Intelligent Data Analysis 17 (2013) 49–69
DOI 10.3233/IDA-120567
IOS Press

Discovering descriptive rules in relational dynamic graphs

Kim-Ngan T. Nguyen[a], Loïc Cerf[b], Marc Plantevit[c] and Jean-François Boulicaut[a,*]
[a]*INSA-Lyon, LIRIS, UMR5205, Villeurbanne Cedex, France*
[b]*Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
[c]*Université Lyon 1, LIRIS, UMR5205, Lyon, France*

Abstract. Graph mining methods have become quite popular and a timely challenge is to discover dynamic properties in evolving graphs or networks. We consider the so-called relational dynamic oriented graphs that can be encoded as n-ary relations with $n \geqslant 3$ and thus represented by Boolean tensors. Two dimensions are used to encode the graph adjacency matrices and at least one other denotes time. We design the pattern domain of multi-dimensional association rules, i.e., non trivial extensions of the popular association rules that may involve subsets of any dimensions in their antecedents and their consequents. First, we design new objective interestingness measures for such rules and it leads to different approaches for measuring the rule confidence. Second, we must compute collections of *a priori* interesting rules. It is considered here as a post-processing of the closed patterns that can be extracted efficiently from Boolean tensors. We propose optimizations to support both rule extraction scalability and non redundancy. We illustrate the added-value of this new data mining task to discover patterns from a real-life relational dynamic graph.

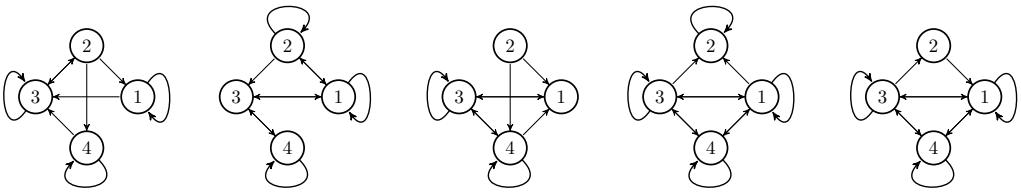
Keywords: Evolving graph, boolean tensor, rule discovery, closed pattern, non redundancy

1. Introduction

Graph mining is a popular topic (see, e.g. [2]). Many researchers have considered knowledge discovery from large collections of graphs while others focus the analysis of one large graph or network. In the latter case, we observe complementary directions of research. On one hand, global properties of graphs are studied like power-law distribution of node degrees or diameters [5,15,24,32,35]. On another hand, it is possible to use data mining algorithms to identify local patterns in the graphs (e.g., frequent subgraphs, clique patterns) [9,10,19,25,30]. Such techniques can indeed benefit from the huge research effort on 0/1 data analysis and the analogy between Boolean matrices with either the definition of bi-partite graphs or graph encoding by means of adjacency matrices.

We investigate pattern discovery from dynamic directed relational graphs, i.e., from a collection of static directed graphs that all share the same set of uniquely identified vertices. In our setting, given a set of vertices, directed edges can change (i.e., appear or disappear) through time and such a dynamic graph can be modeled by a sequence of static graphs. For instance, Fig. 1 depicts a dynamic directed graph involving four nodes through five timestamps. It can be represented as the sequence of its adjacency

*Corresponding author: Jean-François Boulicaut, INSA-Lyon, LIRIS, UMR5205, F-69621, France. Tel.: +33 4 72438905; Fax: +33 4724378713; E-mail: jean-francois.boulicaut@insa-lyon.fr.



	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	a_1	a_2	a_3	a_4
d_1	1		1		1	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	1	1	1	1	1
d_4			1	1			1	1	1		1	1	1		1	1	1		1	1
	t_1				t_2				t_3				t_4				t_5			

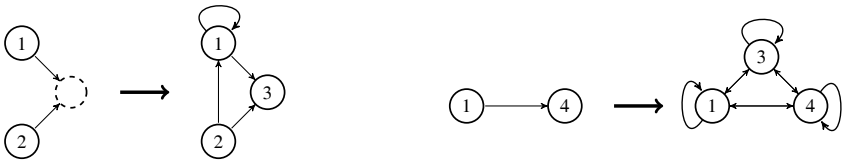
Fig. 1. $\mathcal{R}_E \subseteq \{d_1, d_2, d_3, d_4\} \times \{a_1, a_2, a_3, a_4\} \times \{t_1, t_2, t_3, t_4, t_5\}$.

matrices underneath. It describes the relationship between the departure vertices in $D^1 = \{d_1, d_2, d_3, d_4\}$ and the arrival vertices in $D^2 = \{a_1, a_2, a_3, a_4\}$ at the timestamps in $D^3 = \{t_1, t_2, t_3, t_4, t_5\}$. 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 indicates a directed edge from d_i to a_j at time t_k. Therefore, at least three dimensions are necessary to encode a dynamic relational graph as a Boolean tensor or a ternary relation (the one depicted in Fig. 1 is called $\mathcal{R}_E$). However, more dimensions may be used to encode additional information on edges and/or time aspects.

Studying descriptive rule mining from dynamic graphs is a rather new research topic and most of previous works impose severe restrictions on the form of the rules. The key contribution of this paper is the proposal of a new form of rules which generalizes the inter-dimensional rules from [28]. Our rules may now involve any subset of dimensions in both the left-hand side and the right-hand side.[1] In particular, the temporal dimensions can either explicitly appear in the rules or be used to measure their relevancy (i.e., these dimensions "support" the rules). It provides patterns that describe the graph evolution at a local level. Two examples of inter-dimensional rules are given in Figs 2a and 3a. Figure 2a depicts a rule that is preserved at several timestamps. It intuitively means that if, at a time, the edges from the vertices 1 and 2 go to the same vertices then these vertices are 1 and 3. The rule in Fig. 3a indicates that when we observe that the edges departing from the vertices 1, 3 and 4 have the same arrivals then that usually happens at the times t_4 and t_5. By removing the constraint that a given dimension cannot appear at both sides of the rule, more rules become valid. For example, the rule in Fig. 2b describes the dependency between sub-networks. More precisely, it tells that the sub-network at its body can be confidently enlarged to a clique. This clique turns to be the maximal one involving the sub-network. The rule in Fig. 3b indicates that, at Time t_3, most of the edges from Vertex 3 reach vertices that also are the arrivals of edges leaving this Vertex 3 at the times t_4 and t_5. So, a new type of pattern which is more general is defined: *the multidimensional association rule*.

To assess the relevancy of such rules, we design a straightforward extension of the classical frequency measure and two original and non trivial generalizations of the confidence measure: (a) the *exclusive confidence*, which quantifies the "exclusiveness" of the rules, turns out to be easily interpretable but a threshold on this measure does not allow to prune the rule search space; (b) *the natural confidence*, which relies on a "natural" proportion of elements in the support domain, turns out to be harder to interpret (additional data matching the body but not the head not always decrease the natural confidence) but

[1] In inter-dimensional rules, body and head dimension domains have to be *disjoint.*



(a) $\{d_1, d_2\} \to \{a_1, a_3\}$. (b) $\{d_1\} \times \{a_4\} \to \{d_3, d_4\} \times \{a_1, a_3\}$.

Fig. 2. Example of rules on $\{D^1, D^2\}$ in $\mathcal{R}_E$.

(a) $\{d_1, d_3, d_4\} \to \{t_4, t_5\}$. (b) $\{t_3\} \times \{d_3\} \to \{t_4, t_5\}$.

Fig. 3. Example of rules on $\{D^1, D^3\}$ in $\mathcal{R}_E$.

enforcing a threshold on it provides faster extractions, hence the possibility to mine much larger graphs. Beyond the design of a new semantics for rules in dynamic graphs, we indeed propose an algorithm called PINARD++ that efficiently computes them. It proceeds in three successive steps: (1) it prepares the Boolean tensor to mine; (2) it computes the *frequent* closed sets in that tensor thanks to a clever use of the state-of-the-art algorithm DATA-PEELER [13]; (3) it derives from these patterns the non-redundant rules whose exclusive and natural confidences exceed the user-defined thresholds. This algorithm is an extension of both GEAR [28] and PINARD [29]. Among others, we revisit non-redundancy in this new multidimensional setting. To illustrate the added-value of this new pattern type and the tractability of our extraction method, an experimental study, on a large real-world dynamic network, is reported.

2. A descriptive rule pattern domain

2.1. Preliminary definitions

The proposed semantics for multidimensional association rules (as well as the algorithm listing them all in a given dataset) actually applies to any n-ary relation and thus Boolean tensor. All along the article, the domains (a domain is the set of elements of each dimension) are denoted $D^1, D^2, \ldots, D^n$ and the relation, denoted $\mathcal{R}$, is a subset of $D^1 \times \cdots \times D^n$. Without loss of generality, the domains are supposed disjoint. The set of all domains ($\{D^1, D^2, \ldots, D^n\}$) is denoted $\mathcal{D}$. To emphasize the relevancy of the proposed patterns in dynamic (directed) graphs, the definitions are illustrated on the toy example represented in Fig. 1. It depicts a graph with four vertices and evolving along five timestamps. This graph can be seen as a ternary relation $\mathcal{R}_E$, which relates the departure vertices in D^1 to the arrival vertices in D^2 at the timestamps in D^3.

The patterns of interest only involve dimensions of some of the domains $\mathcal{D}' \subseteq \mathcal{D}$. E.g., in $\mathcal{R}_E$, the analyst may want to focus on subgraph patterns. In this case, $\mathcal{D}'$ involves the two domains D^1 and D^2, i.e., $\mathcal{D}' = \{D^1, D^2\}$. Without loss of generality, the dimensions are assumed ordered such that $\mathcal{D}' = \{D^1, \ldots, D^{|\mathcal{D}'|}\}$.

