Multidimensional Association Rules in Boolean Tensors*

Kim-Ngan T. Nguyen† Loïc Cerf‡ Marc Plantevit§ Jean-François Boulicaut¶

Abstract

Popular data mining methods support knowledge discovery from patterns that hold in binary relations. We study the generalization of association rule mining within arbitrary n-ary relations and thus Boolean tensors instead of Boolean matrices. Indeed, many datasets of interest correspond to relations whose number of dimensions is greater or equal to 3. However, just a few proposals deal with rule discovery when both the head and the body can involve subsets of any dimensions. A challenging problem is to provide a semantics to such generalized rules by means of objective interestingness measures that have to be carefully designed. Therefore, we discuss the need for different generalizations of the classical confidence measure. We also present the first algorithm that computes, in such a general framework, every rule that satisfies both a minimal frequency constraint and minimal confidence constraints. The approach is tested on real datasets (ternary and 4-ary relations). We report on a case study that deals with analyzing a dynamic graph thanks to rules.

1 Introduction.

Mining binary relations often encoded as Boolean matrices has been extensively studied. For instance, a popular application domain deals with basket data analysis, i. e., mining *Transactions* × *Products* relations. Many (local) pattern discovery techniques from potentially large relations have been proposed. Pattern types can be frequent itemsets (see, e. g., [1, 19]), closed itemsets and formal concepts (see, e. g., [23, 4]), association rules (see, e. g., [1]) or their generalizations towards, for instance, the use of negated items (see, e. g., [19, 2]) or a multi-relational setting (see, e. g., [7, 8, 15]). Thanks to decades of research, many efficient algorithms have been designed for large binary relation analysis.

It is however clear that many datasets correspond to n-ary relations where $n > 2$ and thus Boolean tensor analysis. For example, in the general setting where we have *Properties* that describe *Objects*, we may also know when (*Dates*) and where (*Places*) this holds. In other terms, we would like to discover patterns in subsets of *Objects* × *Properties* × *Dates* × *Places* (i. e., 4-ary relations) instead of losing information because of enforced projections or aggregations. A quite interesting special case of Boolean tensor corresponds to dynamic directed graph encoding where two dimensions denote the vertices (input and output ones) while other dimensions are used to introduce temporal dimensions (see the case study in Sect. 5).

Our goal is to generalize the association rule mining task [1] within a Boolean tensor setting. This is however surprisingly difficult. The two main subproblems to address are (a) the semantic specification of the patterns of interest, and (b) their efficient computation. The point (a) is about defining the pattern language and the measures of their objective interestingness. When generalized to n-ary relations, association rules may involve subsets of several of the n dimensions. In this context, what does it mean for a rule to be frequent or to have enough confidence? How to generalize other relevancy concepts such as, for example, non redundancy? Once these declarative issues revisited in the context of n-ary relations, (b) scalable methods must be designed to extract the patterns that satisfy the specification. When possible, correct and complete algorithms remain preferable. By definition, such methods list all solution patterns and only them. Performance issues are important: a good algorithm must scale in the number of dimensions, in the size (number of values) of each of these dimensions, and in the number of tuples in the relation (true values in the associated tensor).

Our contribution is threefolds. First, defining the semantics of the new type of rules in arbitrary n-ary relations has been much harder than expected. The previous work (see Sect. 6) on multidimensional association rules severely constrain the form of the rules. For instance, several approaches only consider rules involving at most one element per dimension. To the best of our knowledge, this proposal currently is the most

*This research is partly funded by the ANR Bingo2 project (2007-2011) and by a Vietnam government scholarship

†Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

‡Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

§Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France

¶Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

general extension of association rule mining [1] towards multidimensional contexts. To design the objective interestingness measures, a difficulty arises when both the body and the head can involve subsets of any dimensions. A key contribution is the proposal of the so-called *natural* and *exclusive confidence* measures. Their relevance is empirically validated, i. e., minimal thresholds on these measures support the discovery of interesting patterns in real datasets. Our second contribution concerns the design and the implementation of the first complete algorithm, namely PINARD[1], that exhaustively lists a priori interesting rules. Its enumeration principles, inspired by the closed pattern mining algorithm from [6], provide an excellent scalability. Finally, beside the empirical validation on a typical basket-like real dataset derived from the `DistroWatch` Web site[2], we report a case study on a dynamic graph analysis thanks to our multidimensional rules. This appears as a promising application domain for pattern discovery from large Boolean tensors.

Section 2 provides the formalization of our new rule pattern domain. Section 3 introduces the first algorithm that computes a priori interesting rules. Section 4 provides experimental results on a real-life ternary relation. Section 5 reports on the analysis of a real dynamic graph thanks to discovered rules. Section 6 discusses the related work. Section 7 briefly concludes.

2 Specifying a New Rule Pattern Domain.

2.1 Preliminary Definitions. The semantics of our patterns applies to arbitrary n-ary relations (or Boolean tensor). For instance, the arity, n, can be five and none of the dimensions has to be specific (e. g., temporal). These dimensions simply are n finite and disjoint sets $\{D^1, \ldots, D^n\} = \mathcal{D}$ and $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$ denotes the relation in which rules are to be discovered.

The definitions are illustrated on a toy ternary relation $\mathcal{R}_E$ (see Table 1). It relates products in $D^1 = \{p_1, p_2, p_3, p_4\}$ bought along seasons in $D^2 = \{s_1, s_2, s_3, s_4\}$ by customers in $D^3 = \{c_1, c_2, c_3, c_4, c_5\}$. Every '1', in Table 1, is at the intersection of three elements $(p_i, s_j, c_k) \in D^1 \times D^2 \times D^3$, which form a 3-tuple present in $\mathcal{R}_E$. For instance, p_1 is bought during s_1 by c_1 and bought by c_4 during s_4 only. The patterns of interest only involve some of the attribute domains $\mathcal{D}' \subseteq \mathcal{D}$. E. g., given $\mathcal{R}_E$, the analyst may want to focus on patterns involving products and seasons in which case $\mathcal{D}' = \{D^1, D^2\}$. Without loss of generality, the dimensions are assumed ordered such that $\mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\}$.

[1] PINARD Is N-ary Association Rule Discovery.
[2] `http://www.distrowatch.com`

Table 1: $\mathcal{R}_E \subseteq \{p_1, p_2, p_3, p_4\} \times \{s_1, s_2, s_3, s_4\} \times \{c_1, c_2, c_3, c_4, c_5\}$

	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4
c_1	1	1	1		1	1	1		1		1		1	1		1
c_2	1	1		1	1	1				1	1	1			1	1
c_3	1	1			1					1	1	1			1	1
c_4		1		1			1	1			1		1	1	1	
c_5												1				1
	s_1				s_2				s_3				s_4			

DEFINITION 2.1. (ASSOCIATION)
$\forall \mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\} \subseteq \mathcal{D}$, $\times_{i=1..|\mathcal{D}'|} X^i$ is an association on $\mathcal{D}'$ iff $\forall i = 1..|\mathcal{D}'|$, $X^i \neq \emptyset \wedge X^i \subseteq D^i$.

By convention, the only association on an empty set (i. e., $\mathcal{D}' = \emptyset$) is denoted $\emptyset$. Given an arbitrary association on $\mathcal{D}'$, $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$ is its *support domain*, hence generalizing the "classical" binary case. Indeed, in a *Transactions* $\times$ *Products* setting, the support domain of an association rule involving products is the set of transactions [1]. In our running example, D^3 is the support domain of every association on $\{D^1, D^2\}$. The *support* of an association is a subset of the support domain. Its definition uses concatenation denoted as '$\cdot$'. For instance, $(p_2, s_1) \cdot (c_2) = (p_2, s_1, c_2)$.

DEFINITION 2.2. (SUPPORT)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, let X be an association on $\mathcal{D}'$. Its support is $s(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \forall x \in X,\ x \cdot t \in \mathcal{R}\}$.

Let us mention some special cases. An association involving the n domains ($\mathcal{D}' = \mathcal{D}$) is either true (every n-tuple it contains is in $\mathcal{R}$), or false (at least one n-tuple it contains is absent from $\mathcal{R}$). By using the convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$ (where ϵ is the empty word), Def. 2.2 reflects that: every association on $\mathcal{D}$ either has zero or one element, ϵ, in its support. The opposite extreme case is the support of the empty association, $s(\emptyset)$, which is $\mathcal{R}$. The support of an association generalizes that of an *itemset* in a binary relation (i. e., when $n = 2$ and $\mathcal{D}' = \{D^1\}$). The cardinality of the support quantifies the frequency of an association, like it does for itemsets.

