Mining Multi-Relational Gradual Patterns
Nhathai Phan, Dino Ienco, Donato Malerba, Pascal Poncelet, Maguelon Teisseire

To cite this version:
Nhathai Phan, Dino Ienco, Donato Malerba, Pascal Poncelet, Maguelon Teisseire. Mining Multi-Relational Gradual Patterns. SDM: SIAM Data Mining, Apr 2015, Vancouver, Canada. pp.9, 10.1137/1.9781611974010.95 . lirmm-01239187

HAL Id: lirmm-01239187
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01239187
Submitted on 7 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Mining Multi-Relational Gradual Patterns

NhatHai Phan∗ Dino Ienco† Donato Malerba‡ Pascal Poncelet§
Maguelonne Teisseire¶

Abstract
Gradual patterns highlight covariations of attributes of the form “The more/less X, the more/less Y”. Their usefulness in several applications has recently stimulated the synthesis of several algorithms for their automated discovery from large datasets. However, existing techniques require all the interesting data to be in a single database relation or table. This paper extends the notion of gradual pattern to the case in which the co-variations are possibly expressed between attributes of different database relations. The interestingness measure for this class of “relational gradual patterns” is defined on the basis of both Kendall’s τ and gradual supports. Moreover, this paper proposes two algorithms, named τRGP Miner and gRGP Miner, for the discovery of relational gradual rules. Three pruning strategies to reduce the search space are proposed. The efficiency of the algorithms is empirically validated, and the usefulness of relational gradual patterns is proved on some real-world databases.

1 Introduction
Nowadays, most of information systems are based on the relational database technology. The logical models of the data are sets of relations or tables possibly linked by foreign key constraints. This contrasts with the usual practice in Data Mining of organizing data in a single relation when they are analyzed. Relational data mining approaches [6] are characterized by both their direct applicability to “multi-relational data” (MRD) and their capability of looking for patterns which involve multiple database relations.

Most of the studies on relational data mining focus on relational patterns at the tuple level, i.e., they express relationships between tuples of different database relations. Relational association rules [4], relational naïve Bayesian classifiers [7], relational regression models [1] and relational subgroups [17], all express patterns as either SQL queries or first-order logic clauses with constraints between tuples or facts. Similarly, the probabilistic relational models [8] define a distribution over a set of instances of a schema, and consider the structure at the level of attribute values.

In this paper, we focus on patterns expressing the relational structure at the attribute level. We consider relational extensions of the class of gradual dependencies which express covariations of attributes of the form “The more/less X, the more/less Y” [10]. This class of patterns has a wide range of applications [3, 5, 18, 19, 20]. For instance, with reference to the financial database in Figure 1, the gradual pattern “the higher the average salary in a district, the bigger the deals made by inhabitants living in the district” could be useful for business planning. While the pattern “the smaller the district, the longer the duration of the loans” provides financial promoters with some useful insights. These two examples show gradual dependencies between attributes from different relations (i.e., the average salary, the size of districts, the deal, and the loan). Hence they could be discovered only in a MRD setting.

To discover this kind of patterns, we first introduce the concept of multi-relational gradual pattern (multi-rgp), and its associated support measures based on Kendall’s τ [3] and gradual support [5]. Then we propose an algorithm, named RGP-Miner, to discover multi-rgps directly from MRDs. Mining the complete set of multi-rgps is a non-trivial task since the size of the search space is exponential in the number of numerical attributes of the multi-relational database. To tame computational complexity, we design three efficient rules named Complementary Pruning, Apriori Pruning, and Backward Pruning to shrink the search space and prevent unnecessary computations. Experiments conducted on real datasets demonstrate the pattern meaning, effectiveness, and efficiency of our proposed approaches.

2 Problem Statement
In this section, we formalize the data model, the notion of relational gradual patterns, and its examples.

∗University of Oregon. haiphan@cs.uoregon.edu
†Irstea Montpellier, France. dino.ienco@teledetection.fr
‡University of Bari, Italy. donato.malerba@uniba.it
§University of Montpellier 2, France. pascal.poncelet@lirmm.fr
¶Irstea Montpellier, France. teisseire@teledetection.fr
2.1 Multi-Relational Data. In this work, we assume that a database \( db \) consists of a set of tables, \( db = \{ T^1, \ldots, T^n \} \), each of which has a schema \( S(T^i) = (PK^i, FK^i, A^i) \) consisting of a primary key \( PK^i \), possibly some foreign keys (set \( FK^i \)), and at least one attribute (set \( A^i \)). For instance, with reference to Figure 1, the Loan table has a key loan-ID, a foreign key account-ID and attributes date, amount, duration, etc.

Foreign keys define the only possible joins between two tables. Without loss of generality, we assume that any pair of tables has at most one foreign key linking them. Indeed, databases can always be losslessly recoded such that this assumption holds.

Moreover, we assume that all attributes have a numerical domain, since we are not interested in finding patterns among categorical attributes. The numerical domain of the attribute \( A^i_j \) is denoted as \( Dom(A^i_j) \). Similarly, the domain of table \( T^i \) is denoted by \( Dom(T^i) \), and corresponds to the Cartesian product of all domains involved, i.e.,

\[
Dom(T^i) = \prod_{FK^i_j \in FK^i} Dom(FK^i_j) \prod_{A^i_j \in A^i} Dom(A^i_j)
\]

Each table is also associated with a set of tuples which also called the table extension. Henceforth, the distinction between the table and its extension is blurred, and we use the notation \( t \in T^i \) to indicate that a tuple \( t \) is in extension of table \( T^i \).

