A Theoretical Framework for Reasoning about Frequent Itemsets

Toon Calders* Jan Paredaens

Universiteit Antwerpen
Departement Wiskunde-Informatica,
Universiteitsplein 1,B-2610 Wilrijk, Belgium
{calders,pareda}@uia.ua.ac.be

Technical report TR006, June 2000

Abstract

In data mining association rules are very popular. Most of the algorithms in the literature for finding association rules start by searching for frequent itemsets. In this paper we consider frequent set expressions. A frequent set expression is a pair containing an itemset and a frequency indicating that the frequency of that itemset is greater than or equal to the given frequency. A system of frequent sets is a collection of such expressions. We give and prove an axiomatization for these systems. This axiomatization characterizes *complete systems*. A system is complete when it explicitly contains all information that it logically implies. Every system of frequent sets has a unique completion. We show that this completion is computable. We proof that deciding completeness is in coNP. Finally we also study some special cases.

*Research Assistant of the Fund for Scientific Research - Flanders (Belgium)(F.W.O. - Vlaanderen).

Contents

1 Introduction

Association rules are one of the most studied topics in data mining. They have many applications [3]. Since their introduction, many algorithms have been proposed to find association rules [1][2][5].

We first give the formal definition of the association mining problem as stated in [1]: Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ be a set of literals, 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 itemsets that are contained in at least a fraction s of the transactions. To optimize the search for frequent itemsets, most algorithms use the following monotonicity principle [6]:

> 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*.

Although the monotonicity of frequency is commonly used, there is to our knowledge no previous work that discusses whether this rule is *complete*, in the sense that it tells us everything we can derive from a set of given frequencies. 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, $A :: 0.6, B :: 0.6, AB :: 0.1, \phi :: 0.5$ is a system of frequent sets. In this system, $A :: 0.6$ expresses the knowledge that itemset A has a frequency of at least 0.6. In this case, the system can be improved. Indeed: from the system we can conclude that $AB :: 0.2$ since $A :: 0.6$ and $B :: 0.6$ and there must be an overlap of at least a 0.2-fraction between the transactions containing A and the transactions containing B. we can also improve $\phi :: 0.5$, because $\phi :: 1$ always holds. Therefore, this system is called incomplete. When a system cannot be improved, it is complete.

In this paper we give three rules **F1**, **F2**, and **F3**, for complete systems of frequent sets. A system is complete iff it satisfies **F1**, **F2**, **F3**. After a small modification of **F3**, we conclude that the question whether a system is complete, is decidable. We also show that for every system there is a unique equivalent system that is complete. We prove that this unique *completion* of a system is computable.

To facilitate the notations in the proofs, we introduce *rare sets*. A rare set expression $K : p_K$ expresses that at most a p_K-fraction of the transactions does not contain at least one item of K.

We proof that deciding completeness is in coNP.

Finally we study some special cases, in which not all frequencies are given. We call such systems *sparse*. For example, the system $S = \{A :: 0.5, B ::$

$0.3, C :: 0.5, AB :: 0.3\}$ is a sparse system, because there are no rare set expressions for AC, AB, ABC and ϕ. We proof that for sparse systems, completeness is still decidable and computable, and we give an axiomatization for sparse systems.

Although the results in this paper cannot directly be used to improve existing algorithms, we strongly believe that a theoretical framework on the implications between frequent itemsets is an interesting and important topic in data mining.

The structure of the paper is as follows: in Section 2 we formally define a system of frequent sets. In Section 3, an axiomatization for complete systems of frequent sets is given. In Section 4, decidability and computability issues are studied. In Section 5, some limited complexity results of deciding whether a system is complete are given. In section 6 we study some special cases. Section 7 concludes the paper.

Acknowledgment

We would like to thank Prof. Dirk Van Gucht and Prof. Ed Robertson from Indiana University, for their preliminary thoughts and reflections on the topic of this paper.

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 $A :: 0.5$, $AB :: 0.3$, $DEF :: 0.1$
R does not satisfy $A :: 0.8$, $ABC :: 0.4$, $DEF :: 0.3$

Figure 1: A matrix together with some frequent set expressions

2 Complete System of Frequent Sets

We formally define a system of frequent sets. We also define what it means for a system to be complete.

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 one in the (i, j)-entry if transaction i contains item j, else this entry is zero. When R is a matrix where the 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 one in every column of K. ◁

Example 1 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[1]. Note that, because R is a matrix, R can have identical rows. ◁

Definition 2 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 $0 \leq p_K \leq 1$.
- A matrix R over I *satisfies* $K :: p_K$ iff $freq(K, R) \geq p_K$. Hence itemset K has frequency at least p_K.

[1] DEF denotes the set $\{D, E, F\}$

- A *system of frequent sets* over I is a collection
$$\left\{_{K \subseteq I}\ K :: p_K$$
of frequent set expressions, with one expression for each $K \subseteq I$.
- A matrix R over I *satisfies* the system $\left\{_{K \subseteq I}\ K :: p_K\right.$ iff R satisfies all $K :: p_K$.

◁

Example 2 In Fig. 1, the matrix R satisfies $A :: 0.6$, because the frequency of A in R is bigger than 0.6. The matrix does not satisfy $B :: 0.7$, because the frequency of B is lower than 0.7. ◁

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

- 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 $S = \left\{_{K \subseteq I}\ K :: p_K\right.$ is *complete* iff for each $K :: p$ logically implied by S, $p \leq p_K$ holds.

◁

Example 3 Let $I = \{A, B, C, D, E, F\}$. Consider the following system: $S = \left\{_{K \subseteq I}\ K :: p_K\right.$, 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 $I = \{A, B, C\}$:
$\{\phi :: 1, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, BC :: 0.6, ABC :: 0.4\}$. This system is complete. We prove this by showing that for every subset K of I, there exists a matrix R_K that satisfies S, and $freq(K, R_K)$ is exactly p_K. These matrices then *prove* that for all K, we cannot further improve on K; i.e. make p_K larger. These proof-matrices are very important in the proof of the axiomatization that is given in the next section. In Fig. 2, the different proof-matrices are given. ◁

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

First of all, we need to prove a rather technical lemma. It may seem trivial, but its importance to the rest of this paper cannot be overestimated! Without this lemma, we would have no guarantee that there always exists a closed system.