Let us now provide some useful definitions to design our rule pattern domain.

DEFINITION 2.3. (PROJECTION π)
$\forall \mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\} \subseteq \mathcal{D}$, let $X = X^1 \times \cdots \times X^{|\mathcal{D}'|}$ be an association on $\mathcal{D}'$. $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X)$ is X^i if $D^i \in \mathcal{D}'$, $\emptyset$ otherwise.

DEFINITION 2.4. (UNION $\sqcup$)
$\forall \mathcal{D}_X \subseteq \mathcal{D}$ and $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, let X (resp. Y) be an association on $\mathcal{D}_X$ (resp. on $\mathcal{D}_Y$). $X \sqcup Y$ is the association on $\mathcal{D}_X \cup \mathcal{D}_Y$ for which $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X \sqcup Y) = \pi_{D^i}(X) \cup \pi_{D^i}(Y)$.

Definition 2.5. (Complement $\setminus$)
$\forall \mathcal{D}_X \subseteq \mathcal{D}$ and $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, let X (resp. Y) be an association on $\mathcal{D}_X$ (resp. on $\mathcal{D}_Y$). $Y \setminus X$ is the association on $\{D^i \in \mathcal{D}_Y \mid \pi_{D^i}(Y) \not\subseteq \pi_{D^i}(X)\}$ for which $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(Y \setminus X) = \pi_{D^i}(Y) \setminus \pi_{D^i}(X)$.

Definition 2.6. (Inclusion $\sqsubseteq$)
$\forall \mathcal{D}_X \subseteq \mathcal{D}$ and $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, let X (resp. Y) be an association on $\mathcal{D}_X$ (resp. on $\mathcal{D}_Y$). X is included in Y, denoted $X \sqsubseteq Y$, iff $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X) \subseteq \pi_{D^i}(Y)$.

With this straightforward generalization of the inclusion, the *anti-monotonicity* of the frequency (i. e., of the support cardinality), that is well known in itemset mining, still holds with associations. The proof is given in annex.

Theorem 2.1. (Frequency anti-monotonicity)
$\forall \mathcal{D}_X \subseteq \mathcal{D}$ and $\forall \mathcal{D}_Y \subseteq \mathcal{D}$, let X (resp. Y) be an association on $\mathcal{D}_X$ (resp. on $\mathcal{D}_Y$), $X \sqsubseteq Y \Rightarrow |s(X)| \geq |s(Y)|$.

In $\mathcal{R}_E$, $\{p_1, p_2\} \times \{s_1\}$ and $\{p_1, p_2\} \times \{s_1, s_2\}$ are two associations on $\{D^1, D^2\}$, whereas $\{p_1, p_2\}$ is an association on $\{D^1\}$ ($\pi_{D^2}(\{p_1, p_2\}) = \emptyset$). We have:

- $s(\{p_1, p_2\} \times \{s_1\}) = \{c_1, c_2, c_3\}$;
- $s(\{p_1, p_2\} \times \{s_1, s_2\}) = \{c_1, c_2\}$;
- $s(\{p_1, p_2\}) = \{(s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_4, c_1), (s_4, c_4)\}$.

Because $\{p_1, p_2\} \sqsubseteq \{p_1, p_2\} \times \{s_1\} \sqsubseteq \{p_1, p_2\} \times \{s_1, s_2\}$, Th. 2.1 holds. Indeed, $|s(\{p_1, p_2\})| \geq |s(\{p_1, p_2\} \times \{s_1\})| \geq |s(\{p_1, p_2\} \times \{s_1, s_2\})|$.

2.2 Multidimensional Association Rules. Given an n-ary relation $\mathcal{R}$ on $\mathcal{D}$ and the user-defined domains of interest $\mathcal{D}' \subseteq \mathcal{D}$, a *multidimensional association rule* on $\mathcal{D}'$ is a couple of associations whose union is an association on $\mathcal{D}'$. It is simply called a rule when it is clear from the context.

Definition 2.7. (Rule)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, $X \to Y$ is a multidimensional association rule on $\mathcal{D}'$ iff $X \sqcup Y$ is an association on $\mathcal{D}'$.

In $\mathcal{R}_E$, $\{p_1, p_2\} \to \{s_1, s_2\}$ and $\{p_4\} \times \{s_3, s_4\} \to \{p_3\}$ are two rules on $\{D^1, D^2\}$. $\{p_1\} \to \{p_2\}$ is not a rule on $\{D^1, D^2\}$ because no element in D^2 appears in its *body* (the association on the left hand side of '$\to$') or in its *head* (the association on the right hand side of '$\to$'). It is a rule on $\{D^1\}$.

In the binary case (i. e., $n = 2$), the classical semantics of association rules is based on two measures: a frequency and a confidence. A priori interesting rules are defined as those whose both measures exceed user-specified thresholds [1]. A rule is frequent if it is supported by enough objects. A rule can be trusted, i. e., the analysts can be confident in it, if there is a high enough conditional probability to observe the head when the body holds.

In the context of n-ary relations, it turns out that a natural definition of rule frequency exists. On the contrary, it is hard to define a confidence measure for general rules. More precisely, the difficulty arises for any rule whose head involves some dimension that is not in its body.

2.3 Rule Frequency. The (relative) frequency of an association rule is a proportion of elements in the support domain of the union of its body and its head.

Definition 2.8. (Frequency)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ a rule on $\mathcal{D}'$. Its frequency is:

$$f(X \to Y) = \frac{|s(X \sqcup Y)|}{|\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i|} \ .$$

In $\mathcal{R}_E$, we have:

- $f(\{p_1, p_2\} \to \{s_1, s_2\}) = \frac{|s(\{p_1, p_2\} \times \{s_1, s_2\})|}{|D^3|} = \frac{|\{c_1, c_2\}|}{|\{c_1, c_2, c_3, c_4, c_5\}|} = \frac{2}{5}$;
- $f(\{p_4\} \times \{s_3, s_4\} \to \{p_3\}) = \frac{|s(\{p_3, p_4\} \times \{s_3, s_4\})|}{|D^3|} = \frac{2}{5}$.

2.4 Rule Confidence.

2.4.1 The Problem. Is it possible and useful to directly generalize the confidence measure of association rules in binary relations to n-ary relations? Doing so, the confidence of a rule $X \to Y$ would be $\frac{|s(X \sqcup Y)|}{|s(X)|}$. If X and $X \sqcup Y$ are associations on the same domain(s) (they have the same support domain), this definition is intuitive: the confidence is a proportion of elements in a same support domain. For instance, in $\mathcal{R}_E$, the confidence of $\{p_4\} \times \{s_3, s_4\} \to \{p_3\}$ would be: $\frac{|s(\{p_3, p_4\} \times \{s_3, s_4\})|}{|s(\{p_4\} \times \{s_3, s_4\})|} = \frac{|\{c_2, c_3\}|}{|\{c_2, c_3, c_5\}|} = \frac{2}{3}$. It is a proportion of customers and it means that the customers who buy p_4 during both s_3 and s_4 also tend to buy p_3 during these seasons.

Nevertheless, this semantics is not satisfactory for any rule whose head involves some dimension that is not in its body. Indeed, in this case, $s(X \sqcup Y)$ and $s(X)$ are incomparable sets and the ratio of their cardinalities does not make any sense. For instance, in $\mathcal{R}_E$, consider the rule $\{p_1, p_2\} \to \{s_1, s_2\}$. $s(\{p_1, p_2\} \times \{s_1, s_2\}) = \{c_1, c_2\}$ is a set of customers, whereas $s(\{p_1, p_2\})$ is not. It contains couples such as (s_1, c_1) or (s_2, c_1). As a result, there is a need for a new confidence measure that would make sense for any multidimensional association

rule $X \to Y$. This measure should be equal to $\frac{|s(X \sqcup Y)|}{|s(X)|}$ when X and $X \sqcup Y$ are defined on the same domain(s).

2.4.2 Exclusive Confidence. Computing the confidence of a rule $X \to Y$ on $\mathcal{D}'$ is problematic if X is defined on a set $\mathcal{D}_X$ *strictly* included in $\mathcal{D}'$. However, it is possible to introduce a factor such that $|s(X)|$ and $|s(X \sqcup Y)|$ become comparable. The idea is to multiply $|s(X \sqcup Y)|$ by the cardinalities of its projections in the domains that are absent from $\mathcal{D}_X$.

