Mining Disjunctive Rules in Dynamic Graphs

Kim-Ngan T. Nguyen
INSA-Lyon
LIRIS, UMR5205, F-69621, France
thi-kim-ngan.nguyen@insa-lyon.fr

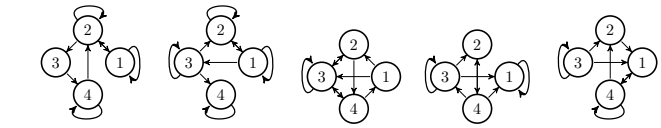
Marc Plantevit
Université Lyon 1
LIRIS, UMR5205, F-69622, France
marc.plantevit@univ-lyon1.fr

Jean-François Boulicaut
INSA-Lyon
LIRIS, UMR5205, F-69621, France
jean-francois.boulicaut@insa-lyon.fr

***Abstract*—Recently, a generalization of association rules that hold in n-ary Boolean tensors has been proposed. Moreover, preliminary results concerning their application to dynamic relational graph analysis have been obtained. We build upon such a formalization to design more expressive local patterns in this special case of dynamic graph where the set of vertices remains unchanged though edges that connect them may appear or disappear at the different timestamps. To design the pattern domain of the so-called disjunctive rules, we have to design (a) the pattern language, (b) interestingness measures which serve as the counterpart of the popular support and confidence measures in standard association rules, and (c) an efficient algorithm that may compute every rule that satisfies some primitive constraints like minimal frequencies or minimal confidences. The approach is tested on real datasets and we discuss the expressivity and the relevancy of some computed disjunctive rules.**

I. Introduction

Relational graphs are quite common: we can often collect large volumes of interactions between identified entities. Mining such graphs to discover interaction patterns has been studied a lot. For instance, many researchers have designed efficient solutions to compute collections of subgraph patterns in static graphs (e.g., looking for dense subgraphs like cliques or quasi-cliques, frequent subgraph mining). Far less work has concerned the analysis of dynamic ones, i.e., when interactions may appear of dissapear at the different timestamps for which the relational graph is available. The popular asssociation rule mining task has many applications since its definition in [1]. It has been extended in various directions, including towards multidimensional data (see, e.g., [2], [3]). To the best of our knowledge, the formalization in [4] is the most general ones: it describes the semantics of conjunctive descriptive rules in arbitrary n-ary relations. Both the body and the head of the rules can involve subsets of any dimensions. Furthermore, the method in [4] computes many redundant rules and this problem has been studied in [5]. Notice that in these two papers, preliminary applications to relational oriented dynamic graph analysis have been reported. In [6], the authors have also introduced graph-evolution rules to discover patterns that describe local changes occurring in an evolving graph. However, in these previous works, the rules are always conjunctive rules. We want to increase the expressivity of the discovered rules thanks to disjunction in rule heads and to consider their added-value for dynamic graph analysis. More precisely, we investigate relational directed dynamic graph mining: such graphs can be represented by means of the collection of their adjacency matrices at each considered timestamp. In other terms, their sets of vertices are fixed and only edges can appear or disappear. For example, Fig. 1 depicts such a dynamic graph: 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' value 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. Two examples of disjunctive rules that we want to extract are given in Fig. 2. The rule in Fig. 2a indicates that, at a time, if the edges from Vertices 3 and 4 converge then they tend to converge to Vertex 3 or both Vertex 1 and Vertex 2. The rule in Fig. 2b means that, at a time, if the graph contains the sub-graph including the edge from Vertex 1 to Vertex 2 then it can contains the sub-graph including the edges from Vertices 1 and 2 to Vertices 1 and 2 or the sub-graph including the edges from Vertices 1 and 3 to Vertices 2 and 3. To the best of our knowledge, none available method can support the discovery of such rules. Notice however that the formalization in [4] provides a sound basis for an extension towards disjunctive rules. The second contribution in this paper concerns the design and the implementation of the algorithm CIDRE[1], that exhaustively lists a priori interesting rules. It builds upon an efficient algorithm that computes closed patterns from Boolean tensors [7] and our previous work on non-redundant multidimensional association rule discovery [5].



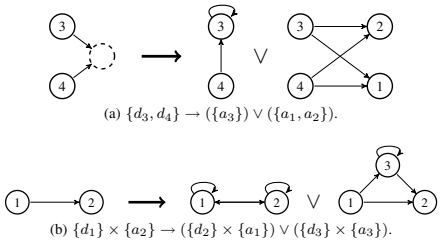
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
d_1	1	1			1	1	1	1		1	1		1							1
d_2	1	1	1		1	1					1	1				1	1			
d_3				1		1	1			1	1		1	1	1		1	1	1	
d_4		1		1				1	1		1		1	1	1		1	1		1
	t_1				t_2				t_3				t_4				t_5			

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

II. A Disjunctive Rule Pattern Domain

The proposed semantics for disjunctive rules (as well as the algorithm listing them all in a given dataset) actually applies to any n-ary relation and thus arbitrary Boolean tensors. The domains (a domain is the set of elements of each

[1]CIDRE Is a Disjunctive Rule Extractor.



(a) $\{d_3, d_4\} \rightarrow (\{a_3\}) \vee (\{a_1, a_2\})$.

(b) $\{d_1\} \times \{a_2\} \rightarrow (\{d_2\} \times \{a_1\}) \vee (\{d_3\} \times \{a_3\})$.

