systematic approach is written down in Fig. 5. We prove that for every system these two meta-steps are all there is needed to reach the completion.

Definition 23 *Let I be a set of items, $J \subseteq I$, and $S = \left\{_{K \subseteq I}\ K{:}p_K$ a system of rare sets over I. The* projection of S on J, *denoted $Proj(S, J)$, is the system $S' = \left\{_{K \subseteq J}\ K{:}p_K$.*

Lemma 24 *Let I be a set of items, $J \subseteq I$, and $S = \left\{_{K \subseteq I}\ K{:}p_K$ a system of rare sets over I.*

(1) If S is complete, then also $Proj(S, J)$ is complete.
(2) if S satisfies $\boldsymbol{\mathcal{R}_2}$, then $Proj(C(S), J) = C(Proj(S, J))$.

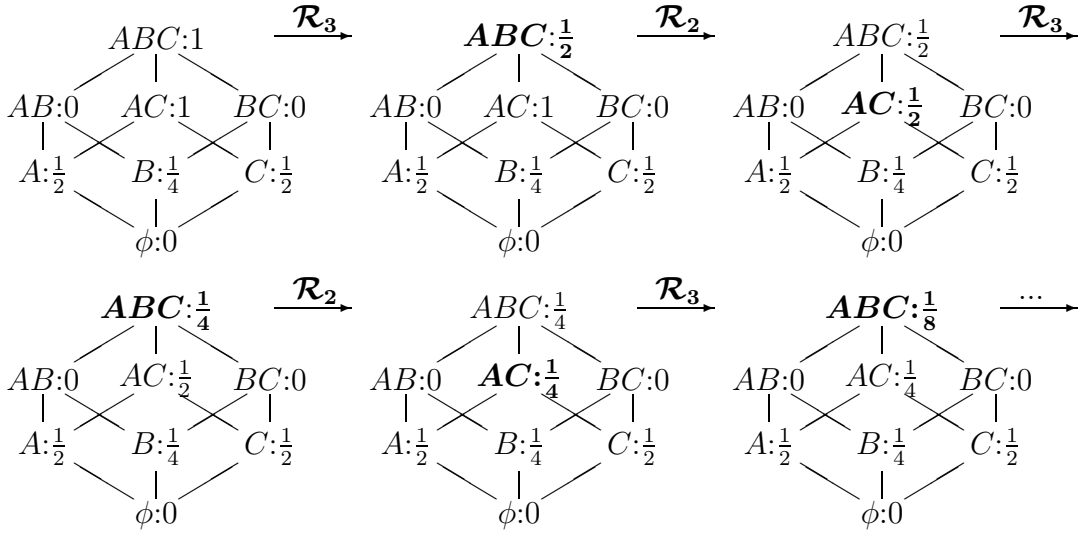

Fig. 3. "Random" application of the rules can lead to infinite loops

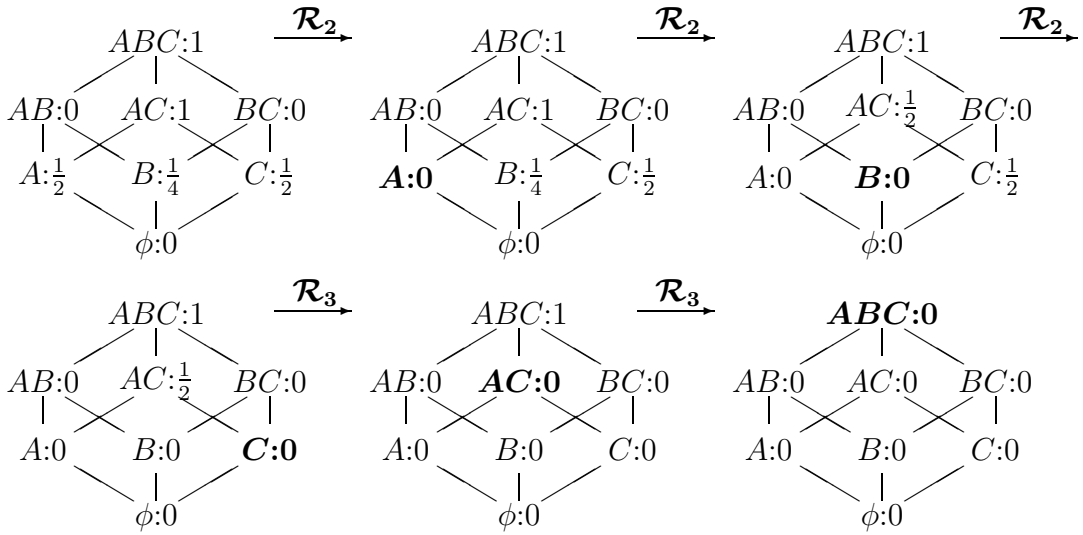

Fig. 4. Systematic application of the rules avoids infinite computations

PROOF. (1) is straightforward.
(2) Let $C(Proj(S,J)) = \left\{_{K\subseteq J}\right. K{:}p_K$. Then, for every $K \subseteq J$, we can construct a proof-matrix R_K, such that $rare(K, R_K) = p_K$, and for all $L \subseteq J$, $rare(L, R_K) \leq p_L$.[5] We will now extend this matrix R_K over J to a proof-matrix $\widehat{R_K}$ of K over I. $\widehat{R_K}$ contains the same number of rows as R_K, and is formed by adding $r(i) = 1$ to each row $r \in R_K$, for all $i \in I - J$. $\widehat{R_K}$ satisfies S, since it is constructed in such a way that for all $L \subseteq I$ holds that

[5] The existence of this proof-matrix can easily be derived from the proof of the completeness of $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$.