B, C, BC, ABC

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

A, AB, AC

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

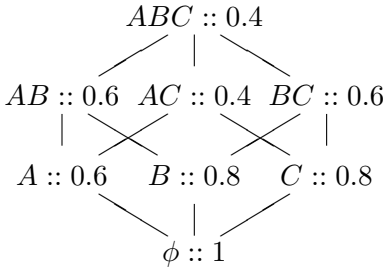

Figure 2: Proof-matrices for a system of frequent sets

Lemma 1 *Let S be a system of frequent sets, and $P \subseteq [0,1]$. If for all $p \in P$ holds that $S \models K :: p$, then also $S \models K :: supremum(P)$.*

Proof. Let R be an arbitrary matrix that satisfies S. R is by definition finite; say $|R| = n$. Therefore, the frequencies of the itemsets in R can only take values in $F = \{\frac{i}{n} \mid 0 \leq i \leq n\}$ (the frequencies are defined as ratios). Let $m := max\{f \in F \mid f < sup(P)\}$ (m is well-defined; F is finite). There must exist an element p in P such that $m < p \leq supremum(P)$. Because $K :: p$ holds in R, $freq(K,R) \geq p > m$. We can conclude that $freq(K,R) \geq sup(P)$, and thus R satisfies $K :: sup(P)$. Since R was arbitrary, the lemma holds. ◁

Proposition 1 *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, $S = \left\{_{K \subseteq I}\ K :: p_K\right.$. 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 $m_K = max\{p_K \mid S \models K :: p_K\}$ (well-defined, see Lemma 1). The system $\left\{_{K \subseteq I}\ K :: m_K\right.$ is clearly the unique completion of S. ◁

Example 4 $I = \{A, B, C\}$. The system $\{\phi :: 1, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, \mathbf{BC :: 0.6}, ABC :: 0.4\}$ is the unique completion of the system $\{\phi :: 0.8, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, \mathbf{BC :: 0.4}, ABC :: 0.4\}$. $BC :: 0.6$ is implied by the second system, since there is an overlap of at least 0.6 between the rows having a one on B and the rows having a one on C. ◁

Remark that when a system is complete, it is not necessary that there exists one matrix such that for all itemsets the frequency is exactly the frequency given in the system. Consider for example the following system: $\{\phi :: 1, A :: 0.5, B :: 0.5, C :: 0.1, AB :: 0, AC :: 0, BC :: 0, ABC :: 0\}$. 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 one in A or in B. So, every row having a one in C has also a one in A or in B, and thus violates respectively $freq(AC) = 0$, or $freq(BC) = 0$.

3 Axiomatizations

We give an axiomatization for frequent sets. An axiomatization in this context is a set of rules that are satisfied by the system iff it is complete. In order to simplify the notation we first introduce rare sets.

3.1 Rare Sets

Definition 4 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 zero in at least one column of K.
- A *rare set expression* over I is an expression $K : p_K$ with $K \subseteq I$ and $0 \leq p_K \leq 1$.
- A matrix R over I *satisfies* $K : p_K$ 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 $\left\{_{K \subseteq I}\ K : P_K\right.$ of rare set expressions, with one expression for each $K \subseteq I$.
- A matrix R over I *satisfies* the system $\left\{_{K \subseteq I}\ K : p_K\right.$ 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 $S = \left\{_{K \in I}\ K : p_K\right.$ is *complete* iff for each $K : p$ logically implied by S, $p_K \leq p$ holds.

◁

Example 5 In Fig. 1, the matrix R satisfies $A : 0.4$, because the rareness of A in R is smaller than 0.4. The matrix does not satisfy $B : 0.3$, because the rareness of B is greater than 0.3. Let $I = \{A, B\}$. The system $\{AB : 0.8, A : 0.3, B : 0.4, \phi : 0.4\}$ is not complete. The unique completion of this system is $\{AB : 0.7, A : 0.3, B : 0.4, \phi : 0\}$. ◁

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. The second part of the proposition shows that an axiomatization for rare sets automatically yields an axiomatization for frequent sets.

Proposition 2 *Let $I = \{I_1 \ldots I_n\}$ 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 $K :: 1 - p_K$.*

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 the previous proposition.

3.2 Axiomatization of Rare Sets

Before we give the axiomatization, we first define bags.

Definition 5

- A *bag* over a set S is a total function from S into $\mathbb{N}$.
- Let $\mathbf{K}$ be a bag over S and $s \in S$. We say that *s appears n times in* $\mathbf{K}$ iff $\mathbf{K}(s) = n$.
- If $\mathbf{K}$ and $\mathbf{L}$ are bags over S, then we define the *bag-union of* $\mathbf{K}$ *and* $\mathbf{L}$, notation $\mathbf{K} \bigcup \mathbf{L}$, as follows: for all $s \in S$, $(\mathbf{K} \bigcup \mathbf{L})(s) = \mathbf{K}(s) + \mathbf{L}(s)$.
- Let $\mathbf{K}$ be a bag over the subsets of a set S. Then $\bigcup \mathbf{K}$ denotes the bag $\bigcup_{K \in \mathbf{K}} K$. The degree of an element $s \in S$ in $\mathbf{K}$, denoted $deg(s, \mathbf{K})$ is the number of times s appears in $\bigcup \mathbf{K}$.
- Let $S = \{s_1, s_2, \ldots, s_n\}$. $\{\!\{ c_1's_1, \ldots, c_n's_n \}\!\}$ denotes the bag over S in which s_i appears c_i times for $1 \leq i \leq n$.
- Let S be a set, $\mathbf{K}$ a bag over S. $\sum_{s \in S} \mathbf{K}(s)$ is the *cardinality of* $\mathbf{K}$, and is denoted by $|\mathbf{K}|$.

$\triangleleft$

The next three rules form an axiomatization for complete systems of rare sets in the sense that the complete systems are exactly the ones that satisfy these three rules. The p_K's that appear in the rules, indicate the rareness-values given in the systems for the set K; i.e. $K : p_K$ is in the system.

R1 $p_\phi = 0$

R2 If $K_2 \subseteq K_1$, then $p_{K_2} \leq p_{K_1}$

R3 Let $K \subseteq I$, $\mathbf{M}$ a bag of subsets of K. Then

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

with $k = min_{a \in K}(deg(a, \mathbf{M}))$

Lemma 2 *Given a set of indices I and given a_K, b_K for every non-empty $K \subseteq I$. Consider the following system of inequalities:*

$$\left\{ {}_{K \subseteq I} \; a_K \leq \sum_{i \in K} X_i \leq b_K \right.$$

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

The rather technical proof is given in the appendix.

Theorem 1 *Let S be a system of rare sets over I. The following two statements are equivalent:*

- *S is a complete system.*
- *S satisfies* **R1**, **R2**, *and* **R3**.

We first prove the soundness of the rules **R1**, **R2**, and **R3**.

Proposition 3 *Let S be a system of rare sets over I. If S is complete, then S satisfies* **R1**, **R2**, *and* **R3**.

Proof. **R1** and **R2** are trivial.
R3 Let $S = \{ {}_{K \in I} \; K : p_K$ be a complete system, and let $\mathbf{M}$ be a bag of subsets over $K \subseteq I$. We will prove that $\frac{\sum_{M \in \mathbf{M}} p_M}{k} \geq p_K$ with $k = min_{a \in K}(deg(a, \mathbf{M}))$. Suppose R is a matrix over I, and R satisfies S. Let for all $Z \subseteq I$, D_Z be the set of rows having a zero in at least one column of Z. Then we know that every row in D_K, appears in at least k of the following sets: $\{\!\{ D_M \mid M \in \mathbf{M} \}\!\}$, because $t \in D_K$ implies that there is an $a \in K$, such that $t(a) = 0$. Because $deg(a, \mathbf{M}) \geq k$, there must be at least k sets in $\mathbf{M}$ that contain a. For each set $M \in \mathbf{M}$ with $a \in M$ must $t \in D_M$. Therefore, we can conclude that $k|D_k| \leq \sum_{M \in \mathbf{M}} |D_M| \leq n \sum_{M \in \mathbf{M}} p_M$. S is complete, and in every matrix R that satisfies S, $k\frac{|D_k|}{n} \leq \sum_{M \in \mathbf{M}} p_M$ holds. We can conclude $p_K \leq \frac{\sum_{M \in \mathbf{M}} p_M}{k}$. ◁

Now we will prove the completeness of **R1**, **R2**, and **R3**.