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

dimension) are denoted $D^1, D^2, \ldots, D^n$, these domains are supposed finite and disjoint, we use $\mathcal{D}$ to denote the set of all domains ($\{D^1, D^2, \ldots, D^n\}$). The relation $\mathcal{R}$ is a subset of $D^1 \times \cdots \times D^n$. The idea is that associations between domains make sense on a subset of the domains in $\mathcal{R}$. However, the domains that do not appear in the associations (such a set of domains is denoted $\mathcal{D}_S$) can be used to measure their strength (e.g., counting for frequencies). The Cartesian product of all the domains in $\mathcal{D}_S$ ($\times_{D^i \in \mathcal{D}_S} D^i$) is named the *support domain* of the associations on sub-sets of $\mathcal{D} \setminus \mathcal{D}_S$. For example, in $\mathcal{R}_E$, assuming $\mathcal{D}_S = \{D^3\}$, D^3 is the support domain of the associations on $\{D^1\}$, $\{D^2\}$ and $\{D^1, D^2\}$. Without loss of generality, the domains in $\mathcal{D}_S$ are assumed ordered such that $\mathcal{D}_S = \{D^{n-|\mathcal{D}_S|+1}, D^{n-|\mathcal{D}_S|+2}, ..., D^n\} \subseteq \mathcal{D}$.

Definition 1 (Association)**.** *$\forall \mathcal{D}' \subseteq \mathcal{D} \setminus \mathcal{D}_S$, $\times_{D^i \in \mathcal{D}'} X^i$ is an association on $\mathcal{D}'$ iff $\forall D^i \in \mathcal{D}'$, $X^i \neq \emptyset \wedge X^i \subseteq D^i$. By convention, the only association on $\mathcal{D}' = \emptyset$ is denoted $\emptyset$.*

For example, in $\mathcal{R}_E$, $\{d_1\} \times \{a_2\}$, $\{d_1, d_3\} \times \{a_2, a_3\}$ and $\{d_3, d_4\} \times \{a_3\}$ are associations on $\{D^1, D^2\}$. $\{d_3, d_4\}$ is not, it is an association on $\{D^1\}$.

We use the operators on associations defined in [4]. We assume their semantics can be understood thanks to the given examples on $X_e = \{d_2, d_4\}$ (an association on $\{D^1\}$) and $Y_e = \{d_3, d_4\} \times \{a_3\}$ (an association on $\{D^1, D^2\}$).

- *Projection of an association on a dimension (π)*: $\pi_{D^1}(X_e) = \{d_2, d_4\}$, $\pi_{D^2}(X_e) = \emptyset$, $\pi_{D^3}(X_e) = \emptyset$, $\pi_{D^1}(Y_e) = \{d_3, d_4\}$, $\pi_{D^2}(Y_e) = \{a_3\}$, and $\pi_{D^3}(Y_e) = \emptyset$.
- *Union of two associations ($\sqcup$)*: $X_e \sqcup Y_e$ is an association on $\{D^1, D^2\}$ ($= \{D^1\} \cup \{D^1, D^2\}$), $X_e \sqcup Y_e = (\pi_{D^1}(X_e) \cup \pi_{D^1}(Y_e)) \times (\pi_{D^2}(X_e) \cup \pi_{D^2}(Y_e)) = (\{d_2, d_4\} \cup \{d_3, d_4\}) \times (\{\emptyset\} \cup \{a_3\}) = \{d_2, d_3, d_4\} \times \{a_3\}$.
- *Complement of two associations ($\setminus$)*: $Y_e \setminus X_e$ is an association on $\{D^1, D^2\}$, $Y_e \setminus X_e = (\pi_{D^1}(Y_e) \setminus \pi_{D^1}(X_e)) \times (\pi_{D^2}(Y_e) \setminus \pi_{D^2}(X_e)) = (\{d_3, d_4\} \setminus \{d_2, d_4\}) \times (\{a_3\} \setminus \{\emptyset\}) = \{d_3\} \times \{a_3\}$. In contrast, $X_e \setminus Y_e$ is an association on $\{D^1\}$ only and $X_e \setminus Y_e = \pi_{D^1}(X_e) \setminus \pi_{D^1}(Y_e) = \{d_2, d_4\} \setminus \{d_3, d_4\} = \{d_2\}$.
- *Inclusion of associations ($\sqsubseteq$)*: $\{d_1\} \times \{a_2\} \sqsubseteq \{d_1, d_3\} \times \{a_2, a_3\}$ and $\{d_3, d_4\} \sqsubseteq \{d_3, d_4\} \times \{a_3\}$.

The support domain of an association on $\mathcal{D}' \subseteq \mathcal{D} \setminus \mathcal{D}_S$ is $\times_{D^i \in \mathcal{D}_S} D^i$, e.g., D^3 is the support domain of associations on sub-sets of $\{D^1, D^2\}$. The *support* of an association is the set of elements included in the support domain and "connected" with this association. In its definition, the concatenation is denoted as '$\cdot$'. For instance, $(d_2, a_3) \cdot (t_1) = (d_2, a_3, t_1)$.