The database as whole should satisfy referential integrity, i.e., foreign-key values in a tuple refer to existing tuples in the table for which this foreign key refers to. In formal, if \( t \in T^i \) and \( S(T^i) = (PK^i, FK^i, A^i) \), then for each \( FK^i_j \in FK^i \) whose domain is \( Dom(PK^i) \), the following condition must hold: there exists a tuple \( t' \in T^i \) such that \( \pi_{FK^i_j}(t) = \pi_{PK^i}(t') \), where \( \pi \) is the usual projection operator of relational algebra.

2.2 Multi-Relational Gradual Pattern. In the case of a (single) table, an itemset boils down to a selection [11] [15]. However, in relational gradual pattern context, a selection on an attribute cannot reflect the attribute graduality. The main reason is that the value of a specific attribute \( A^i_j \) (resp. set of attributes) must change following a regular rule (e.g., “\( A^i_j \) is increasing or decreasing”) over the tuples. Therefore, a common selection in which attribute values are selected by some simple operations (e.g., \( >, <, = \)) cannot present the attribute graduality in our context.

To address the issue, we propose the definition of relational gradual patterns in which the graduality is covered. Let us denote the increase and decrease of an attribute \( A^i_j \) as \( A^i_j \) and \( A^i_j \). Let \( E^i \) be the set of all single gradual attributes which are generated by \( A^i \) from table \( T^i \), \( E^i \) can be defined as follows:

\[
E^i = \bigcup_{j=1}^{\mid A^i \mid} \{ A^i_j \} \cup \{ A^i_j \}
\]

In a single table \( T^i \), a relational gradual pattern \( q^i \) is a combination of gradual attributes in \( E^i \). The set of all potential gradual patterns \( q^i \) are called \( Q^i \). In addition, “proper” relational gradual patterns could be generated from gradual attributes which are from different tables. If the gradual attributes in a pattern are from a single table \( T^i \), then the pattern is called a mono-rgp. If the gradual attributes are from two tables \( T^i \) and \( T^j \) then the pattern is called a bi-rgp. Finally, a multi-rgp can contain gradual attributes from multiple tables in MRDs. The definition of relational gradual patterns can be defined as follows:

**Definition 2.1. (Relational Gradual Pattern).** Let \( db = \{ T^1, \ldots, T^n \} \) be a database for which each table \( T^i \) has a schema \( S(T^i) = (PK^i, FK^i, A^i) \).

1. Mono-relational gradual pattern (mono-rgp):
   \( p = q^i \) where \( q^i \in Q^i \)

2. Bi-relational gradual pattern (bi-rgp) between \( T^i, T^j \):
   a. if \( T^i.T^j \) is a 1:N relation then
      \( p = q^i q^j \) where \( q^i \in Q^i, q^j \in Q^j \) and \( PK^i \subseteq FK^j \)
   b. if \( T^i.T^j \) is a M:N relation then
      \( p = q^i q^j \) where \( q^i \in Q^i, q^j \in Q^j \) and \( \exists T^c \in db \)
         s.t. \( (PK^i \subseteq PK^c, PK^j \subseteq PK^c) \) or \( (PK^i \subseteq FK^c, PK^j \subseteq FK^c) \)

3. Multi-relational gradual pattern (multi-rgp). \( p = q^i q^j \ldots q^m \) where \( \forall j \in \{ i, \ldots, m-1 \} : q^j \in Q^j \)
   and \( q^j q^{j+1} \) is a bi-rgp

In the Definition 2.1, we define the patterns that can be extracted in a single relation, in two relations, and in multi-relations. The combination of two monorgps from two relations become a bi-rgp. The relations from \( T^i \) to \( T^j \) in 2.a and 2.b respectively are 1:N and M:N relationships. Note that, in 2.b, the table \( T^c \) is a connection relation between \( T^i \) and \( T^j \). Finally, a multi-rgp is a generalization of the bi-rgp so that every consecutive pairs of mono-rgps is a bi-rgp.

For instance, in Figures 1 and 2, potential interest-pairs are from a single table \( T^i \), then the pattern is called a mono-rgp. If the gradual attributes are from two tables \( T^i \) and \( T^j \) then the pattern is called a bi-rgp. Finally, a multi-rgp can contain gradual attributes from multiple tables in MRDs. The definition of relational gradual patterns can be defined as follows:

**Definition 2.1. (Relational Gradual Pattern).** Let \( db = \{ T^1, \ldots, T^n \} \) be a database for which each table \( T^i \) has a schema \( S(T^i) = (PK^i, FK^i, A^i) \).