Definition 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 $\mathcal{D}' = \emptyset$ is denoted $\emptyset$.*

For example, in $\mathcal{R}_E$, $\{d_1, d_2\} \times \{a_1\}$ and $\{d_1, d_2\} \times \{a_1, a_3\}$ are associations on $\{D^1, D^2\}$, $\{d_1, d_2\}$ is an association on $\{D^1\}$ and $\{a_1, a_2\}$ is an association on $\{D^2\}$.

Given an arbitrary association on $\mathcal{D}'$, $\times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$ is its *support domain*. Like with binary relations or binary matrices, the support domain is a set of elements that are counted to provide a frequency interestingness measure. For instance, in the popular *Transactions* $\times$ *Products* setting, the support domain of an association rule involving products is the set of transactions [4]. In our running example, we see that D^3 is the support domain of associations 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, $(d_1, a_3) \cdot (t_1) = (d_1, a_3, t_1)$.

Definition 2 (Support). *$\forall \mathcal{D}' \subseteq \mathcal{D}$, let X be an association on $\mathcal{D}'$. Its support is $s(X) = \{u \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \forall x \in X, x \cdot u \in \mathcal{R}\}$.*

Let us mention some special cases. An association involving the n domains ($\mathcal{D}' = \mathcal{D}$) is either false (at least one n-tuple it contains is absent from $\mathcal{R}$) or true (every n-tuple it contains is in $\mathcal{R}$). By using the convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$ (where ϵ is the empty word), Definition 2 reflects that every possible 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 $\mathcal{D}' = \{D^1\}$). For example, in $\mathcal{R}_E$: $s(\{d_1, d_2\} \times \{a_1\}) = \{t_1, t_2, t_5\}$, $s(\{d_1, d_2\} \times \{a_1, a_3\}) = \{t_1, t_2\}$ and $s(\{d_1, d_2\}) = \{(a_1, t_1), (a_3, t_1), (a_1, t_2), (a_2, t_2), (a_3, t_2), (a_2, t_4), (a_1, t_5)\}$.

Let us now introduce some operators to manipulate associations. Their definitions are illustrated on $X_e = \{a_1, a_2\}$ and $Y_e = \{d_1, d_2\} \times \{a_1, a_3\}$.

Definition 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 $\mathcal{D}'$. $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X) = X^i$ if $D^i \in \mathcal{D}'$, $\emptyset$ otherwise.*

For example, $\pi_{D^1}(X_e) = \emptyset$, $\pi_{D^2}(X_e) = \{a_1, a_2\}$, $\pi_{D^3}(X_e) = \emptyset$, $\pi_{D^1}(Y_e) = \{d_1, d_2\}$, $\pi_{D^2}(Y_e) = \{a_1, a_3\}$, and $\pi_{D^3}(Y_e) = \emptyset$.

Definition 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 an association on $\mathcal{D}_X \cup \mathcal{D}_Y$ for which $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X \sqcup Y) = \pi_{D^i}(X) \cup \pi_{D^i}(Y)$.*

For example, $X_e \sqcup Y_e$ is an association on $\{D^2\} \cup \{D^1, D^2\} = \{D^1, D^2\}$, $X_e \sqcup Y_e = (\pi_{D^1}(X_e) \cup \pi_{D^1}(Y_e)) \times (\pi_{D^2}(X_e) \cup \pi_{D^2}(Y_e)) = (\emptyset \cup \{d_1, d_2\}) \times (\{a_1, a_2\} \cup \{a_1, a_3\}) = \{d_1, d_2\} \times \{a_1, a_2, a_3\}$.

Definition 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 an association on $\{D^i \in \mathcal{D}_Y \mid \pi_{D^i}(Y) \not\subseteq \pi_{D^i}(X)\}$ for which $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(Y \setminus X) = \pi_{D^i}(Y) \setminus \pi_{D^i}(X)$.*

For example, $Y_e \setminus X_e$ is an association on $\{D^1, D^2\}$, $Y_e \setminus X_e = (\pi_{D^1}(Y_e) \setminus \pi_{D^1}(X_e)) \times (\pi_{D^2}(Y_e) \setminus \pi_{D^2}(X_e)) = (\{d_1, d_2\} \setminus \emptyset) \times (\{a_1, a_3\} \setminus \{a_1, a_2\}) = \{d_1, d_2\} \times \{a_3\}$. In contrast, $X_e \setminus Y_e$ is an association on $\{D^2\}$ only and $X_e \setminus Y_e = \pi_{D^2}(X_e) \setminus \pi_{D^2}(Y_e) = \{a_1, a_2\} \setminus \{a_1, a_3\} = \{a_2\}$.

Definition 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 $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X) \subseteq \pi_{D^i}(Y)$.*

For example, there are inclusions between three of the four associations illustrating Definition 1: $\{d_1, d_2\} \sqsubseteq \{d_1, d_2\} \times \{a_1\} \sqsubseteq \{d_1, d_2\} \times \{a_1, a_3\}$.

With this generalized inclusion, the *anti-monotonicity* of the support cardinality, that is well known in itemset mining, still holds. The proof is given in the Technical Annex at the end of the paper.

Theorem 1 (Support 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 $\mathcal{D}_Y$), $X \sqsubseteq Y \Rightarrow |s(X)| \geqslant |s(Y)|$.*

For example, considering the double inclusion illustrating Definition 6, one can verify that $|s(\{d_1, d_2\})| \geqslant |s(\{d_1, d_2\} \times \{a_1\})| \geqslant |s(\{d_1, d_2\} \times \{a_1, a_3\})|$, i.e., Theorem 1 holds.

2.2. Non redundant multidimensional association rules

2.2.1. Multidimensional association rules

Given the n-ary relation $\mathcal{R}$ 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 7 (Multidimensional association 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 $\mathcal{R}_E$, $\{d_1, d_2\} \to \{a_1, a_3\}$ and $\{d_3\} \times \{a_2\} \to \{a_1, a_3, a_4\}$ are two rules on $\{D^1, D^2\}$. $\{d_1\} \to \{d_2\}$ 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 '$\to$') nor 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 [4]. 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 fairly hard to define a confidence measure for general rules.

2.2.2. Rule frequency

The (relative) frequency of a rule is, in the support domain, the proportion of elements supporting the union of its body and its head.

Definition 8 (Rule frequency). *$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ a rule on $\mathcal{D}'$. Its frequency is:*

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

Let us give two examples of rules whose frequencies are $\frac{2}{5}$ in $\mathcal{R}_E$: r_e denotes $\{d_1, d_2\} \to \{a_1, a_3\}$ and r_s denotes $\{d_3\} \times \{a_2\} \to \{a_1, a_3, a_4\}$.

- $f(r_e) = \frac{|s(\{d_1,d_2\} \sqcup \{a_1,a_3\})|}{|D^3|} = \frac{|s(\{d_1,d_2\} \times \{a_1,a_3\})|}{|D^3|} = \frac{|\{t_1,t_2\}|}{|\{t_1,t_2,t_3,t_4,t_5\}|}$;
- $f(r_s) = \frac{|s(\{d_3\} \times \{a_2\} \sqcup \{a_1,a_3,a_4\})|}{|D^3|} = \frac{|s(\{d_3\} \times \{a_1,a_2,a_3,a_4\})|}{|D^3|} = \frac{|\{t_4,t_5\}|}{|\{t_1,t_2,t_3,t_4,t_5\}|}$.

2.2.3. Rule confidence

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 domains (so 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 $\{d_3\} \times \{a_2\} \to \{a_1, a_3, a_4\}$ would be: $\frac{|s(\{d_3\} \times \{a_2\} \sqcup \{a_1,a_3,a_4\})|}{|s(\{d_3\} \times \{a_2\})|} = \frac{|s(\{d_3\} \times \{a_1,a_2,a_3,a_4\})|}{|s(\{d_3\} \times \{a_2\})|} =$

$\frac{|\{t_4,t_5\}|}{|\{t_1,t_4,t_5\}|} = \frac{2}{3}$. It is a proportion of timestamps and it means that every time the graph involves an edge from d_3 to a_2 then it also tends to involve the edges from d_3 to a_1, a_3 and a_4.

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 disjoint sets and the ratio of their cardinalities does not make any sense. For instance, in $\mathcal{R}_E$, consider the rule $\{d_1, d_2\} \rightarrow \{a_1, a_3\}$. $s(\{d_1, d_2\} \times \{a_1, a_3\}) = \{t_1, t_2\}$, is a set of timestamps, whereas $s(\{d_1, d_2\})$ is not. It contains couples such as (a_1, t_1) or (a_3, t_2). As a result, there is a need for a new confidence measure that would make sense for any multidimensional association rule $X \rightarrow Y$. We expect however that this measure will be equal to $\frac{|s(X \sqcup Y)|}{|s(X)|}$ when X and $X \sqcup Y$ are defined on the same domain(s).

Exclusive confidence: Computing the confidence of a rule $X \rightarrow Y$ 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 9 (Exclusive confidence)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \rightarrow Y$ be a rule on $\mathcal{D}'$ and $\mathcal{D}_X$ the domains on which X is defined. The exclusive confidence of this rule is:*

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

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 above, if X is an association on $\mathcal{D}'$, the exclusive confidence of $X \rightarrow Y$ is $\frac{|s(X \sqcup Y)|}{|s(X)|}$ under the convention $\times_{D^i \in \emptyset} \pi_{D^i}(Y) = \{\epsilon\}$.

