Information gain-based selection of sequential patterns extracted from partial unimodal probabilistic bases of sequences

Nicolas Méger, Tuan Nguyen, Christophe Rigotti, Catherine Pothier, Emmanuel Trouvé

▶ **To cite this version:**

Nicolas Méger, Tuan Nguyen, Christophe Rigotti, Catherine Pothier, Emmanuel Trouvé. Information gain-based selection of sequential patterns extracted from partial unimodal probabilistic bases of sequences. 2024-1, Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC) - Polytech Annecy-Chambéry. 2024, pp.85. hal-04472843

HAL Id: hal-04472843

https://hal-lara.archives-ouvertes.fr/hal-04472843

Submitted on 22 Feb 2024

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.

Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

LISTIC laboratory Université Savoie Mont Blanc

Report #2024-1

February 20, 2024

Information gain-based selection of sequential patterns extracted from partial unimodal probabilistic bases of sequences

Nicolas Méger[1*], Tuan Nguyen[1†], Christophe Rigotti[2†], Catherine Pothier[3†], Emmanuel Trouvé[1†]

[1*]Université Savoie Mont Blanc, Polytech Annecy-Chambéry, LISTIC Laboratory, B.P. 80439, Annecy Cedex, F-74944, France.
[2]Univ Lyon, INSA Lyon, CNRS, INRIA, UCBL, LIRIS, UMR5205, Villeurbanne, F-69621, France.
[3]Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, Villeurbanne, F-69621, France.

*Corresponding author(s). E-mail(s): nicolas.meger@univ-smb.fr;
Contributing authors: hoang-viet-tuan.nguyen@univ-smb.fr; christophe.rigotti@insa-lyon.fr; catherine.pothier@insa-lyon.fr; emmanuel.trouve@univ-smb.fr;
†These authors contributed equally to this work.

Abstract

The uncertainty of symbolic data can be represented by probability mass functions. Numerous work adopt this approach to characterize the uncertainty of the events forming a probabilistic base of sequences and extract sequential patterns under the possible worlds semantics. To our knowledge, there is no method for selecting sequential patterns from probabilistic bases of sequences whose probability mass functions are unimodal and for which only the probabilities of the modes are available. Since this situation arises for several kinds of data, a method for selecting sequential patterns extracted from partial unimodal probabilistic bases of sequences is thus proposed in this paper. Using an information gain approach, it outputs informative patterns whose occurrences tend to describe the dataset in a complementary way. Experiments on synthetic and real datasets show that

the method is scalable and that selected patterns, beside being informative and complementary, help end-users to complete their knowledge.

Keywords: Uncertain data, Sequential pattern mining, Probabilistic databases, Information gain

1 Introduction

Uncertain data are becoming increasingly available with the rise of new sensor technologies (wireless sensors, MEMS[1] sensors) and the development of forecasting, imputation or privacy-preserving techniques (Aggarwal, 2009; Aggarwal and Yu, 2009). Such data need to be dealt with to deliver applications ranging from environmental surveillance to mobile tracking (Zhao et al, 2014; Qian et al, 2020). In this context, probabilistic databases (Suciu and Dalvi, 2005; Green and Tannen, 2006) can be used to represent uncertain data. As recalled informally in Aggarwal and Yu (2009), a probabilistic database is "a finite probability space whose outcomes are all possible database instances consistent with a given schema". This model has been adopted in various data mining workflows, such as frequent itemset mining and frequent sequential pattern mining, by adapting existing deterministic concepts and methods to that probabilistic case.

1.1 Frequent itemset mining in uncertain data

Frequent itemset mining, a pattern mining task originally proposed to analyze deterministic item transactions (Agrawal et al, 1993), can benefit from the concept of probabilistic databases as illustrated by the numerous works referenced in Leung (2011). Basically, these works take into account the *uncertainty about the presence of each one of the items* contained within each transaction, each transaction containing one or more items. More precisely, each item uncertainty is simply expressed by an

[1] Microelectromechanical systems.

existential probability (Leung, 2011) quantifying to which extent the item is likely to be present in the transaction. All item probabilities are further exploited under the *possible worlds semantics* as defined in Abiteboul et al (1987) to mine frequent itemsets over all possible worlds (Chui et al, 2007), i.e., all possible database instances that might be generated by considering either the presence or the absence of each item of each transaction. An itemset is said to be frequent if its expected support, i.e., the expected number of transactions in which it occurs, is greater or equal to a user-defined threshold. A drawback of this approach is that it does not bring any information about the confidence with which an itemset is frequent within each one of the possible worlds. It is thus proposed in Bernecker et al (2009) to assess the confidence of an itemset to be frequent, which can be approximated as "the percentage of possible worlds in which [an itemset] is frequent" (Leung, 2011). It relies on the probability mass function of the support of each itemset to extract the so-called *probabilistic* frequent patterns, i.e. itemsets whose probabilities to be frequent are greater or equal to a user-defined thresold.

Both frequent itemsets and probabilistic frequent itemsets can be extracted efficiently by assuming that items are all independent from each other, and by mobilizing the antimonotonicity properties of the expected support and frequentness probability constraints (Chui et al, 2007; Bernecker et al, 2009; Leung, 2011). Big data extensions based on the MapReduce programming model have also been proposed (Leung et al, 2014). Complementary strategies can be adopted to lower resource consumption such as using light-weight data structures, e.g., trees in Leung et al (2008), designing efficient algorithms, e.g., dynamic programming in Bernecker et al (2009) or Sun et al (2010), sampling the dataset (Calders et al, 2010), or considering additional user-specified constraints (Leung, 2011). In order to focus on the most interesting frequent itemsets, top-k approaches have been designed (Zhang et al, 2008; Bernecker

et al, 2009; Cormode et al, 2009), and it has been proposed by Bonchi et al (2011) to select the ones that well compress the probabilistic database under a compression scheme based on the Minimum Description Length (MDL) principle. The techniques for mining frequent itemsets in probabilistic databases have been extended to mine streams of uncertain data by taking into account the unbounded, continuous and variable nature of these streams (Leung and Hao, 2009; Leung and Jiang, 2011).

1.2 Frequent sequential pattern mining in uncertain data

Nevertheless, itemsets do not express any temporal order. As explained by Agrawal and Srikant (1995), if deterministic item transactions are time-stamped, i.e., each itemset present in a transaction describes an event, then the database can be modeled as a base of sequences from which frequent sequential patterns can be extracted. Muzammal and Raman (2010) extended such a deterministic approach to the probabilistic case to deal with uncertainties. In more detail, they first define a deterministic base of sequences as a set of event sequences, each event being time-stamped and described by an itemset, and each sequence being associated with a unique source, i.e., a unique sequence identifier. Then, two exclusive uncertainty contexts have been studied. They relate to either the *source uncertainties* or the *event uncertainties*. The latter are modeled by associating each event with its existential probability and its unique source, whilst, for source uncertainties, a probability mass function over all the possible sources is given for each event. As for itemsets (Chui et al, 2007; Bernecker et al, 2009), Muzammal and Raman (2010) defined the frequent sequential patterns under the possible world semantics by relying on expected supports, and the probabilistic frequent sequential patterns by considering frequentness probabilities. The computation of the expected support measure was shown to be tractable whatever the type of uncertainties that is considered. Regarding frequentness probability, the authors demonstrated that it can be computed for the event uncertainties case

whilst it is reported to be intractable when considering source uncertainties. Indeed, since a same event has to be reported for multiple sources, it has to be represented by multiple events that are fully dependant from each other, which leads to a complexity burden (Muzammal and Raman, 2010). Consequently, probabilistic frequent sequential patterns have been abandoned by Muzammal et Raman, and various efficient algorithms have been proposed in Muzammal and Raman (2011) and Muzammal and Raman (2015) to mine frequent itemsets in a source uncertainties context. They rely on the pattern candidate generation (breadth or depth-first explorations in Agrawal and Srikant (1995) and Zaki (2001) respectively) and the prefix growth (Pei et al, 2004) frameworks exploiting the antimonotonicity property of the expected support constraint. As for deterministic sequential pattern mining, the prefix growth framework is shown to be more scalable, which is not the case when mining frequent itemsets from uncertain data (Tong et al, 2012).

A refinement of the event uncertainty context is proposed in Hooshsadat et al (2012) by processing existential probabilities characterizing each of one the items forming an event to extract frequent sequential pattern. An alternative to the event uncertainty context is proposed in Zhao et al (2014). Each event is described with a single item chosen in the set of all possible items where each item is associated with the probability of characterizing the event. The case where no item is chosen is also considered. In other words, the information content of an event can be represented by a discrete random variable whose probability mass function is defined over the set of all possible items and a special item indicating that nothing is observed. When resorting to such probability mass functions, each possible world is envisaged by considering all possible combinations of possible events. Such an *element-level uncertain data model* can therefore be exploited under a probabilistic frequentness scheme (Zhao et al,

2014) to extract probabilistic frequent sequential patterns using a prefix growth-based algorithm. They also propose to extract such patterns under the *sequence-level uncertain data model* where each sequence is described by a random variable whose probability mass function is defined over the set of all possible sequences including the empty one. The problem of mining frequent sequential patterns from a single uncertain event sequence data using single item-based events coming along with their respective existential probabilities has also been addressed in Wan et al (2013) under the probabilistic frequentness scheme. Finally, the temporal uncertainties of the event timestamps can also be taken into account under the possible world semantics as done in Ge et al (2015) and Ge et al (2017) where each timestamp is uncertain and is represented by a random variable defined using either a uniform distribution or a probability mass function approximating the distribution of any arbitrary shaped uncertainty. Using a prefix growth-based algorithm, such an uncertain probabilistic base of sequences is mined to extract probabilistic frequent sequential patterns. It is worth noting that the majority of the works about probabilistic base of sequences reported in this section achieves good extraction performances because they assume that events are independent from each other.

1.3 Contribution

Whatever the uncertainty data model that is employed to mine itemsets or sequential patterns, it appears that the information characterizing uncertainty is complete and thus sufficient to consider each one of the possible world. Indeed, each element (item, event, source, sequence, timestamp, etc.) for which uncertainty is provided can be associated with a probability mass function that is fully known. For instance, such as source uncertainties in Muzammal and Raman (2015) or timestamp uncertainties in Ge et al (2017) are explicitly modeled using probability mass functions. Though not explicitly defined as such, the element-level and the sequence-level uncertain data

model proposed in Zhao et al (2014) can also be expressed using probability mass functions.

1.3.1 Partial unimodal probabilistic base of sequences

Nevertheless, in some cases, some probabilities are missing and the probability complement rule can not be exploited to derive them. The case of *partial unimodal probabilistic base of sequences* is exposed and addressed in this paper. In such a base, each event is meant to exist, its time span is certain, but it is described by a single item that is uncertain. This uncertainty is expressed through a unimodal probability mass function taking into account all possible items that may describe an event and for which the only available probability is the one of its mode. In other words, each event is described by a random variable characterized by the item that is the more likely to be reported (the mode) and its probability. As so long as the variable domain contains more than two different items, missing probabilities can not be estimated without assuming some type of distribution. In this work, distributional assumptions are avoided for the sake of generality and to favour knowledge discovery. The base of sequences that is dealt with is therefore not a probabilistic one, but a *partial unimodal* probabilistic one.

Such a base can be encountered in various applications such as RFID or GPS tracking where uncertain localizations (Chen et al, 2010; Pelekis et al, 2010; Teng et al, 2014; Qian et al, 2020) can be reported using the most probable ones and their probabilities. In that case, each time an event occurs, it is represented by a random variable ranging over all possible locations, e.g., a set of points of interest. Its probability mass function is only known for its mode, i.e., the most probable location of the event, where the associated probability comes from the measurement process. One could also think about a customer survey where each respondent is asked for the product or the service he/she is the most likely to buy for each one of the years to come while giving associated probabilities. In this application, for each customer and for each year,

the choice is represented by a random variable ranging over the set of all available products/services, and its probability mass function is unknown except for its mode, i.e., the most probable product/service. In Section 7, experiments on a real dataset are reported, where the sequences represent a displacement field time series. In this application, each event denotes a displacement with a corresponding random variable ranging over the displacement magnitude levels. The probability associated to the mode being then derived from a confidence measure provided by the displacement detection process.

1.3.2 Information gain-based selection of sequential patterns

With the aim of exploiting such a database, we propose to extract sequential patterns such as frequent ones from the only deterministic base of sequences that can be instantiated from it, i.e. the most likely possible world. Such patterns can be numerous. In order to select a reduced set of sequential patterns, while keeping its ability to summarize the database, different techniques have been proposed for the deterministic case such as considering maximal sequential patterns (Raissi et al, 2006; García-Hernández et al, 2006; Luo and Chung, 2004), closed sequential patterns (Yan et al, 2003b; Wang et al, 2007; Yu et al, 2012) or optimal compression patterns (Tatti and Vreeken, 2012; Lam et al, 2014a; Ibrahim et al, 2016). Such a reduced set of sequential patterns can also be obtained by assessing them with swap randomization techniques (Méger et al, 2015, 2019). Of these approaches, the optimal compression one, which selects a reduced set of patterns that best compresses the dataset, is a promising direction. Indeed, selected patterns tend to be frequent and also complement each other very well in terms of data locations. The final set of selected patterns is thus less redundant and also allows for easier interpretation. As previously mentioned (see Section 1.1), such an approach has been proposed in Bonchi et al (2011) to select itemsets that well compress a probabilistic database. To our knowledge, these compression-based approaches, though inspiring, can not be directly used to

select sequential patterns by taking into account mode probabilities only.

A method for selecting sequential patterns extracted from partial unimodal probabilistic bases of sequences is therefore proposed and detailed in this paper. Preliminary evidences of the interest of such mining were reported early in a conference paper (Nguyen et al, 2018b). Using an information gain approach, the method outputs informative patterns whose occurrences tend to describe the dataset in a complementary way. More precisely, events are described by independent discrete random variables whose probability mass functions, also termed distributions, are initially supposed to be uniform. Once revealed to end-users, the pattern occurrences and their corresponding mode probabilities allow to constrain and refine initial event distributions. Finally, patterns are selected in greedy way by selecting, for each iteration, the pattern leading to the highest information gain knowing the patterns selected during the previous iterations.

After having defined the partial unimodal probabilistic bases of sequences and adapted the concept of frequent sequential patterns to this context in Section 2, the principle of the proposed method is detailed in Section 3. The pattern occurrence constraints and their combinations are formalized in Section 4 and Section 5 presents the pattern occurrence constraint-based minimization of the Kullback-Leibler divergence that is employed to refine event distributions. Section 6 gives a toy example illustrating the operation of the proposed method and Section 7 provides experiments on synthetic and real datasets showing that the method is scalable and that selected pattern are informative and complementary, thus allowing end-users to complete their knowledge.

2 Partial unimodal probabilistic bases of sequences and frequent sequential patterns

Inspired by the seminal works regarding the bases of sequences (Agrawal and Srikant, 1995) and the probabilistic bases of sequences (Muzammal and Raman, 2010), a *partial unimodal probabilistic event* is defined as an event that exists, whose occurrence date is certain and for which only the event type having the highest probability of being observed is supplied along with this probability. In other words, the uncertainty about the nature of the event is expressed through a unimodal probability mass function defined over all possible event types that may describe an event, and for which the only available probability is the one of its unique mode. Such an event differs from the ones defined in Agrawal and Srikant (1995) and Muzammal and Raman (2010) since the latter are defined as sets of items. In addition, it also differs from the different kind of probabilistic events that can be found in the sequential pattern mining literature (Muzammal and Raman, 2010, 2015; Hooshsadat et al, 2012; Zhao et al, 2014; Wan et al, 2013), were corresponding probability mass functions are fully specified.

More formally, the event types, their probabilities and the partial unimodal probabilistic events are defined as follows:

Definition 1 (event type and event) Let $\mathbb{E} = \{e_1, e_2, \ldots, e_n\}$ be a set containing all possible *event types* of the base. An occurrence of an event type at a date t is termed an *event*. All events are mutually exclusive, i.e., two events cannot correspond to the same date.

Definition 2 (event probability) The *event probability* $\rho^{t,e}$, with $t \in \mathbb{N}$, $e \in \mathbb{E}$ and $\rho^{t,e} \in [0, 1]$, is the probability of observing event type e at date t.

Notice that for any date t, by definition 1, $\sum_{e \in \mathbb{E}} \rho^{t,e} = 1$.

Definition 3 (partial unimodal probabilistic event) A *partial unimodal probabilistic event* is a triple $\langle t, e, \rho^{t,e} \rangle$, with $t \in \mathbb{N}$ and $e \in \mathbb{E}$ such that $\forall e' \in \mathbb{E} \setminus \{e\}$, $\rho^{t,e} > \rho^{t,e'}$. For the sake of clarity, when clear from the context, a partial event $\langle t, e, \rho^{t,e} \rangle$ will be written $\langle t, e, \rho \rangle$.