Definition 2.9. (Exclusive Confidence)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ a rule on $\mathcal{D}'$ and $\mathcal{D}_X$ the domains on which X is defined. Its exclusive confidence is:

$$c_{exclusive}(X \to Y) = \frac{|s(X \sqcup Y)| \times |\times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} \pi_{D^i}(Y)|}{|s(X)|}$$

Roughly speaking, the remedial factor $|\times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} \pi_{D^i}(Y)|$, applied to $|s(X \sqcup Y)|$, allows to count the elements at the numerator of the fraction "in the same way" as those at the denominator. As desired (see Sect. 2.4.1), if X is an association on $\mathcal{D}'$, the exclusive confidence of $X \to Y$ is $\frac{|s(X \sqcup Y)|}{|s(X)|}$ under the convention $\times_{D^i \in \emptyset} \pi_{D^i}(Y) = \{\epsilon\}$.

For example, consider the rule $\{p_1, p_2\} \to \{s_1, s_2\}$ in $\mathcal{R}_E$ and let us name *transaction* a customer's purchase during a specific season. There are two customers, c_1 and c_2, who buy both products p_1 and p_2 during both seasons s_1 and s_2, i. e., we have $|\{c_1, c_2\}| \times |\{s_1, s_2\}| = 4$ transactions. Consider now the body of the rule, i. e., $\{p_1, p_2\}$. Seven transactions, (s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_4, c_1) and (s_4, c_4), involve both p_1 and p_2. Thus, $c_{\text{exclusive}}(\{p_1, p_2\} \to \{s_1, s_2\})$ is:

$$\frac{|s(\{p_1, p_2\} \times \{s_1, s_2\})| \times |\{s_1, s_2\}|}{|s(\{p_1, p_2\})|} = \frac{4}{7}$$

The customer c_3 buys both products p_1 and p_2 during the season s_1, whereas he/she does not buy them together during the season s_2. This actually lowers the confidence in the fact that customers like buying both products during the seasons s_1 and s_2. Notice also that the customer c_1 buying these two products during season s_4 lowers the confidence as well. In fact, the exclusive confidence $c_{\text{exclusive}}(\{p_1, p_2\} \to \{s_1, s_2\})$ indicates to what extent the products p_1 and p_2 are bought together during the seasons s_1 and s_2 *only*. This *exclusivity* explains the chosen name. If $c_{\text{exclusive}}(\{p_1, p_2\} \to \{s_1, s_2\})$ was 1, every customer who buys p_1 and p_2 together would *always* do so during both seasons s_1 and s_2 (and *never* during another season).

This exclusivity aspect makes sense for the discovery of interesting association rules. Indeed, it penalizes the rules with "non-maximal" heads. For instance, the rule $\{p_1, p_2\} \to \{s_2\}$ is supported by the customers c_1 and c_2, who also buy the product p_1 and p_2 during season s_1. That is why the exclusive confidence of this rule, $\frac{2}{7}$, is lower than that of $\{p_1, p_2\} \to \{s_1, s_2\}$, $\frac{4}{7}$. The difference of two transactions at the numerator directly relates with the two customers c_1 and c_2, who also buy the products p_1 and p_2 during season s_1.

Unfortunately, this exclusivity also makes the function $X \mapsto c_{\text{exclusive}}(X \to Y \setminus X)$ (with $X \sqsubseteq Y$) not increase w.r.t. $\sqsubseteq$. For example, consider the rules $\{s_3\} \to \{p_2, p_3, p_4\}$ and $\{s_3\} \times \{p_3\} \to \{p_2, p_4\}$ in $\mathcal{R}_E$. We observe that $\{s_3\} \sqsubseteq \{s_3\} \times \{p_3\}$, however $c_{\text{exclusive}}(\{s_3\} \times \{p_3\} \to \{p_2, p_4\}) < c_{\text{exclusive}}(\{s_3\} \to \{p_2, p_3, p_4\})$ ($\frac{2}{4} < \frac{6}{10}$). This absence of property prevents the sound use of anti-monotonic pruning to efficiently list every rule having an exclusive confidence greater than a user-defined threshold. Let us now consider an alternative definition for the confidence.

2.4.3 Natural Confidence. To define the confidence of $X \to Y$, a straightforward generalization of the binary case is problematic when the support domain of X is different from that of $X \sqcup Y$. "Forcing" the support of X to be a subset of the support domain $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$ of $X \sqcup Y$ allows to define a confidence measure that is a *natural* proportion, i. e., a proportion of elements in a same support domain. The cost of such a natural confidence is the need for a new definition of the support when applied to rule bodies.

Definition 2.10. (Natural support of bodies)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ be a rule on $\mathcal{D}'$. The natural support of X is:

$$s_{\mathcal{D} \setminus \mathcal{D}'}(X) = \{t \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \exists u \in \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} D^i \text{ such that } \forall x \in X,\ x \cdot u \cdot t \in \mathcal{R}\}\ ,$$

where $\mathcal{D}_X$ is the set of domains on which X is defined. For $x \cdot u \cdot t$ to possibly be in $\mathcal{R}$, the domains in $\mathcal{D}_X$ must appear first, i. e., the domain index may have to be changed.

Definition 2.11. (Natural confidence)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ be a rule on $\mathcal{D}'$. Its natural confidence is:

$$c_{natural}(X \to Y) = \frac{|s(X \sqcup Y)|}{|s_{\mathcal{D} \setminus \mathcal{D}'}(X)|}\ .$$

As desired (see Sect. 2.4.1), if X is an association on $\mathcal{D}'$, the natural confidence of $X \to Y$ is $\frac{|s(X \sqcup Y)|}{|s(X)|}$ under the convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$.

Once again, consider the rule $\{p_1, p_2\} \to \{s_1, s_2\}$ in $\mathcal{R}_E$. The customers who buy the products p_1 and p_2

together (during at least one season) are c_1, c_2, c_3, and c_4. Among them, only c_1 and c_2 buy p_1 and p_2 during both seasons s_1 and s_2. Thus, the natural confidence $c_{\text{natural}}(\{p_1, p_2\} \to \{s_1, s_2\})$ is:

$$\frac{|s(\{p_1, p_2\} \times \{s_1, s_2\})|}{|s_{\{D^3\}}(\{p_1, p_2\})|} = \frac{|\{c_1, c_2\}|}{|\{c_1, c_2, c_3, c_4\}|} = \frac{2}{4} \ .$$

It means that half of the customers buying both p_1 and p_2 during a same season do so during both seasons s_1 and s_2. Now, the customers who support the rule can buy both p_1 and p_2 during another season and that does not "lower" the natural confidence, whereas it does lower the exclusive one (see Sect. 2.4.2). Moreover, the natural confidence can give rise to pruning during the rule enumeration process.

THEOREM 2.2. (PRUNING CRITERION) *Let $X \to Y \setminus X$ and $X' \to Y \setminus X'$ be two rules on $\mathcal{D}'$, we have: $X \sqsubseteq X' \sqsubseteq Y \Rightarrow c_{natural}(X \to Y \setminus X) \leq c_{natural}(X' \to Y \setminus X')$.*

The proof is given in the technical annex. In $\mathcal{R}_E$, $\{p_1, p_2\} \to \{s_1, s_2\}$ and $\{p_1, p_2\} \times \{s_1\} \to \{s_2\}$ are two rules on $\{D^1, D^2\}$. The natural confidence of the first rule is $\frac{2}{4}$ (see above). The natural confidence of the second one is $\frac{|s(\{p_1,p_2\}\times\{s_1,s_2\})|}{|s_{D^3}(\{p_1,p_2\}\times\{s_1\})|} = \frac{|\{c_1,c_2\}|}{|\{c_1,c_2,c_3\}|} = \frac{2}{3}$. It illustrates Th. 2.2. Indeed, $\{p_1, p_2\} \sqsubseteq \{p_1, p_2\} \times \{s_1\} \sqsubseteq \{p_1, p_2\} \times \{s_1, s_2\}$ and $c_{\text{natural}}(\{p_1, p_2\} \to \{s_1, s_2\}) \leq c_{\text{natural}}(\{p_1, p_2\} \times \{s_1\} \to \{s_2\})$. In Sect. 3, this theorem is used to prune the search space where no rule can satisfy a minimal natural confidence constraint.

2.5 Canonical Rules.

DEFINITION 2.12. (SYNTACTIC EQUIVALENCE)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, the rules $X \to Y$ and $X \to Z$ on $\mathcal{D}'$ are syntactically equivalent iff $X \sqcup Y = X \sqcup Z$.

Proving the following lemma is straightforward.

LEMMA 2.1. *Syntactically equivalent rules have the same frequency, the same exclusive confidence and the same natural confidence.*

DEFINITION 2.13. (CANONICAL RULE)
$\forall \mathcal{D}' \subseteq \mathcal{D}$, a rule $X \to Y$ on $\mathcal{D}'$ is canonical iff $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X) \cap \pi_{D^i}(Y) = \emptyset$.