1. Mono-relational gradual pattern (mono-rgp):
   \( p = q^i \) where \( q^i \in Q^i \)

2. Bi-relational gradual pattern (bi-rgp) between \( T^i, T^j \):
   a. if \( T^i.T^j \) is a 1:N relation then
      \( p = q^i q^j \) where \( q^i \in Q^i, q^j \in Q^j \) and \( PK^i \subseteq FK^j \)
   b. if \( T^i.T^j \) is a M:N relation then
      \( p = q^i q^j \) where \( q^i \in Q^i, q^j \in Q^j \) and \( \exists T^c \in db \)
         s.t. \( (PK^i \subseteq PK^c, PK^j \subseteq PK^c) \) or \( (PK^i \subseteq FK^c, PK^j \subseteq FK^c) \)

3. Multi-relational gradual pattern (multi-rgp). \( p = q^i q^j \ldots q^m \) where \( \forall j \in \{ i, \ldots, m-1 \} : q^j \in Q^j \)
   and \( q^j q^{j+1} \) is a bi-rgp

In the Definition 2.1, we define the patterns that can be extracted in a single relation, in two relations, and in multi-relations. The combination of two monorgps from two relations become a bi-rgp. The relations from \( T^i \) to \( T^j \) in 2.a and 2.b respectively are 1:N and M:N relationships. Note that, in 2.b, the table \( T^c \) is a connection relation between \( T^i \) and \( T^j \). Finally, a multi-rgp is a generalization of the bi-rgp so that every consecutive pairs of mono-rgps is a bi-rgp.

For instance, in Figures 1 and 2, potential interesting multi-rgps are as follows:

**Bi-rgp from a 1:N relation.** In table Loan, we have a mono-rgp such as \( q^2 = \{ Amount >, Duration > \} \). Additionally, we can have a bi-rgp by connecting table Loan to table Account over the foreign key account-ID (i.e. Account.PK = Loan.FK). A potential bi-rgp is \( p = q^1 q^2 = \{ Frequency >, Amount >, Duration > \} \) where
tuple pair can be defined as follows:

\[ \text{gradual in respect to } q \]

The occurrences of relational gradual patterns can thus contain sets of gradual tuple pairs.

**Definition 3.1. (Gradual Tuple Pair).** Given a mono-rgp \( p = \{A_1^{i_1}, \ldots, A_n^{i_n}\} \in Q \) with \( i_1, \ldots, i_n \in \{<,>,=\} \). A pair of distinct tuples \((t, t')\) s.t. \( t, t' \in T_i \) is gradual in respect to \( q^i \) if:

\[ \forall A_i^{i_i} \in q^i : \pi_{A_i}(t) \neq \pi_{A_i}(t') \]

Given a multi-rgp \( p = q^1 q^2 \ldots q^m \), a pair of tuples \((t, t')\) s.t. \( t, t' \in (T^1 \bowtie T^2 \bowtie \ldots \bowtie T^m) \), is gradual in respect to \( p \) if:

\[ \forall q^i_j \in p : (\pi_{T^i}(t), \pi_{T^i}(t')) \text{ respects } q^j \]

**Definition 3.2. (Pattern Occurrence).** Let \( db = \{T^1, \ldots, T^n\} \) be a database. Given a pattern \( p = q^1 q^2 \ldots q^m \), the occurrence of \( p \), denoted \( \text{occ}(p) \), is a set of gradual tuple pairs \((t, t')\) in respect to \( p \). Note that \( t, t' \in (T^1 \bowtie T^2 \bowtie \ldots \bowtie T^m) \).

It is illustrative to consider what the domain of such patterns is. That is, “what does an instance look like?” In essence, the definition of these domains follows the structure of the pattern occurrence.

**Definition 3.3. (Pattern Occurrence Domain).** Let \( db = \{T^1, \ldots, T^n\} \) be a database. Given a multi-rgp \( p = q^1 q^2 \ldots q^m \), the domain of \( p \)’s occurrences, denoted \( \text{DomOcc}(p) \), is given by

\[ \text{DomOcc}(p) = (T^1 \bowtie T^2 \bowtie \ldots \bowtie T^m)^2 \]

For brevity, we only report the primary key and foreign keys in a pattern occurrence. For instance, given a mono-rgp \( q^1 = \{\text{Frequency}>\} \in Q_{\text{Account}} \) (Figure 3). One of the occurrences of \( q^1 \) is \((\text{account}_{1D}, \text{account}_{1D})\) = \((10, 11)\). The occurrences of \( q^1 \), denoted \( \text{occ}(\text{Frequency}) \), are illustrated in Figure 3. In similar, we have the occurrences of \( q^2 = \{\text{Amount}>\} \in Q_{\text{Loan}} \), denoted \( \text{occ}(\text{Amount}) \). If we have a bi-rgp \( p = q^1 q^2 \) then an occurrence of \( p \) is \((t_1, t_2) = (10, 11), (t_1', t_2') = (30, 32)\) where \( t_1, t_2 = (10, 11) \in \text{occ}(q^1) \), \( t_1', t_2' = (30, 32) \in \text{occ}(q^2) \) and \( \pi_{\text{account}_{1D}}(t_1) = \pi_{\text{account}_{1D}}(t_1') = 10 \), \( \pi_{\text{account}_{1D}}(t_2) = \pi_{\text{account}_{1D}}(t_2') = 11 \). Obviously, we can compute the occurrences of \( p \) which are illustrated as the links between \( \text{occ}(\text{Frequency}) \) and \( \text{occ}(\text{Amount}) \).
Kendall’s will be used to define the support measure. Let us first propose a set of operations which for multi-rgps which contain a series of bi-rgps. (i.e., 1:N or M:N). It becomes even more complicated given a multi-rgp. Given a bi-rgp, the number of "all possible cases" depends on the types of the relations (i.e., 1:N or M:N). It becomes even more complicated for multi-rgps which contain a series of bi-rgps.