Such an event is *partial* since only the probability $\rho^{t,e}$ of observing e at date t is available in the base. The probabilities of the other event types, i.e. those belonging to $\mathbb{E} \setminus \{e\}$, are unknown when considering the same occurrence date t. These probability are not given in the base, but it is assumed that they are strictly lesser than $\rho^{t,e}$: such an event is said to be *unimodal*. Thus, since $\sum_{e \in \mathbb{E}} \rho^{t,e} = 1$ and $\forall e' \in \mathbb{E} \setminus \{e\}$, $\rho^{t,e} > \rho^{t,e'}$ we know that $\rho^{t,e} \in]\frac{1}{|\mathbb{E}|}, 1]$.

Once the partial unimodal probabilistic events defined, the concepts of *partial unimodal probabilistic event sequence* and *partial unimodal probabilistic base of sequences* can be formalized as follows:

Definition 4 (partial unimodal probabilistic event sequence) For a source s, the *partial unimodal probabilistic sequence* of s is a pair (sid, seq), where sid is the source (or sequence) identifier and seq is a tuple of partial unimodal probabilistic events, of the form $seq = \langle \langle t_1, e_1, \rho_1 \rangle, \langle t_2, e_2, \rho_2 \rangle, ..., \langle t_n, e_n, \rho_n \rangle \rangle$ containing the n partial unimodal probabilistic events of s, where $t_1 < t_2, \ldots, < t_n$.

Definition 5 (partial unimodal probabilistic base of sequences) A *Partial Unimodal Probabilistic Base of Sequences* (*PUPBoS*) is then a set of partial unimodal probabilistic event sequences having different *sid*.

A toy PUPBoS is given as Example 1.

Example 1 Let $\mathbb{E} = \{A, B, C\}$ be a set of three symbols. A partial unimodal probabilistic base of sequences $\mathcal{B}$ is given hereafter. It contains four partial unimodal probabilistic event

sequences, each sequence containing four partial unimodal probabilistic events occurring at dates t_1, t_2, t_3 and t_4.

$$\mathcal{B} = \{(1, \langle\langle t_1, B, 0.4\rangle, \langle t_2, C, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, C, 0.5\rangle\rangle),$$
$$(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle),$$
$$(3, \langle\langle t_1, B, 0.9\rangle, \langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle),$$
$$(4, \langle\langle t_1, C, 0.8\rangle, \langle t_2, A, 0.7\rangle, \langle t_3, B, 0.8\rangle, \langle t_4, A, 0.7\rangle\rangle)\}$$

This base could for instance give the most probable carbon dioxide levels, denoted by the symbols available in $\mathbb{E}$, supplied by four indoor air quality sensors (i.e., four sources) that also provide level probabilities depending on the operating conditions (e.g., temperature, humidity) and expressing to which extent supplied levels can be trusted[2] Back to the examples given in Section 1.3, for the GPS or RFID tracking application, Example 1 would give the most probable locations of four mobile objects over time, expressed with symbols coming from $\mathbb{E}$ and denoting points of interest. These symbols would be reported at dates t_1, t_2, t_3 and t_4 along with their probabilities that depend on the measurement conditions (number of RFID/GPS signals, signal strength). Regarding, the customer application, the products that are likely to be bought by four customers for years t_1, t_2, t_3 and t_4 would be denoted using symbols in $\mathbb{E}$, while probabilities would be directly estimated and supplied by customers themselves.

In such a base, events are assumed to be independent from each other. This assumption has been adopted by the very vast majority of the works about probabilistic sequential pattern mining (see Section 1). As for bases of sequences or probabilistic bases of sequences, we propose to mine partial unimodal probabilistic bases of sequences with the aim of extracting *frequent sequential patterns* as originally defined in Agrawal and Srikant (1995). Formally, in our context where an event is described by a single event type, a sequential pattern and its occurrence are defined as follows:

[2]Notice that the method presented in this paper can also handle more general cases, where sequences have different lengths and events are not synchronous.

Definition 6 (sequential pattern) A *sequential pattern* α is a tuple $\langle \alpha_1, \alpha_2, \ldots, \alpha_m \rangle$ where $\alpha_1, \ldots, \alpha_m$ are event types in $\mathbb{E}$ and m is the *length* of α. Such a pattern is also denoted as $\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$.

Definition 7 (occurrence) Let $\mathcal{B}$ be a PUPBoS and $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ be a sequential pattern. Then the pair $(sid, \langle \langle t_1, \alpha_1, \rho_1 \rangle, \langle t_2, \alpha_2, \rho_2 \rangle, \ldots, \langle t_m, \alpha_m, \rho_m \rangle \rangle)$, where $t_1 < t_2 < \ldots < t_m$, is an *occurrence* of α in $\mathcal{B}$ if there exists $(sid, seq) \in \mathcal{B}$ such that $\forall i \in \{1, \ldots, m\}, \langle t_i, \alpha_i, \ \rho_i \rangle \in seq$.

In the proposed approach, patterns are selected according to the amount of information that is conveyed by their occurrences. In order to derive as much information as possible from the knowledge of pattern occurrences, we propose to rely on a more informative type of occurrences, namely the *earliest minimal occurrences*. Inspired by the *minimal occurrences* that were first proposed in Mannila et al (1997), minimal occurrences are defined in the context of PUPBoS as follows:

Definition 8 (minimal occurrence) Let $\mathcal{B}$ be a PUPBoS and $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ be a sequential pattern. Then the occurrence $(sid, \langle \langle t_1, \alpha_1, \rho_1 \rangle, \langle t_2, \alpha_2, \rho_2 \rangle, \ldots, \langle t_m, \alpha_m, \rho_m \rangle \rangle)$ of α in $\mathcal{B}$ is a *minimal occurrence* of α in $\mathcal{B}$ if there does not exist any other occurrence $(sid, \langle \langle t'_1, \alpha_1, \rho'_1 \ \rangle, \langle t'_2, \alpha_2, \ \rho'_2 \rangle, \ldots, \langle t'_m, \alpha_m, \ \rho'_m \rangle \rangle)$ of α in $\mathcal{B}$ such that $t'_1 > t_1 \wedge t'_m < t_m$ or $t'_1 > t_1 \wedge t'_m = t_m$ or $t'_1 = t_1 \wedge t'_m < t_m$.

An occurrence of pattern α is therefore minimal if it does not spread fully over another occurrence of α whose time span is shorter. In Example 1, the minimal occurrences of sequential pattern $A \rightarrow A$ in sequence 3 are $(3, \langle \langle t_2, A, 0.9 \rangle, \langle t_3, A, 0.7 \rangle \rangle)$ and $(3, \langle \langle t_3, A, 0.7 \rangle, \langle t_4, A, 0.8 \rangle \rangle)$. Occurrence $(3, \langle \langle t_2, A, 0.9 \rangle, \langle t_4, A, 0.8 \rangle \rangle)$ is not minimal since it contains them. The partial unimodal probabilistic events forming a minimal occurrence do not need to be contiguous. For instance, in sequence 4,

occurrence $(4, \langle\langle t_2, A, 0.7\rangle, \langle t_4, A, 0.7\rangle\rangle)$ is minimal.

An earliest minimal occurrence is then defined as follows:

Definition 9 (earliest minimal occurrence) Let $\mathcal{B}$ be a PUPBoS, $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ be a sequential pattern and $(sid, seq) \in \mathcal{B}$. The triple $(sid, \langle\langle t_1, \alpha_1, \rho_1\rangle, \langle t_2, \alpha_2, \rho_2\rangle, \ldots, \langle t_m, \alpha_m, \rho_m\rangle\ \rangle, \rho_{min})$ is an *earliest minimal occurrence* of α in $\mathcal{B}$ if $(sid, \langle\langle t_1, \alpha_1, \rho_1\rangle, \langle t_2, \alpha_2, \rho_2\rangle, \ldots, \langle t_m, \alpha_m, \rho_m\rangle\ \rangle)$ is a minimal occurrence of α in $\mathcal{B}$ and if there does not exist any other minimal occurrence $(sid, \langle\langle t_1, \alpha_1, \rho_1\rangle, \langle t'_2, \alpha_2, \rho'_2\rangle, \ldots, \langle t_m, \alpha_m, \rho_m\rangle\ \rangle)$ of α in $\mathcal{B}$ for which there exists $i \in \{2, \ldots, m-1\}$ such that $t'_i < t_i$. The last element of the earliest minimal occurrence, i.e., ρ_{min}, reflects the lower bound of the quality of this part of the sequence. It is defined as the minimum probability in seq between t_1 and t_m: $\rho_{min} = \min\{\rho \mid \langle t, e, \rho\rangle \in seq \wedge t_1 \leq t \leq t_m\}$.

Consequently, if there exists several minimal occurrences of a pattern starting and ending at the same times in the same sequence, then the one formed by the earliest partial unimodal probabilistic events is selected. In addition, the minimum probability, ρ_{min}, of the partial unimodal probabilistic events occurring in the same sequence and the same time interval is supplied. Back to Example 1, two minimal occurrences of sequential pattern $B \rightarrow C \rightarrow A$ can be found in sequence 2: $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle)$ and $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A,\ 0.5\rangle\rangle)$. Only the former can be used to form the earliest minimal occurrence $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle)$ by adding the minimum probability observed between its starting date t_1 and its ending date t_4. It should be noticed that such an occurrence can bring supplementary information. For instance, here, it implies that sequence 2 cannot contain an event type A at date t_3, otherwise the occurrence would not be minimal.

By counting the number of partial unimodal probabilistic event sequences in which a pattern occurs at least once, it is possible to focus on frequent sequential patterns by adapting this concept to PUPBoS:

Definition 10 (support) Let $\mathcal{B}$ be a PUPBoS and $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ be a sequential pattern. A partial unimodal probabilistic event sequence (sid, seq) of $\mathcal{B}$ supports α if there exists an earliest minimal occurrence $(sid, \langle\langle t_1, \alpha_1, \rho_1\rangle, \langle t_2, \alpha_2, \rho_2\rangle, \ldots, \langle t_m, \alpha_m, \rho_m\rangle \rangle, \rho_{min})$ of α in $\mathcal{B}$. The set of the partial unimodal probabilistic event sequences supporting α is noted $S(\alpha)$. The *support* of α in $\mathcal{B}$, denoted by $support(\alpha)$, is simply the number of partial unimodal probabilistic event sequences in $\mathcal{B}$ that support α, i.e. $|S(\alpha)|$.

Definition 11 (frequent sequential pattern) Let $\sigma \in [0, 1]$ be a *relative support threshold*. A sequential pattern α is a *frequent sequential pattern* in the PUPBoS $\mathcal{B}$ if $support(\alpha)/|\mathcal{B}| \geq \sigma$

Example 2 Let us consider $\mathcal{B}$, the PUPBoS given by Example 2. If $\sigma = \frac{1}{2}$, then sequential patterns are required to occur in two partial unimodal probabilistic event sequences at least. The frequent sequential patterns are given in Table 1 along with the sequence/source identifiers of the sequences supporting them and their corresponding support measures.

Table 1: The frequent sequential patterns in $\mathcal{B}$

pattern	sequence/source id	support
A	$1, 2, 3, 4$	4
B	$1, 2, 3, 4$	4
C	$1, 2, 4$	3
$A \rightarrow A$	$3, 4$	2
$B \rightarrow A$	$1, 2, 3, 4$	4
$B \rightarrow C$	$1, 2$	2
$C \rightarrow A$	$1, 2, 4$	3
$C \rightarrow C$	$1, 2$	2
$B \rightarrow C \rightarrow A$	$1, 2$	2
$B \rightarrow C \rightarrow C$	$1, 2$	2

Note that considering minimal occurrences (see Definition 8) or occurrences (see Definition 7), instead of earliest minimal occurrences, leads to the same frequent

sequential patterns (but not the same occurrences) since the support measure is established by counting the number of partial unimodal probabilistic event sequences in which they occur at least once. The standard deterministic approach (Agrawal and Srikant, 1995) also considers that a sequential pattern is frequent as long as it occurs in a sufficient number of event sequences, whatever the occurrence type that is considered. Thus, the frequent patterns specified by Definition 11 can be obtained using standard algorithms to search for frequent sequential patterns. So, in this paper, we propose a two steps process: (1) Use any existing algorithm to find the frequent sequential patterns; (2) Handle the probabilistic aspect in a following step, by selecting informative and complementary patterns among the frequent ones. It should be noted that the antimonotonicity property of the support constraint, on which existing extraction methods rely, can be directly exploited within either the pattern candidate generation, breadth or depth-first explorations (Agrawal and Srikant, 1995; Zaki, 2001), or the prefix growth (Pei et al, 2004) framework.

In addition to the frequency constraint, other constraints such as those focusing on maximal patterns (Luo and Chung, 2005; Raissi et al, 2006; García-Hernández et al, 2006; Luo and Chung, 2004), or closed patterns (Yan et al, 2003a,b; Wang et al, 2007; Yu et al, 2012), can be used to focus on smaller output collections. Nevertheless, even if such constraints are considered, end-users generally face large output collections of sequential patterns and automatic methods for guiding them towards the most promising ones are needed. This is for example achieved by selecting the patterns that best compress the dataset (Lam et al, 2014b), or by assessing the patterns with swap randomization techniques (Méger et al, 2015, 2019). To our knowledge, no selection method exploiting the probabilities available in partial unimodal probabilistic bases of sequences is available. Such a method is therefore proposed and is the main contribution of this paper. Its principle is presented in the following section.

3 Information gain-based selection of sequential patterns : principle of the method

In this section, we propose an original method for finding a set of informative and complementary patterns that takes into account the mode probabilities available in PUPBoS. Let us consider a partial base of sequences containing N partial unimodal probabilistic events that are assumed to be independent from each other (see Section 2). Its information content can be expressed as the entropy of a set of N independent random variables, noted $\mathcal{V} = \{X_1, X_2, \ldots, X_N\}$, such that each random variable is associated with a partial unimodal probabilistic event. Let us suppose that our knowledge of the PUPBoS is partial. The key intuition of the method is that if we are given the earliest minimal occurrences of a pattern α, then this knowledge provides additional information about the probability mass functions[3] of the variables in $\mathcal{V}$ and thus reduces the entropy of $\mathcal{V}$. The larger this reduction is, the more informative the pattern is. In the following, this information gain, provided by α, with respect to the current partial information we have about the PUPBoS is noted $\Delta(\alpha)$. Finding an optimal set of patterns with respect to an entropy criterion is in general NP-hard (Lam et al, 2014b). Inspired by the *SeqKrimp* algorithm (Lam et al, 2014b) that was designed to find sets of sequential patterns that compress datasets, we propose a greedy suboptimal algorithm for selecting the most informative and complementary sequential patterns in an iterative way. It is given as Algorithm 1.

Algorithm 1 takes the following elements as inputs: k, the number of patterns to be selected, $\mathcal{V}$, the set of random variables representing the information content of the PUPBoS and P, a set of sequential patterns extracted from the PUPBoS. Set P can contain any kind of sequential patterns such as the frequent, the closed or the maximal ones. The algorithm starts, line 2, by initializing the distributions of all variables

[3] For the sake of simplicity, these functions are also referred to as *distributions* in the following.

Algorithm 1 Selection of k informative and complementary sequential patterns

Input: k, the number of patterns to be selected, P, the set of sequential patterns extracted from the PUPBoS, $\mathcal{V}$, the set of random variables representing the information content of the PUPBoS

Output: Φ a set of informative and complementary sequential patterns

1: $\Phi \leftarrow \emptyset$
2: $\forall X \in \mathcal{V}$, set the distribution of X to the uniform distribution
3: **while** ($|\Phi| < k$ and $|P| > 0$) **do**
4: $\alpha^* \leftarrow \text{argmax}_{\alpha \in P}(\Delta(\alpha))$, i.e., the pattern maximizing the information gain w.r.t. the current knowledge of the distributions of the variables in $\mathcal{V}$
5: $\Phi \leftarrow \Phi \cup \{\alpha^*\}$
6: $P \leftarrow P \setminus \{\alpha^*\}$
7: For all $X \in \mathcal{V}$, update the distribution of X according to the occurrences of α^*
8: **end while**
9: **return** Φ

in $\mathcal{V}$ to a uniform distribution defined over domain $\mathbb{E}$. This setting is assumed to begin with a state corresponding to a maximum of entropy, i.e., the user knows nothing except domain $\mathbb{E}$. As long as patterns are available in P, the algorithm iterates until k informative and complementary sequential patterns have been found and added to Φ which is finally returned. At each iteration, line 4, the sequential pattern α^* whose earliest minimal occurrences lead to the highest information gain obtained by refining the distributions of the corresponding variables in $\mathcal{V}$ is selected. Then, the current sets of patterns Φ and P are updated accordingly lines 5 and 6. Finally, line 7, the distributions of the random variables are updated according to the occurrences of α^*.