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

Note that in the extreme case $\mathcal{D}' = \mathcal{D}$ (i. e., $\mathcal{D}_S = \emptyset$), by convention $\times_{D^i \in \emptyset} D^i = \{\epsilon\}$ (where ϵ is the empty word), the support of an association on $\mathcal{D}$ has either zero element (at least one n-tuple it contains is absent from $\mathcal{R}$) or one element, ϵ, (every n-tuple it contains is in $\mathcal{R}$). The support of the empty association, $s(\emptyset)$, is $\times_{D^i \in \mathcal{D}_S} D^i$. This definition of the support is a slight evolution of the definition given in [4]. It clearly generalizes that of an *itemset* in a Boolean matrix (i.e., when $n = 2$ and $\mathcal{D}_S = \{D^2\}$). For example, assuming that D^3 is the support domain in $\mathcal{R}_E$: $s(\{d_1\} \times \{a_2\}) = \{t_1, t_2, t_3\}$, $s(\{d_1, d_3\} \times \{a_2, a_3\}) = \{t_2, t_3\}$, $s(\{d_3, d_4\} \times \{a_3\}) = \{t_3, t_4\}$, $s(\{d_3, d_4\}) = \{t_1, t_3, t_4, t_5\}$.

The *anti-monotonicity* of the support cardinality, that is well known in itemset mining, still holds here.

Theorem 1 (Support Anti-Monotonicity)**.** *$\forall \mathcal{D}_X \subseteq \mathcal{D} \setminus \mathcal{D}_S$ and $\forall \mathcal{D}_Y \subseteq \mathcal{D} \setminus \mathcal{D}_S$, let X (resp. Y) be an association on $\mathcal{D}_X$ (resp. on $\mathcal{D}_Y$), $X \sqsubseteq Y \Rightarrow s(X) \supseteq s(Y)$.*

For example, we can check that $s(\{d_1\} \times \{a_2\}) \supseteq s(\{d_1, d_3\} \times \{a_2, a_3\})$ and $s(\{d_3, d_4\}) \supseteq s(\{d_3, d_4\} \times \{a_3\})$, i. e., that Theorem 1 holds.

Given a relation $\mathcal{R}$ and the user-defined support domain $\mathcal{D}_S \subseteq \mathcal{D}$, a *disjunctive rule* on $\mathcal{D} \setminus \mathcal{D}_S$ is of the form $X \rightarrow \vee\mathcal{Y}$, that is disjunctions can appear on its head. It is simply called a rule when it is clear from the context.

Definition 3 (Disjunctive Rule)**.** *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, $X \rightarrow \vee\mathcal{Y}$ is a disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$ iff X is an association on a sub-set of $\mathcal{D} \setminus \mathcal{D}_S$ and $\mathcal{Y}$ is a set of associations on sub-sets of $\mathcal{D} \setminus \mathcal{D}_S$ such that $\forall Y \in \mathcal{Y}$, $X \sqcup Y$ is an association on $\mathcal{D} \setminus \mathcal{D}_S$.*

For example, with D^3 as the support domain in $\mathcal{R}_E$, $\{d_1, d_3\} \times \{a_2\} \rightarrow \{a_3\}$, $\{d_3, d_4\} \rightarrow (\{a_3\}) \vee (\{a_1, a_2\})$ and $\{d_1\} \times \{a_2\} \rightarrow (\{d_2\} \times \{a_1\}) \vee (\{d_3\} \times \{a_3\})$ are three disjunctive rules on $\{D^1, D^2\}$.

In the binary case (i. e., $n = 2$), the semantics of association rules [1], even when generalized to disjunctive terms [8], [9] are based on the frequency and the confidence measures. *A priori* interesting rules are defined as those whose both measures exceed user-specified thresholds. In the context of n-ary relations, extensions of these measures have been proposed (see, e.g., the natural and the exclusive confidence measures on multidimensional association rules [4]). We have to adapt such measures to our disjunctive rule mining setting.

To illustrate the definitions, given the support domain D^3 in $\mathcal{R}_E$, we use r_1 to denote $\{d_3, d_4\} \rightarrow (\{a_3\}) \vee (\{a_1, a_2\})$ and r_2 to denote $\{d_1\} \times \{a_2\} \rightarrow (\{d_2\} \times \{a_1\}) \vee (\{d_3\} \times \{a_3\})$.

Definition 4 (Association Frequency)**.** *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $X \rightarrow \vee\mathcal{Y}$ be a disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$. $\forall Y \in \mathcal{Y}$, the association frequency of $X \rightarrow Y$ is $f_a(X \rightarrow Y) = \frac{|s(X \sqcup Y)|}{|\times_{D^i \in \mathcal{D}_S} D^i|}$.*

For example, considering r_1, $f_a(\{d_3,d_4\} \rightarrow \{a_3\}) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_3\}))|}{|D^3|} = \frac{|s(\{d_3,d_4\}\times\{a_3\})|}{|D^3|} = \frac{2}{5}$, and $f_a(\{d_3,d_4\} \rightarrow \{a_1,a_2\}) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_1,a_2\}))|}{|D^3|} = \frac{2}{5}$.

Definition 5 (Association Confidence). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $X \rightarrow \vee\mathcal{Y}$ be a disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$. $\forall Y \in \mathcal{Y}$, the association confidence of $X \rightarrow Y$ is $c_a(X \rightarrow Y) = \frac{|s(X \sqcup Y)|}{|s(X)|}$.*

For example, considering again r_1, $c_a(\{d_3,d_4\} \rightarrow \{a_3\}) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_3\}))|}{|s(\{d_3,d_4\})|} = \frac{2}{4}$, and $c_a(\{d_3,d_4\} \rightarrow \{a_1,a_2\}) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_1,a_2\}))|}{|s(\{d_3,d_4\})|} = \frac{2}{4}$. It means that, in rule r_1, the confidence of the conjunction between the body and any association in the head is 0.5.