To breakthrough the challenge, we propose a novel Kendall’s τ-based support to measure the graduality of multi-rgps. Let us first propose a set of operations which will be used to define the support measure.

- **distinct(PKi, Tj)**
  Given two tables Ti and Tj s.t. PKi ⊆ FKj: 1) distinct(PKi, Tj) returns a set of “distinct” values pk ∈ PKi s.t. ∃t’ ∈ Tj : pk ∈ πFKj(t’).
  - **count(pk, Tj)** returns the number of tuples t’ ∈ Tj s.t. pk ∈ πFKj(t’).
  - **distinct(PKmj, p)**
    Given a multi-rgp p = q1...qm, distinct(PKmj, p) is a set of “distinct” PKmj values which are involved in the occurrence domain of p.

In a single table context, |all possible cases| of a mono-rgp p = qi actually is the number of randomly chosen pairs of tuples (t, t’) s.t. t, t’ ∈ Ti and (t, t’) possibly respects the pattern p. Indeed, the |all possible cases| = \( \binom{|Ti|}{2} \) and not |Ti|(|Ti| − 1) since it is not possible that (t, t’) and (t’, t) support p at the same time. The support of mono-rgps can be defined as follows:

### Definition 4.1. (Kendall’s τ support of mono-rgps)
Given a MRD db = \{ T1, ..., Tn \}, and a mono-rgp p = qi. The pattern support can be defined as follows:

\[
\sigma_\tau (p) = \frac{|occ(p)|}{\binom{|T1|}{2}}
\]

### Definition 4.2. (Kendall’s τ support of bi-rgps for 1:N relations)
Given a MRD db = \{ T1, ..., Tn \}, and a bi-rgp p = qi qj s.t. T1.Tj is 1:N. The pattern support can be defined as follows:

\[
\sigma_\tau (p) = \frac{|occ(p)|}{\binom{|distinct(PK1,Tj)|}{2}} = \frac{|occ(p)|}{\sum_{pk \in distinct(PK1,Tj)} \binom{count(pk,Tj)}{2}}
\]

Regarding to bi-rgp p = qi qj with the relation Ti : Tj is 1:N, the |all possible cases| intuitively is \( \binom{|distinct(PK1,Tj)|}{2} \). This is because we need to join two tables via PK1. In addition, only the values of PK1 which exist in table Tj can support p by respecting the mono-rgp qj. The number of these primary key values is |distinct(PK1, Tj)|. However, it is different from mono-rgps, there are some cases needed to be eliminated from \( \binom{|distinct(PK1,Tj)|}{2} \). In mono-rgps, the values of PK1 are distinct meanwhile they are not distinct after joining two tables Ti and Tj in bi-rgps. If two tuples in \( T_i \bowtie T_j \) have the same value of PK1 then the combinations of them cannot respect the mono-rgp qj in p. Therefore, we need to eliminate these combinations from the |all possible cases|.

All of these combinations can be computed as \( \sum_{pk \in distinct(PK1,Tj)} \binom{count(pk,Tj)}{2} \).
The support of multi-rgps can be considered as a generalization of bi-rgps. Given a multi-rgp $p = q^1 \ldots q^m$, $p$ can be rewrote as $p = (q^i \ldots q^{i−1})q^i$. In fact, $(q^1 \ldots q^{m−1})$ can be considered as a mono-rgp whose domain is $T^1 \times \ldots \times T^{m−1}$ and $q^m$ is another mono-rgp. So, the [all possible cases] clearly depends on the relationship between two tables $T^{m−1}$ and $T^m$. As a result, the Definitions 4.2, 4.3 can be utilized to compute the support for multi-rgps. What we need is to identify values of the primary key $PK^{m−1}$ which exist in the $occ(q^1 \ldots q^{m−1})$. The set of these values are denoted as $PK_T = dist(PK^{m−1}, p')$ where $p' = q^1 \ldots q^{m−1}$. The support of multi-rgps can be defined as follows:

**Definition 4.4.** (Kendall’s $τ$ support of multi-rgps). Let $db = \{T^1, \ldots, T^n\}$ be a database, the support of multi-rgps $p = q^1 \ldots q^{m−1}q^m$ can be defined as follows:

1) if $T^{m−1} \times T^m$ is $1:N$

$$\sigma_p = \frac{|occ(p)|}{\left(\binom{|dist(PK^+, T^m)|}{2} - \sum_{pk \in dist(PK^+, T^m)} \binom{count(pk, T^m)}{2}\right)}$$

2) if $T^{m−1} \times T^m$ is $M:N$ and $T^c$ is the connection table

$$\sigma_p = \frac{|occ(p)|}{\left(\binom{|dist(PK^+, T^m)|}{2} \times \binom{|dist(PK^+, T^c)|}{2}\right)}$$

In practice, to reduce the memory consumption we index only the tuples exist in the occurrences of $(q^1 \ldots q^{n−1})$. The following theorem gives the expected $σ_r$, denoted $E[σ_r]$, in the case of statistically independent attributes. This expected support can be used as a reference point to assess the quality of a multi-rgp.

**Theorem 4.1.** ($E[σ_r]$). Given a set of statistically independent gradual attributes $s = \{A_1, A_2, \ldots, A_n\}$, $P_s = \{p_1, \ldots, p_n\}$ is the set of all possible independent multi-rgps generated from $s$. The expected support of a pattern $p \in P_s$ is $E[σ_r](p) = \frac{1}{(2^n−1+3^n−1+1)}$.