Let $\Delta_{occ}(o)$ denote the information gain obtained by revealing an earliest minimal occurrence o. In a partial unimodal probabilistic event sequence (sid, s) supporting α, we consider that only the *best* occurrence of α, i.e. the one leading to the highest information gain, is revealed to the user. Thus, the gain coming from α in (sid, s) is $\Delta((sid, s), \alpha) = \max_{o \in \mathcal{O}}\{\Delta_{occ}(o)\}$, where $\mathcal{O}$ is the set of occurrences of α in (sid, s). Finally, the measure of the gain $\Delta(\alpha)$, which assesses the interest of this pattern over all sequences it covers, is defined as:

$$\Delta(\alpha) = \frac{\sum_{(sid,s)\in S(\alpha)} \Delta((sid,s),\alpha)}{|S(\alpha)|}$$

with $S(\alpha)$ the set of partial unimodal probabilistic event sequences supporting α. The intuition for this normalization, i.e., dividing by $|S(\alpha)| = support(\alpha)$, is that a pattern supported by many partial unimodal probabilistic event sequences with occurrences bringing few information is likely to be less desirable than a pattern supported by less sequences but relying on very informative occurrences.

Computing $\Delta_{occ}(o)$, the information gain associated to an earliest minimal occurrence o, is a crucial step. It can be achieved by exploiting the event types and the mode probabilities provided by the partial unimodal probabilistic events forming o. Unveiling a partial unimodal probabilistic event $\langle t, e, \rho\rangle$ indeed allows to refine the current distribution p of its corresponding random variable X into a new distribution q by imposing that $Pr(X = e) = \rho$. The kind of occurrence, i.e., the earliest minimal occurrences as defined in Section 2, also yields additional constraints that can be used to refine the distributions of the random variables associated to the partial unimodal probabilistic events occurring in-between the events forming pattern occurrences. The different possible constraints and their combinations are defined and studied in Section 4.

More generally, let us consider one of the random variable X and its current known distribution p. The information regarding X and conveyed by an occurrence o is expressed with a constraint ξ. In order to quantify the minimum information quantity brought by o, we aim to find a distribution q for X that is the closest to p (i.e., adding the smallest amount of information) but satisfying ξ (that must hold for the real distribution). We here rely on the *Kullback-Leibler* divergence which is defined as follows:

$$D(q||p) = \sum_{e\in\mathbb{E}} q(e)\log_2\left(\frac{q(e)}{p(e)}\right)$$

where $p(e)$ (resp. $q(e)$) is $Pr(X = e)$ in distribution p (resp. q). The measure $D(q||p)$ quantifies *the information gained* if the distribution p is replaced by the real distribution q (Rényi, 1961). The distribution q we are looking for is then a distribution satisfying constraint ξ and minimizing $D(q||p)$, so as to be as close as possible to p (from an information content perspective). It is obtained by a constraint-based optimization scheme which is described in Section 5. The value of $\Delta_{occ}(o)$ is then the sum of the minimized Kullback-Leibler divergences, over all the pairs (p, q) for the random variables in $\mathcal{V}$ that are involved in the constraints derived from occurrence o. This minimization procedure is applied by Algorithm 1 at line 7 to update the distributions of the random variables affected by the constraints implied by the earliest minimal occurrences of α^*.

4 Occurrence constraints

Revealing pattern earliest minimal occurrences formed by partial unimodal probabilistic events allows to constrain the distributions of the random variables representing the information content of a PUPBoS. Section 4.1 inventories the types of constraints that can be imposed and Section 4.2 presents a study of their combinations.

4.1 Constraint types

Let us consider a partial base of sequences containing N partial unimodal probabilistic events. Its information content can be expressed as the entropy of a set of N random variables, noted $\mathcal{V} = \{X_1, X_2, \ldots, X_N\}$, each random variable being associated to a partial unimodal probabilistic event through a bijective mapping. Let ξ be a constraint that is revealed and applied to a random variable $X \in \mathcal{V}$ representing the information content of a partial unimodal probabilistic event $\langle t, e, \rho \rangle$. Constraint ξ is formalized as follows:

- by giving $\mathbb{E}_C$, $\mathbb{E}_C \subset \mathbb{E}$, the set of possible values for e, terms *candidate* event types;

- and by specifying some restrictions on $Pr(X = e_c)$ for the e_c in $\mathbb{E}_C$.

4.1.1 Constraint type #1: direct constraints

This type of constraint is the most stringent one. It occurs when a partial unimodal probabilistic event $\langle t, e, \rho \rangle$ forming an pattern earliest minimal occurrence is revealed and used to refine the distribution of its corresponding random variable X. In that case, such a constraint is written:

$$\xi = \begin{cases} \mathbb{E}_C = \{e\} \\ \Pr(X = e) = \rho \end{cases}$$

For example, upon finding an occurrence $o = (sid, \langle\langle t_{i_1}, e_{i_1}, \rho_{i_1}\rangle, \langle t_{i_2}, e_{i_2}, \rho_{i_2}\rangle, \ldots, \langle t_{i_m},\ e_{i_m},\ \rho_{i_m}\rangle\rangle)$, for each one of the events $\langle t_{i_k}, e_{i_k}, \rho_{i_k}\rangle$, with $1 \leq k \leq m$, its event type and its probability are disclosed. This knowledge is the most informative one that can be obtained about these events. A constraint of type #1 is therefore the strongest constraint.

4.1.2 Constraint type #2: strong propagation constraints

Beside using directly the partial unimodal probabilistic events forming revealed pattern earliest minimal occurrences, it is possible to propagate their information to also constrain the distributions of the random variables associated to the events occurring in-between the events forming pattern earliest minimal occurrences. This is made possible by taking into account the nature of the earliest minimal occurrences.

Let us consider a pattern $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_m$ and $\langle sid, \langle\langle t_1, \alpha_1, \rho_1\rangle\rangle, \langle t_2, \alpha_2, \rho_2\rangle, \ldots\ , \langle t_m, \alpha_m, \rho_m\rangle\rangle, \rho_{min}\rangle$, a minimal earliest occurrence of α. For all $j \in \{2, \ldots, m\}$ and for each event $\langle t_u, e_u, \rho_u\rangle$ of sid such that $t_{j-1} < t_u < t_j$, its corresponding variable random variable X is such that $\mathbb{E}_C = \mathbb{E} \setminus \alpha_j$ since α_j can not

occur by definition of the earliest minimal occurrences (see Section 2, Definition 9). In addition, by definition of the minimal occurrences (see Section 2, Definition 8), event type α_1 can not occur at t_u such that $t_1 < t_u < t_2$ and must also be removed from the set $\mathbb{E}_C$ of the corresponding random variable. And, finally, by definition of the earliest minimal occurrences, for each one of the variables corresponding to an event whose occurrence date is in $[t_1, t_m]$, there exists one event type belonging to $\mathbb{E}_C$ such that its probability is greater or equal to ρ_{min}. This reduction of $\mathbb{E}_C$ and the associated probability constraint thus allow to constrain the distributions of the random variables corresponding to the events occurring in-between the events forming a pattern occurrence. This technique is termed as *propagation*. In Example 1, the earliest minimal occurrence $(2, \langle\langle t_1, B,\ 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle)$ of pattern $B \rightarrow C \rightarrow A$ imposes that the event type at date t_3 can not be A and that its probability is greater or equal to 0.4.

The propagation is *strong* if $|\mathbb{E}| = 2$. Indeed, in that case, $\mathbb{E}_C$ is always reduced to a singleton, i.e., $\mathbb{E}_C = \{e_{remainder}\}$. In addition, by definition of the earliest minimal occurrences, $Pr(X = e_{remainder}) \geq \rho_{min}$. The propagation can also be strong if $|\mathbb{E}| = 3$ and $\alpha_1 \neq \alpha_2$. In that case, for each event $\langle t_u, e_u, \rho_u\rangle$ of *sid* such that $t_1 < t_u < t_2$, the constraint imposed on the associated variable distribution is $\mathbb{E}_C = \mathbb{E} \setminus \{\alpha_1, \alpha_2\} = \{e_{remainder}\}$ and $Pr(X = e_{remainder}) \geq \rho_{min}$. The constraint resulting from a strong propagation is a constraint of type #2. It is weaker than a constraint of type #1 and it is formalized as follows:

$$\xi = \begin{cases} \mathbb{E}_C = \{e_{remainder}\} \\ \Pr(X = e_{remainder}) \geq \rho_{min} \end{cases}$$

For instance, in Example 1, in sequence 1, the earliest minimal occurrence $\langle 1, \langle \langle t_1, B, 0.4 \rangle, \langle t_3, A, 0.7 \rangle \rangle, 0.4 \rangle$ of pattern $B \rightarrow A$ imposes that the event type of the partial unimodal probabilistic event occurring at date t_2 is C since B and A can not occur. In addition, its minimum probability is 0.4.

4.1.3 Constraint type #3: soft propagation constraints

If propagation is performed without being able to reduce $\mathbb{E}_C$ to a single candidate event type, then the resulting constraint is of type #3, i.e. it is a *soft propagation constraint*. Such a constraint is weaker than a constraint of type #1 or #2 and it is written:

$$\xi = \begin{cases} \mathbb{E}_C \subset \mathbb{E}, \text{with } \ |\mathbb{E}_C| \geq 2 \\ \exists e_c \in \mathbb{E}_C \text{ such that } \Pr(X = e_c) \geq \rho_{min} \end{cases}$$

Note that $\mathbb{E}_C$ is included into $\mathbb{E}$ strictly. Indeed, a soft propagation constraint always originates from a pattern occurrence imposing that at least one event type belonging to $\mathbb{E}$ can not be candidate.

An example can be found with the earliest minimal occurrence $\langle 2, \langle \langle t_1, B, 0.5 \rangle, \langle t_2, C, 0.5 \rangle, \langle t_4, A, 0.5 \rangle \rangle, 0.4 \rangle$ imposes the following constraint on the event occurring at date t_3 in sequence 2:

$$\xi = \begin{cases} \mathbb{E}_C = \{B, C\} \\ \exists e \in \mathbb{E}_C \text{ such that } \Pr(X = e) \geq 0.4 \end{cases}$$

4.2 Constraint combinations

Algorithm 1 starts by considering that all random variable distributions are uniform. At each iteration, the selection of the most informative pattern α^* (line 4) is performed

by refining the random variable distributions using the constraints imposed by the earliest minimal occurrences of the patterns left in P and by assessing the associated information gain for each pattern separately.

Each distribution refinement of a variable X is simply obtained by computing a constraint ξ that combines the constraint $\xi^{current}$ produced by an earliest minimal occurrence on X with $\xi^{previous}$, the constraint on X that was established during the previous iterations.

Once the most informative pattern α^* has been selected, then (line 7) the distribution of the variables $X \in \mathcal{V}$ are refined according to the earliest minimal occurrences of α^*.

Since each random variable is supposed to be uniform, a constraint imposing the uniform distribution is build for each random variable before the first iteration. Contrary to other constraint types, this kind of constraint is not imposed by a pattern occurrence. It is termed as a *uniform* constraint and written:

$$\xi_U = \begin{cases} \mathbb{E}_C = \mathbb{E} \\ \Pr(X = e_c) = \frac{1}{|\mathbb{E}|}, \forall e_c \in \mathbb{E}_C \end{cases}$$

Therefore, as soon as the first occurrence constraint $\xi^{current}$ is discovered for a variable X, it replaces the initial constraint $\xi_{uniform}$. Any constraint differing from a uniform one is indeed preferred since it differs from the initial knowledge, thus providing more information on the realization of X. If $\xi^{previous}$ is not a uniform constraint then:

1. If one of the two constraints, $\xi^{previous}$ or $\xi^{current}$, is a type #1 constraint, the resulting constraint ξ is the same as the type #1 constraint since the latter is always the strongest constraint that can be imposed on each event. Moreover, when two type #1 constraints affect the same random variable, they are identical since

they correspond to the discovery of the same partial probabilistic event forming an earliest minimal occurrence.

2. For the constraints of type #2 or #3, $\xi^{previous}$ and $\xi^{current}$ are combined so as to get the strongest possible constraint and gain as much information as possible. To this aim, their respective sets of candidate event types are intersected and the highest probability lower bound is retained.

More formally, since $\xi_{previous}$ can be of type #2 or #3, $\xi_{previous}$ is written:

$$\xi_{previous} = \begin{cases} \mathbb{E}_C^{previous} \subset \mathbb{E} \\ \exists e_c \in \mathbb{E}_C \text{ such that } \Pr(X = e_c) \geq \rho_{min}^{previous} \end{cases}$$

The same holds for $\xi_{current}$:

$$\xi_{current} = \begin{cases} \mathbb{E}_C^{current} \subset \mathbb{E} \\ \exists e_c \in \mathbb{E}_C \text{ such that } \Pr(X = e_c) \geq \rho_{min}^{current} \end{cases}$$

Finally, ξ is obtained by combining $\xi_{previous}$ and $\xi_{current}$ as follows:

$$\xi = \begin{cases} \mathbb{E}_C = \mathbb{E}_C^{previous} \cap \mathbb{E}_C^{current} \\ \exists e_c \in \mathbb{E}_C \mid \Pr(X = e_c) \geq \rho_{min} \; with \; \rho_{min} = \max\{\rho_{min}^{previous}, \rho_{min}^{current}\} \end{cases}$$

Table 2 summarizes the possible combinations and the resulting constraint types that can be obtained. Note that:

1. Combining constraints of type #3 can produce constraints of type #2 thanks to the intersection of the candidate sets.

2. Combining a constraint of type #2 with a constraint of type #3 always leads to a constraint of type #2.

Table 2: Constraint type combinations and resulting constraint types

$\xi_{previous}$ type	$\xi_{current}$ type	ξ type
uniform	1	1
uniform	2	2
uniform	3	3
1	1, 2, 3	1
2	1	1
2	2, 3	2
3	1	1
3	2	2
3	3	2, 3

5 Kullback-Leibler divergence minimization under occurrence constraints for distribution refinement

Once a new constraint ξ on a random variable is produced, it is exploited to refine the distribution p of that variable into a new distribution q such that the *Kullback-Leibler divergence (KL divergence)* (Kullback and Leibler, 1951) between p and q is minimum. This allows us to quantify the guaranteed amount of information brought by ξ. Let I be a set of symbols. The KL divergence, also termed *relative entropy*, is defined as follows:

$$D(q||p) = \sum_{\lambda \in I} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right)$$

By convention, $0 \log_2 \frac{0}{0} = 0$, $0 \log_2 \frac{0}{p} = 0$ and $p \log_2 \frac{q}{0} = \infty$. This measure is always positive and equals 0 if and only if distributions p and q are identical. It is not a distance measure as it is not symmetric and it does not obey the triangle inequality. According to Rényi (1961), $D(q||p)$ quantifies the *obtained information* if distribution

p is replaced by distribution q.

The distribution q replacing distribution p and minimizing the KL divergence under constraint ξ is established with Theorem 1. The latter favors distributions such that the candidate event types all have the same probability and such that the event types that are not candidates all have the same probability as well. Minimizing the KL divergence ensure that a maximum of entropy is preserve and thus that the distribution that will be chosen is one corresponding to the minimum guarantied gain of information.