Consider the rule $\{d_1, d_2\} \rightarrow \{a_1, a_3\}$ in $\mathcal{R}_E$. To visualize the computation of its exclusive confidence, Fig. 4 depicts the dynamic graph in Fig. 1 though it only keeps a selection of its pairs of edges ("pairs" because two vertices are at the body of the rule) (a) having 1 *and* 2 as departure vertices and (b) going to a *same* vertex at the *same* time. There are seven, $|s(\{d1, d2\})| = 7$, such pairs (half the number of edges in Fig. 1). It is the denominator of the exclusive confidence. Among these pairs of edges, four go to a_1 *and* a_3 (the head of the rule) at the *same* time (they are thick in Fig. 4) and this is our numerator:

$$c_{exclusive}(\{d_1, d_2\} \rightarrow \{a_1, a_3\}) = \frac{|s(\{d_1, d_2\} \sqcup \{a_1, a_3\})| \times |\{a_1, a_3\}|}{|s(\{d_1, d_2\})|} = \frac{4}{7} \ .$$

At t_5, the pair of edges, with 1 and 2 as departure vertices, goes to 1 only (no analog pair goes to 3). That is why it is only counted at the denominator of the exclusive confidence measure. This pair somehow lowers the confidence in the fact that the edges departing from the vertices 1 and 2 converge to *both* the vertices 1 and 3. More interestingly, at t_2, despite the presence of the two pairs satisfying the rule, the fact that there is an additional one going to 2 also lowers the confidence in the fact that the edges departing from the vertices 1 and 2 *exclusively* converge to the vertices 1 and 3. This *exclusivity* explains the chosen name. In fact, for $c_{exclusive}(\{d_1, d_2\} \rightarrow \{a_1, a_3\})$ to be 1, i.e., the maximal possible value, every time the edges departing from the vertices 1 and 2 concur, they must do so towards *both* the vertices 1 and 3 and *only them*.

Unfortunately, this exclusivity also makes the function $X \mapsto c_{exclusive}(X \rightarrow Y \setminus X)$ (with $X \sqsubseteq Y$) not increase w.r.t. $\sqsubseteq$. For example, consider the rules $\{d_4\} \rightarrow \{d_3\} \times \{a_1, a_3, a_4\}$ and $\{d_4\} \times \{a_3\} \rightarrow \{d_3\} \times \{a_1, a_4\}$ in $\mathcal{R}_E$, $c_{exclusive}(\{d_4\} \rightarrow \{d_3\} \times \{a_1, a_3, a_4\}) = \frac{9}{13}$ and $c_{exclusive}(\{d_4\} \times \{a_3\} \rightarrow$

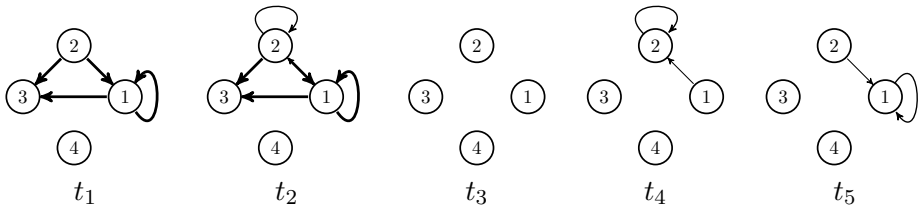


Fig. 4. Computing the exclusive confidence of $\{d_1, d_2\} \to \{a_1, a_3\}$.

$\{d_3\} \times \{a_1, a_4\}) = \frac{3}{5}$. We observe that $\{d_4\} \sqsubseteq \{d_4\} \times \{a_3\} \sqsubseteq \{d_3, d_4\} \times \{a_1, a_3, a_4\}$, however $c_{exclusive}(\{d_4\} \to \{d_3\} \times \{a_1, a_3, a_4\})$ is greater than $c_{exclusive}(\{d_4\} \times \{a_3\} \to \{d_3\} \times \{a_1, a_4\})$. This prevents to efficiently list every rule having an exclusive confidence greater than a user-defined threshold.

The exclusive confidence measure actually penalizes a rule whose elements in its support domain individually allow to conclude on other elements than those at its head. In this way, a minimal exclusive confidence threshold favors the discovery of multidimensional association rules with "maximal" heads. Let us now consider an alternative definition for the confidence measure.

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 $X \sqcup Y$. "Enforcing" 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 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:*

$$s_{\mathcal{D} \setminus \mathcal{D}'}(X) = \{u \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i \mid \exists w \in \times_{D^i \in \mathcal{D}' \setminus \mathcal{D}_X} D^i \text{ such that } \forall x \in X,\ x \cdot w \cdot u \in \mathcal{R}\} \ ,$$

where $\mathcal{D}_X$ is the set of domains on which X is defined. For $x \cdot w \cdot u$ 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 11 (Natural confidence)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ be a rule on $\mathcal{D}'$. Its natural confidence is:*

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

Notice that 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 $\{d_1, d_2\} \to \{a_1, a_3\}$ in $\mathcal{R}_E$. To compute its natural confidence, the initial selection of the relevant pairs of edges is similar to the one presented for the computation of the exclusive confidence (and illustrated by Fig. 4). However, this number of pairs is not the denominator of the natural confidence. Instead, the denominator is the number of timestamps (the support domain of $\{d_1, d_2\} \sqcup \{a_1, a_3\}$) where at least one pair of edges was selected, i.e., four (the only snapshot of the graph where no pair was selected is at time t_3). Among these four timestamps, two support the rule, i.e., have pairs of edges going to both vertices 1 and 3. This number $|s(\{d_1, d_2\} \sqcup \{a_1, a_3\})|$ is the numerator of the natural confidence measure:

$$c_{natural}(\{d_1, d_2\} \to \{a_1, a_3\}) = \frac{|s(\{d_1, d_2\} \sqcup \{a_1, a_3\})|}{|s_{\{D^3\}}(\{d_1, d_2\})|} = \frac{|\{t_1, t_2\}|}{|\{t_1, t_2, t_4, t_5\}|} = \frac{2}{4} \ .$$

At Time t_4 (or t_5), which does not support the rule, selecting a pair of edges decreases the natural confidence. Notice however that selecting more pairs in *one* such snapshot does not decrease the measure, i.e., the natural confidence only captures an information about the presence of *at least one* selected pair. At Time t_1 (or t_2), which does support the rule, selecting more pairs of edges (than those necessary to the satisfaction of the rule) does not decrease the natural confidence either. In particular, at Time t_2, the pair of edges going from the vertices 1 and 2 to the Vertex 2 has no influence on the natural confidence. This unintuitive behavior makes the natural confidence difficult to interpret: the body of the rule must be understood "in the light" of the additional dimensions at the head of the rule. On the positive side, and contrary to the exclusive confidence, the natural confidence simply is a proportion of elements in the support domain (in this case, the temporal dimension). Furthermore, this measure has a monotonicity property, which the exclusive confidence misses. It enables the efficient discovery of multidimensional association rules in large datasets.

Theorem 2 (Pruning criterion)**.** *Let $X \to Y \setminus X$ and $X' \to Y \setminus X'$ be two rules on $\mathcal{D}'$. We have: $X \sqsubseteq X' \sqsubseteq Y \Rightarrow c_{natural}(X \to Y \setminus X) \leqslant c_{natural}(X' \to Y \setminus X')$.*

The proof is given in the Technical Annex. In $\mathcal{R}_E$, $\{d_1, d_2\} \to \{a_1, a_3\}$ and $\{d_1, d_2\} \times \{a_1\} \to \{a_3\}$ 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 $\frac{|s(\{d_1,d_2\}\times\{a_1\}\sqcup\{a_3\})|}{|s_{D^3}(\{d_1,d_2\}\times\{a_1\})|} = \frac{|s(\{d_1,d_2\}\times\{a_1,a_3\})|}{|s_{D^3}(\{d_1,d_2\}\times\{a_1\})|} = \frac{|\{t_1,t_2\}|}{|\{t_1,t_2,t_5\}|} = \frac{2}{3}$. It illustrates Theorem 2. Indeed, $\{d_1, d_2\} \sqsubseteq \{d_1, d_2\} \times \{a_1\} \sqsubseteq \{d_1, d_2\} \times \{a_1, a_3\}$ and $c_{natural}(\{d_1, d_2\} \to \{a_1, a_3\}) \leqslant c_{natural}(\{d_1, d_2\} \times \{a_1\} \to \{a_3\})$. In Section 3, this theorem is used to prune the search space where no rule can satisfy a minimal natural confidence constraint.

2.2.4. Canonical and non-redundant rules

Definition 12 (Syntactic equivalence of rules)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D}$, 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 1. *Syntactically equivalent rules have the same frequency, the same exclusive confidence and the same natural confidence.*

Definition 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 on frequency and/or confidences can be condensed, without any loss of information, into its canonical rules only. Indeed, given a canonical rule $X \to Y$ in the collection, Lemma 1 entails that all syntactically equivalent rules necessary are in the collection as well. Moreover constructing them is easy: they are the rules $X \to Y \sqcup Z$ with $Z \sqsubseteq X$. For example, in $\mathcal{R}_E$, let us consider the following rules:

- r_1: $\{d_3\} \times \{a_2\} \to \{a_1, a_3, a_4\}$ ($f : 0.4$, $c_{natural} : 0.67$, $c_{exclusive} : 0.67$),
- r_2: $\{d_3\} \times \{a_2, a_3\} \to \{a_1\}$ ($f : 0.4$, $c_{natural} : 0.67$, $c_{exclusive} : 0.67$),
- r_3: $\{d_4\} \to \{d_3\} \times \{a_1, a_3, a_4\}$ ($f : 0.6$, $c_{natural} : 0.6$, $c_{exclusive} : 0.69$),
- r_4: $\{d_4\} \times \{a_3\} \to \{d_3\} \times \{a_1, a_4\}$ ($f : 0.6$, $c_{natural} : 0.6$, $c_{exclusive} : 0.6$),

They all are canonical and have their frequencies, their exclusive confidences and their natural confidences respectively exceeding 0.4, 0.6, and 0.6. In this regard, they *individually* satisfy this aspect of interestingness. Nevertheless, *altogether*, they provide redundant information. For instance, r_2 is more

specific than r_1 because it requires more conditions to apply (to match the body of the rule, a graph must additionally have the edge from the Vertex 3 to itself) and its conclusion is less informative (it does not tell anything about a_4). However this specialization does not grant r_2 a greater frequency or greater confidences than r_1. Therefore, r_2 is said redundant. Similarly the interestingness measures of r_3 make the more specific rule r_4 redundant. Since the analyst would not find any added-value in the rules r_2 and r_4, they should not be returned. In other terms, the concept of non-redundant rule [38] is to be revisited in our extended setting.