The proof of Theorem 4.1 is available at [https://sites.google.com/site/ihaiphan/](https://sites.google.com/site/ihaiphan/). Until now, we have proposed the definition of Kendall’s τ support of multi-rgps. In [5], Di-Jorio et. al. propose a gradual support measure to emphasize the consecutiveness in changing of attributes over the values. Instead of a pair of tuples, the gradual support concerns on a list of attributes $\mathcal{L} = \{t_i, t_{i+1}, \ldots, t_k\}$ in which $\forall t_i, t_{i+1} \in \mathcal{L} : (t_i, t_{i+1})$ is a gradual tuple pair (Def. 3.1). In the gradual support, $|occ(p)|$ becomes the size of the maximal list of tuples $\mathcal{L}$ which supports $p$ in the MRD $db$. The [all possible cases] becomes the possible longest $\mathcal{L}$ in the $db$. Based on this idea, we also propose the gradual support, denoted $σ_g(p)$, for multi-rgps.

### 5 Multi-Relational Gradual Pattern Miner

Extracting the complete set of multi-rgps is a non-trivial task. At first glance, the number of potential multi-rgps is exponential, i.e., approximately $2^{2\times|A_{db}|}$ where $A_{db}$ is a set of all numerical attributes in the MRD $db$.

**Key idea.** Facing the huge potential search space, we propose an approach, named RGP Miner. In RGP Miner, we first extract mono-rgps from single tables by applying GRITE algorithm [5]. Then the MRD $db$ is transformed into a graph $G = (V, E)$ where $V$ is a set of vertices and $E$ is a set of edges (Figure 4). Each vertex $v_i \in V$ stands for $T^i \in db$ and each edge $e_{ij}$ can be considered as a relationship between two tables $T^i$ and $T^j$. Next, we apply a depth-first search on the set of vertices following $E$ to combine mono-rgps together to be multi-rgps. To avoid unnecessary computations and searches, we propose three pruning rules which are Complementary Pruning, Apriori Pruning, and Backward Pruning. Complementary Pruning is used to avoid the support computation for complement patterns. Then, Apriori Pruning is used to eliminate infrequent patterns. Finally, Backward Pruning avoids computing redundant patterns which have been traversed before.

Before presenting our algorithm, we first give definitions of complement and frequent patterns. Given two gradual attributes $A^>_j$ and $A^<_j$ which are from table $T^i$, we have $σ_r(A^>_j) = σ_r(A^<_j)$. This is because $\forall t, t' : (t, t') \in occ(A^>_j)$ then $(t', t) \in occ(A^<_j)$ and vice versa. $A^>_j$ and $A^<_j$ are called complement each other. Two patterns $p$ and $p'$ are called complement each other if for all gradual attributes $A^*_j$ in $p$ there is a complement attribute of $A^*_j$ in $p'$ and vice versa. We also have $σ_r(p) = σ_r(p')$. Therefore, we can infer the support of a pattern from its complement.

In fact, the number of multi-rgps is exponential. So we only focus on frequent multi-rgps and not all of them. Let us define the frequent multi-rgps as follows:

**Definition 5.1.** (Frequent Pattern). Given a predefined minimal support threshold $σ_{min}$, and a multi-rgp $p = q^1q^2 \ldots q^m$. $p$ is an frequent pattern if:

$$\forall j \in \{i_1, \ldots, m\} : σ_r(q^i \ldots q^j) ≥ σ_{min}$$

In order to start the depth-first search process, vertices in $V$ are collected into two different categories, source vertices and leaf vertices. The source vertices are not pointed from the other vertices while the leaf vertices do not point to any other nodes. The search process is started with source vertices. Note that the vertex $v_{m+1}$ can be added into the node $r = \{v_1, \ldots, v_m\}$ if there is an edge from $v_m$ to $v_{m+1}$.

Let us define some useful notations which will be used in our proposed algorithm as follows:
• A set of frequent multi-rgps $Q^r$.

  Given a current node $r = \{v_i, v_{i+1}, \ldots, v_n\}$, $Q^r$ returns the set of all frequent multi-rgps which are extracted from combining frequent mono-rgps in $Q^{v_i}, \ldots, Q^{v_n}$.

• Merging two nodes, $C^{r \cup r'} = (Q^r \bowtie Q^{r'}) \cup Q^r \cup Q^{r'}$.

  Given two nodes $r$ and $r'$, $C^{r \cup r'} = (Q^r \bowtie Q^{r'}) \cup Q^r \cup Q^{r'}$ contains all potential multi-rgps by merging $r$ and $r'$.

• Attaching two multi-rgps, $p \bullet p'$.

  Given two multi-rgps $p = q^i, \ldots, q^m$ and $p' = q^q, \ldots, q^n$, we have $p \bullet p' = q^i, \ldots, q^q, q^q, \ldots, q^n$.

**Property 5.1. (Complementary Pruning).** Given two complement pattern $p$ and $p'$, we have: $\sigma(p) = \sigma(p')$.

**Proof.** We have $\forall (t_1, t_2) \in occ(p) \text{ then } (t_2, t_1) \in occ(p')$. Thus, $|occ(p)| = |occ(p')|$. In addition, $DomOcc(p) = DomOcc(p')$. So the all the cases of $p$ and $p'$ are the same. Consequently, $\sigma_p(p) = \sigma_p(p')$.