Definition 6 (Disjunctive Frequency). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $X \rightarrow \vee\mathcal{Y}$ be a disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$. The disjunctive frequency of $X \rightarrow \vee\mathcal{Y}$ is $f_d(X \rightarrow \vee\mathcal{Y}) = \frac{|\cup_{Y\in\mathcal{Y}} s(X \sqcup Y)|}{|\times_{D^i\in\mathcal{D}_S} D^i|}$.*

We have $f_d(r_1) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_3\}))\cup s((\{d_3,d_4\})\sqcup(\{a_1,a_2\}))|}{|D^3|} = \frac{|\{t_3,t_4,t_5\}|}{5} = \frac{3}{5}$, and $f_d(r_2) = \frac{|s((\{d_1\}\times\{a_2\})\sqcup(\{d_2\}\times\{a_1\}))\cup s((\{d_1\}\times\{a_2\})\sqcup(\{d_3\}\times\{a_3\}))|}{|D^3|} = \frac{|\{t_1,t_2,t_3\}|}{5} = \frac{3}{5}$.

Definition 7 (Disjunctive Confidence). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $X \rightarrow \vee\mathcal{Y}$ be a disjunctive rule on $\mathcal{D}\setminus\mathcal{D}_S$. The disjunctive confidence of $X \rightarrow \vee\mathcal{Y}$ is $c_d(X \rightarrow \vee\mathcal{Y}) = \frac{|\cup_{Y\in\mathcal{Y}} s(X \sqcup Y)|}{|s(X)|}$.*

For example, $c_d(r_1) = \frac{|s((\{d_3,d_4\})\sqcup(\{a_3\}))\cup s((\{d_3,d_4\})\sqcup(\{a_1,a_2\}))|}{|s(\{d_3,d_4\})|} = \frac{3}{4}$, and $c_d(r_2) = \frac{|s((\{d_1\}\times\{a_2\})\sqcup(\{d_2\}\times\{a_1\}))\cup s((\{d_1\}\times\{a_2\})\sqcup(\{d_3\}\times\{a_3\}))|}{|s(\{d_1\}\times\{a_2\})|} = \frac{3}{3}$. Rule r_1 indicates that, at a time, if the outer edges from Vertex 3 and Vertex 4 go to the same nodes then they tend to converge to Vertex 3 or Vertices 1 and 2 ($c_d(r_1) = 0.75$). In half of cases, when the edges from Vertex 3 and Vertex 4 converge, they converge to Vertex 3 ($c_a = 0.5$). However, these outer edges can also go to both Vertex 1 and Vertex 2 with $c_a = 0.5$. Rule r_2 means that, at a time, if the graph contains the sub-graph including the edges from Vertex 1 to Vertex 2 then it is sure ($c_d = 1$) that the graph contains the sub-graph including the edges from Vertices 1 and 2 to Vertices 1 and 2 or the sub-graph including the edges from Vertices 1 and 3 to Vertices 2 and 3. Enabling disjunctions within the heads of the rules provides rules that convey more information than conjunctive rules.

The association confidence has a monotonicity property. In Sec. III, it is used to prune the search space where no rule can satisfy a minimal association confidence constraint.

Theorem 2 (Pruning criterion). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let X and X' be associations on sub-sets of $\mathcal{D} \setminus \mathcal{D}_S$ and Y be an association on $\mathcal{D} \setminus \mathcal{D}_S$. We have $X' \sqsubseteq X \sqsubseteq Y \Rightarrow c_a(X' \rightarrow Y \setminus X') \leq c_a(X \rightarrow Y \setminus X)$.*

Definition 8 (Canonical Rule). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, a disjunctive rule $X \rightarrow \vee\mathcal{Y}$ on $\mathcal{D}\setminus\mathcal{D}_S$ is a canonical iff $\forall Y \in \mathcal{Y}$ and $\forall D^i \in \mathcal{D}$, $\pi_{D^i}(X) \cap \pi_{D^i}(Y) = \emptyset$.*

In $\mathcal{R}_E$, let us consider the following rules:

- r_3: $\{d_3\} \times \{a_1\} \rightarrow (\{a_2,a_3\}) \vee (\{d_4\} \times \{a_2\})$ $(f_d : 0.4, c_d : 1)$,
- r_4: $\{d_3\} \times \{a_1,a_2\} \rightarrow (\{a_3\})$ $(f_d : 0.4, c_d : 1)$,
- r_5: $\{d_3\} \times \{a_3\} \rightarrow (\{d_1\} \times \{a_2\}) \vee (\{a_1,a_2\}) \vee (\{a_2\}) \vee (\{d_4\})$ $(f_d : 0.8, c_d : 1)$,
- r_6: $\{d_3\} \times \{a_3\} \rightarrow (\{d_1\} \times \{a_2\}) \vee (\{a_1\}) \vee (\{a_1,a_2\}) \vee (\{a_2\}) \vee (\{d_4\})$ $(f_d : 0.8, c_d : 1)$.