Complete(S)
  $p_\phi = 0$
  $T{-}Down(S)$
  $B{-}Up(S)$

$T{-}Down(S)$
  for $i = n$ **downto** 1 **do**
    for all itemsets K of cardinality i **do**
      make $p_K = min_{K \subseteq L}(p_L)$

$B{-}Up(S)$
  for $i = 1$ **to** n **do**
    for all itemsets K of cardinality i **do**
      make $p_K = min_{\mathcal{K}, \text{ minimal } k\text{-cover of } K}\left(\frac{\sum_{K' \in \mathcal{K}} p_{K'}}{k}\right)$

Fig. 5. Algorithm Complete for finding the completion of the system $S = \left\{_{K \subseteq I}\ K{:}p_K \text{ over } I = \{I_1, \ldots, I_n\}\right.$

$rare(L, \widehat{R_K}) = rare(L \cap K, R_K) \leq p_{L \cap K} \leq^{(\boldsymbol{\mathcal{R}_2})} p_L$. Therefore, $C(S)$ must contain $K{:}p_K$, since $\widehat{R_K}$ is a proof-matrix for $K{:}p_K$. □

Theorem 25 *The algorithm in Fig. 5 computes the completion of the system of rare sets S.*

PROOF. We will prove this theorem by induction on $|I|$. In the base case $|I| = 0$ the condition is trivially fulfilled. Suppose the theorem holds for $1, \ldots, |I| - 1$. $B{-}Up(Proj(T{-}Down(S), J)) = Proj(B{-}Up(T{-}Down(S)), J)$ and $T{-}Down(Proj(T{-}Down(S), J)) = Proj(T{-}Down(S), J)$ with $J \subseteq I$. Therefore, for all $J \subset I$ holds:

$$\begin{aligned}
Proj(C(S), J) &= Proj(C(T{-}Down(S)), J) && (S \models T{-}Down(S)) \\
&= C(Proj(T{-}Down(S), J)) && \text{(Lemma 24)} \\
&= B{-}Up(Proj(T{-}Down(S), J)) && \text{(Induction hypothesis)} \\
&= Proj(B{-}Up(T{-}Down(S)), J) && (\boldsymbol{\mathcal{R}_3} \text{ only uses subsets})
\end{aligned}$$

We only need to show now that the rareness value for I in $B{-}Up(T{-}Down(\mathrm{S}))$ equals the rareness value in $C(S)$. This equality is straightforward, since all other rareness values between these two systems are equal, and I can only be adjusted by the bottom-up rule $\boldsymbol{\mathcal{R}_3}$, and this bottom-up rule is applied in the last step of $B{-}Up(.)$. □

6 Sparse Systems

Definition 26 *Let I be a set of items.*

- *A* sparse system *of rare sets is a collection*

$$\left\{_{K \in P}\ K{:}p_K\right.$$

of rare set expressions, with $P \subseteq 2^I$. Hence, not every subset of I has to be present in the system.

- *A matrix R over I* satisfies *a sparse system S iff R satisfies all $K{:}p_K$, $K \in P$.*
- *A sparse system S* logically implies *a rare set expression $K{:}p$, iff every matrix that satisfies S, also satisfies $K{:}p$.*
- *A sparse system $\left\{_{K\in P}\ K{:}p_K$ is* complete *if for all $K{:}p$ with $K \in P$, that are logically implied by the system, $p_K \leq p$ holds.*

The following proposition says that ever complete sparse system can be extended to a full system.

Proposition 27 *Let I be a set of items, and $S = \left\{_{K\in P}\ K{:}p_K$ be a sparse system. The following two statements are equivalent:*

- *S is complete*
- *There exists a complete full system $\overline{S} = \left\{_{K\subseteq I}\ K{:}\widehat{p_K}$, such that for all $K \in P$, $p_K = \widehat{p_K}$ holds.*

PROOF. ($\Rightarrow$) Let R be an arbitrary system satisfying S. Then R satisfies the system $\widehat{S} = \left\{_{K\subseteq I}\ p{:}q_K$, with $q_K = p_K$ if $K \in P$, and $q_K = 1$ else. Hence, R satisfies the complete system $\overline{S} = C(\widehat{S}) = \left\{_{K\subseteq I}\ K{:}c_K$. Therefore, R satisfies the sparse system $\left\{_{K\in P}\ K{:}c_K$. This system has to be equal to S, because S is complete, and $c_K \leq p_K$ for all $K \subseteq I$.
($\Leftarrow$) $\overline{S}$ is complete. Therefore, for every $K \in P$, there exists a proof-matrix R_K such that R_K satisfies $\overline{S}$, and $\mathit{rare}(K, R_K) = \widehat{p_K}$. Since R_K also satisfies S, S must be complete. □

The proposition leads to the following algorithm for computing the completion of the sparse system $S = \left\{_{K\in P}\ K{:}c_K$

(1) Let $F = \left\{_{K\subseteq I}\ K{:}p_K$, with $p_K = c_K$ if $K \in P$, else $p_K = 1$.
(2) Compute the completion $C(F) = \left\{_{K\subseteq I}\ K{:}p'_K$ of F with the methods in Section 5.
(3) Let $C(S) = \left\{_{K\in P}\ K{:}p'_K$.