Proposition 4 *Let S be a system of rare sets over I. If S satisfies* **R1**, **R2**, *and* **R3**, *then S is complete.*

Proof. Let $S = \{ {}_{K \in I} \; K : p_K$ be a system that satisfies **R1**, **R2**, and **R3**. We will proof that S is complete. Therefore, we will prove that for every $K : p_K$ in the system, there exists a matrix R over I, such that R satisfies S, and the rareness of K in R is exactly p_K [2].

Let β_Z be the exact fraction of rows in R that have a zero in every column of Z, and a one everywhere else. Then the rareness of a set L becomes: $\sum_{Z \subseteq L} \beta_Z$. We will construct a matrix that satisfies S and has rareness exactly p_K for K,

[2] Remark the similarity with Armstrong-relations in functional dependency theory [4]

by specifying all β_Z's. We can construct such a matrix if the following system of inequalities has a solution:

$$\begin{cases} \forall Z \subseteq I : 0 \leq \beta_Z \leq 1 & (1) \\ \sum_{Z \subseteq I} \beta_Z = 1 & (2) \\ p_K = \sum_{Z \subseteq K} \beta_Z & (3) \\ \forall L \subseteq Z, L \neq K : p_L \geq \sum_{Z \subseteq L} \beta_Z & (4) \end{cases}$$

(1) states that all fractions are between zero and one.
(2) states that the sum of all fractions must be one.
(3) states that the rareness of K must exactly be p_K.
(4) states that for every other subset of I, the rareness must be below the rareness value in the system.

We will even prove that there exists a solution to the following system of inequalities:

$$\begin{cases} \forall a \in I : 0 \leq \beta_a \leq 1 & (1') \\ 0 \leq \beta_0 \leq 1 & (2') \\ (\sum_{a \in I} \beta_a) + \beta_0 = 1 & (3') \\ p_K = \sum_{a \in K} \beta_a & (4') \\ \forall L \subset K : p_L \geq \sum_{a \in L} \beta_a & (5') \end{cases}$$

When this second system has a solution, then also the first system has a solution. Starting from a solution to the second system, we construct a solution for the first one:
$\forall Z \subseteq I : Z \notin \{\{a\} \mid a \in I\} \cup \{\phi\} \Rightarrow \beta_Z = 0$
$\forall a \in I : \beta_{\{a\}} = \beta_a$
$\beta_\phi = \beta_0$

(1),(2), and (3) are trivially fulfilled.
(4) Take $L \neq K$, then $p_L \geq^{(\mathbf{R2})} p_{L \cap K} \geq \sum_{a \in L \cap K} \beta_a = \sum_{Z \subseteq L} \beta_Z$

This system, on its turn, has a solution if the following system has a solution:

$$\begin{cases} \forall L \subseteq K : p_K - p_L \leq \sum_{a \in K} \beta_a - \sum_{a \in L} \beta_a \leq p_K & (1'') \end{cases}$$

1' is ok: choose $L = K - \{a\}$, then $0 \leq^{(\mathbf{R2})} p_K - p_{K-\{a\}} \leq \beta_a \leq p_K \leq 1$
2'+3' are ok: let $\beta_0 = 1 - \sum_{a \in K} \beta_a = 1 - p_K$
4' is ok: choose $L = \phi$, $p_L = 0$ (**R1**), and thus $p_K \leq \sum_{a \in K} \beta_K \leq p_K$
5' is ok: $p_L - p_K \geq \sum_{a \in L} \beta_a - \sum_{a \in K} \beta_a + 4'$.

According to Lemma 2, this third system has a solution iff for all bags $\mathbf{M}$ and $\mathbf{N}$ over the subsets of K , such that $\bigcup \mathbf{M} = \bigcup \mathbf{N}$, $\sum_{M \in \mathbf{M}} p_K - p_{K-M} \leq \sum_{N \in \mathbf{N}} p_N$ holds.
Let $\mathbf{L} = \mathbf{N} \bigcup \{\!\{ K - M \mid M \in \mathbf{M} \}\!\}$.
Then, by **R3** we have that $\frac{\sum_{L \in \mathbf{L}} p_L}{k} \geq p_K$, with $k = min_{a \in K} \#(\{\!\{ N \mid a \in N \wedge N \in \mathbf{N} \}\!\} \bigcup \{\!\{ M \mid M \in \mathbf{M} \wedge a \notin M \}\!\})$.
Because $\#\{\!\{ M \mid M \in \mathbf{M} \wedge a \in M \}\!\} = \#\{\!\{ N \mid N \in \mathbf{N} \wedge a \in n \}\!\}$, $k = \#\mathbf{M}$.
We have: $\sum_{L \in \mathbf{L}} p_L \geq \#\mathbf{M} p_K$.

Since $\sum_{L\in\mathbf{L}} p_L = \sum_{N\in\mathbf{N}} p_N + \sum_{M\in\mathbf{M}} p_{K-M}$
and $\#\mathbf{M}p_K = \sum_{M\in\mathbf{M}} p_K$,
$\sum_{M\in\mathbf{M}} p_K - p_{K-M} \leq \sum_{N\in\mathbf{N}} p_N$ holds. ◁

Example 6 The system $\{\phi : 0.5, A : 0.5, B : 0.25, C : 0.5, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since $\phi : 0.5$ violates **R1**.
The system $\{\phi : 0, A : 0.5, B : 0.25, C : 0.5, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since for example $AB : 0$ and $A : 0.5$ together violate **R2**.
The system $\{\phi : 0, A : 0, B : 0, C : 0, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since $A : 0$, $C : 0$, and $AC : 1$ together violate **R3**.
The system $\{\phi : 0, A : 0, B : 0, C : 0, AB : 0, AC : 0, BC : 0, ABC : 0\}$ is complete, since it satisfies **R1**, **R2**, and **R3**. This system is the unique completion of all systems in this example. ◁

3.3 Axiomatization of Frequent Sets

From Proposition 2, we can now easily derive the following axiomatization for frequent sets.

F1 $p_\phi = 1$

F2 If $K_2 \subseteq K_1$, then $p_{K_2} \geq p_{K_1}$

F3 Let $K \subseteq I$, $\mathbf{M}$ a bag of subsets of K. Then

$$p_K \geq 1 - \frac{\#\mathbf{M} - \sum_{M\in\mathbf{M}} p_M}{k},$$

with $k = min_{a\in K}(deg(a, \mathbf{M}))$

Theorem 2 *Let $S = \left\{_{K\subseteq I}\ K :: p_K$ be a system of frequent sets over I. The following two statements are equivalent:*

- *S is a complete system.*
- *S satisfies* **F1**, **F2**, *and* **F3**.

Proof. The Theorem holds since:

- $K :: p_K$ is equivalent with $K : 1 - p_K$, and
- **R1**, **R2**, and **R3** are sound and complete for rare sets.

◁

4 Decidability and Computability

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

In the previous section we introduced and proved an axiomatization for rare and frequent sets. There is however still one problem with this axiomatization. **R3** states a property that has to be checked for all bags over the subsets of K. This number of bags is infinite. So, we cannot conclude that completeness of a system is decidable. 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 can be computed. Therefore, deciding completeness is decidable.