Any complete collection of rules satisfying constraints of frequency and/or confidences can be condensed, without any loss of information, into its canonical rules only. Indeed, given a canonical association rule $X \to Y$ in the collection, Lemma 2.1 entails that all syntactically equivalent rules necessary are in the collection as well. Moreover constructing them is easy: they are the association rules $X \to Y \sqcup Z$ with $Z \sqsubseteq X$.

3 Computing Rules.

Given an n-ary relation $\mathcal{R} \subseteq \times_{D^i \in \mathcal{D}} D^i$, every *a priori* interesting canonical association rule is to be listed. These rules are defined on a chosen subset $\mathcal{D}' \subsetneq \mathcal{D}$, have their frequencies beyond $\mu \in [0; 1]$, their exclusive confidences beyond $\beta_{\text{exclusive}} \in [0; 1]$, and their natural confidences beyond $\beta_{\text{natural}} \in [0; 1]$. In other terms, the algorithm PINARD computes:

$$\{X \to Y \text{ on } \mathcal{D}' \mid \begin{cases} X \to Y \text{ is canonical} \\ f(X \to Y) \geq \mu \\ c_{\text{exclusive}}(X \to Y) \geq \beta_{\text{exclusive}} \\ c_{\text{natural}}(X \to Y) \geq \beta_{\text{natural}} \end{cases} \} \ .$$

PINARD first constructs a new relation $\mathcal{R}_T$ from $\mathcal{R}$. The support domain of any association rule on $\mathcal{D}'$ is $D^{\text{supp}} = \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$. Let $\mathcal{D}_T = \mathcal{D}' \cup D^{\text{supp}}$. The relation $\mathcal{R}_T$ on $\mathcal{D}_T$ is defined as follows:

$$\begin{array}{ll} & (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, e_{|\mathcal{D}'|+1}, \ldots, e_n) \in \mathcal{R} \\ \Leftrightarrow & (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, (e_{|\mathcal{D}'|+1}, \ldots, e_n)) \in \mathcal{R}_T \ . \end{array}$$

The next step is to compute the frequent associations from which the a priori interesting rules will be derived. This can be formalized as the search for every association T on $\mathcal{D}_T$ that satisfies the four following constraints:

- $\mathcal{C}_{\text{connected}}(T) \equiv T \subseteq \mathcal{R}_T$;
- $\mathcal{C}_{\text{on-}\mathcal{D}'}(T) \equiv \forall D^i \in \mathcal{D}', \pi_{D^i}(T) \neq \emptyset$;
- $\mathcal{C}_{\text{entire-supp}}(T) \equiv \pi_{D^{\text{supp}}}(T) = s(T \setminus \pi_{D^{\text{supp}}}(T))$;
- $\mathcal{C}_{\text{freq}}(T) \equiv \frac{|\pi_{D^{\text{supp}}}(T)|}{|D^{\text{supp}}|} \geq \mu$.

The first and the second constraints relate to the definition of an association: T must cover only tuples present in $\mathcal{R}_T$ and $T \setminus \pi_{D^{\text{supp}}}(T)$ must be an association on $\mathcal{D}'$. The third constraint enforces a "closed" support. Indeed, by definition of the support (Def. 2.2), adding an element $f \in D^{\text{supp}} \setminus \pi_{D^{\text{supp}}}(T)$ to T necessarily violates $\mathcal{C}_{\text{connected}}$. Thus, $\mathcal{C}_{\text{entire-supp}}(T)$ is equivalent to $\forall f \in D^{\text{supp}} \setminus \pi_{D^{\text{supp}}}(T), (T \setminus \pi_{D^{\text{supp}}}(T)) \sqcup \{f\} \not\subseteq \mathcal{R}_T$. The last constraint guarantees that the frequency of every association rule involving all elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(T)$ is greater or equal to μ.

Constraint-based mining of closed associations has been recently studied [6, 14, 16]. It has given rise to an extremely efficient enumeration strategy implemented in the state-of-the-art algorithm DATA-PEELER [6]. Furthermore this extractor can handle a very broad class of constraints including the four above. Building upon these enumeration principles actually motivated

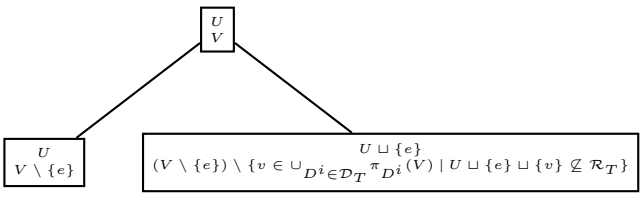


Figure 1: Enumerating the element $e \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V)$.

the constraint-based formalization of our problem. Nevertheless, we do not exactly want the *closed* associations on $\mathcal{D}_T$. Indeed, a frequent association is only closed on the support dimension, D^{supp}, whereas a closed association (called a closed n-set in [6]) is closed on all the dimensions in $\mathcal{D}_T$. Therefore, we adapt the algorithm from [6] to the discovery of every frequent association in $\mathcal{R}_T$. Here we present an abstract view of this process. Technical details can be found in [6].

PINARD recursively partitions the search space into two complementary parts ("divide and conquer"). In this way, a binary tree can represent the search space traversal. At every node of this tree, two associations, namely U and V, are updated. U is, according to $\sqsubseteq$, the smallest association that may be discovered from the node, whereas $U \sqcup V$ is the largest. That is why PINARD is initially called with $U = \emptyset$ and $V = \times_{D^i \in \mathcal{D}_T} D^i$. In an enumeration sub-tree rooted by a left child, an arbitrary element $e \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V)$ is absent from every U association (e is "removed" from V). In the enumeration sub-tree rooted by its sibling node (right child), the same element e is present in every U association (e is "moved" from V to U). Right after an element e is "moved" to U (right child), the constraint $\mathcal{C}_{\text{connected}}$ is enforced. It removes from V every element $v \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V)$ that would violate $\mathcal{C}_{\text{connected}}$ if added to $(U \sqcup \{e\})$, i. e., $U \sqcup \{e\} \sqcup \{v\} \not\sqsubseteq \mathcal{R}_T$. Figure 1 sums up this enumeration process. A left child is traversed first unless the enumerated element is in D^{supp}. This design grants better performance when generating the rules. This is explained later.

An enumeration sub-tree is not explored if at least one of the other three constraints ($\mathcal{C}_{\text{on-}\mathcal{D}'}$, $\mathcal{C}_{\text{entire-supp}}$ or $\mathcal{C}_{\text{freq}}$) is guaranteed to be violated by every U association in it. This guarantee is easily checked at the root of the sub-tree thanks to a generalized anti-monotone property all three constraints satisfy, i. e., if an association violates one of them then every smaller association (w.r.t. $\sqsubseteq$) violates it as well. Thus, if $U \sqcup V$ (the largest association in the sub-tree) violates one of these constraints, the guarantee holds and PINARD aborts the exploration of the related part of the search space. Other anti-monotone constraints can be enforced to enhance the relevance of the associations and provide faster extractions. E. g., minimal numbers of elements can be specified for every dimension of the rule:

$$\mathcal{C}_{(\alpha^i)_{i=1..|\mathcal{D}'|}\text{-sizes}}(T) \equiv \forall D^i \in \mathcal{D}', |\pi_{D^i}(T)| \geq \alpha^i \ .$$

The only other reason for an enumeration node to be a leaf, despite its satisfaction of the constraints, is the actual discovery of a frequent association. It happens when $V = \emptyset$, i. e., when there is no more element to enumerate. Algorithm 3.1 sums up the extraction of every frequent association.