However, it is clear that when the number of sets in P is small, this approach is not very efficient. Suppose that we are given a sparse system with $|P| = m$ rare set expressions over a set with $|I| = n$ items. To compute the completion, we calculate the completion of a system with 2^n expressions, where the input contained m expressions. The following proposition shows that there are more efficient ways to calculate the completion of a sparse system. It shows that we

do not need all subsets of I.

Theorem 28 *The following are equivalent:*

(1) The sparse system $S = \{K_1{:}p_1, \ldots, K_n{:}p_n\}$ *is complete*
(2) S *satisfies*
$\boldsymbol{\mathcal{S}_1}$ $p_\phi = 0$
$\boldsymbol{\mathcal{S}_2}$ *If* $K_2 \subseteq K_1$, *then* $p_{K_2} \leq p_{K_1}$
$\boldsymbol{\mathcal{S}_3}$ *Let* $\mathcal{M}$ *be a minimal* k*-cover of* K_i. *Then*

$$p_{K_i} \leq \frac{\sum_{M \in \mathcal{M}} min_{M \subseteq K_j}(p_{K_j})}{k}.$$

(3) S *satisfies*
$\boldsymbol{\mathcal{S}_1}$ $p_\phi = 0$
$\boldsymbol{\mathcal{S}_2}$ *If* $K_2 \subseteq K_1$, *then* $p_{K_2} \leq p_{K_1}$
$\boldsymbol{\mathcal{X}}$ *Let* $\mathcal{M}$ *be a bag over* $\{K_j \cap K \mid 0 \leq j \leq n\}$ *with minimal degree* k. *Then*

$$p_K \leq \frac{\sum_{M \in \mathcal{M}} min_{M \subseteq K_j}(p_{K_j})}{k} \quad .$$

The proof is in Appendix C.

6.1 Application of Sparse Systems

Suppose only the frequencies for the single-itemsets are given, and we want to derive a lower bound on itemset K. Using a sparse system, the problem is equivalent to finding the completion of the sparse system

$$S = \Big\{_{k \in I} \ \{k\} :: p_k \cup \Big\{ \ K :: 0.$$

It is easy to see that $C(S)$ contains $K :: (\sum_{k \in K} p_k - (k-1))$, since $\langle \{\{k\} \mid k \in K\rangle$ is the only minimal cover of K using the single-itemsets.

7 Related Work

In artificial intelligence literature, probabilistic logic is studied intensively. The link with this paper is that the frequency of an itemset I can be seen as the probability that a randomly chosen transaction from the transaction database satisfies I; i.e., we can consider the transaction database as an underlying probability structure.

Nilsson introduced in [16] the following *probabilistic logic problem*: given a finite set of m logical sentences $S_1, \ldots, S_m$ defined on a set $X = \{x_1, \ldots, x_n\}$ of n boolean variables with the usual boolean operators $\wedge, \vee$, and $\neg$, together with probabilities $p_1, \ldots, p_m$, does there exists a probability distribution on the possible truth assignments of X, such that the probability of S_i being true, is *exactly* p_i for all $1 \leq i \leq m$. *Georgakopoulos et al.* prove in [9] that this problem, they suggest the name *probabilistic satisfiability problem* (PSAT), is NP-complete. This problem, however, does not apply to our framework. In our framework, a system of frequent sets can *always* be satisfied. Indeed, since a system only gives *lower* bounds on the frequencies, the system is always satisfied by a transaction database where each transaction contains every item.

Another, more interesting problem, also stated by *Nilsson* in [16], is that of *probabilistic entailment*. Again a set of logical sentences $S_1, \ldots, S_m$, together with probabilities $p_1, \ldots, p_m$ is given, and one extra logical sentence S_{m+1}, the target. It is asked to find best possible upper and lower bounds on the probability that S_{m+1} is true, given $S_1, \ldots, S_m$ are satisfied with respective probabilities $p_1, \ldots, p_m$. The interval defined by these lower and upper bounds forms the so-called *tight entailment* of S_{m+1}. It is well known that both PSAT and probabilistic entailment can be solved nondeterministically in polynomial time using linear programming techniques. In our framework, a complete system of frequent sets is a system that only contains tight frequent expressions; i.e., the bounds of the frequent expressions in the complete system are the best possible in view of the system, and as such, completeness corresponds to the notion of tight entailment.

For a comprehensive overview of probabilistic logic, probabilistic entailment and various extensions, we refer to [12,13]. Nilsson's probabilistic logic and entailment are extended in various ways, including assigning intervals to logical expressions instead of exact probability values and also considering conditional probabilities [8][15].

In [6], *Fagin et al.* study the following extension. A *basic weight formula* is an expression $a_1 w(\phi_1) + \ldots + a_k w(\phi_k) \geq c$, where $a_1, \ldots, a_k$ and c are integers and $\phi_1, \ldots, \phi_k$ are propositional formulas, meaning that the sum of all a_i times the *weight* of ϕ_i is greater than or equal to c. A *weight formula* is a boolean combination of basic weight formulas. The semantics are introduced by an underlying probability space. The weight of a formula corresponds with the probability that it is true. The main contribution (from the viewpoint of our paper) of [6] is the description of a sound and complete axiomatization for this probabilistic logic. The logical framework in our paper is embedded into the logic in [6]. Indeed, if we introduce a propositional symbol P_i for each item i, the frequent set expression $K{::}p_K$ can be translated as $w(\bigwedge_{i \in K} P_i) \geq p_K$. As such, by results obtained in [6], the implication problem in our framework is guaranteed to be decidable. Satisfiability, and thus also the implication prob-

lem, are NP-complete in Fagin's framework. Our approach differs from Fagin's approach in the sense that we only consider situations where for all expressions a probability is given and that we only consider conjunctive expressions.