We also look at the following problem: when an incomplete system is given, can we compute its completion? We show that the completion is indeed computable. We use **R1**, **R2**, and **R3** as 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.

4.1 Minimal Multi-covers

Definition 6

- A *k-cover* of a set S is a bag $\mathbf{K}$ over the subsets of S such that for all $s \in S$, $deg(s, \mathbf{K}) = k$.
- A bag $\mathbf{K}$ over the subsets of a set S is a *multi-cover* of S if there exists an integer k such that $\mathbf{K}$ is a k-cover of S.
- A k-cover $\mathbf{K}$ of S is *minimal* if it cannot be decomposed as $\mathbf{K} = \mathbf{K}_1 \bigcup \mathbf{K}_2$, with $\mathbf{K}_1$ and $\mathbf{K}_2$ respectively k_1- and k_2-covers of S, $k_1 > 0$ and $k_2 > 0$.

$\triangleleft$

Example 7 Let $K = \{A, B, C, D\}$. $\{\!\{\ 1'AB, 1'BC, 1'CD, 1'AD, 1'ABCD\ \}\!\}$ is a 3-cover of K. It is not minimal, because it can be decomposed into the following two minimal multi-covers of K: $\{\!\{\ 1'AB, 1'BC, 1'CD, 1'AD\ \}\!\}$ and $\{\!\{\ 1'ABCD\ \}\!\}$. $\triangleleft$

The new rule that replaces **R3** states that it is not necessary to check all bags; we only need to check the minimal multi-covers. This gives the following **R3**':

R3' Let $K \subseteq I$, $\mathbf{M}$ a minimal k-cover of K. Then

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

Lemma 3 *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.*

PROOF. We will prove the lemma by induction.

base case $n = 2$. Suppose $\frac{a+c}{b+d} < p$ and $\frac{a}{b} \geq p$ and $\frac{c}{d} \geq p$. a, b, c, d and p are all positive. This yields the following inequalities:

$$\Rightarrow \quad \begin{cases} \frac{a}{b} > \frac{a+c}{b+d} \\ \frac{c}{d} > \frac{a+c}{b+d} \end{cases}$$

$$\Rightarrow \quad \begin{cases} a > b\frac{a+c}{b+d} \\ c > d\frac{a+c}{b+d} \end{cases}$$

$$\Rightarrow \quad a + c > (b + d)\frac{a+c}{b+d}.$$

This is clearly a contradiction.

general case $\frac{a_1+\ldots+a_n}{b_1+\ldots+b_n} < p$, therefore either $\frac{a_1}{b_1} < p$, in which case the lemma is proven, or $\frac{a_2+\ldots+a_n}{b_2+\ldots+b_n} < p$, in which case we can apply the induction hypothesis.

◁

Lemma 4 *Every k-cover $\mathbf{M}$ can be decomposed into a number of minimal multi-covers $\mathbf{M}_1, \ldots, \mathbf{M}_n$, such that $\bigcup_{i=1\ldots n} \mathbf{M}_i = \mathbf{M}$.*

Theorem 3 *Let S be a system of rare sets over I. The following statements are equivalent:*

1. *S is a complete system.*
2. *S satisfies **R1**, **R2**, and **R3**.*
3. *S satisfies **R1**, **R2**, and **R3'**.*

PROOF. $1 \Leftrightarrow 2$ is already established in the previous section. $2 \Rightarrow 3$ is trivial, since **R3**' is more specific than **R3**. Suppose that the system $S = \left\{_{K \subseteq I}\ K : p_K\right.$ satisfies **R1** and **R2**, but does not satisfy **R3**. We will show that it is impossible that it satisfies **R3**'.

There must be a set $K \subseteq I$, and a bag $\mathbf{M}$ over the subsets of K, such that $p_K < \frac{\sum_{m \in \mathbf{M}} p_M}{k}$ with $k = min_{a \in K}(deg(a, \mathbf{M}))$. For each $a \in K$ such that $deg(a, \mathbf{M}) > k$, we replace $deg(a, \mathbf{M}) - k$ of the sets $A \in \mathbf{M}$ that contain a by $A - \{a\}$. In this way, we construct a k-cover $\mathbf{M}'$ of K.
Because S satisfies **R2**, $\sum_{M \in \mathbf{M}} p_M \leq \sum_{M \in \mathbf{M}'} p_M$. The k-cover $\mathbf{M}'$ can be decomposed into different minimal multi-covers $\mathbf{M}_1, \ldots, \mathbf{M}_n$ of K (M_i is a k_i-cover of K). Because $\frac{\sum_{m \in \mathbf{M}'} p_M}{k} = \frac{\sum_{M \in \mathbf{M}_1} p_M + \ldots + \sum_{M \in \mathbf{M}_n} p_M}{k_1 + \ldots + k_n}$, for at least one i, $\frac{\sum_{M \in \mathbf{M}_i}}{k_i} > p_K$ must hold.
Therefore, **R3'** is violated. ◁

Definition 7 A predicate C on $\mathbb{N}^n$, is a *concave predicate* iff for every $a_1 \leq a'_1, \ldots, a_n \leq a'_n$ the following holds: $C(a_1, \ldots, a_n)$ and $C(a'_1, \ldots, a'_n)$ implies that $(a_1, \ldots, a_n) = (a'_1, \ldots, a'_n)$. ◁

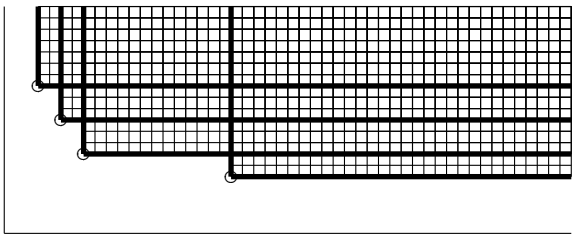

Figure 3: A concave predicate C, together with rectangles, indicating zones that cannot contain points that satisfy C.

In $\mathbb{N}^2$ this definition can be visualized as follows: a predicate is concave iff for every point a such that the predicate on a is valid, there are no other points that satisfy the predicate in the quadrant above and on the right of a. This is illustrated in figure 3. The figure shows the points that fulfill a certain concave predicate C. The grayed area indicates the region that cannot contain points that fulfill the predicate.

Example 8 The following predicates on $\mathbb{N}^2$ are concave:

- $C(x, y)$ iff $x + y = a$, $a \in \mathbb{N}$.
- $C(x, y)$ iff $x.y = a$, $a \in \mathbb{N}$.

The following predicate on $\mathbb{N}^7$ is concave:

- $C(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ iff $x_1.\{a, c\} \bigcup x_2.\{b, c\} \bigcup x_3.\{a, b\} \bigcup x_4.\{a\} \bigcup x_5.\{b\} \bigcup x_6.\{a, b\} \bigcup x_7\{a, b, c\}$ is a minimal 3-cover of $\{a, b, c\}$.