Theorem 1 *Minimization of the Kullback-Leibler divergence under partial probability sum constraints*

Let $I = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ be a set of n symbols. Let I' be a subset of I. Let X and Y be two real numbers in $]0,1[$. Let p and q two distributions over I such that $\forall i \in \{1, \ldots, n\},\ p(\lambda_i) > 0 \wedge q(\lambda_i) > 0$. If p and q satisfy the constraints $\sum_{\lambda \in I'} p(\lambda) = X$ and $\sum_{\lambda \in I'} q(\lambda) = Y$, then the Kullback-Leibler divergence is minimum if and only if p and q obey the following equation:

$$\frac{p(\lambda)}{q(\lambda)} = \begin{cases} \frac{X}{Y}, \ \textit{if } \lambda \in I' \\ \frac{1-X}{1-Y}, \ \textit{if } \lambda \in I \setminus I' \end{cases} \tag{1}$$

In that case, the Kullback-Leibler divergence value is:

$$D(q||p)_{min} = \log_2 \left[\left(\frac{Y}{X} \right)^Y \times \left(\frac{1-Y}{1-X} \right)^{1-Y} \right] \tag{2}$$

Proof Let us suppose that p and q satisfy the two constraints $\sum_{\lambda \in I'} p(\lambda) = X$ and $\sum_{\lambda \in I'} q(\lambda) = Y$. The Kullback-Leibler divergence can be written as:

$$\begin{aligned} D(q||p) &= \sum_{\lambda \in I} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) \\ &= \sum_{\lambda \in I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) + \sum_{\lambda \in I \setminus I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) \end{aligned} \tag{3}$$

If both sides are negated, it becomes:

$$\begin{aligned} -D(q||p) &= -\sum_{\lambda \in I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) - \sum_{\lambda \in I \setminus I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) \\ &= \sum_{\lambda \in I'} q(\lambda) \log_2 \left(\frac{p(\lambda)}{q(\lambda)} \right) + \sum_{\lambda \in I \setminus I'} q(\lambda) \log_2 \left(\frac{p(\lambda)}{q(\lambda)} \right) \end{aligned} \tag{4}$$

Then, if both sides are exponentiated using base 2:

$$2^{-D(q||p)} = \left(\prod_{\lambda \in I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)} \right) \times \left(\prod_{\lambda \in I \setminus I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)} \right) \tag{5}$$

Since Y is in $]0, 1[$, the last equation can be transformed into the following one:

$$2^{-D(q||p)} = \left(\sqrt[Y]{\left(\prod_{\lambda \in I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)} \right)} \right)^{Y} \times \left(\sqrt[1-Y]{\left(\prod_{\lambda \in I \setminus I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)} \right)} \right)^{1-Y} \tag{6}$$

According to the inequality of the weighted arithmetic and geometric means (Kazarinoff, 1961), if non-negative numbers $x_1, x_2, \cdots, x_n$ and non-negative weights $w_1, w_2, \cdots, w_n$ such as $w = w_1 + w_2 + \cdots + w_n > 0$ are considered, then:

$$\sqrt[w]{x_1^{w_1} x_2^{w_2} \cdots x_n^{w_n}} \leq \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w} \tag{7}$$

This inequality becomes equality if and only if $x_1 = x_2 = ... = x_n$.

By using Inequation 7 and identifying the ratios $\frac{p(\lambda)}{q(\lambda)}$ to the numbers x_i and the probabilites $q(\lambda)$ to the weights w_i, we obtain:

$$\sqrt[\sum_{\lambda \in I'} q(\lambda)]{\prod_{\lambda \in I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)}} \leq \frac{\sum_{\lambda \in I'} q(\lambda) \frac{p(\lambda)}{q(\lambda)}}{\sum_{\lambda \in I'} q(\lambda)} \tag{8}$$

and

$$\sqrt[\sum_{\lambda \in I \setminus I'} q(\lambda)]{\prod_{\lambda \in I \setminus I'} \left(\frac{p(\lambda)}{q(\lambda)} \right)^{q(\lambda)}} \leq \frac{\sum_{\lambda \in I \setminus I'} q(\lambda) \frac{p(\lambda)}{q(\lambda)}}{\sum_{\lambda \in I \setminus I'} q(\lambda)} \tag{9}$$

Since the constraint $\sum_{\lambda \in I'} q(\lambda) = Y$ is satisfied, $\sum_{\lambda \in I \setminus I'} q(\lambda) = 1 - Y$ is also satisfied. Consequently, the following inequation can be produced:

$$
\begin{aligned}
2^{-D(q||p)} &\leq \left(\frac{\sum_{\lambda \in I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{Y}\right)^{Y} \times \left(\frac{\sum_{\lambda \in I \setminus I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{1-Y}\right)^{1-Y} \\
\Leftrightarrow 2^{-D(q||p)} &\leq \left(\frac{\sum_{\lambda \in I'} p(\lambda)}{Y}\right)^{Y} \times \left(\frac{\sum_{\lambda \in I \setminus I'} p(\lambda)}{1-Y}\right)^{1-Y}
\end{aligned}
\tag{10}
$$

Since the constraint $\sum_{\lambda \in I'} p(\lambda) = X$ is satisfied, $\sum_{\lambda \in I \setminus I'} p(\lambda) = 1 - X$ is also satisfied and the last inequation finally becomes:

$$
\Leftrightarrow 2^{-D(q||p)} \leq \left(\frac{X}{Y}\right)^{Y} \times \left(\frac{1-X}{1-Y}\right)^{1-Y} \tag{11}
$$

According to the inequality of the weighted arithmetic and geometric means, equality is reached if all the ratios $\frac{p(\lambda)}{q(\lambda)}$ such that $\lambda \in I'$ are equal and if all the ratios $\frac{p(\lambda)}{q(\lambda)}$ such that $\lambda \in I \setminus I'$ are equal. Therefore:

$$
\frac{p(\lambda)}{q(\lambda)} = \begin{cases} \frac{X}{Y}, \text{ if } \lambda \in I' \\ \frac{1-X}{1-Y}, \text{ if } \lambda \in I \setminus I' \end{cases} \tag{12}
$$

In addition, Inequality 10 can be re-written as follows:

$$
D(q||p) \geq \log_2 \left[\left(\frac{Y}{X}\right)^{Y} \times \left(\frac{1-Y}{1-X}\right)^{1-Y}\right] \tag{13}
$$

Consequently, the minimum value of $D(q||p)$ is:

$$
D(q||p)_{min} = \log_2 \left[\left(\frac{Y}{X}\right)^{Y} \times \left(\frac{1-Y}{1-X}\right)^{1-Y}\right] \tag{14}
$$

This value is obtained if and only if distributions p and q obey Equation 12.

□

5.1 Refinement of p into q with a type #1 constraint

Let ξ be a constraint of type #1 imposed by a partial unimodal probabilistic event $\langle t, e, \rho \rangle$. In this case, $\mathbb{E}_C = \{e\}$, $q(e) = \Pr(e) = \rho$. By setting $I = \mathbb{E}$, $I' = \{e\}$, $X = p(e)$ and $Y = \rho$, and by applying Equation 1, it is then possible to determine the probabilities of the remaining event types that minimize the KL divergence between

the previous distribution p and the new one q.

In the case where $q(e)$ =1, Theorem 1 does not apply. The probability of the other event types is then simply set to 0 and the KL divergence is directly calculated using the definition of $D(q||p)$ and the associated conventions.

5.2 Refinement of p into q with a type #2 constraint

Let ξ be a constraint of type #2 imposed by propagation such that $\mathbb{E}_C = \{e_{remainder}\}$ and $\Pr(e_{remainder}) \geq \rho_{min}$. In order to obtain distribution q such that it minimizes the KL divergence, we set $q(e_{remainder}) = \rho_{min}$. The probabilities of the other symbols and the corresponding minimum KL divergence are then determined using the strategy employed for type #1 constraints (see Section 5.1).

Setting $q(e_{remainder})$ to ρ_{min} can be justified as follows: let us consider $I' = \{e_{remainder}\}$ such that $I' \subset I$ with $I = \mathbb{E}$. Then, let us consider $X = \sum_{\lambda \in I'} p(\lambda)) = p(e_{remainder})$ and $Y = \sum_{\lambda \in I'} q(\lambda) = q(e_{remainder})$. Both quantities are in $]0, 1[$[4]. Since obtaining a constraint of type #2 by combining two constraints leads to a situation such that the value of ρ_{min} always increases or remains the same, then $Y \geq X$. Let us consider $\Theta(Y)$, the minimum value of the KL divergence expressed as a function of Y:

$$\begin{aligned} \Theta(Y) &= \log_2 \left[\left(\frac{1-Y}{1-X} \right)^{1-Y} \times \left(\frac{Y}{X} \right)^{Y} \right] \\ &= (1-Y) \log_2 \left(\frac{1-Y}{1-X} \right) + Y \log_2 \frac{Y}{X} \end{aligned} \tag{15}$$

[4]The extreme values (0 and 1) can be treated separately in a similar way to the special case mentioned in Section 5.1.

Its derivative is:

$$\begin{aligned}\frac{d\Theta(Y)}{dY} &= \log_2\left(\frac{1-X}{1-Y}\right) - (1-Y)\left(\frac{1}{(1-Y)ln2}\right) \\ &+ \log_2\frac{Y}{X} + Y\frac{1}{Y\ln 2} \\ &= \log_2\left(\frac{1-X}{1-Y}\right) - \frac{1}{ln2} + \log_2\frac{Y}{X} + \frac{1}{\ln 2} \\ &= \log_2\left(\frac{1-X}{1-Y}\times\frac{Y}{X}\right)\end{aligned} \tag{16}$$

Thus:

$$\frac{d\Theta(Y)}{dY}\begin{cases} < 0, & \text{if } Y \in]0, X[\\ = 0, & \text{if } Y = X \\ > 0, & \text{if } Y \in]X, 1[\end{cases}$$

Since $Y \geq X$ and $Y = \sum_{\lambda \in I'} q(\lambda) = q(e_{remainder})$, then the minimum value of the KL divergence is obtained for the smallest possible value of $Y = q(e_{remainder})$ satisfying constraint ξ, i.e. $q(e_{remainder}) \geq \rho_{min}$. In other words, we set $q(e_{remainder})$ to ρ_{min}.

5.3 Refinement of p into q with a type #3 constraint

Let ξ be a type #3 constraint such that:

$$\xi = \begin{cases} \mathbb{E}_C \subset \mathbb{E} \; with \; |\mathbb{E}_C| \geq 2 \\ \exists e \in \mathbb{E}_C \; such \; that \; \Pr(e) \geq \rho_{min} \end{cases}$$

Since the element e of $\mathbb{E}_C$ such that $\Pr(e) \geq \rho_{min}$ is unknown, we cannot reuse the strategies employed for constraint types #2 and #3 to obtain, in a deterministic way, the distribution q that minimizes the KL divergence. Therefore, as we aim to establish the minimum gain of information brought by the knowledge of the pattern occurrences, a relaxed version of this constraint is considered. It is noted ξ' and defined as follows:

$$\xi' = \begin{cases} \mathbb{E}_C \subset \mathbb{E} \; with \; |\mathbb{E}_C| \geq 2 \\ \sum_{e \in \mathbb{E} \setminus \mathbb{E}_C} \Pr(e) \leq 1 - \rho_{min} \end{cases}$$

Since the probability is non-negative, if a distribution obeys ξ, then it also obeys ξ'. The converse is not true. Such a relaxed constraint is said to be of type #3'.

For example, let us s consider $\mathbb{E} = \{A, B, C, D\}$ and the following constraint of type #3:

$$\xi = \begin{cases} \mathbb{E}_C = \{A, B\} \\ \exists e \in \mathbb{E}_C \text{ such that } \Pr(e) \geq 0.4 \end{cases}$$

This constraint ξ specifies that the candidate set $\mathbb{E}_C$ contains two symbols A and B, and that the probability of occurrence of the symbol associated with the corresponding partial unimodal probabilistic event[5] is at least equal to 0.4. The relaxed constraint corresponding to ξ will be:

$$\xi' = \begin{cases} \mathbb{E}_C = \{A, B\} \\ \sum_{e \in \{C,D\}} \Pr(e) \leq 0.6 \end{cases}$$

Any distribution that satisfies ξ, e.g., $\Pr(A) = 0.5, \Pr(B) = 0.2, \Pr(C) = 0.1, \Pr(D) = 0.2$, also satisfies ξ'. The reverse is not true. For example, the distribution with the following probabilities $\Pr(A) = 0.2, \Pr(B) = 0.3, \Pr(C) = 0.2, \Pr(D) = 0.3$ satisfies ξ' but does not satisfy ξ'.

Once ξ' is defined, the new distribution that satisfies it and that minimizes the KL divergence can be established along with the value of the minimal KL divergence. Let us consider $I = \mathbb{E}$, and $I' = \mathbb{E}_{rejected} = \mathbb{E} \setminus \mathbb{E}_C$, the set of the *rejected* event types, i.e.

[5] This symbol is among the two candidate symbols.

the event types that can not be candidates. Let $X = \sum_{\lambda \in I'} p(\lambda)$ and $Y = \sum_{\lambda \in I'} q(\lambda)$ be the probability sums of the rejected event types. The constraint of type $\#3'$ then implies $Y \leq 1 - \rho_{min}$. According to the derivative of the function $\Theta(Y)$ expressing the minimal value of the KL divergence (see Section 5.3), the minimum quantity of information is obtained for a value of Y set as follows:

- If $1 - \rho_{min} \geq X$, we set $Y = X$ to reach the minimum value of $\Theta(Y)$ while satisfying $Y \leq 1 - \rho_{min}$.
- If $1 - \rho_{min} < X$, the minimum value of $\Theta(Y)$ that can be reached while satisfying $Y \leq 1 - \rho_{min}$ is obtained for the highest possible value of Y, i.e. $Y = 1 - \rho_{min}$.

Once Y is set, the probability of each one of the event types for the new distribution and the corresponding minimal KL divergence are computed by using Theorem 1.

6 Toy example

This Section is aimed at giving a simple example of the selection method. It relies on Example 1. Let us suppose that we want to select the two most informative patterns from the set of the closed sequential patterns, i.e. $\{A \rightarrow A, B \rightarrow A, C \rightarrow A, B \rightarrow C \rightarrow A, B \rightarrow C \rightarrow C\}$ (cf. Section 2). Let us consider the set of the random variables X_j^i such that $i \in \{1, 2, 3, 4\}$ represents the sequence identifier and such that $j \in \{1, 2, 3, 4\}$ refers to the j^{th} partial unimodal probabilistic event type of that sequence. Initially all random variables are uniform. The information gain of each pattern is established by refining these distributions, which is done by minimizing the KL divergence under the constraints imposed by the earliest minimal occurrences of the pattern. The computation of the information gain is detailed hereafter for each pattern. For the first iteration, column *constraint* contains the constraints due to pattern occurrences. They are directly applied since the uniform constraints representing the initial knowledge are discarded as soon as new constraints are discovered. For constraints of type #3, only the relaxed version

is given. Columns p and q give the probabilities of each distribution in the following order: $p(A), p(B), p(C)$ and $q(A), q(B), q(C)$.

The earliest minimal occurrences of $A \rightarrow A$ are: $\langle 3, \langle\langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.7\rangle$, $\langle 3, \langle\langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle, 0.7\rangle, \langle 4, \langle\langle t_2, A, 0.7\rangle, \langle t_4, A, 0.7\rangle\rangle, 0.7\rangle$. Thus:

$\Delta(A \rightarrow A) = 2.230653311/2 = 1.1153266555$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
3	X_2^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.9, 0.05, 0.05$	1.015966907
	X_3^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
	X_3^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
	X_4^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.8, 0.1, 0.1$	0.663034406
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				not counted	not counted
4	X_2^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
	X_3^4	#3': $\mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.3$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.35, 0.35$	0.003671601
	X_4^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	0.811014803
				$\sum \Delta_{occ}$	2.230653311

Since $A \rightarrow A$ occurs twice in sequence 3, the most informative is be chosen, i.e. the one represented by X_2^3 and X_3^3. The information brought by the other occurrence is thus not counted and its constraints are not saved.