ALGORITHM 3.1. PINARD

Input: (U, V)
Output: Every *a priori* interesting association rule involving every element in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(U)$ and possibly some elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(V)$
if $\mathcal{C}_{\text{on-}\mathcal{D}_T}(U \sqcup V) \wedge \mathcal{C}_{\text{entire-supp}}(U \sqcup V) \wedge \mathcal{C}_{\text{freq}}(U \sqcup V)$ **then**
  if $V = \emptyset$ **then**
    RULES$(U \setminus \pi_{D^{\text{supp}}}(U), \emptyset)$
  else
    Choose $e \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V)$
    if $e \in D^{\text{supp}}$ **then**
      PINARD$(U \sqcup \{e\}, (V \setminus \{e\}) \setminus \{v \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V) \mid U \sqcup \{e\} \sqcup \{v\} \not\sqsubseteq \mathcal{R}_T\})$
      PINARD$(U, V \setminus \{e\})$
    else
      PINARD$(U, V \setminus \{e\})$
      PINARD$(U \sqcup \{e\}, (V \setminus \{e\}) \setminus \{v \in \cup_{D^i \in \mathcal{D}_T} \pi_{D^i}(V) \mid U \sqcup \{e\} \sqcup \{v\} \not\sqsubseteq \mathcal{R}_T\})$
    end if
  end if
end if

RULES (Alg. 3.2) computes *a priori* interesting rules, of the form $B \rightarrow H$, whenever a frequent association A ($= U \setminus \pi_{D^{\text{supp}}}(U)$ in Alg. 3.1) is discovered. It splits *all* elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(A)$ between the body B and the head H, i. e., $B \sqcup H = A$. The candidate rules are, again, structured in a tree. By only looking at the heads, H, of the rules (A and H being given, the body, B, is $A \setminus H$), this tree actually is that of APriori [1]. Nevertheless, RULES traverses it depth-first. The root of the tree is $A \rightarrow \emptyset$. At every level, H grows by an element which is removed from B. An arbitrary total order $\prec$ is chosen for the elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(A)$. At every node, the singletons that are allowed to augment (via $\sqcup$) the head are those greater than any element in the current head (i. e., greater than $\max_{\prec}(H)$ and under the convention specifying that $\max_{\prec}(\emptyset)$ is smaller than any other element). The pruning criterion is the minimal natural confidence constraint. According to Th. 2.2, this pruning is safe, i. e., no rule, with a high enough

natural confidence, is missed. As shown in Sect. 2.4.2, the exclusive confidence is not always decreasing along an enumeration branch. That is why it is computed just before a rule is possibly output. The rule is eventually output if this confidence is greater than $\beta_{\text{exclusive}}$.

Algorithm 3.2. Rules

Input: (B, H)
Output: Every canonical association rule with all elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(B \sqcup H)$, a body smaller than B (according to $\sqsubseteq$), a head larger than H (according to $\sqsubseteq$) and satisfying the minimal confidence constraints
for all $e \succ \max_{\prec}(H)$ **do**
  $(B', H') \leftarrow (B \setminus \{e\}, H \sqcup \{e\})$
  if $c_{\text{natural}}(B' \rightarrow H') \geq \beta_{\text{natural}}$ **then**
    if $c_{\text{exclusive}}(B' \rightarrow H') \geq \beta_{\text{exclusive}}$ **then**
      Output $B' \rightarrow H'$
    end if
    Rules(B', H')
  end if
end for

By storing, in an associative array, the frequency of every frequent association discovered so far, the cost of computing the denominators of the confidence measures can be reduced (at the numerator, $s(B \sqcup H) = s(A)$ is constant all along Rules's computation). Indeed, when a rule $B \rightarrow H$ is derived from a frequent association A, $B \sqsubseteq A$ may have already been discovered and its frequency $|s(B)|$ is retrieved without accessing $\mathcal{R}_T$. To profit as much as possible from this, Rules had better been initially called on increasingly larger (w.r.t. $\sqsubseteq$) associations on $\mathcal{D}'$. This actually is, in Alg. 3.1, the reason for reversing the two Pinard sub-calls depending on the condition "**if** $e \in D^{\text{supp}}$": it first explores the association search space where the enumerated element is absent unless this element is in the support dimension. In this way, and according to Th. 2.1 (larger associations have decreasing supports), when Rules is initially called on A, all associations $A' \sqsubseteq A$ on $\mathcal{D}'$ have been discovered, and treated, earlier. When Rules actually needs to access $\mathcal{R}_T$ for the computation of $s(B)$ (this may only happen if B is not an association on $\mathcal{D}'$) or $s_{\mathcal{D} \setminus \mathcal{D}'}(B)$, their cardinalities are also stored in associative arrays to, again, potentially reduce the cost of computing the confidences of the rules that remained to be discovered.

To enhance the quality of the computed rules, we can enforce other user-defined constraints. For example, non redundancy can be specified. A rule $X \rightarrow Y$ is said redundant iff it exists another rule $X' \rightarrow Y'$ such that $(X' \sqcup Y' = X \sqcup Y) \wedge (X' \sqsubset X) \wedge (c_{\text{natural}}(X' \rightarrow Y') \geq c_{\text{natural}}(X \rightarrow Y)) \wedge (c_{\text{exclusive}}(X' \rightarrow Y') \geq c_{\text{exclusive}}(X \rightarrow Y))$.

4 Empirical Validation.

To analyze the behavior of Pinard, we conducted experiments on a real-world dataset. Every experiment has been performed on a GNU/Linux™ system equipped with an Intel® Core™2 Duo CPU E7300 at 2.66 GHz and 3 GB of RAM. Pinard was implemented in C++ and compiled with GCC 4.2.4.

`DistroWatch`[3] is a Web site gathering a comprehensive information about GNU/Linux™, BSD and Solaris operating systems. Every distribution is described on a separate page. When a visitor loads a page, his/her country is known from the IP address. The logs of the Web server are easily converted into a three dimensional tensor that gives for any time period (13 semesters from early 2004 to early 2010) the number of visits from any country on any page (describing 655 distributions). The countries associated with 2,000 or more consultations in at least one semester were kept. Those are the 96 "most active" countries. Then, the numerical data are normalized so that every couple (semester,country) has the same weight. Finally, a procedure, inspired by the computation of a p value, *locally* chooses the relevant 3-tuples: for every distribution (hence, "locally"), the 3-tuples associated with the greatest normalized valued are kept until their sum reaches 20% of the sum of all normalized values involving the distribution. In this way, a 3-tuple (c, d, s) belongs to the resulting relation, $\mathcal{R}_{\texttt{DistroWatch}}$, when a significant amount of users from country c have been visiting the description of the distribution d during semester s. $\mathcal{R}_{\texttt{DistroWatch}}$ contains 21,033 3-tuples, hence a $\frac{21{,}033}{96 \times 655 \times 13} = 2.6\%$ density.

We analyze the results of the experiments with regard to the following questions: (a) Do the discovered rules make sense? (b) What do the different confidence definitions capture?, and (c) How does the algorithm Pinard behave with respect to parameter settings?

Let us first discuss a qualitative study where we look for rules that involve countries and distributions. These two dimensions form the set $\mathcal{D}'$. With the thresholds $\mu = 0.75$, $\beta_{\text{exclusive}} = 0.6$ and $\beta_{\text{natural}} = 0.8$, Pinard computes 58 canonical rules. Here as some of them:

- {Taiwan} × {Fedora} → {B2D}
  (f : 0.846, c_{natural} : 0.917, $c_{\text{exclusive}}$: 0.917);
- {Japan} × {CentOS} → {Ecuador}
  (f : 0.769, c_{natural} : 0.909, $c_{\text{exclusive}}$: 0.909);
- {Berry, Plamo} → {Japan}
  (f : 0.923, c_{natural} : 1, $c_{\text{exclusive}}$: 0.75);
- {Berry, Momonga, Plamo} → {Japan}
  (f : 0.769, c_{natural} : 1, $c_{\text{exclusive}}$: 1);

[3] `http://www.distrowatch.com`

- {Caixa Mágica} → {Portugal}
  ($f : 0.846$, $c_{\text{natural}} : 1$, $c_{\text{exclusive}} : 1$).

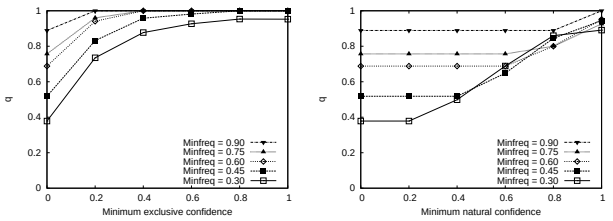
The first rule listed above indicates that when (i. e., the semesters during which) the Taiwanese visitors of `DistroWatch` show interest in *Fedora* then they usually show interest in *B2D* too ($c_{\text{natural}} = c_{\text{exclusive}} = 0.917$). The probability that Ecuadorian people consult *CentOS*, during the semesters Japanese do so, is greater than 90% (the second rule having 0.909 for confidences). Japan is the origin country of *Berry* and *Plamo*, i. e., these distributions are developed by Japanese people. That certainly explains why the visits on the related Web pages almost exclusively come from Japan. Indeed, the natural confidence of the third rule is 1, i. e., whenever (i. e., all semesters during which) both the *Berry* and the *Plamo* pages are loaded, the Japanese people do so. The high exclusive confidence of this rule (0.75) also indicates that visitors from other countries rarely have this behavior. Since the fourth rule adds a third Japanese-developed distribution, *Momonga*, at the body, the resulting exclusive confidence is even higher. It is 1, i. e., outside Japan, no other country frequently consults those three distributions at the same semester. The same interpretation holds for the last rule, i. e., *Caixa Mágica* being developed by and for people in Portugal, it is only visited by them ($c_{\text{natural}} = c_{\text{exclusive}} = 1$).