Definition 14 (Non-redundant rule)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D}$, a rule $X \to Y$ on $\mathcal{D}'$ is non-redundant iff it is canonical and no other canonical rule $X' \to Y'$ is such that:*

$$\begin{cases} (X' \sqcup Y' = X \sqcup Y \wedge X' \sqsubset X) \vee (X' \sqcup Y' \sqsupset X \sqcup Y \wedge X' \sqsubseteq X) \\ f(X' \to Y') \geqslant f(X \to Y) \\ c_{natural}(X' \to Y') \geqslant c_{natural}(X \to Y) \\ c_{exclusive}(X' \to Y') \geqslant c_{exclusive}(X \to Y) \end{cases}.$$

The first condition defines the form of the rules that may make the considered one redundant. Obviously, there exists other more general rules (with less elements) that are not matched. Nevertheless, this definition allows the removal of many redundant rules that are worse that the selected ones in term of frequency (second condition), natural confidence (third condition) and exclusive confidence (fourth condition). For instance, thanks to it, the rules r_2 and r_4 are not presented to the analyst. The choice of the first condition was partly based on procedural considerations: the non-redundant rules, as defined above, can be efficiently derived from closed sets. Before defining the closed sets, let us introduce the relation in which these patterns are extracted. It is obtained from $\mathcal{R}$ by "flattening" the dimensions absent from $\mathcal{D}'$ into a unique support dimension $D^{supp} = \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$. Denoted $\mathcal{R}_A$ until the end of this article, this relation is defined on the domains $\mathcal{D}_A = \mathcal{D}' \cup \{D^{supp}\}$. Assuming that for all $i = 1..n$, e_i is an element of the i^{th} domain, i.e., $e_i \in D^i$, we have to build:

$$\mathcal{R}_A = \{(e_1, e_2, \ldots, e_{|\mathcal{D}'|}, (e_{|\mathcal{D}'|+1}, \ldots, e_n)) | (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, e_{|\mathcal{D}'|+1}, \ldots, e_n) \in \mathcal{R}\}.$$

In this relation, a closed set is an association on $\mathcal{D}_A$ that (a) only covers $|\mathcal{D}_A|$-tuples present in $\mathcal{R}_A$ and (b) cannot be enlarged without violating (a).

Definition 15 (Closed set)**.** *Given a relation $\mathcal{R}_A$ on $\mathcal{D}_A$, X is a closed set in*

$$\mathcal{R}_A \text{ iff } \begin{cases} X \subseteq \mathcal{R}_A \\ \forall D^i \in \mathcal{D}_A, \forall e \in D^i \setminus \pi_{D^i}(X), X \sqcup \{e\} \not\subseteq \mathcal{R} \end{cases}.$$

Considering $\mathcal{R}_E$, if $\mathcal{D}'$ contains two domains, then $\mathcal{R}_A = \mathcal{R}_E$ and $\{d_1, d_2\} \times \{a_1, a_3\} \times \{t_1, t_2\}$ is a closed set. $\{d_1, d_2\} \times \{a_1, a_3\} \times \{t_1, t_2, t_5\}$ is not a closed set because it covers $(d_2, a_3, t_5) \notin \mathcal{R}_A$. $\{d_1, d_2\} \times \{a_3\} \times \{t_1, t_2\}$ is not a closed set either because it can be extended with a_1.

Finally, the following theorem, proved in the Technical Annex, states that the non-redundant rules on $\mathcal{D}'$ are exactly those derivable from the closed sets in $\mathcal{R}_A$ (their elements in $\cup_{D^i \in \mathcal{D}'} D^i$ being split between bodies and heads) and satisfying a second condition pertaining to the confidences of the more general rules sharing the same elements.

Theorem 3 (Closed sets and non-redundant rules)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D}$, let $X \to Y$ be a canonical rule on $\mathcal{D}'$. $X \to Y$ is a non-redundant rule iff $(X \sqcup Y \sqcup s(X \sqcup Y))$ is a closed set in $\mathcal{R}_A$ and $\forall X' \sqsubset X$, $c_{natural}(X' \to (Y \sqcup X) \setminus X') < c_{natural}(X \to Y) \vee c_{exclusive}(X' \to (Y \sqcup X) \setminus X') < c_{exclusive}(X \to Y)$.*

3. Discovering non-redundant rules

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

$$\left\{ X \to Y \text{ on } \mathcal{D}' \mid \begin{cases} X \to Y \text{ is non-redundant} \\ f(X \to Y) \geqslant \mu \\ c_{exclusive}(X \to Y) \geqslant \beta_{exclusive} \\ c_{natural}(X \to Y) \geqslant \beta_{natural} \end{cases} \right\}.$$

PINARD++ proceeds in three successive steps: (1) it constructs the relation $\mathcal{R}_A$ defined at the end of the previous section; (2) it extracts the *frequent* closed sets in $\mathcal{R}_A$; (3) it derives from these patterns the non-redundant rules whose exclusive and natural confidences exceed the user-defined thresholds. The first step is trivial. The second step relies on the state-of-the art algorithm DATA-PEELER for closed set mining under constraints in Boolean tensors. In Section 3.1, we present the constraint to integrate to focus on the *frequent* patterns (i.e., to only discover the closed sets from which frequent enough rules are obtained) as well as other constraints targeted to dynamic graph mining. The derivation of the non-redundant rules from the closed sets is presented in Section 3.2.

3.1. Extracting closed sets under constraints in tensors

Two approaches have been proposed to exhaustively list the closed sets in *ternary* relations, namely CUBEMINER [22] and TRIAS [21]. A third algorithm, DATA-PEELER [13] can compute every closed set in relations of arbitrary arity. Despite its broader scope, it is orders of magnitude faster than both TRIAS and CUBEMINER on ternary relations. Furthermore, DATA-PEELER can efficiently handle an expressive class of constraints. This is particularly appealing in our context. Indeed, Theorem 3 states the link between the non-redundant rules and the closed sets in $\mathcal{R}_A$ but, to be *a priori* interesting, the rules must satisfy constraints. DATA-PEELER can handle some of them directly on the closed sets. This is, in particular, the case of the frequency constraint: the closed sets in $\mathcal{R}_A$, that lead to frequent rules, gather at least a proportion μ of the elements in D^{supp}.

Definition 16 (Frequent closed set)**.** *Given a frequency threshold μ, a closed set C is a frequent closed set if $\frac{|\pi_{D^{supp}}(C)|}{|D^{supp}|} \geqslant \mu$.*

It may also be interesting to specify minimal numbers of elements in the dimensions that the rules involve (i.e., the dimensions in $\mathcal{D}'$).

Definition 17 ($(\alpha^i)_{i=1..|\mathcal{D}'|}$-large closed set)**.** *$\forall (\alpha^i)_{i=1..|\mathcal{D}'|} \in \mathbb{N}^{|\mathcal{D}'|}$, a closed set C is said $(\alpha^i)_{i=1..|\mathcal{D}'|}$-large if $\forall D^i \in \mathcal{D}', |\pi_{D^i}(C)| \geqslant \alpha^i$.*

Other constraints specifically make sense when $\mathcal{R}$ not only is a generic n-ary relation but a collection of graphs or even a dynamic graph (i.e., a timestamped collection of graphs). For instance, a *symmetry* constraint between the set of departure vertices and that of arrival vertices enables the discovery of association rules concluding on *cross-graph closed cliques*.

Definition 18 (Cross-graph closed clique)**.** *A closed set C is a cross-graph closed clique if $\pi_{D^{dep}}(C) = \pi_{D^{arr}}(C)$, where D^{dep} (resp. D^{arr}) is the set of departure (resp. arrival) vertices.*

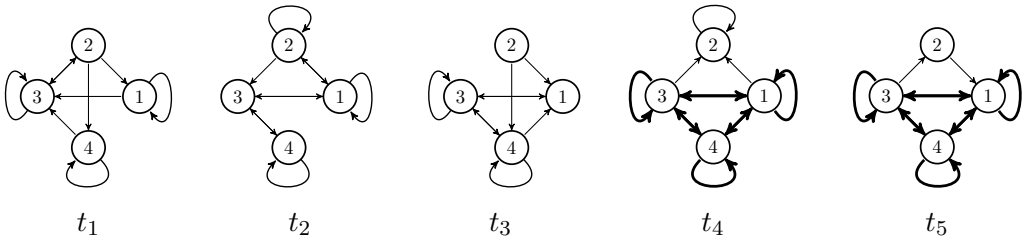


Fig. 5. Maximal clique $\{1, 3, 4\}$ preserved along two timestamps.

Algorithm 1: RULES.

Data: (B, H), i.e., a body and a head
forall the $e \succ \max_{\prec}(H)$ **do**
  if $c_{natural}(B \setminus \{e\} \rightarrow H \sqcup \{e\}) \geqslant \beta_{natural}$ **then**
    $c_e \leftarrow c_{exclusive}(B \setminus \{e\} \rightarrow H \sqcup \{e\})$;
    if $c_e \geqslant \beta_{exclusive} \wedge \neg$REDUNDANT$(B \setminus \{e\}, H \sqcup \{e\}, \epsilon, c_e)$ **then**
      /* ϵ is smaller (w.r.t. $\prec$) than any element */
      output $B \setminus \{e\} \rightarrow H \sqcup \{e\}$;
    RULES$(B \setminus \{e\}, H \sqcup \{e\})$;

In Fig. 5, $\{d_1, d_3, d_4\} \times \{a_1, a_3, a_4\} \times \{t_4, t_5\}$ is such a pattern, i.e., it is not only a closed 3-set but also a cross-graph closed clique (between the vertices 1, 3 and 4; at the times t_4 and t_5). Notice that the *closedness* ensures the maximality of the clique, i.e., it cannot be enlarged into another one that would still hold at both t_4 *and* t_5. It also insures its "maximality on time", i.e., the clique does not appear in any other snapshot of the graph.