They are all canonical and have their association frequencies, their association confidences, their disjunctive frequencies, their disjunctive confidences respectively exceeding 0.4, 0.5, 0.4 and 0.8. Therefore, they may *individually* satisfy this aspect of interestingness. Nevertheless, *all together*, they provide redundant information. For instance, the premise of r_4 is more informative than that of r_3 (to match the body of r_4, a graph must additionally have the edge from the Vertex 3 to Vertex 2), but the conclusion of r_4 is less informative (it does not tell anything about d_4). In addition, this does not provide r_4 a greater frequency or a greater confidence than r_3. Rule r_4 is therefore said redundant. The conclusion of r_6 has more elements than that in the conclusion of r_5. However, in r_6, $\{a_1\} \sqsubset \{a_1,a_2\}$, $f_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1\}) = f_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1,a_2\}) = 0.4$ and $c_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1\}) = c_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1,a_2\}) = 0.5$. Therefore, the appearance of $\{a_1\}$ in the conclusion of r_6 does not provide new insight. $\{a_1\}$ is thus redundant in r_6. In r_5, although $\{a_2\} \sqsubseteq \{a_1,a_2\}$, $\{a_2\}$ is not redundant since $f_a(\{d_3\} \times \{a_3\} \rightarrow \{a_2\}) = 0.8 > f_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1,a_2\}) = 0.4$ and $c_a(\{d_3\} \times \{a_3\} \rightarrow \{a_2\}) = 1 > c_a(\{d_3\} \times \{a_3\} \rightarrow \{a_1,a_2\}) = 0.5$. Since the end-user would not find any added-value in rules r_4 and r_6, these rules must not be returned. In other terms, we have to revisit the concept of non-redundant rule in our setting.

Definition 9 (Non-Redundant Disjunctive Rule). *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, a disjunctive rule $X \rightarrow \vee\mathcal{Y}$ on $\mathcal{D} \setminus \mathcal{D}_S$ is non-redundant iff it is canonical and satisfies the following constraints:*
(1) $\forall Y \in \mathcal{Y}$, $X \rightarrow Y$ is a non-redundant association rule on $\mathcal{D} \setminus \mathcal{D}_S$. It means that, it is canonical and there is no other canonical association rule $X' \rightarrow Y'$, where $X' \sqcup Y'$ is an association on $\mathcal{D} \setminus \mathcal{D}_S$ such that $((X' \sqcup Y' = X \sqcup Y \wedge X' \sqsubset X) \vee (X' \sqcup Y' \sqsupset X \sqcup Y \wedge X' \sqsubseteq X)) \wedge (f_a(X' \rightarrow Y') \geq f_a(X \rightarrow Y)) \wedge (c_a(X' \rightarrow Y') \geq c_a(X \rightarrow Y))$.
(2) It does not exist any rule which is more general than $X \rightarrow \vee\mathcal{Y}$. It means that there is no set of associations $\mathcal{Z}$ defined on sub-sets of $\mathcal{D} \setminus \mathcal{D}_S$ such that $\mathcal{Y} \subset \mathcal{Z}$, $X \rightarrow \vee\mathcal{Z}$ is canonical and satisfies the constraint (1), and $f_d(X \rightarrow \vee\mathcal{Z}) \geq f_d(X \rightarrow \vee\mathcal{Y}) \wedge c_d(X \rightarrow \vee\mathcal{Z}) \geq c_d(X \rightarrow \vee\mathcal{Y})$.

The first condition shows that the non-redundant association rules on $\mathcal{D} \setminus \mathcal{D}_S$, 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 in $\mathcal{D}_S$ into a unique support dimension $D^{\text{supp}} = \times_{D^i\in\mathcal{D}_S} D^i$. Denoted $\mathcal{R}_A$ until the end of this article, this relation is defined on the domains $\mathcal{D}_A = (\mathcal{D} \setminus \mathcal{D}_S) \cup \{D^{\text{supp}}\}$. 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 $\mathcal{R}_A = \{(e_1, e_2, \ldots, e_{|\mathcal{D}\setminus\mathcal{D}_S|}, (e_{|\mathcal{D}\setminus\mathcal{D}_S|+1}, \ldots, e_n))$

such that $(e_1, e_2, \ldots, e_{|\mathcal{D}\setminus\mathcal{D}_S|}, e_{|\mathcal{D}\setminus\mathcal{D}_S|+1}, \ldots, e_n) \in \mathcal{R}\}$.

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 10 (Closed Set)**.** *Given a relation $\mathcal{R}_A$ on $\mathcal{D}_A$, X is a closed set in $\mathcal{R}_A$ iff $(X \subseteq \mathcal{R}_A) \wedge (\forall D^i \in \mathcal{D}_A, \forall e \in D^i \setminus \pi_{D^i}(X), X \sqcup \{e\} \not\subseteq \mathcal{R})$.*

Considering $\mathcal{R}_E$, since $\mathcal{D}_S = \{D^3\}$ contains only one domain, $\mathcal{R}_A = \mathcal{R}_E$. Association $\{d_1, d_3\} \times \{a_2, a_3\} \times \{t_2, t_3\}$ is a closed set. $\{d_1, d_3\} \times \{a_2, a_3\} \times \{t_1, t_2, t_3\}$ is not a closed set because $(d_1, a_3, t_1) \notin \mathcal{R}_A$. $\{d_1, d_3\} \times \{a_2\} \times \{t_2, t_3\}$ is not a closed set because it can be enlarged with a_3.

There is a strong link between closed sets and non-redundant association rules as stated in Theorem 3.

Theorem 3 (Closed Sets and Non-Redundant Association Rules)**.** *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $X \rightarrow Y$ be a canonical association rule such that $X \sqcup Y$ is an association on $\mathcal{D} \setminus \mathcal{D}_S$. $X \rightarrow Y$ is a non-redundant association rule on $\mathcal{D} \setminus \mathcal{D}_S$ iff $(X \sqcup Y \sqcup s(X \sqcup Y))$ is a closed set in $\mathcal{R}_A$ and $\forall X' \sqsubset X$, $c_a(X' \rightarrow (Y \sqcup X) \setminus X') < c_a(X \rightarrow Y)$.*

Non-redundant association rules are key elements to obtain non-redundant disjunctive rules.