By applying Complementary Pruning, we can avoid computing the supports of complement patterns. Even though Complementary Pruning rule can eliminate the computation of a large number of candidates, there are many other candidates need to be ignored to shrink the search space. In essence, if a multi-rgp is not frequent then its extensions are not frequent as well. Therefore, we can eliminate those extensions. Apriori Pruning rule can be defined as follows:

**Property 5.2. (Apriori Pruning).** Given two multi-rgps $p$ and $p'$, the following holds: if $p$ is not frequent then $p \bullet p'$ is not frequent as well.

**Proof.** We have if $p = \{q^i, \ldots, q^m\}$ is not frequent then $\exists j \in \{i, \ldots, m\}$ s.t. $\sigma_q(q^i \ldots q^j) < \sigma_{\min}$. In addition, $p \subseteq p \bullet p'$ and thus $q^i \ldots q^j \subseteq p \bullet p'$. So, $\exists j \in \{i, \ldots, |p \bullet p'|\}$ s.t. $\sigma_q(q^i \ldots q^j) < \sigma_{\min}$. As a result, $p \bullet p'$ is not frequent (Definition 5.1).

For instance, when we add new vertex loan into a current route \{district, account\} (e.g., step 4th in our depth-first search algorithm, Figure 5) we apply Complementary Pruning and Apriori Pruning to avoid unnecessary computations and infrequent patterns. Regarding to the route, there could be redundant routes in the search space. For instance, at step 10th, the multi-rgp candidates in $Q^{\text{district}} \bowtie Q^{\text{account}} \bowtie Q^{\text{disposition}}$ had been evaluated in step 7th. So they are redundant. Let us define Backward Pruning rule to ignore this type of redundant patterns.

**Property 5.3. (Backward Pruning).** Given a current node $r = \{v_i, \ldots, v_n\}$, $r_a = r \setminus \{v_n\}$ is called ancestor of $r$, an already traversed node $r'$. A set of pattern candidates we need to evaluated $F'$ at the step $r$ can be computed as follows:

$$\text{if } \text{dest}(r') = \text{dest}(r) \land \text{common}(r, r') \neq \emptyset \text{ then } F' = (Q^{\text{district}} \bowtie Q^{\text{account}} \cup Q^{\text{disposition}}) \bowtie Q^v$$

**Proof.** We have $C_r = (Q^{r_a} \bowtie Q^{r_v}) \cup Q^{r_a} \cup Q^{r_v}$. Since $Q^{r_a}$ and $Q^{r_v}$ had been evaluated at previous steps, the candidates now become $F_r = Q^{r_a} \bowtie Q^{r_v} = ((Q^{r_a} \bowtie Q^{r_v}) \cup (Q^{r_a} \bowtie Q^{r_v}) \cup Q^{r_v}) \bowtie Q^{r_a} \bowtie Q^{r_a} \bowtie Q^{r_a}$. We have $r_a \cap r_v \subseteq r'$, thus $Q^{r_a} \bowtie Q^{r_v}$ had been evaluated at the step $r'$. Consequently, $F_r = (Q^{r_a} \bowtie Q^{r_v} \cup Q^{r_v} \cup Q^{r_a} \bowtie Q^{r_a}) \bowtie Q^{r_v}$.

In the Backward Pruning rule, the first part of $F'$ (i.e., $Q^{r_a} \bowtie Q^{r_v} \cup Q^{r_v} \cup Q^{r_a} \bowtie Q^{r_a}$) had been evaluated as the steps $r_a$ and $r'$. Therefore, we do not need to reevaluate them at the step $r$.

By applying Backward Pruning, we can discard a number of redundant patterns when there is another route passing the same destination with the current route. For instance, at step 10th, we have the current node $r = \{v_i, v_{i+1}, \ldots, v_n\}$, the set of frequent patterns in $Q^r$ which includes mono-rgps from $Q^{v_n}$ will be reported. For instance, at the step 3rd, only frequent patterns which contain mono-rgps from $Q^{\text{account}}$ will be reported.

All Theorems are still true for the $\sigma_q$ measure. We will omit the proofs for $\sigma_q$ since they are very similar to the reported ones. In addition, the RGP Miner algorithm has two instances, i.e. $\tau$RGP Miner and $g$RGP Miner. The pseudo code of $\tau$RGP Miner is presented in Algorithm 5.1. The main difference in $g$RGP Miner is that we need to navigate a binary matrix of orders for each pattern to obtain the longest list of gradual tuples. In fact, we apply the matrix navigation function presented in the GRITE algorithm [5].
**Algorithm 5.1. rRGP Miner**