From a closed set C, PINARD++ derives multidimensional association rules on $\mathcal{D}'$ that involve all the elements in $\cup_{D^i \in \mathcal{D}'} \pi_{D^i}(C)$. In this way, if $\mathcal{D}'$ contains the dimensions of departure and arrival vertices, specifying a symmetry constraint allows to focus on rules concluding on cliques. In other terms, specifying this constraint allows to look for sets of departure and/or arrival vertices (the bodies of the rules) that *usually* imply larger cliques around them. The thresholds μ, $\beta_{exclusive}$ and β_{natural} are used to specify this in terms of user-defined constraints. When mining dynamic graphs, other useful constraints deal with the time dimension. E.g., the cross graph closed cliques involving almost-contiguous timestamps look interesting [14]. DATA-PEELER's internals and the class of constraints it can efficiently enforce are detailed in [13].

3.2. Deriving non-redundant rules from closed sets

RULES (Algorithm 1) derives *a priori* interesting and non-redundant rules, of the form $B \rightarrow H$, from every frequent closed association A $(= C \setminus \pi_{D^{\text{supp}}}(C))$. 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 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 [4]. Nevertheless, RULES traverses it depth-first. The root of the tree is $A \rightarrow \emptyset$. At every level, H grows by one 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 that $\max_{\prec}(\emptyset)$ is smaller than any other element). Contrary to APRIORI, the pruning

Algorithm 2: REDUNDANT.

Data: (B', H', e', c_e), i.e., a body, a head, the last enumerated element and the exclusive confidence of the tested rule
forall the $f' \in \{f' \in \cup_{D^i \in \mathcal{D}'} \pi_{D^i}(B') \mid f' \succ e'\}$ **do**
  if $c_{natural}(B' \setminus \{f'\} \rightarrow H' \sqcup \{f'\}) = c_{natural}(B' \rightarrow H') \wedge \Big(c_{exclusive}(B' \setminus \{f'\} \rightarrow H' \sqcup \{f'\}) \geqslant c_e \vee \text{REDUNDANT}(B' \setminus \{f'\}, H' \sqcup \{f'\}, f', c_e)\Big)$ **then**
    return true;
return false;

Algorithm 3: PINARD++.

Input: A relation $\mathcal{R}$ on $\mathcal{D}$, $\mathcal{D}' \subsetneq \mathcal{D}$, and $(\mu, \beta_{natural}, \beta_{exclusive}) \in [0, 1]^3$
Output: Every non-redundant and *a priori* interesting rule on $\mathcal{D}'$
$D^{\text{supp}} \leftarrow \times_{D^i \in \mathcal{D} \setminus \mathcal{D}'} D^i$;
$(\mathcal{D}_A, \mathcal{R}_A) \leftarrow (\mathcal{D}' \cup D^{\text{supp}}, \emptyset)$;
forall the $(e_1, e_2, \ldots, e_{|\mathcal{D}'|}, e_{|\mathcal{D}'|+1}, \ldots, e_n) \in \mathcal{R}$ **do**
  $\mathcal{R}_A \leftarrow \mathcal{R}_A \cup (e_1, e_2, \ldots, e_{|\mathcal{D}'|}, (e_{|\mathcal{D}'|+1}, \ldots, e_n))$;
$\mathcal{C} \leftarrow$ DATA-PEELER$(\emptyset, \times_{D^i \in \mathcal{D}_A} D^i)$;
forall the $C \in \mathcal{C}$ **do**
  RULES$(C \setminus \pi_{D^{supp}}(C), \emptyset)$;

criterion is the minimal natural confidence constraint. According to Theorem 2, this pruning is safe, i.e., no rule, with a natural confidence higher than $\beta_{natural}$, is missed. On the opposite, the minimal exclusive confidence and the non-redundancy constraints cannot give rise to search space pruning. That is why they are checked after the constraint on the minimal natural confidence. If both are satisfied then the rule is output. Checking whether the exclusive confidence exceeds $\beta_{exclusive}$ is straightforward. To enforce the non-redundancy, Theorem 3 indicates that, beside the necessity to process a closed set, RULES must check the confidences of the more general rules sharing the same elements. If such a rule has the same natural confidence and a greater or equal exclusive confidence, then the current rule is redundant. That is why the REDUNDANT function (Algorithm 2) browses these more general rules and compare their confidences with that of the current rule. Like RULES, REDUNDANT exploits Theorem 2 to not traverse rules with strictly smaller natural confidence. Finally, PINARD++ (see Algorithm 3) successively (1) constructs $\mathcal{R}_A$, (2) extracts the frequent closed sets in it and (3) derives, from each of these patterns, the non redundant and *a priori* interesting rules.

4. Experimental study

Experiments have 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.

Vélo'v[2] 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. The

[2]http://www.velov.grandlyon.com/.

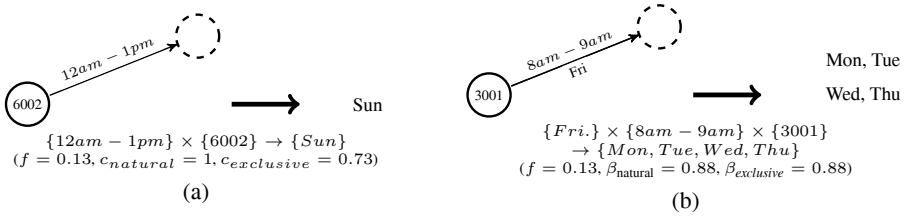


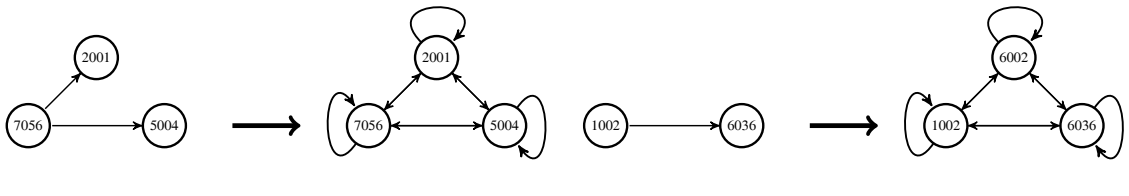
Fig. 6. Example of rules on $\{Departure, Day, Hour\}$.

logs we were granted the access to represent more than 13.1 million rides along 30 months. Vélo'v data can be represented as a dynamic directed graph evolving into two temporal dimensions: the 7 days of the week and the 24 one-hour periods in a day. A significant amount of bicycles (using a local test inspired by the computation of a p-value), that are rented at the (departure) station ds on day d (e.g., Monday) at hour h (e.g., from 1 pm to 2 pm) and returned at the (arrival) station as, translates to an edge from ds to as in the graph timestamped with (d, h). In other terms, (ds, as, d, h) belongs to the relation $\mathcal{R}_{\texttt{Vélov'v}} \subseteq$ *Departure* $\times$ *Arrival* $\times$ *Day* $\times$ *Hour*. $\mathcal{R}_{\texttt{Vélov'v}}$ contains $117,411$ 4-tuples, hence a $\frac{117,411}{7\times24\times327\times327} = 0.7\%$ 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). In other words, the definition of $\mathcal{D}'$ determines the dimensions that may appear in the rules, in the left-hand and/or right-hand sides. Optional additional constraints (e.g., a symmetry constraint) lead to more focused rules taking into account subjective interestingness issues. Let us now discuss a couple of examples.

To study the relationship between stations and their departure times, we discover rules on the dimensions *Departure*, *Day* and *Hour*. As a consequence, the support domain is *Arrival* which contains 327 stations. With $\mu = 0.12$, $\beta_{natural} = 0.8$ and $\beta_{exclusive} = 0.6$, PINARD++ extracts 632 rules. They indicate that preferred departure times are different from one station to another. Figures 6a and 6b report two of them. The rule in Fig. 6a means that the departures from Station 6002 between 12 am and 1pm almost exclusively occur on Sundays ($c_{exclusive} = 0.73$). The natural confidence is 1, i.e., whatever the arrival station, the frequent rides from Station 6002 between midday and 1pm all occur on Sundays. This is consistent with our knowledge of the city because Station 6002 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 indicates that the rides from Station 3001 between 8 am and 9 am usually occur during the working days. This is again consistent with our knowledge that many people living outside Lyon come to work by train and Station 3001 is the closest to the train station in the main working area of the city. It turns out that they then finish their daily trips to work by bicycle.

Let us now consider patterns on graph evolution: we want to look at frequent usage sub-networks (i.e., sub-networks that are often observed) that can confidently be enlarged into cliques? To study such patterns, a rule has to involve $Departure$ and $Arrival$ stations, i.e., $\mathcal{D}' = \{Departure, Arrival\}$. As a result, the support domain is the Cartesian product of the 7 days and the 24 hours. Additional contraints, defined in Section 3.1, are enforced so that PINARD++ processes $(3, 3)$-large cross-graph closed cliques into rules. Moreover we force the body of every rule to be a graph with at least one edge, i.e., it must involve at least one departure station and one arrival station. The non-redundancy of the extracted rules favors the discovery of minimal sub-networks (at the bodies of the rules) that can be confidently (i.e., with a high enough confidence) enlarged into maximal cliques (unions of the bodies and the heads). With $\mu = 0.02$ and $\beta_{natural} = \beta_{exclusive} = 0.7$, 165 rules are discovered. Some of them are reported in Fig. 7.