Also in [8], axioms for a probabilistic logic are introduced. However, the authors are unable to prove completeness of the axioms. For a restricted sub-language (Type-A problems), they prove that their set of axioms is complete. However, this sub-language is not sufficient powerful to express frequent itemset expressions.

On the other side of the spectrum, we have related work within the context of data mining. There have been attempts to prove some completeness results for itemsets in this area. One such attempt is described shortly in [14]. In the presence of constraints on the allowable itemsets, the authors introduce the notion of *ccc-optimality* [6] . ccc-optimality can intuitively be understood as "the algorithm only generates and tests itemsets that still can be frequent, using the current knowledge." Our approach however, is more general, since we do not restrict ourselves to a particular algorithm.

In [3], the author considers a sound rule for inferencing lower bounds. Using this bound, one can discover large frequent patterns without having to consider all subsets first. This observation leads to an efficient algorithm for finding large itemsets, called Max-Miner. No attempt however is known to us in the context of data mining, that studies in a systematic way what we can derive from an arbitrary set of frequent itemsets.

8 Conclusion

We presented an axiomatization for complete systems of frequent sets. As an intermediate stage in the proofs, we introduced the notion of a system of rare sets. The axiomatization for rare sets contains three rules $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$. From these rules we easily derive the axiomatization, $\boldsymbol{\mathcal{F}_1}$, $\boldsymbol{\mathcal{F}_2}$, and $\boldsymbol{\mathcal{F}_3}$ for frequent sets. Because rule $\boldsymbol{\mathcal{R}_3}$ yields a condition that needs to be checked for an infinite number of bags, we replaced $\boldsymbol{\mathcal{R}_3}$ by $\boldsymbol{\mathcal{R}_3}$'. We showed that the completion of a system can be computed by applying $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$' as inference rules. If these rules first are applied top-down, and then bottom-up, the completion is reached within a finite number of steps. We also studied sparse systems, where not for every itemset a rareness was given. We adjusted the rules $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$ such that a more efficient calculation of the completion is possible.

[6] ccc-optimality stands for Constraint Checking and Counting-optimality

References

[1] R. Agrawal, T. Imilienski and A. Swami, Mining association rules between sets of items in large databases, in: P. Buneman and S. Jajodia, eds., *Proceedings ACM SIGMOD* (Washington, D.C, 1993) 207–216.

[2] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, in: J.B. Bocca, M. Jarke and C. Zaniolo, eds., *Proceedings VLDB* (Santiago de Chile, Chile, 1994) 487–499.

[3] R. J. Bayardo, Efficiently Mining Long Patterns from Databases, in: L.M. Haas and A. Tiwary, eds., *Proceedings ACM SIGMOD* (Seattle, Washington, 1998) 85–93.

[4] T. Calders and J. Paredaens, A Theoretical Framework for Reasoning about Frequent Itemsets, Technical Report 00-6, Department Mathematics and Computer Science, Universiteit Antwerpen, Belgium, http://win-www.uia.ac.be/u/calders/download/axiom.ps, 2000

[5] T. Calders and J. Paredaens, Axiomatization of frequent sets, in: J. Van den Bussche and V. Vianu, eds., *Proceedings ICDT* (London, UK, 2001) 204–218.

[6] R. Fagin, J. Halpern and N. Megiddo, A Logic for Reasoning about Probabilities, *Information and Computation* **87(1,2)** (1990) 78–128.

[7] R. Fagin and M. Y. Vardi, Armstrong Databases for Functional and Inclusion Dependencies, *Information Processing Letters* **16(1)** (1983) 13–19.

[8] A. M. Frisch and P. Haddawy, Anytime Deduction for Probabilistic Logic, *Artificial Intelligence* **69(1-2)** (1994) 93–112.

[9] G. Georgakopoulos, D. Kavvadias and C. H. Papadimitriou, Probabilistic Satisfiability, *Journal of Complexity* **4** (1988) 1–11.

[10] G. Hadley, *Linear Programming* (Addison-Wesley, Reading, Mass., 1962).

[11] J. Han, J.Pei and Y. Yin, Mining frequent patterns without candidate generation, in: W. Chen, J.F. Naughton and P.A. Bernstein, eds., *Proceedings ACM SIGMOD* (Dallas, TX, 2000) 1–12.

[12] P. Hansen, B. Jaumard, G.-B. D. Nguets and M. P. de Aragäo. Models and Algorithms for Probabilistic and Bayesian Logic. in: *Proc. IJCAI'95* (Montreal, Canada, 1995) 1862–1868.

[13] P. Hansen, B. Jaumard. Probabilistic Satisfiability. *Les Cahiers du GERAD* **G-96-31** (1996).

[14] L. V.S. Laksmanan, R.T. Ng, J. Han and A. Pang, Optimization of Constrained Frequent Set Queries with 2-variable Constraints, in: A. Delis, C. Faloutsos and S. Ghandeharizadeh, eds., *Proc. ACM SIGMOD* (Philadelphia, Pennsylvania, 1999) 157–168.