In fact, most of the discovered rules of the form *distributions* → *countries* involve countries where the distributions are developed. These rules clearly make sense and validate our semantics. Indeed, distributions that are specifically developed by and for a country (with, often, language specifics taken into account) mainly attract users from this country. The proportion of such rules (among those of the same form) is:

$$q = \frac{|\{D \to P \mid \begin{cases} D \subseteq D^{\text{distributions}} \wedge P \subseteq D^{\text{countries}} \\ \forall p \in P, \exists d \in D \mid \text{origin}(d) = p \end{cases} \}|}{|\{D \to P \mid D \subseteq D^{\text{distributions}} \wedge P \subseteq D^{\text{countries}}\}|}$$

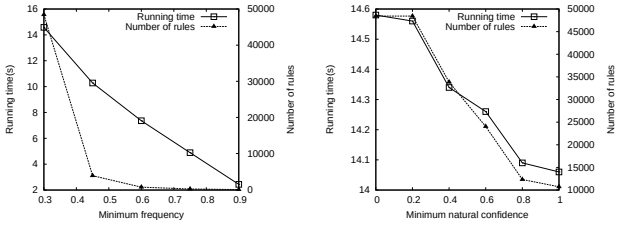
where $origin(d)$ is the origin country of the distribution d. Given our background knowledge, more relevant collections of rules should have higher q values.

Thus, to test whether higher minimal thresholds on the designed measures (i. e., the frequency and the confidences) actually capture more relevant patterns, Fig. 2 plots q in function of these thresholds. We observe that q actually increases w.r.t. every minimal threshold and this empirically corroborates the relevance of our semantics. The measure q increases more quickly with $\beta_{\text{exclusive}}$ than with β_{natural}. This makes sense: a conjunction of distributions that *exclusively* interests visitors from a given country usually involves at least one



(a) q with varying $\beta_{\text{exclusive}}$ ($\beta_{\text{natural}} = 0$). (b) q with varying β_{natural} ($\beta_{\text{exclusive}} = 0$).

Figure 2: Confidence qualitative assessment.



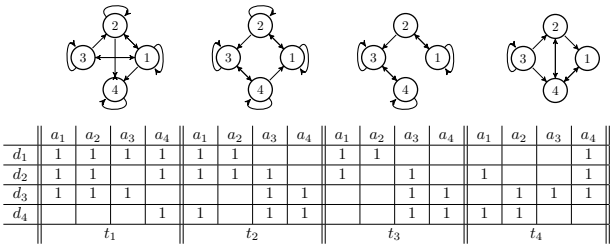
(a) Pruning w.r.t. μ. (b) Pruning w.r.t. β_{natural}.

Figure 3: Pruning effectiveness.

distribution developed by and for people in this country. Finally, it is interesting to understand that, under a given minimal frequency constraint μ, the collections of rules computed with $\beta_{\text{natural}} \leq \mu$ ($\beta_{\text{exclusive}}$ remaining constant) are the same, hence the steps in Fig. 2b. Indeed, the natural confidence is a proportion of elements in the support domain of the rule and the frequency constraint forces the rule to match at least a proportion μ of elements in this domain. As a consequence, no rule can have a natural confidence beneath μ.

We now report a performance study for the extraction of rules involving countries and distributions, i. e., $\mathcal{D}' = \{\text{Countries, Distributions}\}$. When the minimal frequency threshold increases, both the number of frequent rules and the running time decrease (see Fig. 3a obtained with $\beta_{\text{natural}} = \beta_{\text{exclusive}} = 0$). Indeed, PINARD prunes large areas of the search space where every association violates the constraint $\mathcal{C}_{\text{freq}}$. Theorem 2.2 allows to prune the search space too. Indeed, the RULES algorithm does not develop the enumeration sub-trees that only contain rules with too small natural confidences. That is why both the number of rules and the time it takes to extract them decrease when the minimum natural confidence threshold increases (see Fig. 3b). This experiment was performed with $\beta_{exclusive} = 0$, $\mu = 0.3$, and β_{natural} varying between 0 and 1.

PINARD's scalability was tested on the extraction of



	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
d_1	1	1	1	1	1	1			1	1						1
d_2	1	1		1	1	1	1		1		1		1			1
d_3	1	1	1				1	1			1	1		1	1	1
d_4				1	1		1	1			1	1	1	1		
	t_1				t_2				t_3				t_4			

Figure 4: The dynamic graph $\mathcal{R}_G \subseteq \{d_1, d_2, d_3, d_4\} \times \{a_1, a_2, a_3, a_4\} \times \{t_1, t_2, t_3, t_4\}$.

these rules with $\mu = 0.75$ and $\beta_{\text{natural}} = \beta_{\text{exclusive}} = 0$. $\mathcal{R}_{\texttt{DistroWatch}}$ was replicated up to 10 times w.r.t. the timestamps. It turns out that the algorithm scales linearly. More precisely, a linear regression of $R \mapsto \frac{T_R}{T_1}$ (where R is the replication factor; T_R the running time on this replicated dataset) gives $y = 2.27x - 2.91$ with 0.96 as a determination coefficient.

5 A Case Study on Dynamic Graphs.

To illustrate the genericity of our approach, we consider the analysis of a real-life dynamic graph as a case study.

5.1 Dynamic Graphs as n-Ary Relations. Let us investigate rule discovery from dynamic directed graphs, i. e., from collections of static directed graphs that all share the same set of uniquely identified vertices. For instance, Fig. 4 depicts a dynamic directed graph involving four nodes. Four snapshots of this graph are available. The dynamic graph can be represented as the sequence of its adjacency matrices underneath. It describes the relationship between the tail vertices in $D^1 = \{d_1, d_2, d_3, d_4\}$ and the head vertices in $D^2 = \{a_1, a_2, a_3, a_4\}$ at the timestamps in $D^3 = \{t_1, t_2, t_3, t_4\}$. Every '1', in the adjacency matrices, is at the intersection of three elements $(d_i, a_j, t_k) \in D^1 \times D^2 \times D^3$, which indicate a directed edge from d_i to a_j at time t_k. For instance, the edge from Node 3 to Node 2 at the first timestamp is encoded by the tuple (d_3, a_2, t_1) in $\mathcal{R}_G$. In this way, three dimensions are necessary to encode a dynamic graph, which can then be seen as a ternary relation (e. g., $\mathcal{R}_G$ in Fig. 4). However, more dimensions may be useful to encode, for instance, labels on the edges and/or different time granularities.

5.2 Mining the `Vélo'v` Dynamic Network. `Vélo'v`[4] is a bicycle rental service run by the urban community of Lyon, France. 327 `Vélo'v` stations are spread over Lyon and its surrounding area. At any of these stations, the users can take a bicycle and bring it to any other station. Whenever a bicycle is rented or returned, this event is logged. We were granted the access to such a log listing more than 13.1 million rides along 30 months. Those data can be seen as a dynamic directed graph evolving along the 7 days of the week and the 24 one-hour periods in a day, i. e., a collection of graphs timestamped with labels from both time scales. These two temporal dimensions and the (departure and arrival) stations make the four domains of a relation we call $\mathcal{R}_{\texttt{Vélov'v}}$. (ds, as, d, h) belongs to $\mathcal{R}_{\texttt{Vélov'v}}$ (i. e., there is an edge from ds to as in the graph timestamped with (d, h)) when a significant amount of bicycles (local test inspired by the computation of a p-value) are rented at the (departure) station ds on day d (e. g., Monday) at hour h (e. g., from 1pm to 2pm) and returned at the (arrival) station as. $\mathcal{R}_{\texttt{Vélov'v}}$ contains 117, 411 4-tuples, hence a $\frac{117,411}{327\times327\times7\times24} = 0.7\%$ density.

The temporal dimension(s) of such a dynamic network can either appear in the rules (i. e., in $\mathcal{D}'$) or be used to compute the frequency and the confidences of the rules (i. e., in the support domain). A trivial modification of Rules can additionally force some of the dimensions to *only* appear at the bodies (resp. at the heads) of the rules. These different rule templates support the analysis of different questions. Here are some examples.