Theorem 4 (Non-Redundant Association Rules and Non-Redundant Disjunctive Rules)**.** *$\forall \mathcal{D}_S \subseteq \mathcal{D}$, let $\mathcal{P}$ the set of all non-redundant association rules on $\mathcal{D} \setminus \mathcal{D}_S$, $X \rightarrow \vee\mathcal{Y}$ is a non redundant disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$ iff $\mathcal{Y} = \cup_{X \rightarrow Y \in \mathcal{P}} Y$.*

III. Discovering Non-Redundant Rules

Given an n-ary relation $\mathcal{R} \subseteq \times_{D^i \in \mathcal{D}} D^i$, $\mathcal{D}_S \subset \mathcal{D}$, every interesting and non-redundant disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$ has to be enumerated. Such rules satisfy the user-specified thresholds: the association frequency threshold μ_a, the association confidence threshold β_a, the disjunctive frequency threshold μ_d and the disjunctive confidence threshold β_d. In other terms, our algorithm CIDRE computes:

$$\{X \rightarrow \vee\mathcal{Y} \text{ on } \mathcal{D} \setminus \mathcal{D}_S \mid \begin{cases} X \rightarrow \vee\mathcal{Y} \text{ is non-redundant} \\ \forall Y \in \mathcal{Y}, f_a(X \rightarrow Y) \geq \mu_a \\ \forall Y \in \mathcal{Y}, c_a(X \rightarrow Y) \geq \beta_a \\ f_d(X \rightarrow \vee\mathcal{Y}) \geq \mu_d \\ c_d(X \rightarrow \vee\mathcal{Y}) \geq \beta_d \end{cases}.$$

CIDRE is divided into four successive steps: (1) it constructs the relation $\mathcal{R}_A$ defined in Section II, this step is trivial; (2) it extracts the *frequent* closed sets in $\mathcal{R}_A$; (3) it derives from these closed sets the non-redundant association rules satisfying the user-defined thresholds; (4) it computes the non-redundant disjunctive rules whose disjunctive frequencies and disjunctive confidences hold for the user-defined thresholds μ_d and β_d.

A. Extracting Closed Sets under Constraints

Theorem 3 states the link between the non-redundant association rules and the closed sets in $\mathcal{R}_A$ but, to be *a priori* interesting, the association rules must satisfy constraints. Two approaches have been proposed to exhaustively list the closed sets in *ternary* relations, namely CUBEMINER [10] and TRIAS [11]. A third algorithm, DATA-PEELER [7] 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. To guarantee all association rules exceed the user-defined association frequency threshold, in $\mathcal{R}_A$, we only discover the frequent closed sets which are gather at least a proportion μ_a of the elements in D^{supp}. It means that every extracted closed set C must satisfy the constraint $\mathcal{C}_{freq}(C) \equiv \frac{|\pi_{D^{\text{supp}}}(C)|}{|D^{\text{supp}}|} \geq \mu_a$. DATA-PEELER can handle it directly on the closed sets.

It may also be interesting to specify minimal numbers of elements in the exploited dimensions (i. e., the dimensions in $\mathcal{D} \setminus \mathcal{D}_S$). In this case, every extracted closed set C must satisfy $\mathcal{C}_{(\alpha^i)_{i=1..|\mathcal{D}\setminus\mathcal{D}_S|}-min-sizes}(C) \equiv \forall D^i \in \mathcal{D} \setminus \mathcal{D}_S, |\pi_{D^i}(C)| \geq \alpha^i$.

From a closed set C, ASSOCIATION_RULES (Alg. 1) derives non-redundant association rules on $\mathcal{D} \setminus \mathcal{D}_S$ that involve all the elements in $\cup_{D^i \in \mathcal{D}\setminus\mathcal{D}_S} \pi_{D^i}(C)$.

B. Deriving Non-Redundant Rules from Closed Sets

Algorithm 1: ASSOCIATION_RULES.

```
Data: (B, H), i. e., a body and a head
forall e ≻ max_≺(H) do
    if c_a(B \ {e} → H ⊔ {e}) ≥ β_a then
        forall f ∈ ∪_{D^i ∈ D\D_S} π_{D^i}(B \ {e}) do
            if c_a((B \ {e}) \ {f} → H ⊔ {e} ⊔ {f}) =
            c_a(B \ {e} → H ⊔ {e}) then
                goto skip
        output B \ {e} → H ⊔ {e}
        skip: RULES(B \ {e}, H ⊔ {e})
```

ASSOCIATION_RULES derives *a priori* interesting and non-redundant association 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 $\cup_{D^i \in \mathcal{D}\setminus\mathcal{D}_S} \pi_{D^i}(A)$ 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 [1]. Nevertheless, ASSOCIATION_RULES traverses it in a depth-first way. 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 $\cup_{D^i \in \mathcal{D}\setminus\mathcal{D}_S} \pi_{D^i}(A)$. 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). The pruning criterion is the minimal association confidence constraint. According to Theorem 2, this pruning is safe, i. e., no association rule, with an association confidence

higher than β_a, is missed. On the opposite, the non-redundancy constraints cannot give rise to search space pruning. That is why it is checked after the constraint on the minimal association confidence. If it is satisfied then the rule is output. To enforce the non-redundancy, Theorem 3 indicates that, beside the necessity to process a closed set, ASSOCIATION_RULES must check the association confidences of the more general association rules sharing the same elements. If such a rule has the same association confidence then the current rule is redundant.