[15] T. Lucasiewicz, Local probabilistic deduction from taxonomic and probabilistic knowledge-bases over conjunctive events, *Journal of Approximate Reasoning* **21** (1999) 23–61.

[16] N. Nilsson, Probabilistic Logic, *Artificial Intelligence* **28** (1986) 71–87.

[17] J. Paredaens, Axiomatization of frequent sets, Technical Report 99-11, Universiteit Antwerpen, Department of Mathematics and Computer Science, Belgium, 1999.

A Proof of Theorem 14

Lemma 29 *Let $S = \left\{_{K\subseteq I}\ K{:}p_K$ be a system that satisfies $\boldsymbol{\mathcal{R}_1}$ and $\boldsymbol{\mathcal{R}_2}$. If for all $K \subseteq I$, the system*

$$\left\{ p_K - p_L \leq \sum_{a\in K} X_a - \sum_{a\in L} X_a \leq p_K, \quad \forall L \subseteq K \right. \tag{A.1}$$

has a rational solution, then S is complete.

PROOF. Let $K \subseteq I$. We show that there exists a proof-matrix R for K. Let $(\forall a \in K)X_a = \beta_a$ be a solution of (A.1). We have $(\forall a \in K)0 \leq \beta_a \leq 1$, and $\sum_{a\in K} \beta_a = p_K \leq 1$ (from the case $L = \{\}$.) Let R be a matrix satisfying: (a) a fraction β_a of the rows has a 0 in column a, and a 1 elsewhere, with $a \in K$; (b) a fraction $1 - \sum_{a\in K} \beta_a$ of the rows has a 1 in all columns. Because all β_a's are rational, such a matrix exists. R is a proof-matrix for $K{:}p_K$. □

Lemma 30 *Given a set of items I and given a_K, b_K, positive rational numbers, for every non-empty $K \subseteq I$. Consider the following system of inequalities:*

$$a_K \leq \sum_{i\in K} X_i \leq b_K, \quad \forall L \subseteq K$$

This system has a solution $(x_1, \ldots, x_{|I|})$, x_i rational, iff for all $\mathcal{K}$ and $\mathcal{L}$, bags of subsets of I with $\bigcup\mathcal{K} = \bigcup\mathcal{L}$ it is true that $\sum_{K\in\mathcal{K}} a_K \leq \sum_{L\in\mathcal{L}} b_L$.

PROOF. We will use induction on $|I|$.

$|I| = 0$ Trivially fulfilled.
General case Suppose the lemma holds for $1, 2, \ldots, |I| - 1$. Let $1 \in I$, and

$$\begin{aligned} UB &:= \{(\textstyle\sum_{L\in\mathcal{L}} b_L - \sum_{K\in\mathcal{K}} a_K)/\alpha \mid \bigcup\mathcal{K} \cup \langle\alpha \cdot 1\rangle = \bigcup\mathcal{L}\} \\ LB &:= \{(\textstyle\sum_{L\in\mathcal{L}} a_L - \sum_{K\in\mathcal{K}} b_K)/\alpha \mid \bigcup\mathcal{K} \cup \langle\alpha \cdot 1\rangle = \bigcup\mathcal{L}\} \end{aligned} \tag{A.2}$$

We show that $max(LB) \leq min(UB)$. Let $\mathcal{K}, \mathcal{L}, \alpha, \mathcal{K}', \mathcal{L}', \alpha'$ be such that

$$\bigcup \mathcal{K} \cup \{\alpha \cdot 1\} = \bigcup \mathcal{L}, \text{ and } \bigcup \mathcal{K}' \cup \{(\alpha') \cdot 1\} = \bigcup \mathcal{L}'.$$

Then $\bigcup(\alpha' \mathcal{L} \cup \alpha \mathcal{K}') = \bigcup(\alpha \mathcal{L}' \cup \alpha' \mathcal{K})$ is true. Therefore

$$\alpha' \sum_{L \in \mathcal{L}} a_L + \alpha \sum_{K \in \mathcal{K}'} a_K \leq \alpha \sum_{L \in \mathcal{L}'} b_L + \alpha' \sum_{K \in \mathcal{K}} b_K,$$

and thus

$$(\sum_{L \in \mathcal{L}} a_L - \sum_{K \in \mathcal{K}} b_K)/\alpha \leq (\sum_{L \in \mathcal{L}'} b_L - \sum_{K \in \mathcal{K}'} a_K)/\alpha'.$$

Choose now β_1 rational s.t. $max(LB) \leq \beta_1 \leq min(UB)$.