◁

Proposition 5 *If C is a concave predicate over $\mathbb{N}^n$, then the number of points in $\mathbb{N}^n$ that satisfy C is finite.*

Proof. We will proof the proposition by induction on the number of dimensions n:

base case The base case $n = 1$ is trivial.

general case We assume that the proposition is valid if the number of dimensions is strictly less than n. When there is no point satisfying C, the proposition is true. In the other case, we can choose a point p that satisfies C. Starting from this point, we divide $\mathbb{N}^n$ into a finite number of (overlapping) sets:

- $P = \{q \in \mathbb{N}^n \mid \forall i : q_i \geq p_i\}$
- $P_{i=j} = \{q \in \mathbb{N}^n \mid q_i = j\}$, $\forall\ 1 \leq i \leq n\ \forall\ 0 \leq j < p_i$

The number of sets is clearly finite (there are $\sum_{i=1\ldots n} p_i + 1$ sets). Since C is a concave predicate, P contains only one point that satisfies C (the point p). $P_{i=j}$ contains only a finite number of elements that satisfies C by induction. Indeed, we can identify $P_{i=j}$ with $\mathbb{N}^{n-1}$ through the 1-1 function $f_{ij} : P_{i=j} \to \mathbb{N}^{n-1}$ that is defined as follows: $f_{ij}(a) = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$. Now we define C' on $\mathbb{N}^{n-1}$: $C'(b)$ iff $C(f_{ij}^{-1}(b))$. It is an easy verification that C' is a concave predicate on $\mathbb{N}^{n-1}$. By induction, the number of elements in $\mathbb{N}^{n-1}$ that satisfy C' and thus the number of points in $P_{i=j}$ that satisfy C is finite.

◁

Proposition 6 *Let K be a finite set. The number of minimal multi-covers of K is finite.*

PROOF. Let $|K| = n$. Order the subsets of K; i.e. $2^K = \{S_1, \ldots, S_{2^n}\}$. Define now the following predicate C on $\mathbb{N}^{(2^n)}$: $C(c_1, \ldots, c_{2^n})$ iff $\bigcup_{i=1\ldots 2^n} c_i.S_i$ is a minimal multi-cover for K. The predicate C is clearly concave, and is as such finite. Therefore, the number of minimal multi-covers is finite. ◁

Proposition 7 *Let K be a finite set. The minimal multi-covers of K are computable.*

PROOF. Let $f : 2^K \to \{c_1, \ldots, c_{|2^K|}\}$ be an isomorphism between the subsets of K and the set of variables $\{c_1, \ldots, c_{|2^K|}\}$; i.e. to each subset of K, a distinct variable $f(K)$ is assigned. Every $|2^K|$-tuple $(c_1, \ldots, c_{|2^K|})$ represents a bag over the subsets of S, namely $\bigcup_{S \subseteq K} c_S.S$.
This bag is a k-cover of K iff
$\forall a \in K : \sum_{S \subseteq K, a \in S} c_S = k$, ($a$ appears exactly k times in the bag.)
and all c_S's are positive integers.

Together with the requirements $k > 0$ and $\forall S : c_S \geq 0$, we can find at least one solution $(s_1, \ldots, s_{|2^K|})$, using integer programming. We can easily check whether this solution is a minimal multi-cover. Since we are only interested in minimal multi-covers, all other solutions of interest to us have $c_1 < s_1$, or $c_2 < s_2$, ..., or $c_{|2^K|} < s_{|2^K|}$. Therefore, all minimal covers in this case have $c_1 = 0$ or $c_1 = 1$ or $\ldots$ or $c_1 = s_1 - 1$ or $c_2 = 0$ or $c_2 = 1$ or $\ldots$ or $c_2 = s_2 - 1$ or $\ldots$ or $c_{|2^K|} = s_{|2^K|} - 1$. For each of these cases we recursively solve the original system, where in the case of $c_j = l$, all c_j's are replaced by l. These systems have one variable less than the original system. We can proof by induction on the number of variables in the system, that we can find the minimal solutions. ◁

4.2 Computing the Completion of a System

We prove that by applying **R1**, **R2**, and **R3** as rules, we can compute the completion of any given system.

Applying for example rule **R2** 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 $K_1 : p_{K_2}$. It is clear that **R1** can only be applied once; **R2** and **R3** never create situations in which **R1** can be applied again.

R2 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 **R2**, by going top-down; we first try to improve the frequencies of the biggest itemsets, before continuing with the smaller ones.

R3 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 **R3**, by applying the rule bottom-up.

A trivial algorithm to compute the completion of a system is the following: apply **R1**, and then keep applying **R2** and **R3** until a fixpoint is reached. Clearly, the *limit* of this approach yields a complete system, but it is not clear that a fixpoint will be reached within a finite number of steps. Moreover, there are examples of situations in which infinite loops are possible. In Fig. 4, such an example is given. The completion of the first system, is clearly all rareness values equal to zero, because for every matrix satisfying the system, none of the rows have a zero in AB, and none have a zero in BC, so there are no zeros at all in the matrix. When we keep applying the rules as in Fig. 4, we never reach this fixpoint, since in step $2n$, the value for ABC is $\left(\frac{1}{2}\right)^n$. This is however no disaster; we show that when we apply the rules **R2** and **R3** in a systematic way, we always reach a fixpoint within a finite number of steps. This systematic approach is illustrated in Fig. 5. We first apply **R2** top-down until we reach a fixpoint for **R2**, and then we apply **R3** bottom-up until we reach a fixpoint for **R3**. The general systematic approach is written down in Fig. 6. We prove that for every system these two meta-steps are all there is needed to reach the completion.

Definition 8 Let I be a set of items, $J \subseteq I$, and $S = \left\{_{K \subseteq I}\ K : p_K\right.$ a system of rare sets over I. The *projection of S on J*, denoted $proj(S, J)$, is the system $S' = \left\{_{K \subseteq J}\ K : p_K\right.$. ◁

Lemma 5 *Let I be a set of items, $J \subseteq I$, and $S = \left\{_{K \subseteq I}\ K : p_K\right.$ a system of rare sets over I.*

- *If S is complete, then also $proj(S, J)$ is complete.*
- *if S satisfies* **R2***, then $proj(C(S), J) = C(proj(S, J))$.*

Proof. The first statement is trivial.
2) Let $C(proj(S, J)) = \left\{_{K \subseteq J}\ K : p_K\right.$. Then, for every $K \subseteq J$, we can construct a matrix R_K, such that $rare(K, R_K) = p_K$, and for all $L \subseteq J$, $rare(L, R_K) \leq p_L$ [3]. We will now extend this matrix R_K over J to the matrix $\widehat{R_K}$ over I.

[3]This fact can easily be derived from the proof of the completeness of **R1**, **R2** and **R3**.