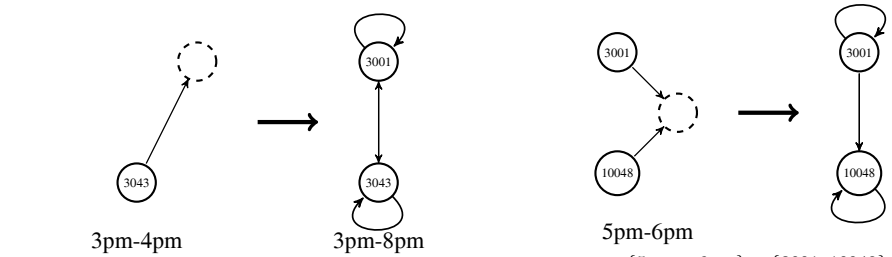
$\{7056\} \times \{5004, 2001\} \rightarrow \{5004, 2001\} \times \{7056\}$
$(f = 0.05, \beta_{\text{natural}} = 0.9, \beta_{\textit{exclusive}} = 0.9)$

(a)

$\{1002\} \times \{6036\} \rightarrow \{6036, 6002\} \times \{1002, 6002\}$
$(f = 0.04, c_{\text{natural}} = 1, c_{\textit{exclusive}} = 1)$

(b)

Fig. 7. Example of rules of the form "sub-network" $\rightarrow$ "maximal clique".



$\{3 - 4pm\} \times \{3043\} \rightarrow \{4 - 8pm\} \times \{3001\} \times \{3043, 3001\}$
$(f = 1, \beta_{\text{natural}} = 1, \beta_{\textit{exclusive}} = 0.74)$

(a)

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

(b)

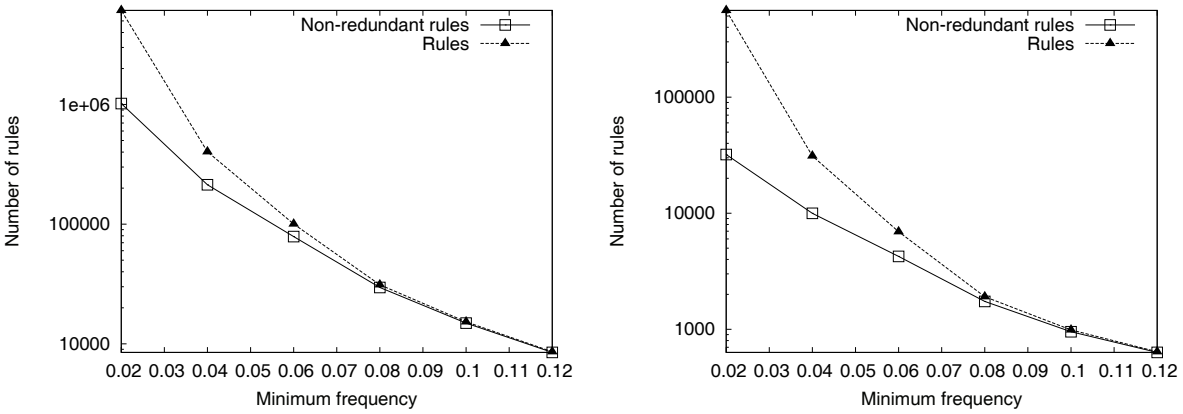
Fig. 8. Example of rules of the form $Hours \times Departures \rightarrow Arrivals$.

The enlarged sub-networks can contain only more edges (see Fig. 7a) or more vertices (see Fig. 7b). These rules explicit diverse mechanisms like auto-regulation and convergence. They can potentially be used to anticipate the effect of a typical breakdown: a station that can only emit (resp. receive) bicycles. If such a station is at the body of a rule, then the other stations in the rules may be overloaded (resp. suffer a shortage). Notice, however, that the discovered rules are descriptive. Using them to make predictions is an interesting perspective.

Here is another interesting question: do some stations exchange many bicycles at favored hours every day? A rule answering it must obviously be defined on $\mathcal{D}' = \{Departure, Arrival, Hour\}$. To focus on rules that hold every day, the minimal frequency threshold is set to 1. With $\beta_{\textit{natural}} = 1$ and $\beta_{\textit{exclusive}} = 0.6$, PINARD++ returns 51 rules involving at least one time period, two departure stations and two arrival stations. Figure 8 depicts two 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.

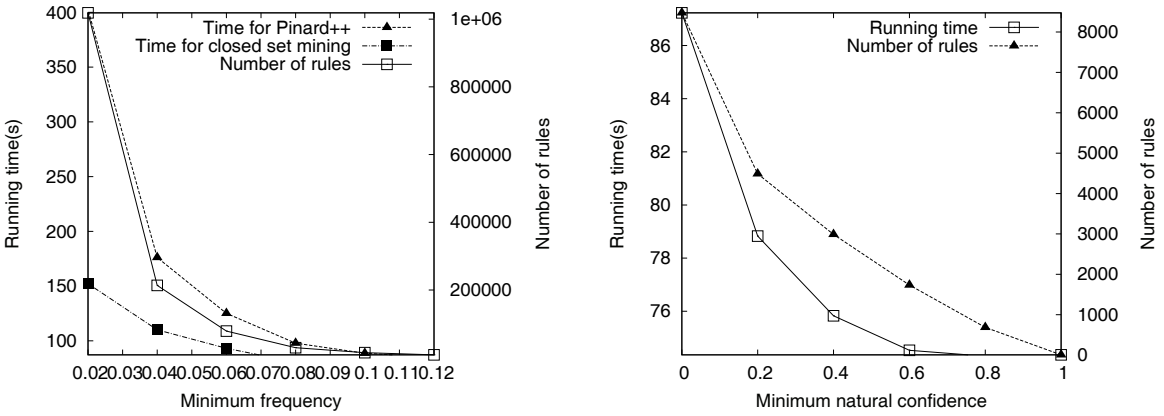
When mining rules that only satisfy the minimum frequency and minimum confidence constraints, many redundant rules are returned although they do not provide new insights. Figure 9 illustrates the proportion of rules that are avoided thanks to our non-redundancy approach (see Section 2.2.4). Obviously, with low minimal frequency constraints, this significantly limits pattern flooding.

Let us finally provide a performance study when mining *a priori* interesting rules in $\mathcal{R}_{\texttt{Vélo'v}}$ with $\mathcal{D}' = \{Departure, Day, Hour\}$. When the minimal frequency threshold increases, both the number of frequent rules and the running time decrease. Figure 10a was obtained with $\beta_{\textit{natural}} = \beta_{\textit{exclusive}} = 0$. PINARD++ prunes large areas of the search space where no association is frequent. The time spent on extracting the closed sets is given as well. It shows that each step contributes to the overall complexity. Theorem 2 enables to deeply prune the search space too. Indeed, the RULES algorithm does not traverse



(a) $\beta_{natural} = \beta_{exclusive} = 0$. (b) $\beta_{natural} = 0.8$, $\beta_{exclusive} = 0.6$.

Fig. 9. Impact of non-redudancy.



(a) Pruning w.r.t. μ. (b) Pruning w.r.t. $\beta_{natural}$.

Fig. 10. Effectiveness of PINARD++ .

the enumeration sub-trees empty of confident rules (w.r.t. $\beta_{natural}$). That is why both the number of rules and the time it takes to extract them decrease when the minimum natural confidence threshold increases. Experiments in Fig. 10b are performed with $\beta_{exclusive} = 0$, $\mu = 0.12$, and $\beta_{natural}$ varying between 0 and 1.

PINARD++'s scalability is tested w.r.t. the size and the density of the data. Starting with the size, rules on $\{Departure, Day, Hour\}$ are mined with $\mu = 0.12$ and $\beta_{natural} = \beta_{exclusive} = 0$ in datasets obtained from $\mathcal{R}_{\texttt{Vélov'v}}$ by replicating, up to ten times, the support dimension, i.e., the arrival stations. It turns out that the algorithm scales linearly: a linear regression of $s \mapsto \frac{T_s}{T_1}$ (where s is the replication factor and T_s the running time on the data with s replications) gives $y = 1.1x + 0.016$ with 0.9997 as a determination coefficient.

To test the PINARD++'s scalability w.r.t. the density of the dataset, synthetic 3-ary relations are generated. The sizes of the domains are constant: $10 \times 50 \times 100$. Here, the only variable is the density, i.e., the ratio between the number of 3-tuples present in the relation and $10 \times 50 \times 100 = 50,000$. We made it increase, 0.02 by 0.02, from 0.1 (for the first dataset) to 0.5 (for the last dataset). The PINARD++'s

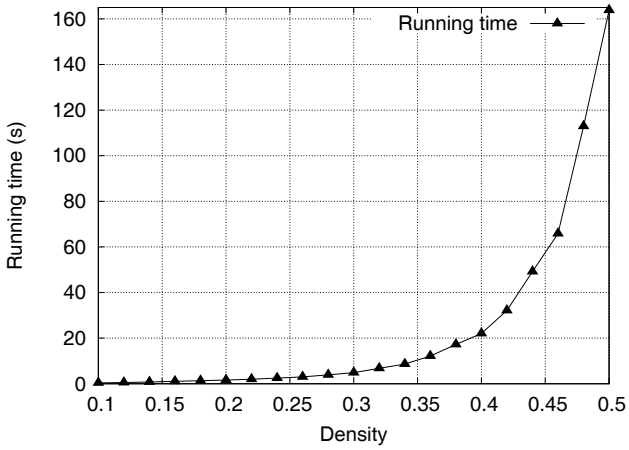


Fig. 11. PINARD++'s scalability w.r.t. the density.

running times are in Fig. 11. As expected, when the density is important, the extraction is much harder. However, let us note that 40% density is already extremely high to be met in practice.

5. Related work

Mining graphs has recently received a lot of attention in the data mining community. Many papers study the evolution of graphs over time with a large variety of techniques. On one hand, several papers have focused on the evolution of macroscopic graph properties [5,15,24,32,35] where some have concerned the dynamical properties of real graphs such as densification laws, shrinking diameters [26], and the evolution of known communities over time [6,24]. On the other hand, some works have studied graph evolution at a local level thanks to local patterns. This section focuses on such methods. Besides, notice that our work differs from tensor factorization based patten mining approaches [1] which is generally targeted towards numerical data analysis and do not consider complete methods for local pattern discovery.