Consider the following system A.3(X_1 has been replaced by β_1), $a'_K = max(a_K, a_{(K \cup \{1\})} - \beta_1)$, and $b'_K = min(b_K, b_{(K \cup \{1\})} - \beta_1)$, for all $K \subseteq I - \{1\}$.

$$\begin{cases} a'_K \leq \sum_{k \in K} X_k \leq b'_K, & \forall K \subseteq I - \{1\} \end{cases} \tag{A.3}$$

We use induction to show this system has a solution. Therefore, we need to show that whenever $\bigcup \mathcal{K} = \bigcup \mathcal{L}$,

$$\sum_{K \in \mathcal{K}} max(a_K, a_{K \cup \{1\}} - \beta_1) \leq \sum_{L \in \mathcal{L}} min(b_L, b_{L \cup \{1\}} - \beta_1) \tag{A.4}$$

holds. Let $\mathcal{K} = \mathcal{K}' \bigcup \mathcal{K}''$, $\mathcal{L} = \mathcal{L}' \bigcup \mathcal{L}''$, where

$$\begin{aligned} \mathcal{K}' &= \langle K \in \mathcal{K} \mid a_K < a_{K \cup \{1\}} - \beta_1 \rangle, \\ \mathcal{L}' &= \langle L \in \mathcal{L} \mid a_L < a_{L \cup \{1\}} - \beta_1 \rangle. \end{aligned} .$$

Suppose $|\mathcal{L}'| > |\mathcal{K}'|$. Then we have

$$\overbrace{\bigcup_{L \in \mathcal{L}'} (L \cup \{1\}) \cup \bigcup \mathcal{L}''}^{\mathcal{N}} = \overbrace{\bigcup_{K \in \mathcal{K}'} (K \cup \{1\}) \cup \bigcup \mathcal{K}''}^{\mathcal{M}} \cup (|\mathcal{L}'| - |\mathcal{K}'| \{1\}).$$

Since $\beta_1 \geq max(LB)$,

$$\beta_1 \geq \frac{\sum_{M \in \mathcal{M}} a_M - \sum_{N \in \mathcal{N}} b_N}{|\mathcal{L}'| - |\mathcal{K}'|}$$

holds. In case $|\mathcal{L}'| < |\mathcal{K}'|$, a similar argument can be used, but with UB instead of LB (A.2). Therefore, (A.4) holds, and by induction the second system has a solution $\beta_2, \ldots, \beta_{|I|}$. It is easy to see that $\beta_1, \ldots, \beta_{|I|}$ is a solution for the original system.

□

Lemma 31 *Let* $S = \left\{_{K \subseteq I} K{:}p_K\right.$. *If* S *satisfies* $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, *and* $\boldsymbol{\mathcal{R}_3}$, *then the system*

$$\left\{ p_K - p_L \leq \sum_{a \in K} X_a - \sum_{a \in L} X_a \leq p_K, \quad \forall L \subseteq K \right. \tag{A.5}$$

has a rational solution.

PROOF. According to Lemma 30, (A.5) has a solution iff for all bags $\mathcal{M}$ and $\mathcal{N}$ over the subsets of K, such that $\bigcup \mathcal{M} = \bigcup \mathcal{N}$,

$$\sum_{M \in \mathcal{M}} p_K - p_{K-M} \leq \sum_{N \in \mathcal{N}} p_N$$

holds. Let $\mathcal{L} = \mathcal{N} \bigcup \langle K - M \mid M \in \mathcal{M} \rangle$.
Then, by $\boldsymbol{\mathcal{R}_3}$ we have that $\sum_{L \in \mathcal{L}} p_L \geq k p_K$, with

$$k = min_{a \in K} \left| \langle N \mid a \in N \wedge N \in \mathcal{N} \rangle \bigcup \langle M \mid M \in \mathcal{M} \wedge a \notin M \rangle \right|.$$

Because $|\langle M \mid M \in \mathcal{M} \wedge a \in M \rangle| = |\langle N \mid N \in \mathcal{N} \wedge a \in n \rangle|$, it follows that $k = |\mathcal{M}|$.
Therefore, $\sum_{L \in \mathcal{L}} p_L \geq |\mathcal{M}|\, p_K$ holds.
Since

$$\sum_{L \in \mathcal{L}} p_L = \sum_{N \in \mathcal{N}} p_N + \sum_{M \in \mathcal{M}} p_{K-M},$$

and $|\mathcal{M}|\, p_K = \sum_{M \in \mathcal{M}} p_K$,
$\sum_{M \in \mathcal{M}} p_K - p_{K-M} \leq \sum_{N \in \mathcal{N}} p_N$ holds. □

Completeness If S satisfies $\boldsymbol{\mathcal{R}_1}$, $\boldsymbol{\mathcal{R}_2}$, and $\boldsymbol{\mathcal{R}_3}$, then

$$\left\{ p_K - p_L \leq \sum_{a \in K} X_a - \sum_{a \in L} X_a \leq p_K, \quad \forall L \subseteq K \right.$$

has a rational solution (Lemma 31.) Therefore, S is complete (Lemma 29.) □

B Proof of Theorem 21 and 22

Lemma 32 *Let $a_1, \ldots a_n, b_1, \ldots, b_n$ be strict positive reals. Then $\frac{a_1+\ldots+a_n}{b_1+\ldots+b_n} < p$ implies that at least for one i, $\frac{a_i}{b_i} < p$ holds.*

Lemma 33 *Every k-cover $\mathcal{M}$ can be decomposed into a number of minimal multi-covers $\mathcal{M}_1, \ldots, \mathcal{M}_n$, such that $\bigcup_{i=1\ldots n} \mathcal{M}_i = \mathcal{M}$.*

Theorem 21 $1 \Rightarrow 2$ is trivial, since $\boldsymbol{\mathcal{R}_3}$' is more specific than $\boldsymbol{\mathcal{R}_3}$. Suppose that the system $S = \left\{_{K \subseteq I} K{:}p_K \right.$ satisfies $\boldsymbol{\mathcal{R}_1}$ and $\boldsymbol{\mathcal{R}_2}$, but does not satisfy $\boldsymbol{\mathcal{R}_3}$. We will show that it is impossible that it satisfies $\boldsymbol{\mathcal{R}_3}$'.