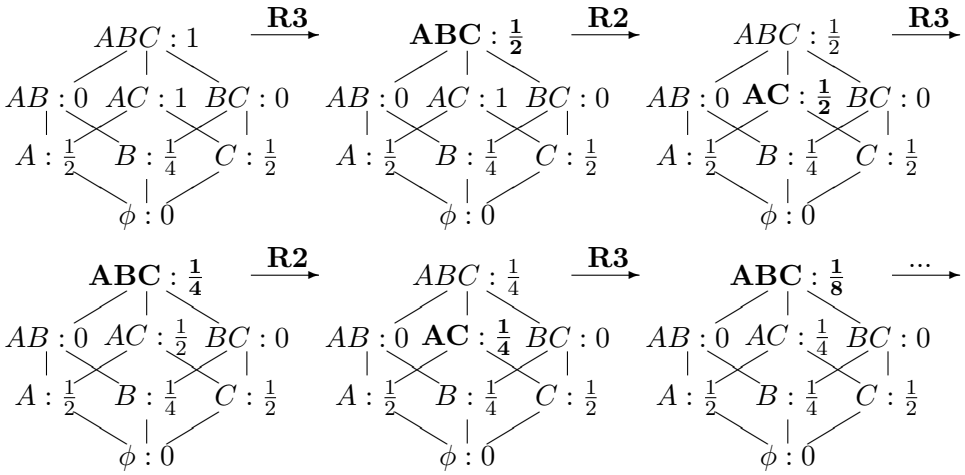

Figure 4: "Random" application of the rules can lead to infinite loops

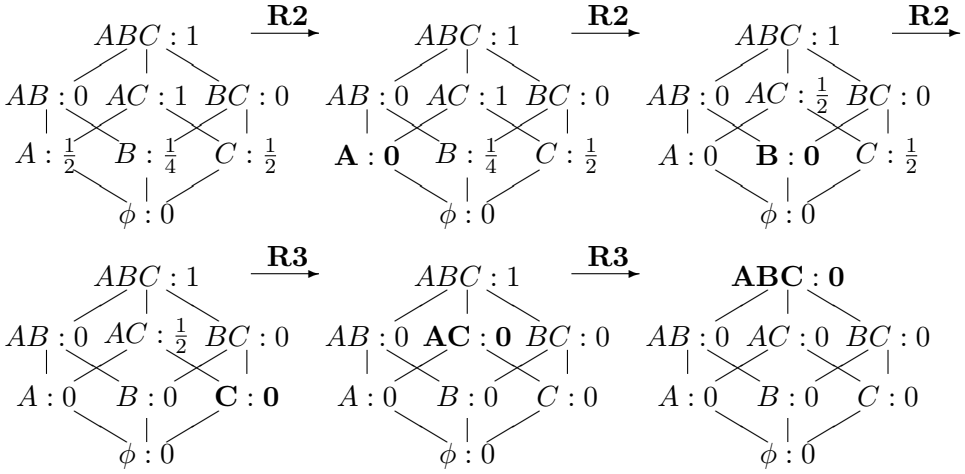

Figure 5: Systematic application of the rules avoids infinite computations

Input: System of rare sets $S = \left\{_{K \subseteq I}\ K : p_K \text{ over } I = \{I_1, \ldots, I_n\}\right.$.
Output: Completion of S.

Close(S)
  $p_\phi = 0$
  TopDown(S)
  BottomUp(S)

TopDown(S)
  $i = n$
  while($i > 0$)
    for all itemsets K of cardinality i **do**
      make $p_K = min_{K \subseteq L}(p_L)$
    end for
    $i = i - 1$
  end while

BottomUp(S)
  $i = 1$
  while($i \leq n$)
    for all itemsets K of cardinality i **do**
      make $p_K = min_{\mathbf{K}}$, minimal k-cover of K $\left(\frac{\sum_{K' \in \mathbf{K}} p_{K'}}{k}\right)$
    end for
    $i = i + 1$
  end while

Figure 6: Algorithm for finding the completion of a system

$\widehat{R_K}$ contains the same number of rows as R_K, and for each row $r \in R_K$, there is a row r' in $\widehat{R_K}$, such that $r'(i) = r(i)$ for all $i \in I$, and $r'(i) = 1$ else. $\widehat{R_K}$ satisfies S, since it is constructed in such a way that for all $L \subseteq I$ holds that $rare(L, \widehat{R_K}) = rare(L \cap K, R_K) \leq p_{L \cap K} \leq^{(\mathbf{R2})} p_L$. Since $\widehat{R_K}$ satisfies S, and $rare(K, \widehat{R_K}) = p_K$, $C(S)$ must contain $K : p_K$; $\widehat{R_K}$ is a proof-matrix for $K : p_K$. Because K was an arbitrary subset of J, the lemma holds. ◁

Theorem 4 *The algorithm in Fig. 6 computes the closure of the system of rare sets S.*

Proof.
It is easy to see that after the top-down step the system satisfies **R1** and **R2**. Now we will proof by induction, that after the i-th iteration in the bottom-up loop, all itemsets with at most i items will satisfy **R1**, **R2**, and **R3**; i.e. we cannot apply any of these rules to change the rareness of an itemset with less than i items. For the base case $i = 0$ this is trivially true. Suppose the induction

hypothesis holds for $1, \ldots, i-1$. Then, when we start the i-th iteration, all itemsets of cardinality less than i will have their final values. Applying the for-loop for itemset K, with $|K| = i$, will then yield the closure of $proj(S, K)$, because after this loop, $proj(S, K)$ will be closed for **R3**, and **R2** cannot be applied, because otherwise the value of at least one subset of K will be changed. Because S satisfies **R2**, K will also have reached its final value (Proposition 5). Therefore after the i-th iteration all itemsets of cardinality up to i will be final. $\triangleleft$

5 Complexity

In this section we give some limited results on the complexity of deciding completeness. Because we only know that the number of minimal multi-covers is finite, and we only have a very naive algorithm to compute them, we have no real idea about the complexity of the proposed algorithm.

We cannot state strong results concerning the complexity of deciding completeness of a system, but the complexity of deciding whether is system is incomplete, is easy to establish:

Proposition 8 *Deciding completeness of a system of rare sets is in coNP.*

Proof. Consider the proof of Theorem 1. We can derive from this proof that if a system is complete, then for every K there must exist a solution to the following system:

$$\begin{cases} \forall Z \subseteq I : 0 \leq \beta_Z \leq 1 \\ \sum_{Z \subseteq I} \beta_Z = 1 \\ p_K = \sum_{Z \subseteq K} \beta_Z \\ \forall L \subseteq Z, L \neq K : p_L \geq \sum_{Z \subseteq L} \beta_Z \end{cases}$$

When all these systems of inequalities can be satisfied for all itemsets K, the system must be complete. When there is one K such that the system of inequalities has no solution, the system of rare sets is incomplete. Thus, we guess an itemset K, and we can check in polynomial time if the corresponding system has a solution, because it is an instance of linear programming. When this system has no solution, the system of rare sets is not complete. ◁

6 Sparse Systems

Definition 9 Let I be a set of items.

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

$$\{_{K \in P}\ K : p_K$$

  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 if it satisfies every rare set expression in the system.

- A sparse system logically implies a rare set expression, if every matrix that satisfies the system, also satisfies the rare set expression.

- A sparse system $\{_{K \in P}\ K : p_K$ 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 9 *Let I be a set of items, and $S = \{_{K \in P}\ K : p_K$ be a sparse system. The following two statements are equivalent:*

- *S is complete*

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

Corollary 1 *Completeness for sparse systems is decidable, and the completion of a sparse system is computable.*

The proposition and the corollary give the following algorithm for computing the completion of a sparse system: make the system full by adding $K : 1$ for all itemsets $K \notin P$ and compute the closure of this full system with the methods in Section 4. Then adjust the values of the sparse system to the values in the full system. If there are no values that can be adjusted in the sparse system,

the sparse system was complete. Otherwise, we computed the completion of the sparse system.