**Input:** MRD db, source nodes S, $G=(V,E)$, $\sigma_{\text{min}}$

**Output:** all frequent multi-rgps

```
begin
  R := \emptyset;
  \text{foreach } v \in S \text{ do }
  \hspace{1em} RGpattern(db, r, v, G, $\sigma_{\text{min}}$);
  \text{RGpattern}(db, r, v, G, $\sigma_{\text{min}}$);
begin
  \text{foreach } node l \in v.edges do
    \hspace{1em} combining(v, r, l);
  \text{output } Q^2; \text{ delete } Q^2; R := R \cup r;
  \text{combining}(route r, node v, node l);
begin
  if type(v, l) = 1; N then
    \hspace{1em} denom := (\lfloor \text{join}(PK^s,l) \rfloor) + 1;
  else
    \hspace{1em} denom := (\lfloor \text{distinct}(PK^s,k) \rfloor) \times (\lfloor \text{distinct}(PK^l,k) \rfloor) + 1;
  temp := 0;
  if pattern q' \in Q^2 do
    if Backward(q', l, r, R) = true then
      \hspace{1em} if exist(db, q') = false then
        \hspace{2em} create(db, q');
      \hspace{1em} if pattern q' \in Q^2 do
        if complement(q', q, temp) = false then
          \hspace{2em} create(db, q');
        \hspace{1em} occ := |q'| \times |q';
        \hspace{1em} \sigma_r(q',q; \sigma_{\text{min}})
        \hspace{1em} if \sigma_r(q',q) \geq \sigma_{\text{min}} then
          \hspace{2em} temp := \text{temp} \cup q';
        \hspace{1em} delete q';
        \hspace{1em} else
          temp := \text{temp} \cup q';
      delete q';
      \text{RGpattern}(db, r, l, G, $\sigma_{\text{min}}$);
end
```

**6 Experimental Results**

A comprehensive experiment study has been conducted on real datasets which are the Financial data\(^1\) (PKDD Cup 99) and the Thrombosis data\(^1\) (PKDD Cup 2001). The Financial database scheme corresponds to the one given in Figure 1. There are 4,500 accounts, 5,369 clients, 5,369 objects in disposition, 6,471 objects in order, 2,500 objects in transaction, 682 objects in loan, 892 cards, and 77 districts. The experiments are carried out on a 2.8GHz Intel Core i7 cpu, 4GB memory.

### 6.1 Patterns

Interesting multi-rgps were discovered in the Financial database and Table 1 illustrates top strongest and longest patterns. We discuss some of them in the following.

**Bases on $\sigma_r$.** The rRGP Miner returns the pattern $p = \{\text{district.Unemploy96}, \text{order.Amount}\}$ with $\sigma_r(p) = 58.32\%$. The pattern could be explained as: “the less unemployed ratio districts are, the more expensive orders the habitants, who live in the districts,

```
\text{tRGP} \quad \text{gRGP} \\
\quad \text{supp} \quad \text{temp} (a)
```

```
\text{tRGP} \quad \text{gRGP} \\
\quad \text{supp} \quad \text{temp} (b)
```

```
\text{tRGP} \quad \text{gRGP} \\
\quad \text{supp} \quad \text{temp} (c)
```

```
\text{tRGP} \quad \text{gRGP} \\
\quad \text{supp} \quad \text{temp} (d)
```

**Figure 6:** (a)(b)-Running time and #patterns on the Financial MRD. (c)(d)-Running time and #patterns on the Thrombosis MRD.

$do$. This pattern is very useful to recommend for business men that the bigger deals are from the less unemployed ratio districts. In addition, by applying the Theorem 4.1, the $E[\sigma_r](p)$ only is $14.3\%$. Thus, the pattern has an attractive support value.

Another pattern related to the average salary is $p = \{\text{district.Avg-Salary}, \text{orders.Amount}\}$ with $\sigma_r(p) = 52.98\%$. The pattern can be explained as: “the lower average salary the districts are, the less expensive orders the habitants, who live in the districts, $do$”.

**Bases on $\sigma_g$.** The gRGP Miner also returns many interesting multi-rgps. Interestingly, the top strongest patterns based on $\sigma_r$ and $\sigma_g$ are very different. For instance, the pattern $p = \{\text{district.#It-500}, \text{loan.Payment}\}$ with $\sigma_g(p) = 44.444\%$. The pattern means: “the more number of municipalities with inhabitans < 499 the districts have, the less the monthly payment is per inhabitant”. This pattern is very interesting since it reveals unknown/unpublished loan policies.

Another interesting pattern is $p = \{\text{district. #It-500}, \text{loan.Duration}\}$ with $\sigma_g(p) = 60\%$. The pattern means: “the more number of municipalities with inhabitans < 499 the districts have, the longer debts the habitants have”. This pattern shows that the habitants who are from the districts which have more number of municipalities with inhabitans < 499 are interested in borrowing money in longer time. Therefore, it could be useful to manage business strategies.

**6.2 Efficiency and Pattern Distribution.** To the best of our knowledge, there is no previous work which addresses multi-rgp issue. Therefore, in the efficiency experiments, we examine the proposed algorithms by varying the support threshold.

\[^1\]\text{http://lisp.vse.cz/challenge/}.
### Table 1: Frequent Multi-RGPs.