There must be a set $K \subseteq I$, and a bag $\mathcal{M}$ over the subsets of K, such that $p_K > \frac{\sum_{m \in \mathcal{M}} p_M}{k}$ with $k = min_{a \in K}(deg(a, \mathcal{M}))$. For each $a \in K$ such that $deg(a, \mathcal{M}) > k$, we replace $deg(a, \mathcal{M}) - k$ of the sets $A \in \mathcal{M}$ that contain a by $A - \{a\}$. In this way, we construct a k-cover $\mathcal{M}'$ of K.
Because S satisfies $\boldsymbol{\mathcal{R}_2}$, $\sum_{M \in \mathcal{M}} p_M \geq \sum_{M \in \mathcal{M}'} p_M$. The k-cover $\mathcal{M}'$ can be decomposed into different minimal multi-covers $\mathcal{M}_1, \ldots, \mathcal{M}_n$ of K, with M_i a k_i-cover of K (Lemma 33). Because $\frac{\sum_{m \in \mathcal{M}'} p_M}{k} = \frac{\sum_{M \in \mathcal{M}_1} p_M + \ldots + \sum_{M \in \mathcal{M}_n} p_M}{k_1 + \ldots + k_n}$, for at least one i, $\frac{\sum_{M \in \mathcal{M}_i}}{k_i} < p_K$ must hold (Lemma 32.) Therefore, $\boldsymbol{\mathcal{R}_3}'$ is violated. □

Lemma 34 *Let C be a positive integer, and let $\mathcal{N}$ be a bag over $\{-C, -C+1, \ldots, -1, 0, 1, \ldots, C-1, C\}$ with $\sum_{n \in \mathcal{N}} n = 0$. If $|\mathcal{N}| > 2C^3$, then there exists $\mathcal{M} \subset \mathcal{N}$ ($\phi \neq \mathcal{M} \neq \mathcal{N}$!), with $\sum_{m \in \mathcal{M}} m = 0$.*

PROOF. If $0 \in \mathcal{N}$, the lemma clearly holds. Assume $0 \notin \mathcal{N}$. $\mathcal{N}^+ = \langle n \in \mathcal{N} \mid n > 0 \rangle$, $\mathcal{N}^- = \langle n \in \mathcal{N} \mid n < 0 \rangle$. At least one of $\mathcal{N}^-$ and $\mathcal{N}^+$ contains more than C^3 elements. Assume that $|\mathcal{N}^+| \geq C^3$. Therefore, there is at least one positive integer p that occurs C times. Because the sum of the elements in $\mathcal{N}^+$ is at least C^3, the sum of the elements in $\mathcal{N}^-$ is at most $-C^3$. Therefore, there are at least C^2 elements in $\mathcal{N}^-$, and thus there is a negative element n such that $deg(n, \mathcal{N}^-) \geq C$. (The same result obtains if $|\mathcal{N}^-| \geq C^3$.) It is also clear that $|n|p = -pn$, and thus the bag $\langle |n| \cdot p, p \cdot n \rangle$ has sum 0, and is a non-empty subbag of $\mathcal{N}$. □

Theorem 22 We will prove this theorem by induction on the size of K.
Base case $|K| = 1$. Trivial, since $\langle K \rangle$ is the only minimal multi-cover of K.
General case. We assume by induction the theorem holds for sets L with size up to $|K|-1$. Thus, the degree and the cardinality of a minimal multi-cover of a set L of cardinality smaller than $|K|$ is bounded, since there is only a finite

number of them. Let d, c be the respective bounds on the degree and the cardinality of the minimal multi-covers of sets of cardinality at most $|K|-1$.

Let $\mathcal{K}$ be a minimal k-cover of K, $a \in K$. It is clear that $\mathcal{L} = proj(\mathcal{K}, K - \{a\}) := \langle S - \{a\} \mid S \in \mathcal{K}\rangle$ is a (not necessarily minimal) multi-cover of $K - \{a\}$. According to Lemma 33, we can split $\mathcal{L} = \mathcal{L}_1 \cup \ldots \cup \mathcal{L}_n$ with $\mathcal{L}_i$ a minimal l_i-cover of $K - \{a\}$. Therefore, we can split $\mathcal{K} = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_n$ with $\mathcal{L}_i = proj(\mathcal{K}_i, K - \{a\})$ a minimal l_i-cover of $K - \{a\}$ (note however that this decomposition of $\mathcal{K}$ is not necessarily unique). By induction, $l_i \leq d$ and $|\mathcal{L}_i| \leq c$. Consider now the bag $\mathcal{M} = \langle l_1 - deg(a, \mathcal{K}_1), \ldots, l_n - deg(a, \mathcal{K}_n)\rangle$. The sum of the bag is 0, since $\sum_{i=1}^{n} l_i = k = \sum_{i=1}^{n} deg(a, \mathcal{K}_i)$. Notice also that $-c \leq l_i - deg(a, \mathcal{K}_i) \leq d \leq c$. Because $\mathcal{K}$ is minimal, for every subbag not equal to $\mathcal{M}$, the sum is not 0, otherwise the union of the $\mathcal{K}_i$'s that correspond to this subbag, would be a multi-cover of K, and thus $\mathcal{K}$ would not be minimal. Therefore, via Lemma 34, the cardinality of $\mathcal{M}$ is bounded by $2c^3$. Thus, $|\mathcal{K}| \leq 2c^4$. Therefore, there are at most $2^{2|K|c^4}$ minimal multi-covers of K and thus the number of minimal multi-covers is finite. □