The earliest minimal occurrences of $B \rightarrow A$ are: $\langle 1, \langle\langle t_1, B, 0.4\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.4\rangle$, $\langle 2, \langle\langle t_1, B, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle, \langle 3, \langle\langle t_1, B, 0.9\rangle, \quad \langle t_2, A, 0.9\rangle\rangle, 0.9\rangle, \quad \langle 4, \langle\langle t_3, B, \quad 0.8\rangle$, $\langle t_4, A, 0.7\rangle\rangle, 0.7\rangle$. Thus:

$\Delta(B \to A) = 3.728284049/4 = 0.93207101225$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
1	X_1^1	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	#2: $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^1	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	0.431695414
2	X_1^2	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	#2 : $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^2	#2: $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.197948814
3	X_1^3	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.9, 0.05$	1.015966907
	X_2^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.9, 0.05, 0.05$	1.015966907
				$\Delta_{occ} = \sum D(q\|\|p)$	2.031933814
4	X_3^4	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.8, 0.2$	0.663034406
	X_4^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				$\sum \Delta_{occ}$	3.728284049

The earliest minimal occurrences of $C \to A$ are: $\langle 1, \langle \langle t_2, C, 0.9 \rangle, \langle t_3, A, 0.7 \rangle \rangle, 0.7 \rangle$, $\langle 2, \langle \langle t_3, C, 0.4 \rangle, \langle t_4, A, 0.5 \rangle \rangle, 0.4 \rangle, \langle 4, \langle \langle t_1, C, 0.8 \rangle, \langle t_2, A, 0.7 \rangle \rangle, 0.7 \rangle$. Thus:

$\Delta(C \to A) = 2.585318922/3 = 0.861772974$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
	X_2^1	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	$1,015966907$
1	X_3^1	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
	X_3^2	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
2	X_4^2	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.098974407
	X_1^4	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.1, 0.8$	0.663034406
4	X_2^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				$\sum \Delta_{occ}$	2.585318922

The earliest minimal occurrences of $B \to C \to A$ are: $\langle 1, \langle \langle t_1, B, 0.4 \rangle, \langle t_2, C, 0.9 \rangle, \langle t_3, A, 0.7 \rangle \rangle, 0.4 \rangle, \langle 2, \langle \langle t_1, B, 0.5 \rangle, \langle t_2, C, 0.5 \rangle, \langle t_4, A, 0.5 \rangle \ \rangle, 0.4 \rangle$. Thus:

$\Delta(B \to C \to A) = 1.688537918/2 = 0.844268959$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
1	X_1^1	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	1.015966907
	X_3^1	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.433650415
2	X_1^2	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	#1 : $\mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
	X_3^2	#3': $\mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.6$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	0
	X_4^2	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.254887503
				$\sum \Delta_{occ}$	1.688537918

The earliest occurrences of pattern $B \to C \to C$ are: $\langle 1, \langle \langle t_1, B, 0.4 \rangle, \langle t_2, C, 0.9 \rangle, \langle t_4, C, 0.5 \rangle \rangle, 0.4 \rangle, \langle 2, \langle \langle t_1, B, 0.5 \rangle, \langle t_2, C, 0.5 \rangle, \langle t_3, C, 0.4 \rangle \ \rangle, 0.4 \rangle$. Thus:

$\Delta(B \rightarrow C \rightarrow C) = 1.298878222/2 = 0.649439111$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
1	X_1^1	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	1.015966907
	X_3^1	#3': $\mathbb{E}_C = \{A, B\}, Pr(C) \leq 0.6$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	0
	X_4^1	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	1.114941314
2	X_1^2	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
	X_3^2	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
				$\Delta_{occ} = \sum D(q\|\|p)$	0.183936908
				$\sum \Delta_{occ}$	1.298878222

Therefore, the first pattern to be selected is $A \rightarrow A$ since its information gain overcomes other pattern information gains. The distributions of the random variables that are affected by the constraints imposed by its occurrences are replaced by the ones computed and selected when estimating its information gain. Its occurrence constraints are also saved to be combined with the occurrence constraints of the remaining patterns during the next iteration. For the latter, column *constraint* contains the newly discovered constraint as well as the constraint due to $A \rightarrow A$ (marked with '*'). The constraint combinations are performed according to their types (see Section 4.2). For this toy example, they always imply a constraint of type #1. Produced constraints are therefore not mentioned since they are the same as the type #1 constraints that are combined. Results are as follows:

$\Delta(B \to A) = 2.259639254/4 = 0.5649098135$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
1	X_1^1	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	#2: $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^1	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	$0, 403671601$
				$\Delta_{occ} = \sum D(q\|\|p)$	0.431695414
2	X_1^2	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	#2 : $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^2	#2: $\mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.197948814
3	X_1^3	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.9, 0.05$	1.015966907
	X_2^3	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$0.9, 0.05, 0.05$	$0.9, 0.05, 0.05$	0
		#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.9$*			
				$\Delta_{occ} = \sum D(q\|\|p)$	1.015966907
4	X_3^4	#1: $\mathbb{E}_C = \{B\}, Pr(B) = 0.8$	$0.3, 0.35, 0.35$	$\frac{6}{65}, 0.8, \frac{7}{65}$	0.614028119
		#3': $\mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.3$*			
	X_4^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$0.7, 0.15, 0.15$	$0.7, 0.15, 0.15$	0
		#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$*			
				$\Delta_{occ} = \sum D(q\|\|p)$	0.614028119
				$\sum \Delta_{occ}$	2.259639254

$\Delta(C \rightarrow A) = 2.181647322/3 = 0.727215774$					
sid	X_j^i	constraint (type: details)	p	q	$D(q\|\|p)$
1	X_2^1	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	$1,015966907$
	X_3^1	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
2	X_3^2	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.098974407
4	X_1^4	#1: $\mathbb{E}_C = \{C\}, Pr(C) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.1, 0.8$	0.663034406
	X_2^4	#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$0.7, 0.15, 0.15$	$0.7, 0.15, 0.15$	0
		#1: $\mathbb{E}_C = \{A\}, Pr(A) = 0.7^*$			
				$\Delta_{occ} = \sum D(q\|\|p)$	$0,663034406$
				$\sum \Delta_{occ}$	2.181647322

The information gain of $B \rightarrow C \rightarrow A$ is still $\Delta(B \rightarrow C \rightarrow A) = 0.844268959$ since it occurs in partial unimodal probabilistic event sequences where $A \rightarrow A$ does not appear. The same holds for pattern $B \rightarrow C \rightarrow C$, with $\Delta(B \rightarrow C \rightarrow C) = 0.649439111$. Thus, the second and last iteration selects $B \rightarrow C \rightarrow A$. At this stage, the set of selected patterns is $\{A \rightarrow A, B \rightarrow C \rightarrow A\}$. Since it contains two patterns and since we asked for the two most informative one, the selection algorithm ends. With this example, it can be observed that the method selects patterns that are complementary since they are supported by distinct partial unimodal probabilistic events.

7 Experiments

As stated in Section 3, we recall as a preamble that the method proposed in this paper can select any type of sequential patterns, e.g., frequent or closed ones, as long

as they are extracted from a PUPBoS. For the sake of clarity, all of our experiments are focused on frequent sequential patterns.

After having described the datasets and the prototypes we rely on in Section 7.1, the scalability of the method is studied in Section 7.2. The information gain and the complementarity of selected patterns are further assessed in Section 7.3. Finally, Section 7.4 provides qualitative results showing that relevant patterns are selected.

7.1 Datasets and prototypes

7.1.1 Synthetic datasets

Inspired by the work of Muzammal and Raman on probabilistic base of sequences (Muzammal and Raman, 2015), synthetic PUPBoS are generated by relying on the IBM Quest synthetic data generator (Agrawal and Srikant, 1995). Its source code can be downloaded from Fournier-Viger's website (Agrawal and Srikant, 2024). This prototype can generate, among other dataset types, base of sequences as defined for the deterministic case. We modified this generator to output sequences of unimodal probabilistic events together with probabilities for modes where these probabilities were sample from *beta* distributions.

Such *beta* distributions are convenient for 1) they are continuous distributions defined on bounded intervals, 2) convex and concave shapes can be considered and 3) the expectation as well as the variance of a *beta* distribution can be simply expressed using shape parameters (Cramer, 1999).

In the following, we stick to the naming conventions adopted by Agrawal and Srikant in Agrawal and Srikant (1995) to mention the original parameters of the IBM Quest generator that are used, i.e., N the number of possible event types, $|C|$ the average number of events per sequence, $|D|$ the number of sequences (in thousands),

N_S the number of potential frequent sequential patterns, and $|S|$ the average length of potential frequent sequential patterns. The potential frequent sequential patterns and their expected relative supports are built from equiprobable event types using constant random seeds. As done in Muzammal and Raman (2015), N_S and $|S|$ are fixed and set to 100 and 7 respectively. Consequently, the potential frequent patterns employed to synthesize datasets are identical whatever the dataset generation settings that are adopted. Remaining IBM Quest parameters are either irrelevant in a PUPBoS context or resort to default values. They are thus not mentioned explicitly hereafter. Table 3 gives the values used to generate synthetic datasets for parameters $|C|$, $|D|$ and N .

Table 3: Experimental settings for parameters $|C|$, $|D|$ and N.

parameter	minimum	maximum	sampling step
$\|C\|$	10	90	20
$\|D\|$	50	250	50
N	3	43	10

Additional new parameters are the *beta* distribution shape parameters, namely α and β. Whatever the value of N, five different shape configurations are adopted as reference to sample probabilities of the modes from *beta* distributions with support $]\frac{1}{N}, 1]$. As an example, Figure 1 represents these five configurations for $N = 3$. The expectation and the variance of a random variable X generated by sampling these distributions are supplied in Table 4. More details about this sampling from these *beta* distributions can be found in Johnson et al (1994) and Walck (1996).

Following the naming convention of Zaki (2001), each synthetic PUPBoS is named *CiDjK* where i is the average number of events per sequence and j gives the number of sequences (in thousands). In addition, each synthetic dataset name is postfixed by *-AaBb-Nn* with a the value of shape parameter α, b the value of shape parameter β,

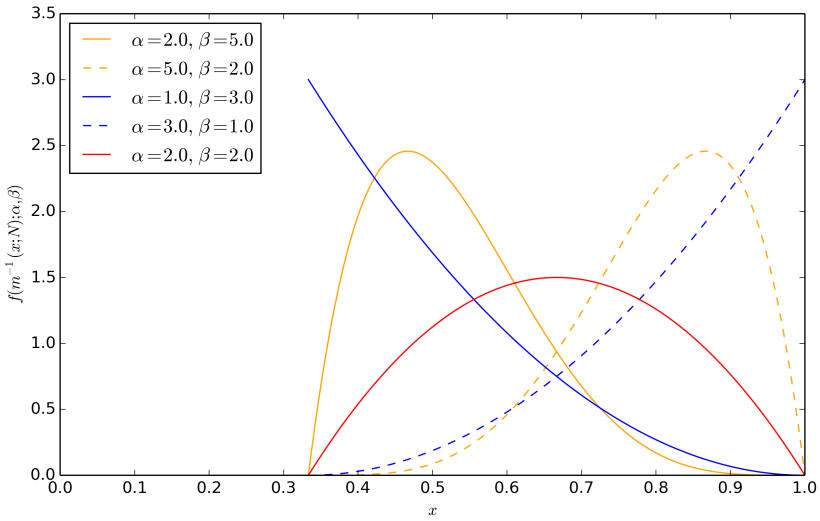

Fig. 1: The five shape configurations of the probability density functions of the reference *beta* distributions for $N = 3$

and n the number of possible event types. For example, the dataset *C10D50K-A2B2-N3* is built using 3 event types, has on average 10 events per sequence and contains 50K sequences. The mode probabilities of its partial unimodal probabilistic events obey a *beta* distribution whose shape parameters α and β have the same value, i.e. 2. These datasets can be freely accessed (Nguyen and Méger, 2024b).

Table 4: Expectation and variance of X for the shape parameters of the five reference *beta* distributions and $N = 3$

α	2	5	1	3	2
β	5	2	3	1	2
$E(X)$	0.524	0.809	0.5	0.833	0.667
$var(X)$	0.011	0.011	0.017	0.017	0.022

7.1.2 Greenland dataset

A symbolic *Displacement Field Time Series (DFTS)* used in Nguyen et al (2018a) to explore the behavior of the Greenland Ice Sheet is transformed into a PPUBoS to assess the method proposed in this paper[6]. This dataset covers a part of the western Greenland Ice Sheet with a square grid of 458×500 cells. Each cell relates to a 240×240 m location and supplies a description of the displacement observed at the Earth surface for that location and for each one of 20 contiguous time periods spanning three decades, from 1985 to 2013. Each location is thus characterized by a sequence of 20 displacement descriptions. Each displacement description provides a period number which is interpreted as a date, a symbol denoting a displacement magnitude level ('1' = low, '2' = medium, '3' = high) and a confidence measure defined over $]0, 1]$ and expressing to which extent the symbol that is supplied is the correct one. Symbols '1', '2' and '3' are equiprobable at the scale of the whole dataset. A special symbol, '0', is also used if no displacement magnitude level is available because of acquisition conditions. In that case, the corresponding period number (date) remains reported while the associated confidence measure is set to a dedicated negative value. For more details about this symbolic DFTS, the reader is referred to Nguyen et al (2018a).

In this paper, the symbol given by a displacement description is assumed to be the most likely one and the corresponding mode probability is obtained by mapping its confidence measure x from $]0, 1]$ to $]\frac{1}{3}, 1]$ with function m defined by Equation 17 and setting $N = 3$.

$$m(x; N) = x\ (1 - \frac{1}{N}) + \frac{1}{N} \tag{17}$$

Displacement descriptions containing symbol '0', and thus negative confidence values, are left unchanged since they represent missing data. In other words, after

[6] We thank the authors of Tedstone et al (2015) for making the original numerical data available to us.

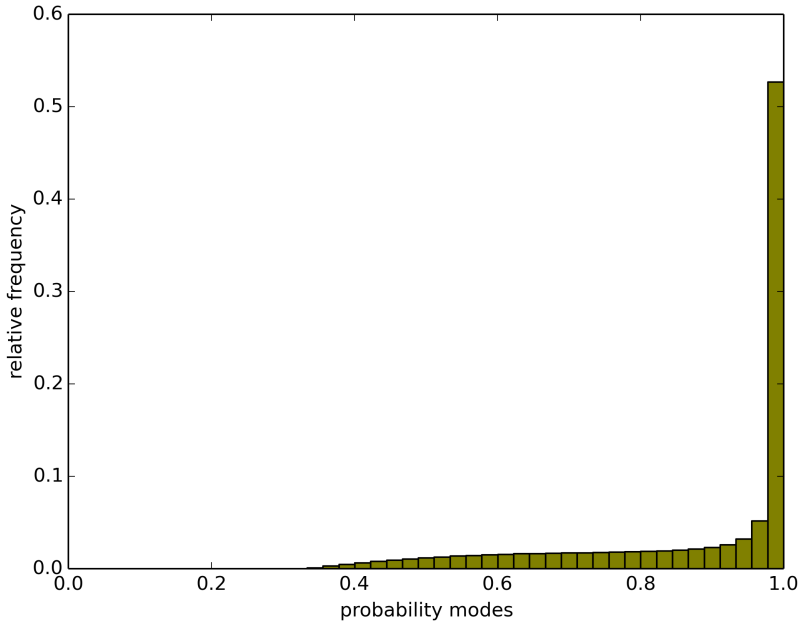

Fig. 2: Relative frequency histogram of the mode probabilities of the Greenland dataset

having replaced the positive confidence measures of the original symbolic DFTS by the probabilities of the modes, we obtain a PUPBoS containing 229,000 (458×500 locations) sequences such that each sequence contains 20 partial unimodal probabilistic events. About 30% of the dataset contains missing data that are discarded when extracting patterns. As for any event, the proposed pattern selection method initially assumes that the probability mass functions of the missing data over symbols '1', '2' and '3' are uniform. The distribution of the mode probabilities of the Greenland dataset is given by the relative frequency histogram depicted in Figure 2. The mean and the variance of the mode probabilities are 0.879 and 0.170 respectively. This dataset can be freely accessed (Nguyen and Méger, 2024b).

7.1.3 Prototypes

Any extractor designed for deterministic bases of sequences can be exploited to mine frequent sequential patterns from a PUPBoS. As explained in Section 2, mining the deterministic base of sequences obtained by removing mode probabilities of a PUPBoS provides a set of deterministic frequent sequential patterns equal to the set of frequent sequential patterns, as defined in Section 2, that can be extracted from the original PUPBoS using the same minimum support threshold. We thus use *DMT4SP (Data Mining Tool 4 Sequential Patterns)* (Rigotti, 2024), a prototype based on the *PrefixGrowth* algorithm (Pei et al, 2007).

Once frequent sequential patterns are extracted using *DMT4SP*, their earliest minimal occurrences, as defined in Section 2, are identified. The patterns are further selected according to their occurrences by applying the information gain-based method proposed in this paper. These two steps are performed using our own implementation, written in C/C++ and named *Complementary and Informative Sequential Patterns (CISP)*. It can be obtained from its URL (Nguyen and Méger, 2024a) and run on Linux platforms. All experiments have been performed on an Intel Xeon 3.5 GHz server running Linux (Ubuntu).

7.2 Scalability

The scalability study of Algorithm 1 is performed according to its worst-case space and time complexities.

An upper bound of the worst-case time complexity is $\mathcal{O}(|P| \times K \times |D| \times |C|_m^2)$, with $|P|$ the number of patterns extracted from a PUPBoS, K the number of patterns to select, $|D|$ the number of sequences forming the PUPBoS, and $|C|_m$ the maximum

number of events forming a sequence. More details about this upper bound can be found in Appendix I. Algorithm 1 can therefore be time-consuming, especially for large datasets, containing long sequences, and large collections of patterns to analyse. However, since for each iteration the information brought by the occurrences of a pattern is evaluated independently of the occurrences of the other patterns, this computation can be easily parallelized. For the sake of the evaluation, *CISP* is run in a single thread that is bound to a single core. Regarding memory consumption, the worst-case space complexity is $\mathcal{O}(|D| \times (|P| + |C|_m))$. The reader is referred to Appendix I for more details about it. All experiments are run by making sure that no memory swaps are triggered.