<table>
<thead>
<tr>
<th>$\sigma_\tau$ (%)</th>
<th>$\sigma_g$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>account.Date &lt; loan.Date</td>
<td>district.#It-500&lt;, loan.Duratio</td>
</tr>
<tr>
<td>district.Unemploy96&lt;, orders.Amount</td>
<td>district.#It-500&lt;, loan.Date</td>
</tr>
<tr>
<td>district.Avg-Salary&lt;, loan.Amount</td>
<td>district.#People&lt;, loan.Duration</td>
</tr>
<tr>
<td>district.#Crime96&lt;, loan.Amount</td>
<td>district.#People&lt;, loan.Payment</td>
</tr>
<tr>
<td>district.#People&lt;, loan.Payment</td>
<td>district.#Crime96&lt;, loan.Duratio</td>
</tr>
<tr>
<td>district.#Crime96&lt;, loan.Amount</td>
<td>district.#Crime96&lt;, loan.Duratio</td>
</tr>
<tr>
<td>district.Unemploy96&lt;, loan.Amount</td>
<td>district.Unemploy96&lt;, loan.Date</td>
</tr>
<tr>
<td>district.Avg-Salary&lt;, orders.Amount</td>
<td>district.#It-500&lt;, loan.Payment</td>
</tr>
<tr>
<td>district.#Entrepreneur&lt;, loan.Payment</td>
<td>district.#People&lt;, loan.Payment</td>
</tr>
</tbody>
</table>

Some of long patterns

<table>
<thead>
<tr>
<th>$\sigma_\tau$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>district.#It-2000&lt;, #Crime95&lt;, account.Date&lt;, loan.Date&lt;</td>
</tr>
<tr>
<td>district.#It-2000&lt;, #Crime95&lt;, #Crime96&lt;, account.Date&lt;, loan.Date&lt;</td>
</tr>
</tbody>
</table>

### Figure 5: Multi-RGP search space of the graph in Figure 4.

### Figure 7: #Redundant patterns in Financial database.

For Financial database, Figure 6a shows that $\tau$RGP Miner seems linear in terms of executing time. Furthermore, with the same support value, $g$RGP Miner is more efficient than $\tau$RGP Miner. This is because, by applying $\sigma_g$, there are less number of extracted patterns than $\sigma_\tau$ (Figure 6b). In essence, $\sigma_g$ is tighter than $\sigma_\tau$ since the gradual tuple pairs are required to satisfy the consecutiveness constraint. However, there are differences in Thrombosis database. In most of cases, $g$RGP Miner is slower than $\tau$RGP Miner (Figure 6c). Since the matrix size (i.e., $32,935 \times 32,935$) which we need to navigate is significantly larger compare with the ones in Financial database (i.e., $6,471 \times 6,471$). The matrix navigation for computing the longest route is expensive.

### Figure 8: Multi-rgp distribution with $\sigma_{min} \geq 0.15$.

However, with low support values (i.e., $\sigma=0.1$), $g$RGP Miner is faster than $\tau$RGP Miner. This is because there are many patterns extracted by applying $\sigma_\tau$ (Figure 6d).

Let us consider the effectiveness of Backward Pruning rule (Figure 7). By applying the Backward Pruning, a large number of redundant patterns can be ignored in Financial database. In fact, the number of redundant patterns is much larger than the number of extracted ones. As a result, the process speeds up.

The distribution of multi-rgps extracted in Financial database is illustrated in Figure 8. We can consider that $\sigma_\tau$ and $\sigma_g$ are converged at low value area ($\sigma_\tau, \sigma_g \leq 0.3$). However, they are diverse at the high value areas. It means the two support measures illustrate the graduality in different angles. They offer us a
powerful tool to discover more number of useful patterns which could benefit different real world applications. As a result, both $\sigma_g$ and $\sigma_\tau$ are useful in mining multi-rgps.

7 Related Work

Most previous work on relational pattern mining can be categorized into methods that generalize ideas from frequent itemset mining to the relational setting and methods that are based on Inductive Logic Programming (ILP). In this section we discuss the differences between these approaches and our approach as well as other works that do not fall into these two categories.

Well known ideas and algorithms from frequent itemset mining can be used for multi-relational data if applied on the join of all tables. This kind of patterns essentially is itemsets such that items are attribute values and transactions are the tuples of the join table [9, 12, 13]. A different approach is taken by Smurflg [15] where the support is measured with respect to every table, as the relative number of keys that the items correspond to. Warmr [4] and Farmer [14] are ILP-based methods. The patterns are logic rules which can be regarded as local models of the database. The support is defined as the relative number of key values of one target table that satisfy the rule. The purpose of these methods is to extract the most frequent rules.

R-KRIMP [12] patterns are similar to the work of [2], who define these patterns as simple conjunctive queries of the form: $[p_0,\ldots,p_n]$. RDB-Krimp [11] is an improvement of R-KRIMP. RDB-Krimp and RMiner [16] are methods for mining relational databases by using information theoretic ideas for the assessment of patterns. RDB-Krimp focus on the total description length of the database joint with the patterns, and patterns are deemed more interesting if they are better at compressing this description length. RMiner insteads deem patterns more interesting if they describe surprising aspects of the database in a concise way.

Compare with the previous work, we are interested in mining patterns at a higher level. This type of patterns is the correlation between attributes from a graduality point of view.

8 Conclusion and Future Directions

In this paper, we propose the relational gradual pattern concept which enables us to examine the correlations between attributes from a graduality point of view in MRDs. To efficiently mine frequent patterns, we also define its associated support measures based on Kendall’s $\tau$ and gradual supports in MRD context. Then, $\tau$RGP and $g$RGP Miner algorithms are proposed to extract all the frequent patterns. One of the future directions is to investigate top-k representative multi-rgp mining approach to avoid extracting huge amount of patterns. Another work is to define relational gradual pattern at tuple level. That helps us better understand the structure of MRDs from a graduality point of view.

References