C. Computing Non-Redundant Disjunctive Rules

$\mathcal{P}$ denotes the set of all non-redundant association rules on $\mathcal{D} \setminus \mathcal{D}_S$ which are extracted in Sec. III-B. According to Theorem 4, we construct non-redundant disjunctive rules of the form $X \rightarrow \vee\mathcal{Y}$ where $\mathcal{Y} = \cup_{X \rightarrow Y \in \mathcal{P}} Y$. Alg. 2 only outputs the non-redundant disjunctive rules whose disjunctive frequencies and disjunctive confidences exceed the user-defined thresholds. CIDRE (see Alg. 2) successively (1) constructs $\mathcal{R}_A$, (2) extracts the frequent closed sets in it, (3) derives, from each of these closed sets, the *a priori* interesting and non-redundant association rules and (4) computes the interesting and non-redundant disjunctive rules.

Algorithm 2: CIDRE.

Input: A relation $\mathcal{R}$ on $\mathcal{D}$, $\mathcal{D}_S \subsetneq \mathcal{D}$, and $(\mu_a, \beta_a, \mu_d, \beta_d) \in [0,1]^4$
Output: Every interesting and non-redundant disjunctive rule on $\mathcal{D} \setminus \mathcal{D}_S$

$D^{\text{supp}} \leftarrow \times_{D^i \in \mathcal{D}_S} D^i$
$(\mathcal{D}_A, \mathcal{R}_A) \leftarrow ((\mathcal{D} \setminus \mathcal{D}_S) \cup D^{\text{supp}}, \emptyset)$
forall $(e_1, \ldots, e_{|\mathcal{D}\setminus\mathcal{D}_S|}, e_{|\mathcal{D}\setminus\mathcal{D}_S|+1}, \ldots, e_n) \in \mathcal{R}$ **do**
  $\mathcal{R}_A \leftarrow \mathcal{R}_A \cup (e_1, \ldots, e_{|\mathcal{D}\setminus\mathcal{D}_S|}, (e_{|\mathcal{D}\setminus\mathcal{D}_S|+1}, \ldots, e_n))$
$\mathcal{C} \leftarrow$ DATA-PEELER$(\emptyset, \times_{D^i \in \mathcal{D}_A} D^i)$
$\mathcal{P} \leftarrow \emptyset$
forall $C \in \mathcal{C}$ **do**
  $\mathcal{P} \leftarrow \mathcal{P} \cup$ ASSOCIATION_RULES$(C \setminus \pi_{D^{supp}}(C), \emptyset)$
forall $X \rightarrow Y \in \mathcal{P}$ **do**
  $\mathcal{Y} \leftarrow Y$
  forall $X \rightarrow Y' \in \mathcal{P}$ *such that* $Y' \neq Y$ **do**
    $\mathcal{Y} \leftarrow \mathcal{Y} \cup Y'$
    delete $X \rightarrow Y'$ from $\mathcal{P}$
  if $(f_d(X \rightarrow \vee\mathcal{Y}) \geq \mu_d) \wedge (c_d(X \rightarrow \vee\mathcal{Y}) \geq \beta_d)$ **then**
    output $X \rightarrow \vee\mathcal{Y}$
  delete $X \rightarrow Y$ from $\mathcal{P}$

IV. 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. CIDRE 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 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 1pm to 2pm) 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.

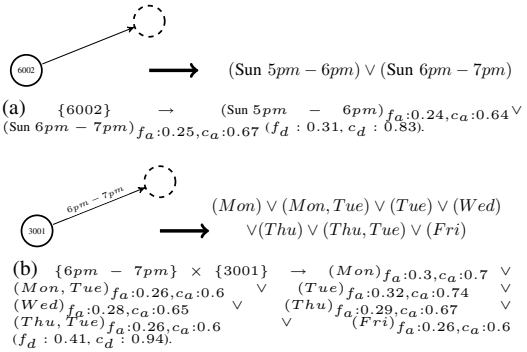


Figure 3: Example of rules on $\{Departure, Day, Hour\}$.

The temporal dimension(s) of such a dynamic network either appear in the rules (i. e., in $\mathcal{D} \setminus \mathcal{D}_S$) or can be used to compute the frequencies and the confidences of the rules (i. e., in $\mathcal{D}_S$). Different templates depend on different mining motivations. Let us now discuss a couple of examples.

To study departure time periods of stations, we discover rules on the dimensions *Departure*, *Day* and *Hour*. As a consequence, the support domain is *Arrival* which contains 327 stations. With $\mu_a = \mu_d = 0.2$, $\beta_a = 0.6$ and $\beta_d = 0.8$ CIDRE extracts 33 rules. They indicate that preferred departure times are different from one station to another. Fig. 3 reports two of them. The rule in Fig. 3a means that the departures from Station 6002, with a high enough association confidence ($c_a \geq 0.6$), almost occur between 5pm and 7pm on Sundays ($c_d = 0.83$). Here, arrival stations of the departures from Station 6002 on Sundays between 5pm and 6pm can be different from those between 6pm and 7pm. Therefore, this rule cannot be extracted given previous approaches in [4], [5]. The rule in Fig. 3b indicates that, with a high enough association confidence (≥ 0.6), the rides from Station 3001 between 6pm and 7pm only occur during the working days. Also because, arrival stations of the rides from Station 3001 between 6pm and 7pm can be different between days, this rule cannot be returned by known techniques.