C Proof of Theorem 28

PROOF. $1 \Leftrightarrow 2$ **Soundness** of the three rules is straightforward.
Completeness. Suppose the sparse system $S = \left\{_{K \in P}\right. K{:}p_K$ satisfies $\boldsymbol{\mathcal{S}_1}$, $\boldsymbol{\mathcal{S}_2}$, and $\boldsymbol{\mathcal{S}_3}$. Let $S' = \left\{_{K \subseteq I}\right. K{:}p'_K$, with $p'_K = p_K$ if $K \in P$, and $p'_K = 1$ else. Suppose $C(S') = \left\{_{K \subseteq I}\right. K{:}q_K$. We will show by contradiction that for all $K \in P$, $p_K = q_K$ holds. Suppose there is a $K \in P$ such that $p_K \neq q_K$. $C(S') = B{-}Up(T{-}Down(S'))$ (Theorem 25.) Since S satisfies $\boldsymbol{\mathcal{S}_1}$ and $\boldsymbol{\mathcal{S}_2}$, the rareness of K in $C(S')$ comes from the bottom-up step, and thus there exists a minimal k-cover $\mathcal{K}$ over the subsets of K, such that $\frac{\sum_{L \in \mathcal{K}} q_L}{k} < p_K$. The q_L's in this step can on their turn be obtained in the top-down step, or in the bottom-up step. If q_L was obtained in the top-down step, then it is easy to see that $q_L = min_{L \subseteq K_i} p_{K_i}$; i.e. the minimum rareness of all supersets of L that were given as input. In the other case, q_L was obtained by a bottom-up step. In that case, there exists a minimal l-cover $\mathcal{L}$ over the subsets of L, such that $q_L = \sum_{L' \in \mathcal{L}} q_{L'}$. We now construct a kl-cover $\mathcal{K}'$ of K as follows: $\mathcal{K}' = (\mathcal{K} - \langle L\rangle) \bigcup \mathcal{L}$. $\mathcal{K}'$ is a kl-cover. In this way we can get rid of all q_L's that were obtained by application of a bottom-up step, because we can iteratively replace each q_L that was obtained by application of $\boldsymbol{\mathcal{R}_3}$, by a sum of $q_{L'}$'s, where all $L' \subset L$. When these L' are obtained by $\boldsymbol{\mathcal{R}_3}$, we can replace them by $q_{L''}$ of even smaller sets L''. Since the singleton sets can only be obtained by $\boldsymbol{\mathcal{R}_2}$, this recursion must stop, and thus there exists an m-cover $\mathcal{M}$ such that $\frac{\sum_{M \in \mathcal{M}} q_M}{m} < p_K$, and all q_M's are obtained by $\boldsymbol{\mathcal{R}_2}$. As such, for all M, $q_M = min_{M \subseteq K_i} p_{K_i}$, and thus $\frac{\sum_{M \in \mathcal{M}} min_{M \subseteq K_i} p_{K_i}}{m} < p_K$. There is still one problem: $\mathcal{M}$ is not necessarily minimal. We can cope with this problem in exactly the same

way as at the end of the proof of Theorem 21.

$2 \Leftrightarrow 3$ Suppose system $S = \{K_1{:}p_1, \ldots, K_n{:}p_n\}$ satisfies $\boldsymbol{\mathcal{S}_1}$, $\boldsymbol{\mathcal{S}_2}$, but does not satisfy $\boldsymbol{\mathcal{S}_3}$. We will show that it also does not satisfy $\boldsymbol{\mathcal{X}}$. Hence, there exists a bag $\mathcal{M}$ with minimal degree k and a set K such that $p_{K_i} > \frac{\sum_{M \in \mathcal{M}} min_{M \subseteq K_j}(p_{K_j})}{k}$. For each $M \in \mathcal{M}$, fix a $K_M \in \{K_1, \ldots, K_n\}$, such that $M \subseteq K_M$, and $p_{K_M} = min_{M \subseteq K_j}(p_{K_j})$. Let $\mathcal{K}$ be the following bag: $\langle K_M \cap K \mid M \in \mathcal{M} \rangle$. The minimal degree of $\mathcal{K}$ is at least k (since $M \subseteq K_M \cap K$ for all $M \in \mathcal{M}$), and hence $p_{K_i} > \frac{\sum_{K \in \mathcal{K}} min_{M \subseteq K_j}(p_{K_j})}{mdeg(\mathcal{K})}$. This inequality is a violation of $\boldsymbol{\mathcal{X}}$.

The other direction is trivial, since $\boldsymbol{\mathcal{S}_3}$ is equivalent to:

Let $\mathcal{M}$ be a bag over the subsets of K with minimal degree k. Then

$$p_K \leq \frac{\sum_{M \in \mathcal{M}} min_{M \subseteq K_j}(p_{K_j})}{k} \quad .$$

Since $\boldsymbol{\mathcal{X}}$ is a specialization of this rule, $\boldsymbol{\mathcal{X}}$ holds whenever $\boldsymbol{\mathcal{S}_3}$ holds. $\square$