Given a frequent sub-network (i. e., a sub-network that is often observed), is it enlargeable with a strong enough confidence? To answer this question, the rules must involve departure and arrival stations, i. e., $\mathcal{D}' = \{Departure, Arrival\}$. The support domain of these rules is the Cartesian product of the 7 days and the 24 hours. The constraint $\mathcal{C}_{(2,2)\text{-sizes}}$ (see Sect. 3) is additionally enforced so that every rule involves at least two departure and two arrival stations. Moreover we constrain the body of every rule to be a graph with at least an edge, i. e., it must involve at least one departure station and one arrival station. Redundant rules are removed. In this way, Pinard discovers the *minimal* sub-networks (at bodies of the rules) that can be confidently enlarged (with the stations at the heads). With $\mu = 0.2$ and $\beta_{\text{natural}} = \beta_{\text{exclusive}} = 0.7$, 84 rules are discovered. Some are reported in Fig. 5. The enlarged sub-networks can contain more nodes (see Rules 5b and 5c) or only more edges (see Rule 5a). These rules suggest diverse phenomena like "auto-regulation" (Rules 5b and 5a) or convergence (Rules 5a and 5c). They can potentially be used to anticipate the effect of a breakdown. For example, if station 1021 fails then station 1002 may soon be saturated since bicycles from stations 2001 and 2024 will converge to the operational station. Notice, however, that the extracted rules are only descriptive (and not predictive). Using them to

[4] `http://www.velov.grandlyon.com/`

support link prediction is an interesting perspective.

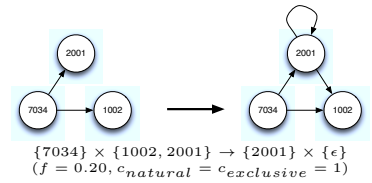
Are stations that emit, at given periods of time, bicycles toward many other stations, typical of some days of the week? To answer this question, the rules must involve time periods and departure stations at their bodies; day information at their heads. With the minimum thresholds $\mu = 0.08$, $\beta_{\text{natural}} = 0.7$, and $\beta_{\text{exclusive}} = 0.6$, 33 such rules are discovered. Fig. 6 reports two of them. The rule in Fig. 6a indicates that most of the departures from station 6002 and between 11am and 12am occur on Sundays ($c_{\text{exclusive}} = 0.71$). This makes sense: this station is at the main entrance of the most popular park, where people like to walk on Sundays and come back home by bicycle, hence the high frequency in terms of number of arrival stations. The rule in Fig. 6b means that we have rare departures from station 1002 between 1am and 3am except on Sundays ($c_{\text{exclusive}} = 0.62$). This makes sense too: this station is located in a district with many pubs, where people like to party at nights between Saturdays and Sundays. Since the public transportation services stop at midnight, `Vélo'v` is a popular way to come back home.

Do some stations exchange many bicycles at favored hours everyday? To answer this question, the mined rules have time periods and departure stations at their bodies; arrival stations at their heads. To discover rules that hold everyday, the minimal frequency threshold is set to 1. With $\beta_{\text{natural}} = 1$ and $\beta_{\text{exclusive}} = 0.8$, PINARD returns 40 rules involving at least one time period, two departure stations and two arrival stations. Fig. 7 depicts one of them. Such rules are valuable for the data owner, who discovers what arrival stations may be impacted by a shortage of bicycles at the stations in the body.

Let us finally provide a performance study of PINARD mining $\mathcal{R}_{\texttt{Vélo'v}}$ for rules such as those depicted in Fig. 5. As expected, Fig. 8 shows that the number of rules and the running time decrease when the minimal frequency threshold (resp. minimal natural confidence threshold) increases. As in Sect. 4, $\mathcal{R}_{\texttt{Vélov'v}}$ was replicated up to ten times w.r.t. its temporal dimensions. With $\mu = 0.1$ and $\beta_{\text{natural}} = \beta_{\text{exclusive}} = 0$, a linear regression of $R \mapsto \frac{T_R}{T_1}$ (where R is the replication factor and T_R the running time on the replicated dataset) gives $y = 0.51x + 0.5$ with 0.97 as a determination coefficient. This low slope highlights the effectiveness of PINARD.

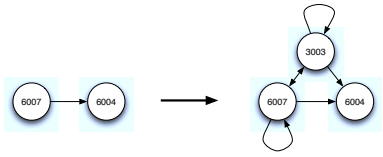
6 Related Work.

Since the seminal paper [1], the discovery, in binary relations, of association rules with high enough supports and confidences has been extensively studied. Many works deal with the generalization of this task towards



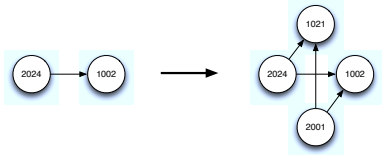
$\{7034\} \times \{1002, 2001\} \rightarrow \{2001\} \times \{\epsilon\}$
$(f = 0.20, c_{natural} = c_{exclusive} = 1)$

(a)



$\{6007\} \times \{6004\} \rightarrow \{3003\} \times \{6007, 3003\}$
$(f = 0.21, c_{\text{natural}} = c_{\text{exclusive}} = 0.73)$

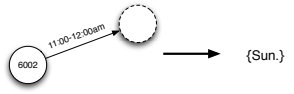
(b)



$\{2024\} \times \{1002\} \rightarrow \{2001\} \times \{1021\}$
$(f = 0.21, c_{\text{natural}} = c_{\text{exclusive}} = 0.77)$

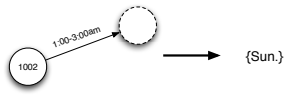
(c)

Figure 5: Example of rules of the form "min. sub-network" → "larger sub-network"



$\{6002\} \times \{11\text{-}12\text{am}\} \rightarrow \{\text{Sun.}\}$
$(f = 0.09, c_{\text{natural}} = 0.91, c_{\text{exclusive}} = 0.71)$

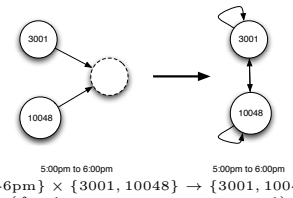
(a)



$\{1002\} \times \{1\text{-}2\text{am},2\text{-}3\text{am}\} \rightarrow \{\text{Sun.}\}$
$(f = 0.09, c_{\text{natural}} = 0.97, c_{\text{exclusive}} = 0.62)$

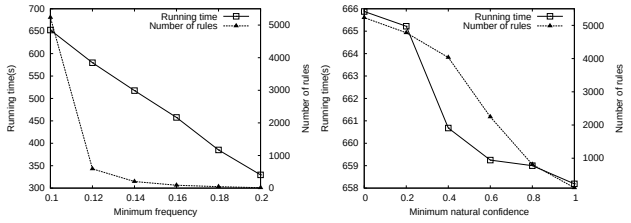
(b)

Figure 6: Example of rules of the form `departures` × `hours` → `days`.



$\{5\text{-}6\text{pm}\} \times \{3001, 10048\} \rightarrow \{3001, 10048\}$
$(f = 1, c_{\text{natural}} = c_{\text{exclusive}} = 1)$

Figure 7: A rule of the form `hours` × `departures` → `arrivals`.



(a) Pruning w.r.t. μ (β_{natural} = $\beta_{\text{exclusive}} = 0$).

(b) Pruning w.r.t. β_{natural} ($\mu = 0.1$ and $\beta_{\text{exclusive}} = 0$).

Figure 8: Pruning effectiveness.

n-ary relations. The rules discovered by these proposals can be classified into three types: intra-dimensional, inter-dimensional and hybrid. In an intra-dimensional rule, all the elements belong to a single dimension. This case has been extremely well studied for binary relations. In [22], the authors propose to discover intra-dimensional association rules in n-ary relations where $n \geq 2$. For each dimension, association rules between its elements are discovered. The Cartesian product of the $n-1$ other dimensions constitutes the support domain. Inter-dimensional association rules were proposed for the discovery of co-occurrences between elements in different dimensions [17, 10, 20]. Their expressiveness is however limited: two elements in the same dimension cannot appear together in a rule. The search for inter-dimensional association rules is guided by a metarule, which contains distinct predicates and enforces a user-defined rule template. The problem of defining the support/frequency out of the transactional framework has also been addressed within a relational database setting, i. e., a multi-relational perspective. [8] proposes the $Warmr$ algorithm that discovers rules over a limited type of Datalog queries. The support of a query is the number of databases for which it gives a non empty answer. In the same way, [11] has recently introduced a support measure based on the key dependencies. Other authors have proposed ad-hoc algorithms to extract hybrid rules in which the repetition of a few dimensions is possible [12, 9, 24]. Given the ability of dynamic graphs to represent real-world phenomena, several researchers have focused on the discovery of association rules in such particular ternary relations (see Sect. 5). With the increasing availability of network data (e. g., social network), it has even become a hot topic in the data mining community. Several works aim at mining local patterns in dynamic graphs [5, 13, 18, 21]. In particular, [18] introduces the periodic subgraph mining problem, i. e., identifying every frequent closed periodic subgraph. The interest and the efficiency of this proposal are empirically demonstrated on several real-world dynamic social networks. A few works tackle the problem of discovering rules from these patterns. [25] and [3] propose to discover descriptive rules to qualify the dynamics of the networks. [25] studies how a graph is structurally transformed through time. The proposed method computes graph rewriting rules that describe the evolution between consecutive graphs. These rules are then abstracted into patterns representing the dynamics of a sequence of graphs. In [3], the authors introduce graph-evolution rules that describe the frequent local changes occurring in a dynamic graph. They discuss what a rule could be in a dynamic graph and how to define its support and the confidence. However, the form of the considered rules is severely restricted. The multi-dimensional association rules we propose in this paper do not suffer from such restrictions. They can involve as many dimensions as desired and each of these dimensions can provide one or more elements to the discovered rules. Furthermore, the repartition of elements between the body and the head of the rules is not constrained. This work is applicable to particular n-ary relations such as dynamic graphs or cross-graph datasets.