In [10], Borgwardt et al. extract local patterns in labeled dynamic graphs. The approach aims at finding subgraphs that are topologically frequent and show an identical dynamic behavior over time, i.e., insertions and deletions of edges occur in the same order. Because this task is computationally hard, the algorithm is not complete. Indeed, computing the overlap-based support measure means solving a maximal independent set problem for which thay propose a greedy algorithm. Inokuchi and Washio introduce a fast algorithm to mine frequent transformation subsequences from a set of dynamic labeled graphs (the labels on vertices and edges can change over time). Assuming that the changes in a dynamic graph are gradual, they propose to succinctly represent the dynamics with a graph grammar: each change between two observed successive graph states is interpolated by axiomatic transformation rules. A significant improvement is proposed in [20]. Motivated by the intractability of their approach on long sequences of large graphs, the same authors define, in [19], induced subgraph subsequence. This novel class of subgraph subsequence enables to efficiently mine frequent patterns from graph sequences containing long sequences and large graphs.

In [37], You et al. study how a graph is structurally transformed through time. They compute graph rewriting rules that describe the evolution of two consecutive graphs. These rules are then abstracted

into patterns representing the dynamics of a sequence of graphs. The main step concerns the computation of maximum common subgraphs between two consecutive graphs. Indeed, this problem is NP-complete. In the case of relational graphs (graphs with unique vertex labels such as the ones tackled by PINARD++), this becomes tractable [11,12]. Indeed, the complexity is then quadratic and graph rewriting rules are efficiently discovered. In [9], the authors focus on detecting clusters of temporal snapshots of an evolving network. These clusters can be interpreted as evolution eras of the dynamic graph. This approach enables to detect periods in which sudden change of behavior appears. Such high-level trends are expressed by sudden increases or decreases of the similarity between the structures of the consecutive graphs. In [25], Lahiri et al. introduce the periodic subgraph mining problem, i.e., identifying every frequent closed periodic subgraph. They empirically study the efficiency and the interest of their proposal on several real-world dynamic social networks. By encoding dynamic graphs as ternary relations [14], describes a constraint-based mining approach to discover maximal cliques that are preserved over almost-contiguous timestamps. The constraints are pushed into the closed pattern mining algorithm DATA-PEELER. Notice that PINARD++ post-processes DATA-PEELER's closed sets to generate *a priori* interesting and non-redundant rules. In [30], Robardet proposes a constraint-based approach too. It studies the evolution of dense and isolated subgraphs defined by two user-parameterized constraints. Associating a temporal event type with each pattern captures the temporal evolution of the identified subgraph, i.e., the formation, dissolution, growth, diminution and stability of subgraphs between two consecutive timestamps. The algorithm incrementally processes the time series of graphs. In [8], the authors introduce the problem of extracting graph evolution rules satisfying minimal support and confidence constraints. It finds isomorphic subgraphs that match the timestamps associated with each edge, and, if present, the properties of the vertices and edges of the dynamic graph. Graph evolution rules are then derived with two different confidence measures. This approach is the closest to ours: it aims at describing a time-evolving graph with descriptive rules. Nevertheless, this work focuses on the dynamic changes in the graph whereas we provide a generic framework to discover multidimensional rules in which the time is either in the rule or in its support.

Considering binary relation mining, since the seminal papers [3,4], the discovery of association rules that satisfy both a minimal support and a minimal confidence constraints has been extensively studied. Many works deal with the generalization of this task towards n-ary relations. These proposals can be classified into different types according to the number of involved dimensions within an association rule [23]. Indeed, three types have been defined: intra-dimensional, inter-dimensional, and hybrid association rules. Concerning intra-dimensional rules, all the elements of a rule belong to a single dimension. This case has been extremely well studied for binary relations. In [31], the authors propose to discover intra-dimensional association rules in a n-ary relation where $n \geqslant 2$. For each dimension, association rules between elements of the dimension domain are discovered. The tuples that belong to the cartesian product of the other dimensions are considered as the transaction domain. Inter-dimensional association rules were proposed to enable the discovery of associations or co-occurrences between elements from different dimensions [7,23]. It should be noticed that dimensions must be distinct and two elements from the same dimension cannot appear together in an inter-dimensional rule. The computation of inter-dimensional association rules is then guided by a metarule. A metarule contains distinct predicates and can be used to focus the data mining search towards rules satisfying the predicates. The absence of repetitive predicates is a limitation on the expressiveness of rules. Other authors have proposed ad-hoc algorithms to extract hybrid rules in which the repetition of few dimensions is enabled [16,18,33].

[17,27,34,36] have studied "How current and past values are related to future values". In [34], Oates et al. look for dependencies in a set of time series (fixed length). They are expressed as rules of the

following form: "If an instance of pattern x begins in the series at time t, then an instance of pattern y will begin at time $t + \delta$ with probability p". Where, a pattern is a set of tokens, each token is described by a token value, a time series which includes this token value and a temporal offset. In [17,27,36], the authors studied the extraction of inter-transaction association rules in which associations are not on the same transaction. The body and the head of a inter-transaction association rule are sets of extended items. Each extended item is described by an attribute item and a point in the m-dimensional space where this attribute item appears. Therefore, in these proposals, the body and the head of a rule have the same dimensions. Furthermore, by construction, these proposed rules do not enable to discover the relation of patterns/items which occur on the same transaction.

This paper is an extension of [28] in which we proposed a restricted form of rules: the inter-dimensional rules where the dimensions at both sides of a rule must be disjoint. In [29], we study multidimensional rules in the general framework of Boolean tensors. However, neither the redundancy issue nor the specifics of dynamic graph mining are addressed in these previous works.

6. Conclusion

We have tackled the problem of describing dynamic graphs via rules that can involve subsets of some arbitrary dimensions (including temporal dimensions) at its body or head. We have proposed a new semantics for multidimensional association rules in dynamic graphs. It relies on relevant objective interestingness measures called the exclusive confidence and the natural confidence. We also revisited non redundancy aspects. We have introduced and implemented PINARD++, an effective solution for computing such rules. Experiments on a real-world dynamic graph demonstrated the interest of our proposal. A timely challenge is to further study primitive constraints that can support more sophisticated knowledge discovery processes in dynamic graphs. Some of these constraints would deal with the temporal dimension(s) (e.g., time contiguity [14]). Other constraints would deal with the "form" of the patterns to discover (e.g., cliques, dense subgraphs, etc.). Using multidimensional association rules for supervised classification is another appealing perspective.

Acknowledgements

This work was partly funded by the ANR project FOSTER (COSINUS 2010), by FAPEMIG, and by a grant from the Vietnamese government.

References

[1] E. Acar, D.-M. Dunlavy and T.-G. Kolda, *Link prediction on evolving data using matrix and tensor factorizations*, ICDM Workshops, IEEE Computer Society, 2009, pp. 262–269.
[2] C.C. Aggarwal and H. Wang, *Managing and mining graph data*, Springer, 2010.
[3] R. Agrawal, T. Imielinski and A.N. Swami, *Mining association rules between sets of items in large databases*, SIGMOD, vol. 22, ACM, 1993, pp. 207–216.
[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, *Fast discovery of association rules*, Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 1996, pp. 307–328.
[5] L. Akoglu, M. McGlohon and C. Faloutsos, *RTM: Laws and a recursive generator for weighted time-evolving graphs*, ICDM, IEEE Computer Society, 2008, pp. 701–706.
[6] L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan, *Group formation in large social networks: membership, growth, and evolution*, KDD, ACM, 2006, pp. 44–54.