It is however clear, that this approach is not very efficient. Suppose that we are given a sparse system with n rare set expressions and over a set with m items. To compute the completion, we calculate the completion of a system with 2^m expressions, where the input only has n expressions. The following proposition shows that there are more efficient ways to calculate the completion of a sparse system.

Proposition 10 *The following rules form an axiomatization for the completeness of the sparse system* $\{K_1 : p_1, \ldots, K_n : p_n\}$

S1 $p_\phi = 0$

S2 *If* $K_2 \subseteq K_1$, *then* $p_{K_2} \leq p_{K_1}$

S3 *Let* $\mathbf{M}$ *be a minimal* k*-cover of* K_i. *Then*

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

Proof.

Soundness of the three rules is trivial.

Completeness. Suppose S satisfies **S1**, **S2**, and **S3**. We proof that there exists a complete full super-system of S. We construct this system as follows: for all itemsets $K \notin P$, add $K : 1$ to the system, and calculate the completion of this system. The resulting system is full and complete. Let $\overline{S} = \left\{_{K \subseteq I}\ K : q_K$ be this system. We will show by contradiction that for all $K \in P$, $p_K = q_K$ holds.
So, suppose there is a $K \in P$ such that $p_K \neq q_K$. $\overline{S}$ is computed by first going top-down, and then bottom-up. Since S satisfies **S1** and **S2**, the rareness of K in $\overline{S}$ comes from the bottom-up step, and thus there exists a minimal k-cover $\mathbf{K}$ over the subsets of K, such that $\frac{\sum_{L \in \mathbf{K}} q_L}{k} < p_K$. 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 $\mathbf{L}$ over the subsets of L, such that $q_L = \sum_{L' \in \mathbf{L}} q_{L'}$. We now construct a kl-cover $\mathbf{K'}$ of K as follows: $\mathbf{K'} = (\mathbf{K} - L) \bigcup \mathbf{L}$. $\mathbf{K'}$ is clearly a kl-cover. 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 **R3**, by a sum of $q_{L'}$'s, were all $L' \subset L$. When these L' are obtained by **R3**, we can replace them by $q_{L''}$ of even smaller sets L''. Since the singleton sets can only be obtained by **R2**, this recursion must stop, and thus there exists a m-cover $\mathbf{M}$ such that $\frac{\sum_{M \in \mathbf{M}} q_M}{m} < p_K$, and all q_M's are obtained by **R2**. As such,

for all M, $q_M = min_{M \subseteq K_i} p_{K_i}$, and thus $\frac{\sum_{M \in \mathbf{M}} min_{M \subseteq K_i} p_{K_i}}{m} < p_K$. There is still one problem: $\mathbf{M}$ is not necessarily minimal. We can cope with this problem in exactly the same way as at the end of the proof of Proposition 3. $\triangleleft$

The proposition learns us that we don't need to compute the completion of a full system in order to see whether a sparse system is complete. This saves us already a considerable amount of work. We would however like to improve this result; the fact that $\mathbf{M}$ is an arbitrary minimal k-cover of K_i still bothers us.

7 Summary

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 contained three rules **R1**, **R2**, and **R3**. From these rules we could easily derive the axiomatization, **F1**, **F2**, and **F3** for frequent sets. By replacing **R3** with **R3'**, we showed that completeness is decidable. We also showed that the completion can be computed, by applying **R1**, **R2**, and **R3** as rules. If these rules are applied first top-down, and then bottom-up, the completion is reached within a finite number of steps. We showed that deciding completeness is in coNP.

We also studied sparse systems. In these systems, not for every itemset a rareness value is given. We proved that also for these sparse systems, completion is decidable and computable. The most obvious way to do this, is to make the sparse system into a full one and then compute the completion. This is however computationally very costly. We showed though, that more efficient algorithms are possible. We gave an axiomatization consisting of three rules **S1**, **S2**, and **S3** for sparse systems.

This paper is to our knowledge the first paper that discussed an axiomatization for frequent itemsets.

References

[1] R. Agrawal, T. Imilienski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD*, 1993

[2] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. VLDB*, 1994

[3] D. S. Associates. The new direct marketing. *Business One Irwin*, 1990

[4] R. Fagin, M. Y. Vardi. Armstrong Databases for Functional and Inclusion Dependencies. In *IPL 16(1): 13-19*, 1983.

[5] J. Han, J.Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. ACM SIGMOD*, 2000

[6] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. In *Data Mining and Knowledge Discovery 1(3)*, 1997.

[7] J. Paredaens. Axiomatization of Frequent Sets. Technical Report TR9911, University of Antwerp, Belgium, Oktober 1999

A Proof of Lemma 2

PROOF. ($\Rightarrow$) Let $(x_1, \ldots, x_{\#I})$ be a solution.
$\sum_{K\in\mathbf{K}} a_K \leq \sum_{K\in\mathbf{K}} \sum_{i\in K} x_i = \sum_{L\in\mathbf{L}} \sum_{j\in L} x_j \leq \sum_{L\in\mathbf{L}} b_L$
($\Leftarrow$) We proof by induction that there exist $c_1, \ldots, c_{\#I}$, such that for all positive integers α_i, β_i holds that

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\leq i\leq \#I} \alpha_i\{i\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\leq i\leq \#I} \beta_i\{i\}$$

implies

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\leq i\leq \#I} \alpha_i c_i \leq \sum_{L\in\mathbf{L}} b_L + \sum_{1\leq i\leq \#I} \beta_i c_i \ .$$

Base case: We remark that

$$\bigcup_{K\in\mathbf{K}} K = \bigcup_{L\in\mathbf{L}} L$$

implies

$$\sum_{K\in\mathbf{K}} a_K \leq \sum_{L\in\mathbf{L}} b_L \ .$$

General case: Suppose that we already chose $c_1, \ldots, c_j$ such that

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\leq i\leq j} \alpha_i\{i\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\leq i\leq j} \beta_i\{i\}$$

implies

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\leq i\leq j} \alpha_i c_i \leq \sum_{L\in\mathbf{L}} b_L + \sum_{1\leq i\leq j} \beta_i c_i \ .$$

We will chose c_{j+1} such that

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\leq i\leq j} \alpha_i\{i\} \cup \alpha_i\{j+1\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\leq i\leq j} \beta_i\{i\} \cup \beta_i\{j+1\}$$

implies

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\leq i\leq j} \alpha_i c_i + \alpha_{j+1} c_{j+1} \leq \sum_{L\in\mathbf{L}} b_L + \sum_{1\leq i\leq j} \beta_i c_i + \beta_{j+1} c_{j+1} \qquad (*)$$

Consider $\mathbf{M}$, γ_i, $\mathbf{N}$, δ_i, $\mathbf{M'}$, γ'_i, $\mathbf{N'}$, δ'_i arbitrary such that

$$\bigcup_{M\in\mathbf{M}} M \cup \bigcup_{1\leq i\leq j} \gamma_i\{i\} = \bigcup_{N\in\mathbf{N}} N \cup \bigcup_{1\leq i\leq j} \delta_i\{i\} \cup \delta_{j+1}\{j+1\}$$

$$\bigcup_{M'\in\mathbf{M'}} M' \cup \bigcup_{1\leq i\leq j} \gamma'_i\{i\} = \bigcup_{N'\in\mathbf{N'}} N' \cup \bigcup_{1\leq i\leq j} \delta'_i\{i\} \cup ({\delta_{j+1}}')\{j+1\}$$