7 Conclusion.

Designing new methods to discover patterns in arbitrary n-ary relations (or Boolean tensors) is a timely challenge. Recently, such methods were proposed for the extraction of closed patterns [16, 14, 6] and multi-dimensional rules were defined in more or less restricted ways. This paper generalizes the popular association rule mining task. Contrary to the related work, our rules do not suffer from severe form constraints: any subsets of any dimensions can appear at their heads and/or their bodies. First, we have defined relevant objective interestingness measures and thus given a semantics to the rules. Then, we have designed and implemented a complete though scalable algorithm that computes them. We have used real-life datasets (a 3-ary relation and a 4-ary relation encoding a dynamic graph) in which truly relevant rules have been discovered. Generalizing important properties of "classical" association rules (e. g., non redundancy) to our framework is an interesting topic we may soon tackle.

Acknowledgements. We want to thank Ladislav Bodnar for sharing with us the `Distrowatch.com` logs.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[2] M.-L. Antonie and O. R. Zaïane. Mining positive and negative association rules: an approach for confined rules. In *ECML/PKDD*, pages 27–38. Springer, 2004.
[3] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis. Mining graph evolution rules. In *ECML PKDD*, pages 115–130. Springer, 2009.
[4] M. Boley, T. Gärtner, and H. Grosskreutz. Formal concept sampling for counting and threshold-free local pattern mining. In *SDM*, pages 177–188. SIAM, 2010.
[5] K.-M. Borgwardt, H.-P. Kriegel, and P. Wackersreuther. Pattern mining in frequent dynamic subgraphs. In *ICDM*, pages 818–822. IEEE Computer Society, 2006.
[6] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Closed patterns meet n-ary relations. *ACM Trans. on Knowledge Discovery from Data*, 3(1):1–36, 2009.
[7] L. Dehaspe and L. De Raedt. Mining association rules in multiple relations. In *ILP*, pages 125–132. Springer, 1997.
[8] L. Dehaspe and H. Toivonen. Discovery of frequent DATALOG patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36, 1999.
[9] G. Dong, J. Han, J.-M.-W. Lam, J. Pei, and K. Wang. Mining multi-dimensional constrained gradients in data cubes. In *VLDB*, pages 321–330. VLDB Endowment, 2001.
[10] L. Feng, J. X. Yu, H. Lu, and J. Han. A template model for multidimensional inter-transactional association rules. *The VLDB Journal*, 11(2):153–175, 2002.
[11] B. Goethals, W. Le Page, and M. Mampaey. Mining interesting sets and rules in relational databases. In *SAC*, pages 997–1001. ACM Press, 2010.
[12] T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: generalizing association rules. *Data Mining and Knowledge Discovery*, 6(3):219–257, 2002.
[13] A. Inokuchi and T. Washio. A fast method to mine frequent subsequences from graph sequence data. In *ICDM*, pages 303–312. IEEE Computer Society, 2008.
[14] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Trias–an algorithm for mining iceberg tri-lattices. In *ICDM*, pages 907–911. IEEE Computer Society, 2006.
[15] T.-Y. Jen, D. Laurent, and N. Spyratos. Mining all frequent projection-selection queries from a relational table. In *EDBT*, pages 368–379. ACM Press, 2008.
[16] L. Ji, K.-L. Tan, and A. K. H. Tung. Mining frequent closed cubes in 3D data sets. In *VLDB*, pages 811–822. VLDB Endowment, 2006.
[17] M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *KDD*, pages 207–210. AAAI Press, 1997.
[18] M. Lahiri and T.-Y. Berger-Wolf. Mining periodic behavior in dynamic social networks. In *ICDM*, pages 373–382. IEEE Computer Society, 2008.
[19] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *KDD*, pages 189–194. AAAI Press, 1996.
[20] R. Ben Messaoud, S. Loudcher Rabaséda, O. Boussaid, and R. Missaoui. Enhanced mining of association rules from data cubes. In *DOLAP*, pages 11–18. ACM Press, 2006.
[21] C. Robardet. Constraint-based pattern mining in dynamic graphs. In *ICDM*, pages 950–955. IEEE Computer Society, 2009.
[22] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Data Science and Classification*, pages 261–270. Springer, 2006.
[23] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering*, 42(65):189–222, 2002.
[24] H.-C. Tjioe and D. Taniar. Mining association rules in data warehouses. *International Journal of Data Warehousing and Mining*, 1(3):28–62, 2005.
[25] C.-H. You, L. B. Holder, and D. J. Cook. Learning patterns in the dynamics of biological networks. In *KDD*, pages 977–986. ACM Press, 2009.

Appendix

Proof. [Theorem 2.1] According to Def. 2.6 and 2.2:

- $X \sqsubseteq Y \Rightarrow \begin{cases} \mathcal{D}_X \subseteq \mathcal{D}_Y \\ \forall D^i \in \mathcal{D},\ \pi_{D^i}(X) \subseteq \pi_{D^i}(Y) \end{cases}$;
- $s(Y) = \{t \in \times_{D^i \in \mathcal{D}\setminus\mathcal{D}_Y} D^i \mid \forall y \in Y,\ y \cdot t \in \mathcal{R}\}$;
- $s(X) = \{w \in \times_{D^i \in \mathcal{D}\setminus\mathcal{D}_X} D^i \mid \forall x \in X,\ x \cdot w \in \mathcal{R}\}$
  $= \{u \cdot t \,|\, u \in \times_{D^i \in \mathcal{D}_Y\setminus\mathcal{D}_X} D^i, t \in \times_{D^i \in \mathcal{D}\setminus\mathcal{D}_Y} D^i$
  and $\forall x \in X,\ x \cdot u \cdot t \in \mathcal{R}\}$.

Let $\pi_{\mathcal{D}\setminus\mathcal{D}_Y} s(X) = \{t \in \times_{D^i \in \mathcal{D}\setminus\mathcal{D}_Y} D^i \mid \exists u \in \times_{D^i \in \mathcal{D}_Y\setminus\mathcal{D}_X} D^i$ such that $\forall x \in X,\ x \cdot u \cdot t \in \mathcal{R}\}$.
Then, $\begin{cases} s(Y) \subseteq \pi_{\mathcal{D}\setminus\mathcal{D}_Y} s(X) \\ |\pi_{\mathcal{D}\setminus\mathcal{D}_Y} s(X)| \leq |s(X)| \end{cases}$,
and $|s(Y)| \leq |\pi_{\mathcal{D}\setminus\mathcal{D}_Y} s(X)| \leq |s(X)|$.

Proof. [Theorem 2.2] Using Def. 2.10, we have $X \sqsubseteq X' \Rightarrow s_{\mathcal{D}\setminus\mathcal{D}'}(X') \subseteq s_{\mathcal{D}\setminus\mathcal{D}'}(X)$.

Because $X \sqsubseteq X' \sqsubseteq Y$, according to Definition 2.11:

$$\begin{cases} c_{\text{natural}}(X \to Y \setminus X) = \frac{|s(Y)|}{|s_{\mathcal{D}\setminus\mathcal{D}'}(X)|} \\ c_{\text{natural}}(X' \to Y \setminus X') = \frac{|s(Y)|}{|s_{\mathcal{D}\setminus\mathcal{D}'}(X')|} \end{cases}$$

$\Rightarrow c_{\text{natural}}(X \to Y \setminus X) \leq c_{\text{natural}}(X' \to Y \setminus X')$.