7.2.1 Dataset impact

Resources consumption is first evaluated by varying two parameters of the IBM Quest synthetic data generator: $|D|$, the number of sequences and $|C|$, the average number of events forming sequences. Their values are given in Section 7.1.1 by Table 3. Notice that $|C|_m$ is likely to increase with higher values of $|C|$. Since the number of event types N and the distribution of mode probabilities have no impact on performances, N is arbitrarily set to 3 symbols while the shape parameters of mode probabilities, α and β (see Section 7.1.1), are both set to 2. Finally, processing parameters are fixed: K is set to 20 and P, the set of sequential patterns extracted from the dataset, is set to the collection of potential frequent sequential patterns being used to synthesize datasets. This set is the same whatever the dataset that is considered. Indeed, it always contains $|P| = N_S = 100$ patterns of average length $|S| = 7$ that are generated from $N = 3$ equiprobable event types using the same random seeds (see Section 7.1.1). Such a choice allows to assess the impact of the dataset parameters only and avoid wasting energy with the extraction of frequent

sequential patterns from each one of the datasets.

According to Figure 3 and Figure 4, execution times linearly vary with respect to $|D|$, if $|C|$ is fixed, while they quadratically increase with $|C|$, if $|D|$ is fixed. Figure 5 and Figure 6 show that memory consumption scales quite linearly with $|D|$, if $|C|$ is fixed, and vice versa. These results are in agreement with the worst-case time and space complexities of Algorithm 1.

7.2.2 Processing parameters impact

The assessment of the processing parameters impact is first performed using dataset *C10D50K-A2B2-N3* as reference (see Section 7.1.1 for naming conventions). Since this evaluation is carried out independently of dataset properties, one of the smallest dataset is chosen to spare useless computations and thus resources. Frequent sequential patterns are extracted from it by varying the support threshold σ to obtain sets whose sizes differ. Each one of these sets is alternately used to initialize P, the set of sequential patterns extracted from the dataset. Table 5 gives the size of these collections w.r.t. σ, the minimum support threshold. Finally, K values range from 20 to 2020 with a sampling step of 500. All configurations such that $K > |P|$ are obviously discarded.

Table 5: $|P|$ vs. σ for *C10D50K-A2B2-N3*

$\|P\|$	125	1218	2233	3269	4238	5215
σ	16%	2%	1.1%	0.8%	0.5%	0.4%

As it can be observed with Figure 7 and Figure 8, execution times are inline with the worst-case time complexity upper bound: they vary less than linearly with respect to K, if $|P|$ is fixed, and vice versa. Regarding memory consumption, as expected and

shown by Figure 9 and Figure 10, K has no impact while the consumption increases less than linearly with $|P|$, which is consistent with the worst-case space complexity. As shown by Figures 11, 12, 13, and 14, all of these observations are also valid for the Greenland dataset for which Table 6 and the same values of K are considered.

Table 6: $|P|$ vs. σ for the Greenland dataset

$\|P\|$	120	1298	2308	3202	4255	5223
σ	30.6%	13.1%	9.6%	7.9%	6.5%	5.7%

7.3 Information gain and complementarity

The information gain of selected patterns is assessed with respect to datasets and processing parameters by simply averaging the information gain obtained for each one of the selected patterns. It is thus expressed as follows:

$$\Delta(\Phi) = \frac{\sum_{\alpha \in \Phi} \Delta(\alpha)}{|\Phi|} \tag{18}$$

Set Φ contains the patterns selected by Algorithm 1, and $\Delta(\alpha)$ denotes the information gain of a sequential pattern α as defined in Section 3. Such an average is chosen to allow comparisons between experiments for which the number of selected patterns differ.

A measure of complementary is also employed to indicate to which extent partial unimodal probabilistic events unveiled by selected patterns are affected by occurrence constraints of a single selected pattern or several selected patterns. Since the overlap of pattern occurrences is not controlled by the Quest generator, this measure is used to assess the impact of processing parameters only. It is denoted $\tau(\Phi)$ and expressed as follows:

$$\tau(\Phi) = \frac{events_{\Phi}}{\sum_{\alpha \in \Phi} events_{\alpha}} \tag{19}$$

Numerator $events_{\Phi}$ indicates the number of partial unimodal probabilistic events affected by at least one constraint introduced by the occurrences of the patterns belonging to Φ, and $events_{\alpha}$ corresponds to the number of events affected by constraints introduced by the occurrences of a pattern α. The value of $\tau(\Phi)$ is in the range $[\frac{1}{|\Phi|}, 1]$. It equals $\frac{1}{|\Phi|}$ if each affected event must satisfy constraints imposed by occurrences of each one of the selected patterns. In that case, no complementarity between the occurrences of the selected pattern is observed. Conversely, it equals 1 if each affected event is refined by constraints originating from a single pattern occurrence. In other words, selected patterns are fully complementary. In order to compare results obtained when varying the number of patterns to select, a rescaled version of $\tau(\Phi)$ is preferred and defined on $[0, 1]$ as follows:

$$T(\Phi) = \frac{\tau(\Phi) - \frac{1}{|\Phi|}}{1 - \frac{1}{|\Phi|}} \tag{20}$$

7.3.1 Dataset impact

As for Section 7.2.1, the dataset impact on the information gain is assessed using synthetic datasets and fixed processing parameters. For the latter, once again, K, the number of pattern to select, is set to 20 while P, the set of patterns extracted from a PUPBoS, is initialized with the 100 potential frequent sequential patterns used to synthesize datasets. Firstly the effect of the mode probabilities is studied by sampling alternately the five *beta* distributions of reference listed in Table 4. This allows us to control the expectation of the mode probabilities. Since the support of these distributions is $]\frac{1}{N}, 1]$ (see Section 2 and Section 7.1), N, the number of event types is also varied according to Table 3. Parameters C and D are set to 10 and 50 respectively to

focus on the smallest possible datasets and avoid useless computations.

Figure 15 shows that $\Delta(\Phi)$ increases with the expectation of the mode probabilities and the number of event types. As a matter of fact, the information gain associated with a pattern quantifies the refinement, imposed by its occurrences, of the distributions of the events it affects. As this gain is established by relying on the Kullbach-Leibler divergence (see Section 3), the more refined distributions differ from original ones, the higher is the gain. Consequently, since event distributions are all assumed to be uniform ones before any refinement, the information gain of a pattern is generally expected to increase with the mode probabilities of the events forming its occurrences. At a larger scale, the information gain of selected patterns, $\Delta(\Phi)$, is thus likely to increase with the mode probabilities of the events forming the occurrences of the patterns belonging to Φ. In addition, the higher N, the higher is the number of event types whose probabilities are refined. The Kullbach-Leibler divergence between initial and refined distribution thus tends to increase with N, and so does $\Delta(\Phi)$.

The impact of the dataset size is evaluated by varying $|C|$, the average number of events per sequence, and $|D|$, the number of sequences. This is performed according to Table 3. For each dataset, the number of event types N is arbitrarily set to 3, and the *beta* distribution whose shape parameters are $\alpha = 2$ and $\beta = 2$ is considered. According to Figure 16, $\Delta(\Phi)$ is independent of $|D|$, whatever the value of $|C|$. The average information gain, $\Delta(\Phi)$, indeed remains stable since 1) it is normalized by the pattern support (see Section 3) and 2) the expected relative supports of the patterns used to synthesize datasets are identical whatever the dataset size (see Section 7.1). Finally, the impact of $|C|$ is not assessed on its own since the spread and the overlap of the pattern occurrences is not controlled by the Quest generator while synthesising them.

7.3.2 Processing parameters impact

As for Section 7.2.2, the assessment of the processing parameters impact is performed using dataset *C10D50K-A2B2-N3* and the Greenland dataset. The higher the number of ranked patterns K, the more it is difficult to find informative, and thus complementary patterns. Algorithm 1 indeed selects the most informative pattern at each iteration by taking account the information already brought by the patterns selected during the previous iterations, i.e. by making sure that selected patterns are complementary from an informational perspective. This behavior is clearly illustrated for both datasets by Figure 17, Figure 18, Figure 19, and Figure 20, where the average information gain and the complementarity are plotted against K. These figures are obtained for all possible K values up to 2020 to detail this behavior at the largest possible scale. Computing these results for each value of K can be performed efficiently by collecting the measures observed for each one the patterns that are ranked during a single experiment for which $K = 2020$. This is performed for the largest set P, to maximize the chance of finding informative patterns. The latter assumption holds for both datasets as shown by Figure 21, Figure 22, Figure 23 and Figure 24 where $\Delta(\Phi)$ and $\tau(\Phi)$ increase with P whatever K. This behavior is best observed for $K = 20$ since the average information gain and the complementarity measure sharply drop towards 0 when increasing K.

Finally, all of the quantitative results advocate for selecting *few* patterns among the largest possible set of patterns since 1) the most informative and complementary ones are the first ones, 2) the space and time resources needed for selecting patterns increase with K, and 3) the chance of finding the most informative and complementary patterns grows with the size of P.

7.4 Qualitative results

7.4.1 Synthetic datasets

Since synthetic datasets are all generated using the same set of 100 potential frequent sequential patterns and the same expected relative supports (see Section 7.1), the same amount of patterns can be extracted from any synthetic dataset by considering the same relative support. In order to avoid useless computations, only the $C10D50 - AaBb - N3$ datasets are retained in this section. For these datasets, if the relative support is set to 2%, i.e. 1000 event sequences, the same 1218 frequent sequential patterns are extracted. Among them, 44 patterns belong to the set of 100 potential frequent sequential patterns generated by Quest to synthesize datatsets, and are the same whatever the dataset. Their median relative support is 8.4% and higher than that of the 1218 patterns whose median relative support is 3.8%.

All 1218 patterns are then supplied to Algorithm 1 to select $K = 100$ patterns. Finally, for each dataset, the number of patterns belonging to both the set of selected patterns and the set of the 44 Quest frequent sequential patterns that were extracted is computed. Table 7 reports this number, termed $n_{\cap}$, for each beta mode distribution. As it can be observed, this number is about a fourth of the 44 Quest patterns. The selection of patterns by Algorithm 1 is thus relevant with respect to the generation of synthetic datasets and avoids concentrating on the most frequent patterns. The latter is confirmed by the median relative support of the 100 patterns selected by Algorithm 1, mrs_{SeqKL}, also reported in Table 7, that is always lower than the one of the 44 Quest patterns.

Table 7: Number of the selected Quest patterns and median relative support of the 100 selected patterns for each one of the $C10D50 - AaBb - N3$ datasets

α β	2 5	5 2	1 3	3 1	2 2
$n_\cap$	11	10	11	10	13
mrs_{SeqKL}	4.5	4.5	4.8	4.4	4.8

As explained in Section 3, the information gain of a pattern is normalized by its support to avoid selecting frequent but not so informative patterns and focus on patterns whose occurrences, even if not numerous, are very informative. Figure 25 and Figure 26 confirms that this behavior is actually adopted by showing that the rank of selected patterns and the rank of the selected Quest patterns is not correlated with their support whatever the dataset that is considered.

7.4.2 Greenland dataset

Regarding the Greenland dataset (see Section 7.1), after having extracted 1298 frequent sequential patterns by setting the relative support threshold to 13.1%, $K = 100$ patterns were selected using Algorithm 1. These patterns are built with 3 symbols denoting displacement magnitude levels ('1' = low, '2' = medium, '3' = high). Among them, pattern $p1 : \ 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 1$ and pattern $p2 : \ 3 \rightarrow 1 \rightarrow 1 \rightarrow 3$ are for example considered of interest. Pattern $p1$ is ranked #1 and exhibits a progressive slowdown of the Greenland Ice Sheet that was reported in Tedstone et al (2015). Pattern $p2$ is ranked #99 and unveils an unknown and more sudden temporary slowdown. As for synthetic datasets, Figure 27 confirms that the information gain of selected patterns is not driven by their supports simply.

These patterns are further located both in space and in time. For each pattern, a map is built by considering the square grid covering the part of the western Greenland Ice Sheet that is analyzed (see Section 7.1). According to the color scale depicted in Figure 30, each affected sequence, and thus each affected grid cell, is associated with a color indicating the ending date of the pattern occurrence. By relying on a Landsat 7 image acquired during the time period for which the dataset has been built, unaffected cells are filled with gray-levels representing the radiometric response of the corresponding locations.

Figure 28 and Figure 29 depict the maps of pattern $p1$ and $p2$ respectively. As it can be observed, they are quite complementary, both spatially and temporally ($T(\Phi) = 0.037$ for the 100 selected patterns). Figure 28 highlights the Nordenskjöld glacier (1) for which a slowdown along one longitudinal transect was exhibited in Tedstone et al (2015). Additionally, it shows that this deceleration pattern can be observed for the Polonia (2) and the Alangordliup Sermia (3) glaciers where, as for the Nordenskjöld glacier, it terminates at the end of series, between 2010 and 2013 (violet and magenta areas). Regarding Figure 29, it shows that pattern $p2$ especially affects the Sarqardliup Sermia glacier (4) where it ends either between 1997 and 1999 (green area) or 2008 and 2009 (blue area). These different timings might be associated with the subglacial relief that is reported in Thorning and Hansen (1987). Figure 29 also points out that $p2$ affects most of the coastal parts of the Greenland Ice Sheet where the glaciers drop down into the ocean. Consequently, beside extracting new knowledge with pattern $p2$, pattern $p1$ allows to complete the knowledge about the already known gradual slowdown by giving its full spatial extent. These selected patterns are thus relevant when analyzing such a real dataset.

8 Conclusion

This work considers the concept of partial unimodal probabilistic bases of sequences. In such a base, each event is certain and its most probable description, i.e. its event type, is provided along with its probability. The probability of all other possible event types are unavailable. This implies that a single possible world is reachable, the most probable one. It is here proposed to mine it to extract frequent sequential patterns from which a subset is selected using an original information gain-based algorithm that outputs complementary and informative patterns. More precisely, events are described by independent discrete random variables whose probability mass functions are initially supposed to be uniform. Once extracted, pattern occurrences and their corresponding mode probabilities are used to constrain and refine initial event distributions. This refinement is performed by minimizing the Kullback-Leibler divergence, between the initial and new distributions, under the revealed constraints, which allows us to quantify and guaranteed the minimum information gain due to the pattern occurrences. This process is achieved in a greedy way by selecting, at each iteration, the pattern leading to the highest information gain knowing the patterns selected during the previous iterations.

Extensive experiments have been run to assess the behavior of the proposed algorithm against the datasets and parameters settings, by using both synthetic and real datasets. They shown that the algorithm scales well while providing informative complementary patterns whose selection is not simply driven by their support. The most informative and complementary patterns being discovered at the very beginning of the process, end-users can ask for a limited set of patterns, say 100 as a maximum for the datasets we used so far, which facilitates their interpretation. Regarding synthetic datasets, a subset of the generative patterns used to form them are selected, which

indicates the interest of the approach. This is corroborated with the real dataset for which a gradual slowdown of the Greenland Ice Sheet is selected as number one. This behavior was known by the experts for a limited set of positions while we here provided its full spatiotemporal extent. In addition to confirming and enriching the user's knowledge with that pattern, an example of an unknown complementary pattern was provided and associated with the subglacial relief through its spatiotemporal behavior. The prototype for selecting patterns and all datasets used in this paper can be downloaded from Nguyen and Méger (2024a,b).

Future works includes generalizing this approach to itemsets and other pattern types, taking into account the spatial and/or temporal autocorrelation, handling several mode probabilities for a single event, avoiding any post-processing by extracting informative and complementary patterns directly, and targeting other applications.

I Worst-case complexity analysis of Algorithm 1

Following the naming conventions adopted by Agrawal and Srikant (1995), a PUPBoS is described by:

- $|C|_m$, the maximum number of events forming an event sequence,
- $|D|$, the number of sequences forming the PUPBoS.