[7] R. Ben Messaoud, S. Loudcher Rabaséda, O. Boussaid and R. Missaoui, *Enhanced mining of association rules from data cubes*, DOLAP, ACM, 2006, pp. 11–18.
[8] M. Berlingerio, F. Bonchi, B. Bringmann and A. Gionis, *Mining graph evolution rules*, ECML/PKDD, vol. 5781, Springer, 2009, pp. 115–130.
[9] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale and D. Pedreschi, *As time goes by: Discovering eras in evolving social networks*, PAKDD, vol. 6118, Springer, 2010, pp. 81–90.
[10] K.M. Borgwardt, H.-P. Kriegel and P. Wackersreuther, *Pattern mining in frequent dynamic subgraphs*, ICDM, IEEE Computer Society, 2006, pp. 818–822.
[11] B. Bringmann and S. Nijssen, *What is frequent in a single graph?* PAKDD, vol. 5012, Springer, 2008, pp. 858–863.
[12] T. Calders, J. Ramon and D. Van Dyck, *Anti-monotonic overlap-graph support measures*, ICDM, IEEE Computer Society, 2008, pp. 73–82.
[13] L. Cerf, J. Besson, C. Robardet and J.-F. Boulicaut, Closed patterns meet n-ary relations, *ACM Trans Knowl Discov Data* **3**(1) (2009), 1–36.
[14] L. Cerf, T.B.N. Nguyen and J.-F. Boulicaut, *Mining constrained cross-graph cliques in dynamic networks*, Inductive Databases and Constraint-based Data Mining, Springer, 2010, pp. 199–228.
[15] Y. Chi, S. Zhu, X. Song, J. Tatemura and B.L. Tseng, *Structural and temporal analysis of the blogosphere through community factorization*, KDD, ACM, 2007, pp. 163–172.
[16] G. Dong, J. Han, J.-M.-W. Lam, J. Pei and K. Wang, *Mining multi-dimensional constrained gradients in data cubes*, VLDB, Morgan Kaufmann, 2001, pp. 321–330.
[17] L. Feng, J.X. Yu, H. Lu and J. Han, *A template model for multidimensional inter-transactional association rules*, *VLDB Journal* **11**(2) (2002), 153–175.
[18] T. Imielinski, L. Khachiyan and A. Abdulghani, Cubegrades: Generalizing association rules, *Data Min Knowl Discov* **6**(3) (2002), 219–257.
[19] A. Inokuchi and T. Washio, *Mining frequent graph sequence patterns induced by vertices*, SDM, SIAM, 2010, pp. 466–477.
[20] A. Inokuchi and T. Washio, *Gtrace2: Improving performance using labeled union graphs*, PAKDD, vol. 6119, Springer, 2010, pp. 178–188.
[21] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter and G. Stumme, TRIAS – *an algorithm for mining iceberg tri-lattices*, ICDM, IEEE Computer Society, 2006, pp. 907–911.
[22] L. Ji, K.-L. Tan and A.K.H. Tung, *Mining frequent closed cubes in 3D data sets*, VLDB, VLDB Endowment, 2006, pp. 811–822.
[23] M. Kamber, J. Han and J.Y. Chiang, *Metarule-guided mining of multi-dimensional association rules using data cubes*, KDD, AAAI, 1997, pp. 207–210.
[24] R. Kumar, J. Novak and A. Tomkins, *Structure and evolution of online social networks*, KDD, ACM, 2006, pp. 611–617.
[25] M. Lahiri and T.Y. Berger-Wolf, *Mining periodic behavior in dynamic social networks*, ICDM, IEEE Computer Society, 2008, pp. 373–382.
[26] J. Leskovec, J. Kleinberg and C. Faloutsos, *Graphs over time: densification laws, shrinking diameters and possible explanations*, KDD, ACM, 2005, pp. 177–187.
[27] H. Lu, J. Han and L. Feng, *Stock movement prediction and n-dimensional inter-transaction association rules*, SIGMOD Workshop DMKD, 1998, pp. 1–7.
[28] K.-N.T. Nguyen, L. Cerf, M. Plantevit and J.-F. Boulicaut, *Discovering inter-dimensional rules in dynamic graphs*, DYNAK, CEUR Workshop Proceedings, 2010, pp. 5–16.
[29] K.-N.T. Nguyen, L. Cerf, M. Plantevit and J.-F. Boulicaut, *Multidimensional association rules in boolean tensors*, SDM, SIAM/Omnipress, 2011, pp. 570–581.
[30] C. Robardet, *Constraint-based pattern mining in dynamic graphs*, ICDM, IEEE Computer Society, 2009, pp. 950–955.
[31] C. Schmitz, A. Hotho, R. Jäschke and G. Stumme, *Mining association rules in folksonomies*, Data Science and Classification, Springer, 2006, pp. 261–270.
[32] J. Sun, S. Papadimitriou, P.S. Yu and C. Faloutsos, *Graphscope: Parameter-free mining of large time-evolving graphs*, KDD, ACM, 2007, pp. 687–696.
[33] H.C. Tjioe and D. Taniar, Mining association rules in data warehouses, *Int Journal of Data Warehousing and Mining* **1**(3) (2005), 28–62.
[34] T. Oates and P.-R. Cohen, *Searching for structure in multiple streams of data*, ICML, Morgan Kaufmann, 1996, pp. 346–354.
[35] H. Tong, S. Papadimitriou, J. Sun, P.S. Yu, and F. Christos, *Colibri: fast mining of large static and dynamic graphs*, KDD, ACM, 2008, pp. 686–694.
[36] A.K.H. Tung, H. Lu, J. Han and L. Feng, *Breaking the barrier of transactions: mining inter-transaction association rules*, KDD, ACM, 1999, pp. 297–301.

[37] C.-H. You, L.B. Holder and D.J. Cook, *Learning patterns in the dynamics of biological networks*, KDD, ACM, 2009, pp. 977–986.
[38] M. J. Zaki, Mining non-redundant association rules, *Data Min Knowl Discov* **9**(3) (2004), 223–248.

Technical annex

Theorem 1. According to the Definitions 6 and 2:

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

Let $\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X) = \{\mathrm{u} \in \times_{D^i \in \mathcal{D} \setminus \mathcal{D}_Y} D^i \mid \exists \mathrm{v} \in \times_{D^i \in \mathcal{D}_Y \setminus \mathcal{D}_X} D^i$ such that $\forall x \in X, x \cdot \mathrm{v} \cdot \mathrm{u} \in \mathcal{R}\}$.

Then, $\begin{cases} s(Y) \subseteq \pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X) \\ |\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X)| \leqslant |s(X)| \end{cases}$ and $|s(Y)| \leqslant |\pi_{\mathcal{D} \setminus \mathcal{D}_Y} s(X)| \leqslant |s(X)|$. □

Theorem 2. Using Definition 10, we have $X \sqsubseteq X' \Rightarrow s_{\mathcal{D} \setminus \mathcal{D}'}(X') \subseteq s_{\mathcal{D} \setminus \mathcal{D}'}(X)$.

Because $X \sqsubseteq X' \sqsubseteq Y$ and according to Definition 11:

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

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

Theorem 3. We first prove that if $X \to Y$ is a non-redundant rule then $X \sqcup Y \sqcup s(X \sqcup Y)$ is a closed set and $\forall X' \sqsubset X$, $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{natural}}(X \to Y) \vee c_{\text{exclusive}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{exclusive}}(X \to Y)$.

- By Definition 2, $X \sqcup Y \sqcup s(X \sqcup Y) \subseteq \mathcal{R}_A$. As a consequence and by definition of the support of $X \sqcup Y \sqcup \{e\}$, $s(X \sqcup Y) \subseteq s(X \sqcup Y \sqcup \{e\})$. It follows, assuming, by contradiction, that $X \sqcup Y \sqcup s(X \sqcup Y)$ is not a closed set (see Definition 15) means it can extended by an element e while still only covering tuples in $\mathcal{R}_A$. By definition of the support of $X \sqcup Y$, this element e cannot be in the support dimension. Therefore $\exists e \in \cup_{D^i \in \mathcal{D}'}(D^i \setminus \pi_{D^i}(X \sqcup Y)) \mid X \sqcup Y \sqcup \{e\} \sqcup s(X \sqcup Y) \subseteq \mathcal{R}_A$. As a consequence, and by definition of the support of $X \sqcup Y \sqcup \{e\}$, $s(X \sqcup Y) \subseteq s(X \sqcup Y \sqcup \{e\})$. It follows from the definitions of the interestingness measures that:

$$\begin{cases} f(X \to Y \sqcup \{e\}) \geqslant f(X \to Y) \\ c_{\text{natural}}(X \to Y \sqcup \{e\}) \geqslant c_{\text{natural}}(X \to Y) \\ c_{\text{exclusive}}(X \to Y \sqcup \{e\}) \geqslant c_{\text{exclusive}}(X \to Y) \end{cases}.$$

  This contradicts the non-redundancy of $X \to Y$ (Definition 14). Therefore $X \sqcup Y \sqcup s(X \sqcup Y)$ is a closed set.
- According to Definition 14, $X \to Y$ non-redundant implies that, for any $X' \sqsubset X$, at least one of these three assertions is true:

1. $f(X' \to (Y \sqcup X) \setminus X') < f(X \to Y)$;
2. $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{natural}}(X \to Y)$;
3. $c_{\text{exclusive}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{exclusive}}(X \to Y)$.

However, by Definition 8, $f(X' \to (Y \sqcup X) \setminus X') = f(X \to Y)$, i.e., the first assertion is always false. Furthermore, Theorem 2 tells that $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') \leqslant c_{\text{natural}}(X \to Y)$. All in all, it can be simply written that $X \to Y$ non-redundant implies $\forall X' \sqsubset X$, $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{natural}}(X \to Y) \vee c_{\text{exclusive}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{exclusive}}(X \to Y)$.

We now prove that if $X \sqcup Y \sqcup s(X \sqcup Y)$ is a closed set and $\forall X' \sqsubset X$, $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{natural}}(X \to Y) \vee c_{\text{exclusive}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{exclusive}}(X \to Y)$ then $X \to Y$ is non-redundant. By contradiction, assume that $X \to Y$ is redundant. By Definition 14, one of the two following cases occurs.

Case 1 There exists a rule $X' \to Y'$ such that

$$\begin{cases} X' \sqcup Y' \sqsupset X \sqcup Y \\ f(X' \to Y') \geqslant f(X \to Y) \end{cases}.$$

According to Theorem 1, $X' \sqcup Y' \sqsupset X \sqcup Y \Rightarrow s(X' \sqcup Y') \sqsubseteq s(X \sqcup Y)$. However given the second condition (and the mere definition of the frequency), we also have $|s(X' \sqcup Y')| \geqslant |s(X \sqcup Y)|$. As a consequence $s(X' \sqcup Y') = s(X \sqcup Y)$ and $X \sqcup Y \sqcup s(X \sqcup Y) \sqsubset X' \sqcup Y' \sqcup s(X' \sqcup Y') \subseteq \mathcal{R}_A$. This contradicts the closedness of $X \sqcup Y \sqcup s(X \sqcup Y)$.

Case 2 There exists a rule $X' \to Y'$ such that:

$$\begin{cases} X' \sqcup Y' = X \sqcup Y \wedge X' \sqsubset X \\ c_{\text{natural}}(X' \to Y') \geqslant c_{\text{natural}}(X \to Y) \\ c_{\text{exclusive}}(X' \to Y') \geqslant c_{\text{exclusive}}(X \to Y) \end{cases}.$$

According to Theorem 2, the first condition implies $c_{\text{natural}}(X' \to Y') \leqslant c_{\text{natural}}(X \to Y)$. As a consequence, the second condition can be rewritten as $c_{\text{natural}}(X' \to Y') = c_{\text{natural}}(X \to Y)$. Since $Y' = (Y \sqcup X) \setminus X'$, this contradicts the assumption that $\forall X' \sqsubset X$, $c_{\text{natural}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{natural}}(X \to Y) \vee c_{\text{exclusive}}(X' \to (Y \sqcup X) \setminus X') < c_{\text{exclusive}}(X \to Y)$.

Therefore $X \to Y$ is non-redundant. □