On graph evolution: if a sub-network is frequent, then

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

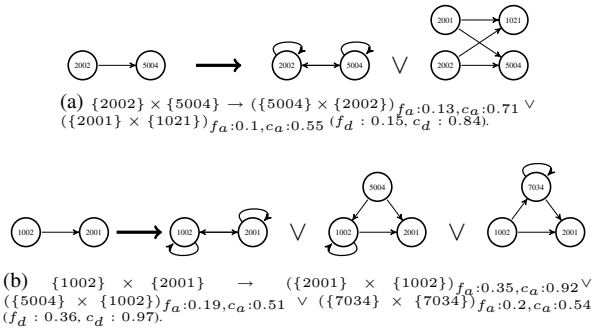
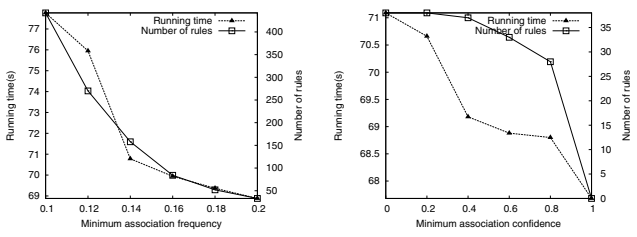


Figure 4: Example of rules of the form "sub-network" → "larger sub-networks".



(a) Pruning w.r.t. μ_a. (b) Pruning w.r.t. β_a.

Figure 5: Effectiveness of CIDRE.

to which sub-networks it can be enlargeable with strong enough confidences. To study such patterns, a rule has to involve $Departure$ and $Arrival$ stations, i. e., $\mathcal{D} \setminus \mathcal{D}_S = \{Departure, Arrival\}$. As a result, the support domain is the Cartesian product of the 7 days and the 24 hours. Additional contraints, the constraint $\mathcal{C}_{(2,2)\text{-min-sizes}}$ (see Sect. III) is additionally enforced so that every rule must involve at least two departure stations and two arrival stations. Moreover we force 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. 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 high enough confidences) enlarged (with the stations at the heads). With $\mu_a = \mu_d = 0.1$, $\beta_a = 0.5$ and $\beta_d = 0.8$, 228 rules are discovered. The larger sub-networks can contain more nodes or only more edges. Some of them are reported in Fig. 4. These rules explicit diverse mechanisms like auto-regulation and convergence. They are much more informative than multidimensional association rules and can be used to anticipate the effect of a typical breakdown: a station that can only emit (resp. receive) bicycles. If such a station at the body of a rule is fail, then the other stations in the rule may be overloaded (resp. suffer a shortage).

Let us finally provide a performance study when mining interesting and non-redundant disjunctive rules in $\mathcal{R}_{\texttt{Vélo'v}}$ with $\mathcal{D}_S = \{Arrival\}$. When the minimal association frequency threshold increases, CIDRE prunes large areas of the search space where no association is frequent, as consequently both the number of frequent rules and the running time decrease. Fig. 5a was obtained with $\mu_d = \mu_a$, $\beta_a = 0.6$ and $\beta_d = 0.8$. 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 ASSOCIATION_RULES algorithm does not traverse the enumeration sub-trees empty of confident rules (w.r.t. β_a). That is why both the number of rules and the time it takes to extract them decrease when the minimum association confidence threshold increases. Experiments in Fig. 5b are performed with $\mu_a = \mu_d = 0.2$, $\beta_d = 0.8$ and β_a varying between 0 and 1. On the contrary, the search space cannot be pruned thanks to the thresholds of disjunctive frequency and disjunctive confidence. Indeed, CIDRE must consider every association rule when computing disjunctive ones.

V. CONCLUSION

We have studied the problem of describing relational dynamic graphs via disjunctive association rules that can involve subsets of any dimension (including temporal dimensions). We have proposed a semantics for such rules and we introduced CIDRE, an efficient solution for computing non redundant disjunctive rules. Experiments on a real-world dynamic network demonstrated the interest of our proposal.

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

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD*, vol. 22. ACM, 1993, pp. 207–216.

[2] M. Kamber, J. Han, and J. Y. Chiang, "Metarule-guided mining of multi-dimensional association rules using data cubes," in *KDD*. AAAI, 1997, pp. 207–210.

[3] H. C. Tjioe and D. Taniar, "Mining association rules in data warehouses," *Int. Journal of Data Warehousing and Mining*, vol. 1, no. 3, pp. 28–62, 2005.

[4] K.-N. T. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut, "Multi-dimensional association rules in boolean tensors," in *SDM*. SIAM / Omnipress, 2011, pp. 570–581.

[5] ——, "Mining descriptive rules in dynamic graphs," *Intelligent Data Analysis*, 2011, special Issue on Dynamic Network Analysis, 30 p. In Press.

[6] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis, "Mining graph evolution rules," in *ECML/PKDD*, vol. 5781. Springer, 2009, pp. 115–130.

[7] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut, "Closed patterns meet n-ary relations," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1–36, 2009.

[8] A. A. Nanavati, K. P. Chitrapura, S. Joshi, and R. Krishnapuram, "Mining generalised disjunctive association rules," in *CIKM*. ACM, 2001, pp. 482–489.

[9] M.-L. Antonie and O. R. Zaïane, "Mining positive and negative association rules: An approach for confined rules," in *ECML/PKDD*, vol. 3202. Springer, 2004, pp. 27–38.

[10] L. Ji, K.-L. Tan, and A. K. H. Tung, "Mining frequent closed cubes in 3D data sets," in *VLDB*. VLDB Endowment, 2006, pp. 811–822.

[11] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme, "TRIAS–an algorithm for mining iceberg tri-lattices," in *ICDM*. IEEE Computer Society, 2006, pp. 907–911.