Appendix I.A Time complexity

The information gain brought by the knowledge of an earliest minimal occurrence is computed by processing all the events occurring within its time window. For each event, this gain is measured by combining its current constraint with the new constraint imposed by the occurrence. If this calculation is considered as elementary operation, then there are as many operations as the number of events forming the occurrence for which the information gain is established. For each sequence covered by

a pattern β, an upper bound on the number of earliest minimal occurrence can be obtained by considering that 1) there are as many occurrences as there are events, 2) each occurrence starts with a different event, and 3) each occurrence ends at the end of the sequence. There will therefore be a maximum of $|C|_m$ occurrences containing $|C|_m, |C|_m - 1, |C|_m - 2, \ldots$, and finally 1 events. The number of events to browse is thus:

$$|C|_m + |C|_m - 1 + |C|_m - 2 + \cdots + 1 = \frac{|C|_m \times (|C|_m + 1)}{2} \tag{21}$$

For each pattern β, the maximum support is $|D|$. Let $|P|$ be the number of patterns in P, the set of patterns extracted from a PUPBoS. To select the best pattern at the first iteration, in the worst case, the number of operations that are performed is:

$$|P| \times |D| \times \frac{|C|_m \times (|C|_m + 1)}{2} \tag{22}$$

For the second iteration, since the most informative pattern is excluded from P, the maximum number of operations will be :

$$(|P| - 1) \times |D| \times \frac{|C|_m \times (|C|_m + 1)}{2} \tag{23}$$

For the i^{th} iteration:

$$(|P| - i + 1) \times |D| \times \frac{|C|_m \times (|C|_m + 1)}{2} \tag{24}$$

In total, to select K patterns, in the worst case, the number of operations that are performed is:

$$\begin{aligned} &\sum_{i=1\ldots K} (|P| - i + 1) \times |D| \times \frac{|C|_m \times (|C|_m + 1)}{2} \\ &= \frac{(2|P| - K + 1) \times K}{2} \times |D| \times \frac{|C|_m \times (|C|_m + 1)}{2} \end{aligned} \tag{25}$$

Since $K > 0$, then $2|P| - K > 2|P|$ and thus[7] an upper bound of the worst-case time complexity of the algorithm is:

$$\mathcal{O}(|P| \times K \times |D| \times |C|_m^2) \tag{26}$$

Appendix I.B Space complexity

Each partial unimodal probabilistic event of the considered PUPBoS must be stored along with its current constraint. If such a record is considered as an elementary memory unit, then the maximum amount of memory needed is $|D| \times |C|_m$.

In addition, the occurrences of each pattern in P can be stored during the first iteration to avoid the cost of finding these occurrences again during the next iterations. By assuming that an elementary memory unit can also store such an information, then the maximum amount of memory is obtained by considering that each pattern occurs in each sequence and equals $|D| \times |P|$.

Consequently, the worst-case space complexity is simply:

$$\mathcal{O}(|D| \times (|P| + |C|_m)) \tag{27}$$

[7] Moreover, it can be noticed that in most settings we should have $K \ll P$ and so $2|P| - K \approx 2|P|$.

References

Abiteboul S, Kanellakis P, Grahne G (1987) On the representation and querying of sets of possible worlds. In: Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data. ACM, New York, NY, USA, SIGMOD '87, pp 34–48, https://doi.org/10.1145/38713.38724

Aggarwal CC (2009) Managing and Mining Uncertain Data. Springer Publishing Company, Incorporated

Aggarwal CC, Yu PS (2009) A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering 21(5):609–623. https://doi.org/10.1109/TKDE.2008.190

Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan, pp 3–14, https://doi.org/10.1109/ICDE.1995.380415

Agrawal R, Srikant R (2024) IBM Quest Synthetic Data Generator, accessed February 20, 2024. URL http://www.philippe-fournier-viger.com/spmf/datasets/IBM_Quest_data_generator.zip

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. SIGMOD Rec 22(2):207–216. https://doi.org/10.1145/170036.170072

Bernecker T, Kriegel HP, Renz M, et al (2009) Probabilistic frequent itemset mining in uncertain databases. In: In: SIGKDD'09, pp 119–128, https://doi.org/10.1145/1557019.1557039

Bonchi F, van Leeuwen M, Ukkonen A (2011) Characterizing uncertain data using compression. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA, pp 534–545, https://doi.org/10.1137/1.9781611972818.46

Calders T, Garboni C, Goethals B (2010) Efficient pattern mining of uncertain data with sampling. In: Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I. Springer-Verlag, Berlin, Heidelberg, PAKDD'10, pp 480–487, https://doi.org/10.1007/978-3-642-13657-3_51

Chen H, Ku WS, Wang H, et al (2010) Leveraging spatio-temporal redundancy for rfid data cleansing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, New York, NY, USA, SIGMOD '10, p 51–62, https://doi.org/10.1145/1807167.1807176

Chui C, Kao B, Hung E (2007) Mining frequent itemsets from uncertain data. In: Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp 47–58, https://doi.org/10.1007/978-3-540-71701-0_8

Cormode G, Li F, Yi K (2009) Semantics of ranking queries for probabilistic data and expected ranks. In: Proceedings of the 2009 IEEE International Conference on Data Engineering. IEEE Computer Society, Washington, DC, USA, ICDE '09, pp 305–316, https://doi.org/10.1109/ICDE.2009.75

Cramer H (1999) Mathematical Models for Statistics. Princeton University Press, Princeton

García-Hernández RA, Martínez-Trinidad JF, Carrasco-Ochoa JA (2006) A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. In: Computational Linguistics and Intelligent Text Processing. Springer, Berlin,

Heidelberg, pp 514–523, https://doi.org/10.1007/11671299_53

Ge J, Xia Y, Wang J (2015) Towards efficient sequential pattern mining in temporal uncertain databases. In: Cao T, Lim EP, Zhou ZH, et al (eds) Advances in Knowledge Discovery and Data Mining. Springer International Publishing, Cham, pp 268–279, https://doi.org/10.1007/978-3-319-18032-8_21

Ge J, Xia Y, Wang J, et al (2017) Sequential pattern mining in databases with temporal uncertainty. Knowl Inf Syst 51(3):821–850. https://doi.org/10.1007/s10115-016-0977-1

Green TJ, Tannen V (2006) Models for incomplete and probabilistic information. In: Grust T, Höpfner H, Illarramendi A, et al (eds) Current Trends in Database Technology – EDBT 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 278–296, https://doi.org/10.1007/11896548_24

Hooshsadat M, Bayat S, Naeimi P, et al (2012) Uapriori: An algorithm for finding sequential patterns in probabilistic data. In: Proceedings of the 10th International FLINS Conference. Kahraman C, Bozbura FT, Kerre EE (eds) UMKEDM. World Scientific, pp 907–912, https://doi.org/10.1016/j.jss.2013.03.105

Ibrahim A, Sastry S, Sastry PS (2016) Discovering compressing serial episodes from event sequences. Knowledge and Information Systems 47(2):405–432. https://doi.org/10.1007/s10115-015-0854-3

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous Univariate Distributions - Volume 2. John Wiley and Sons, New York

Kazarinoff N (1961) Analytic Inequalities. Holt, Rinehart and Winston, New York

Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22(1):79–86

Lam HT, Mörchen F, Fradkin D, et al (2014a) Mining Compressing Sequential Patterns. Statistical Analysis and Data Mining: The ASA Data Science Journal 7(1):34–52. https://doi.org/10.1002/sam.11192

Lam HT, Mörchen F, Fradkin D, et al (2014b) Mining compressing sequential patterns. Stat Anal Data Min 7(1):34–52. https://doi.org/10.1002/sam.11192

Leung CKS (2011) Mining uncertain data. Wiley Int Rev Data Min and Knowl Disc 1(4):316–329. https://doi.org/10.1002/widm.31

Leung CKS, Hao B (2009) Mining of frequent itemsets from streams of uncertain data. In: Proceedings of the 2009 IEEE International Conference on Data Engineering. IEEE Computer Society, Washington, DC, USA, ICDE '09, pp 1663–1670, https://doi.org/10.1109/ICDE.2009.157

Leung CKS, Jiang F (2011) Frequent itemset mining of uncertain data streams using the damped window model. In: Proceedings of the 2011 ACM Symposium on Applied Computing. ACM, New York, NY, USA, SAC '11, pp 950–955, https://doi.org/10.1145/1982185.1982393

Leung CKS, Mateo MAF, Brajczuk DA (2008) A tree-based approach for frequent pattern mining from uncertain data. In: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Springer-Verlag, Berlin, Heidelberg, PAKDD'08, pp 653–661, https://doi.org/10.1007/978-3-540-68125-0_61

Leung CKS, MacKinnon RK, Jiang F (2014) Reducing the search space for big data mining for interesting patterns from uncertain data. In: 2014 IEEE International Congress on Big Data, pp 315–322, https://doi.org/10.1109/BigData.Congress.

2014.53

Luo C, Chung SM (2004) A scalable algorithm for mining maximal frequent sequences using sampling. In: Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on. IEEE, pp 156–165, https://doi.org/10.1109/ICTAI.2004.16

Luo C, Chung SM (2005) Efficient mining of maximal sequential patterns using multiple samples. In: Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005, Newport Beach, CA, USA. SIAM, pp 415–426, https://doi.org/10.1137/1.9781611972757.37

Mannila H, Toivonen H, Inkeri Verkamo A (1997) Discovery of frequent episodes in event sequences. Data Mining Knowledge Discovery 1(3):259–289. https://doi.org/10.1023/A:1009748302351

Méger N, Rigotti C, Pothier C (2015) Swap randomization of bases of sequences for mining satellite image times series. In: Proceedings, Part II, of the European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9285. Springer-Verlag New York, Inc., New York, NY, USA, ECML PKDD 2015, pp 190–205, https://doi.org/10.1007/978-3-319-23525-7_12

Méger N, Rigotti C, Pothier C, et al (2019) Ranking evolution maps for Satellite Image Time Series exploration: application to crustal deformation and environmental monitoring. Data Mining and Knowledge Discovery 33(1):131–167. https://doi.org/10.1007/s10618-018-0591-9

Muzammal M, Raman R (2010) On probabilistic models for uncertain sequential pattern mining. In: Proceedings of the 6th International Conference on Advanced Data Mining and Applications: Part I. Springer-Verlag, Berlin, Heidelberg, ADMA'10, pp

60–72, https://doi.org/10.1007/978-3-642-17316-5_6

Muzammal M, Raman R (2011) Mining sequential patterns from probabilistic databases. In: Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II. Springer-Verlag, Berlin, Heidelberg, PAKDD'11, pp 210–221, https://doi.org/10.1007/978-3-642-20847-8_18

Muzammal M, Raman R (2015) Mining sequential patterns from probabilistic databases. Knowl Inf Syst 44(2):325–358. https://doi.org/10.1007/s10115-014-0766-7

Nguyen T, Méger N (2024a) CISP: Complementary and Informative Sequential Patterns, accessed February 20, 2024. URL https://sites.google.com/view/cisp-prototype/

Nguyen T, Méger N (2024b) Partial unimodal probabilistic bases of sequences, accessed February 20, 2024. URL https://sites.google.com/view/cisp-prototype/

Nguyen T, Méger N, Rigotti C, et al (2018a) A pattern-based method for handling confidence measures while mining satellite displacement field time series: Application to greenland ice sheet and alpine glaciers. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(11):4390–4402. https://doi.org/10.1109/JSTARS.2018.2874499

Nguyen T, Méger N, Rigotti C, et al (2018b) Finding complementary and reliable patterns in displacement field time series of alpine glaciers. In: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp 4209–4212, https://doi.org/10.1109/IGARSS.2018.8518969

Pei J, Han J, Mortazavi-Asl B, et al (2004) Mining sequential patterns by pattern-growth: The prefixspan approach. Knowledge and Data Engineering, IEEE

Transactions on 16(11):1424–1440. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1339268

Pei J, Han J, Wang W (2007) Constraint-based sequential pattern mining: the pattern-growth methods. J Intell Inf Syst 28(2):133–160. https://doi.org/10.1007/s10844-006-0006-z

Pelekis N, Kopanakis I, Kotsifakos E, et al (2010) Clustering uncertain trajectories. Knowledge and Information Systems 28:117–147. https://doi.org/10.1007/s10115-010-0316-x

Qian W, Lauri F, Gechter F (2020) A probabilistic approach for discovering daily human mobility patterns with mobile data. In: Lesot MJ, Vieira S, Reformat MZ, et al (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer International Publishing, Cham, pp 457–470, https://doi.org/10.1007/978-3-030-50146-4_34

Raissi C, Poncelet P, Teisseire M (2006) Speed: Mining maximal sequential patterns over data streams. In: International Conference on Intelligent Systems-ICIS. IEEE, pp 1–8, https://doi.org/10.1109/IS.2006.348478

Rényi A (1961) On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol 1: Contributions to the Theory of Statistics. The Regents of the University of California, pp 547–561

Rigotti C (2024) DMT4SP: Data Mining Tool 4 Sequential Patterns, accessed February 20, 2024. URL https://perso.liris.cnrs.fr/christophe.rigotti/dmt4sp.html

Suciu D, Dalvi N (2005) Foundations of probabilistic answers to queries. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.

ACM, New York, NY, USA, SIGMOD '05, pp 963–963, https://doi.org/10.1145/1066157.1066303

Sun L, Cheng R, Cheung DW, et al (2010) Mining uncertain data with probabilistic guarantees. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, KDD '10, pp 273–282, https://doi.org/10.1145/1835804.1835841

Tatti N, Vreeken J (2012) The long and the short of it: summarising event sequences with serial episodes. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 462–470, https://doi.org/10.1145/2339530.2339606

Tedstone AJ, Nienow PW, Gourmelen N, et al (2015) Decadal slowdown of a land-terminating sector of the Greenland Ice Sheet despite warming. Nature 526(7575):692–695. https://doi.org/10.1038/nature15722

Teng S, Chung T, Chuang K, et al (2014) Toward mining user traversal patterns in the indoor environment. In: Peng WC, Wang H, Zhou ZH, et al (eds) Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2014 International Workshops. Springer Verlag, Germany, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 677–688, https://doi.org/10.1007/978-3-319-13186-3_60

Thorning L, Hansen E (1987) Electromagnetic reflection survey 1986 at the inland ice margin of the pâkitsoq basin, central west greenland. Rapport Grønlands Geologiske Undersøgelse 135:87–95. https://doi.org/10.34194/rapggu.v135.8005

Tong Y, Chen L, Cheng Y, et al (2012) Mining frequent itemsets over uncertain databases. Proc VLDB Endow 5(11). https://doi.org/10.14778/2350229.2350277,

URL https://doi.org/10.14778/2350229.2350277

Walck C (1996) Hand-book on statistical distributions for experimentalists. Universitet Stockholms

Wan L, Chen L, Zhang C (2013) Mining frequent serial episodes over uncertain sequence data. In: Proceedings of the 16th International Conference on Extending Database Technology. ACM, New York, NY, USA, EDBT '13, pp 215–226, https://doi.org/10.1145/2452376.2452403

Wang J, Han J, Li C (2007) Frequent Closed Sequence Mining without Candidate Maintenance. IEEE Transactions on Knowledge and Data Engineering 19(8):1042–1056. https://doi.org/10.1109/TKDE.2007.1043

Yan X, Han J, Afshar R (2003a) Clospan: Mining closed sequential patterns in large databases. In: Proceedings of the Third SIAM International Conference on Data Mining, pp 166–177, https://doi.org/10.1137/1.9781611972733.15

Yan X, Han J, Afshar R (2003b) CloSpan: Mining: Closed Sequential Patterns in Large Datasets. In: Proceedings of the 2003 SIAM International Conference on Data Mining. Proceedings, Society for Industrial and Applied Mathematics, p 166–177, https://doi.org/10.1137/1.9781611972733.15

Yu D, Wu W, Zheng S, et al (2012) Bide-based parallel mining of frequent closed sequences with mapreduce. In: Proceedings of the 12th International Conference on Algorithms and Architectures for Parallel Processing - Volume Part II. Springer-Verlag, Berlin, Heidelberg, ICA3PP'12, p 177–186, https://doi.org/10.1007/978-3-642-33065-0_19

Zaki MJ (2001) SPADE: an efficient algorithm for mining frequent sequences. Machine Learning 42(1/2):31–60. https://doi.org/10.1023/A:1007652502315

Zhang Q, Li F, Yi K (2008) Finding frequent items in probabilistic data. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, New York, NY, USA, SIGMOD '08, pp 819–832, https://doi.org/10.1145/1376616.1376698

Zhao, Yan, Wilfred (2014) Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases. IEEE Transactions on Knowledge and Data Engineering 26(5):1171–1184. https://doi.org/10.1109/TKDE.2013.124