Then

$$\bigcup_{M\in\mathbf{M}} \delta'_{j+1} M \cup \bigcup_{1\leq i\leq j} \delta'_{j+1}\gamma_i\{i\} = \bigcup_{N\in\mathbf{N}} \delta'_{j+1} N \cup \bigcup_{1\leq i\leq j} \delta'_{j+1}\delta_i\{i\} \cup (\delta'_{j+1}\delta_{j+1})\{j+1\}$$

and

$$\bigcup_{M'\in\mathbf{M}'} \delta_{j+1}M' \cup \bigcup_{1\le i\le j} \delta_{j+1}\gamma'_i\{i\} = \bigcup_{N'\in\mathbf{N}'} \delta_{j+1}N' \cup \bigcup_{1\le i\le j} \delta_{j+1}\delta'_i\{i\} \cup (\delta_{j+1}\delta'_{j+1})\{j+1\}$$

Hence,

$$\bigcup_{M\in\mathbf{M}} \delta'_{j+1}M \cup \bigcup_{1\le i\le j} \delta'_{j+1}\gamma_i\{i\} \cup \bigcup_{N'\in\mathbf{N}'} \delta_{j+1}N' \cup \bigcup_{1\le i\le j} \delta_{j+1}\delta'_i\{i\}$$

$$= \bigcup_{M'\in\mathbf{M}'} \delta_{j+1}M' \cup \bigcup_{1\le i\le j} \delta_{j+1}\gamma'_i\{i\} \cup \bigcup_{N\in\mathbf{N}} \delta'_{j+1}N \cup \bigcup_{1\le i\le j} \delta'_{j+1}\delta_i\{i\}$$

and thus, by induction hypothesis we see that:

$$\sum_{M\in\mathbf{M}} \delta'_{j+1}a_M + \sum_{N'\in\mathbf{N}'} \delta_{j+1}a_{N'} + \sum_{1\le i\le j} (\delta'_{j+1}\gamma_i + \delta_{j+1}\delta'_i)c_i \le$$

$$\sum_{M'\in\mathbf{M}'} \delta_{j+1}b_{M'} + \sum_{N\in\mathbf{N}} \delta'_{j+1}b_N + \sum_{1\le i\le j} (\delta_{j+1}\gamma'_i + \delta'_{j+1}\delta_i)c_i$$

and thus

$$\delta'_{j+1}(\sum_{M\in\mathbf{M}} a_M - \sum_{N\in\mathbf{N}} b_N + \sum_{1\le i\le j} (\gamma_i - \delta_i)c_i) \le \delta_{j+1}(\sum_{M'\in\mathbf{M}'} b_{M'} - \sum_{N'\in\mathbf{N}'} a_{N'} + \sum_{1\le i\le j} (\gamma'_i - \delta'_i)c_i)$$

Since $\mathbf{M}$, γ_i, $\mathbf{N}$, δ_i, $\mathbf{M}'$, γ'_i, $\mathbf{N}'$, and δ'_i are arbitrary chosen, we can conclude that for every $\mathbf{M}$, γ_i, $\mathbf{N}$, δ_i with

$$\bigcup_{M\in\mathbf{M}} M \cup \bigcup_{1\le i\le j} \gamma_i\{i\} = \bigcup_{N\in\mathbf{N}} N \cup \bigcup_{1\le i\le j} \delta_i\{i\} \cup \delta_{j+1}\{j+1\}$$

there is a c_{j+1} such that

$$\sum_{M\in\mathbf{M}} a_M + \sum_{1\le i\le j} \gamma_i c_i - \sum_{N\in\mathbf{N}} b_N - \sum_{1\le i\le j} \delta_i c_i \le \delta_{j+1}c_{j+1}$$

and

$$\delta_{j+1}c_{j+1} \le \sum_{M\in\mathbf{M}} b_M + \sum_{1\le i\le j} \gamma_i c_i - \sum_{N\in\mathbf{N}} a_N - \sum_{1\le i\le j} \delta_i c_i$$

Now that we have chosen c_{j+1} we will proof (*).
Let

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\le i\le j} \alpha_i\{i\} \cup \alpha_{j+1}\{j+1\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\le i\le j} \beta_i\{i\} \cup \beta_{j+1}\{j+1\}$$

There are three cases:

1. $\alpha_{j+1} = \beta_{j+1}$
   Then
   $$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\le i\le j} \alpha_i\{i\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\le i\le j} \beta_i\{i\},$$

hence by induction:

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le j} \alpha_i c_i \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le j} \beta_i c_i,$$

and thus

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le j} \alpha_i c_i + \alpha_{j+1} c_{j+1} \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le j} \beta_i c_i + \beta_{j+1} c_{j+1} \ .$$

2. $\alpha_{j+1} < \beta_{j+1}$
   Then

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\le i\le j} \alpha_i\{i\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\le i\le j} \beta_i\{i\} \cup (\beta_{j+1} - \alpha_{j+1})\{j+1\},$$

   hence:

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le j} \alpha_i c_i - \sum_{L\in\mathbf{L}} b_L - \sum_{1\le i\le j} \beta_i c_i \le (\beta_{j+1} - \alpha_{j+1}) c_{j+1}$$

   so

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le j} \alpha_i c_i + \alpha_{j+1} c_{j+1} \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le j} \beta_i c_i + \beta_{j+1} c_{j+1}$$

3. $\alpha_{j+1} > \beta_{j+1}$
   Then

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\le i\le j} \alpha_i\{i\} \cup (\alpha_{j+1} - \beta_{j+1})\{j+1\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\le i\le j} \beta_i\{i\},$$

   hence:

$$(\alpha_{j+1} - \beta_{j+1}) c_{j+1} \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le j} \beta_i c_i - \sum_{K\in\mathbf{K}} b_K - \sum_{1\le i\le j} \alpha_i c_i$$

   so

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le j} \alpha_i c_i + \alpha_{j+1} c_{j+1} \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le j} \beta_i c_i + \beta_{j+1} c_{j+1}$$

Finally we take $\#I$ for j and hence

$$\bigcup_{K\in\mathbf{K}} K \cup \bigcup_{1\le i\le \#I} \alpha_i\{i\} = \bigcup_{L\in\mathbf{L}} L \cup \bigcup_{1\le i\le \#I} \beta_i\{i\}$$

implies

$$\sum_{K\in\mathbf{K}} a_K + \sum_{1\le i\le \#I} \alpha_i c_i \le \sum_{L\in\mathbf{L}} b_L + \sum_{1\le i\le \#I} \beta_i c_i \ .$$

$c_1, \ldots, c_{\#I}$ is a solution. Indeed:
Let $\mathbf{K} = \{\!\{ K \}\!\}, \mathbf{L} = \phi, \alpha_i = 0, \beta_i = 0 \Leftrightarrow i \notin K, \beta_i = 1 \Leftrightarrow i \in K$
then
$a_K \le \sum_{i\in K} c_i$
Let $\mathbf{K} = \phi, \mathbf{L} = \{\!\{ K \}\!\}, \alpha_i = 0, \beta_i = 0 \Leftrightarrow i \notin K, \alpha_i = 1 \Leftrightarrow i \in K$
then
$\sum_{i\in K} c_i \le b_K$
This concludes the proof. ◁