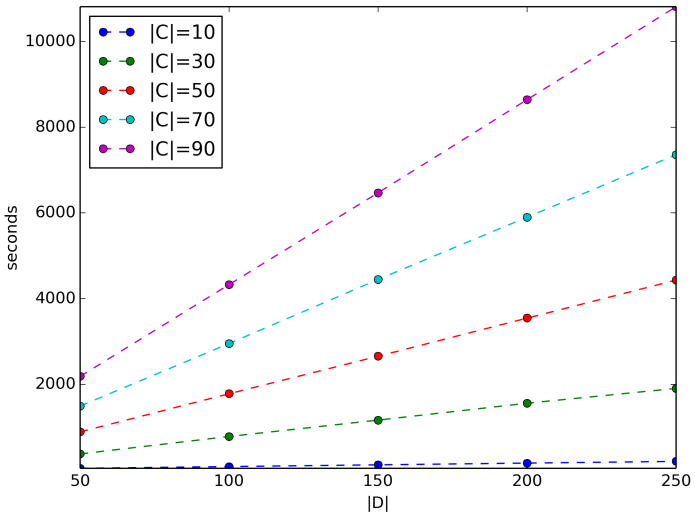

Fig. 3: Execution times vs. $|D|$

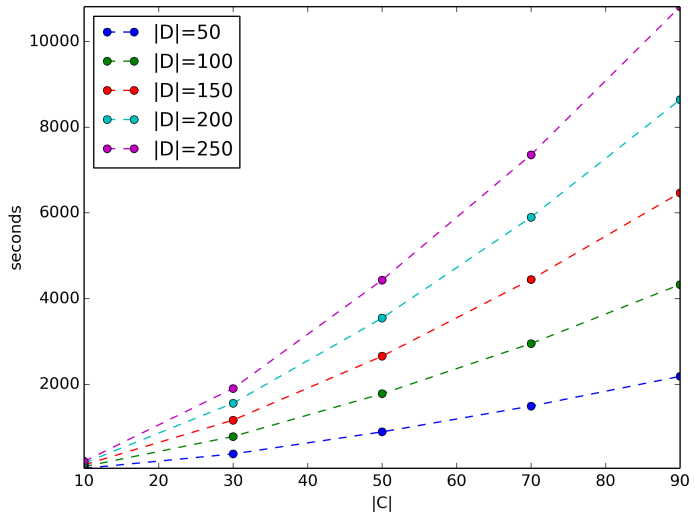

Fig. 4: Execution times vs. $|C|$

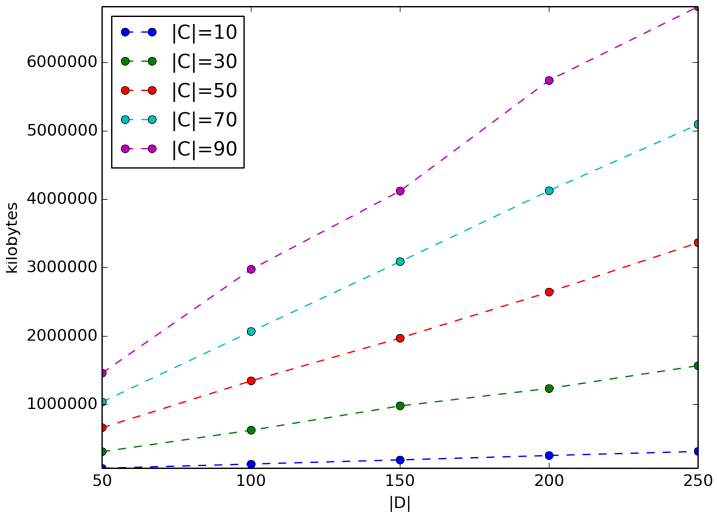

Fig. 5: Maximum memory used vs. $|D|$

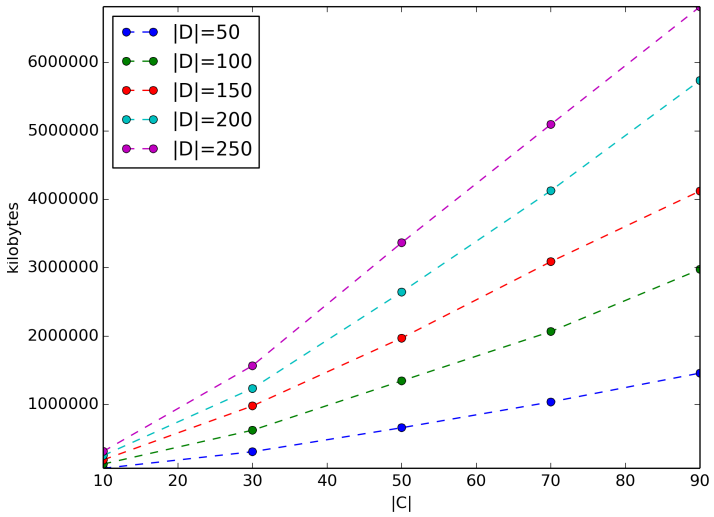

Fig. 6: Maximum memory used vs. $|C|$

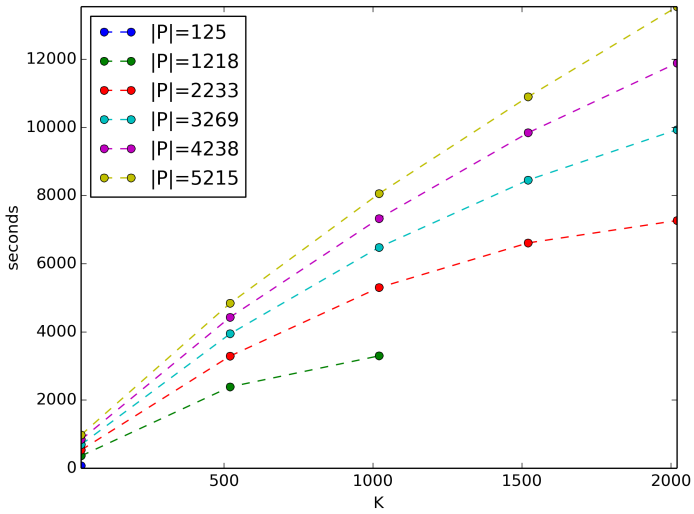

Fig. 7: Execution times vs. K for *C10D50K-A2B2-N3*

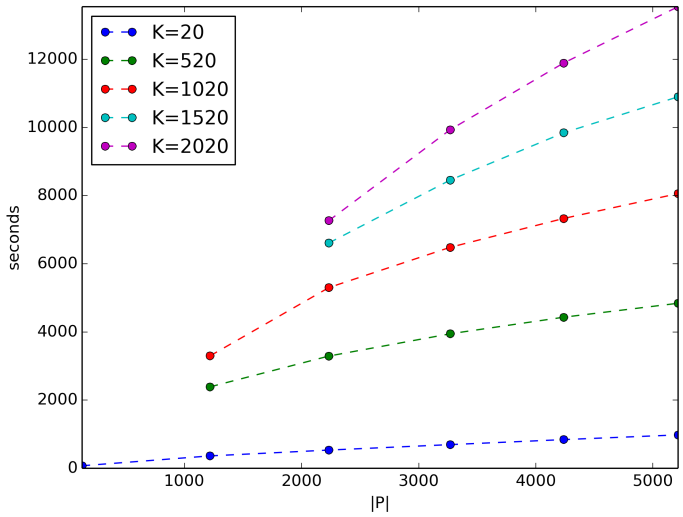

Fig. 8: Execution times vs. $|P|$ for *C10D50K-A2B2-N3*

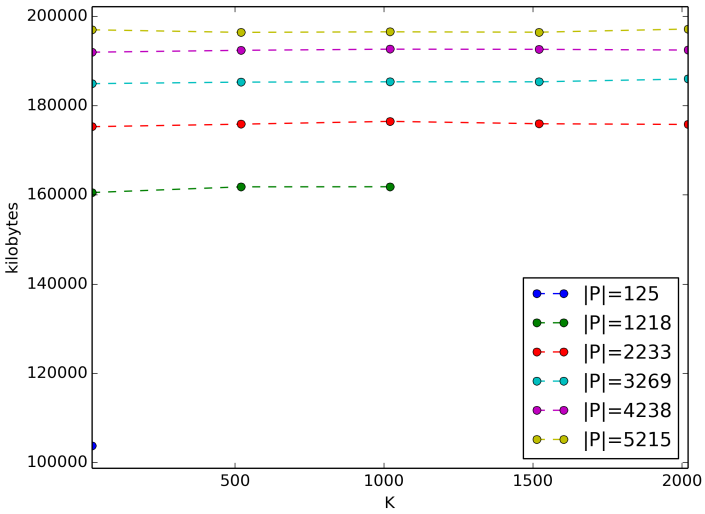

Fig. 9: Maximum memory used vs. K for *C10D50K-A2B2-N3*

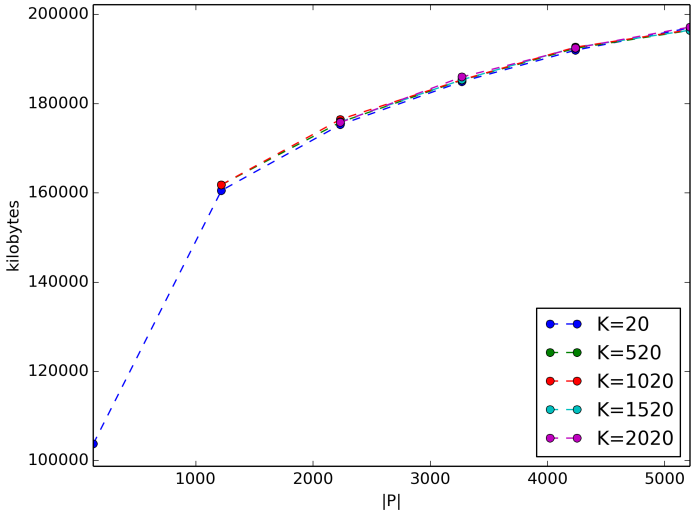

Fig. 10: Maximum memory used vs. $|P|$ for *C10D50K-A2B2-N3*

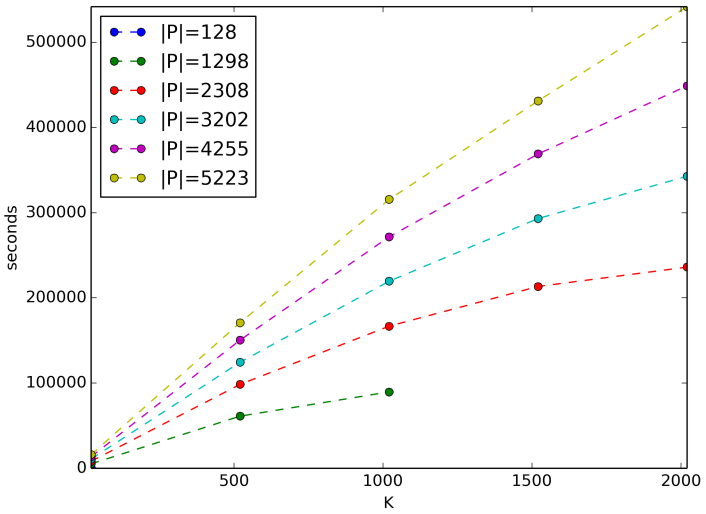

Fig. 11: Execution times vs. K for the Greenland dataset

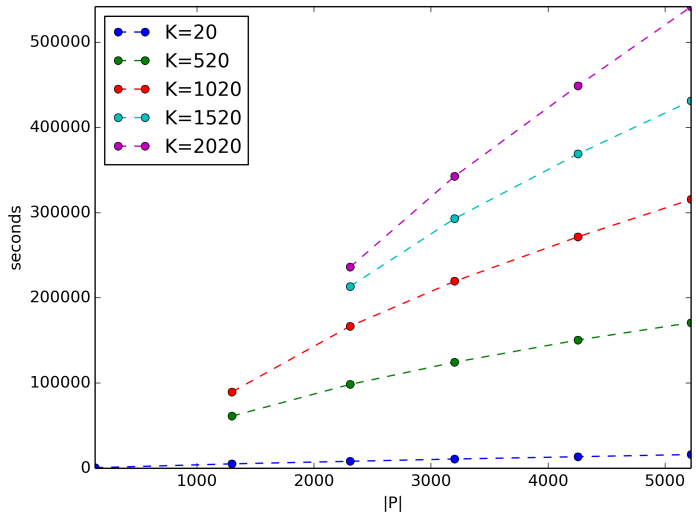

Fig. 12: Execution times vs. $|P|$ for the Greenland dataset

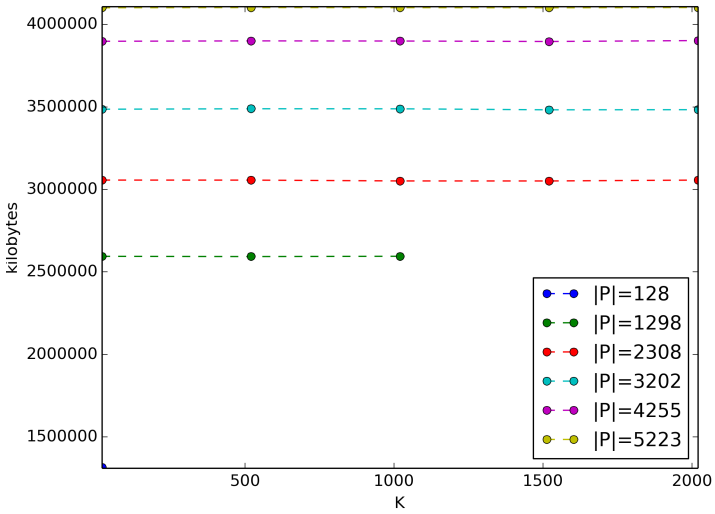

Fig. 13: Maximum memory used vs. K for the Greenland dataset

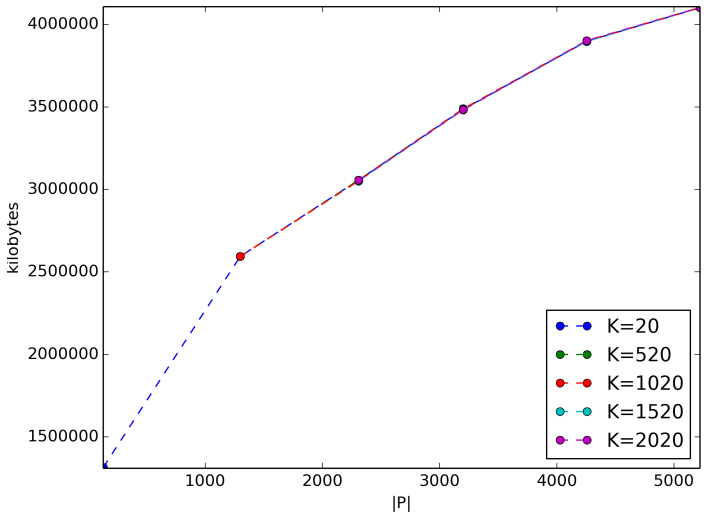

Fig. 14: Maximum memory used vs. $|P|$. Greenland dataset

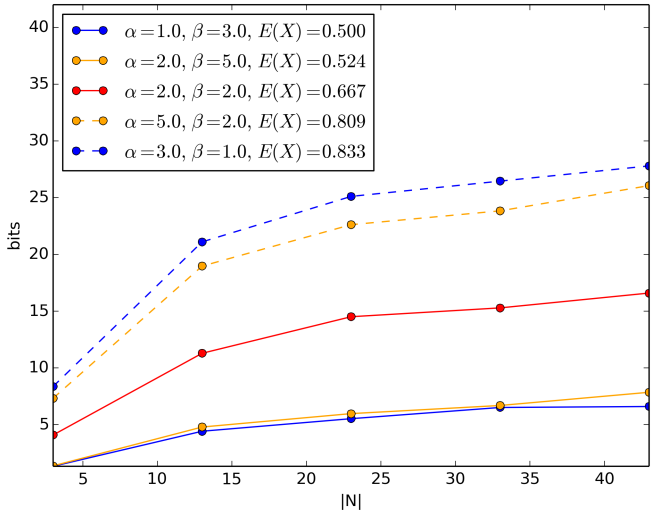

Fig. 15: $\Delta(\Phi)$ vs. N

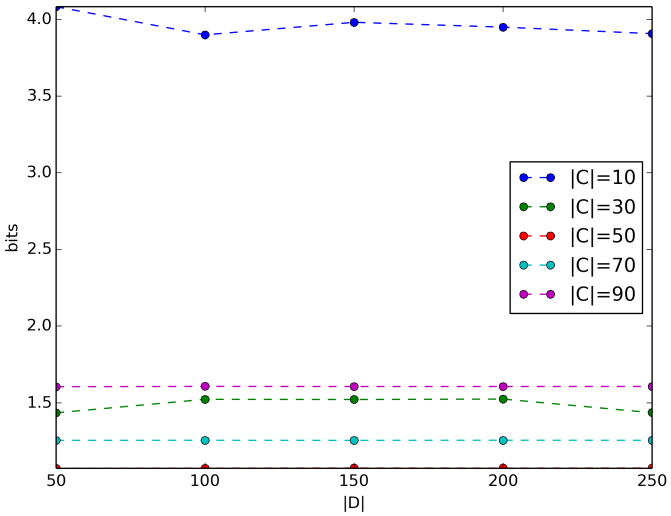

Fig. 16: $\Delta(\Phi)$ vs. $|D|$

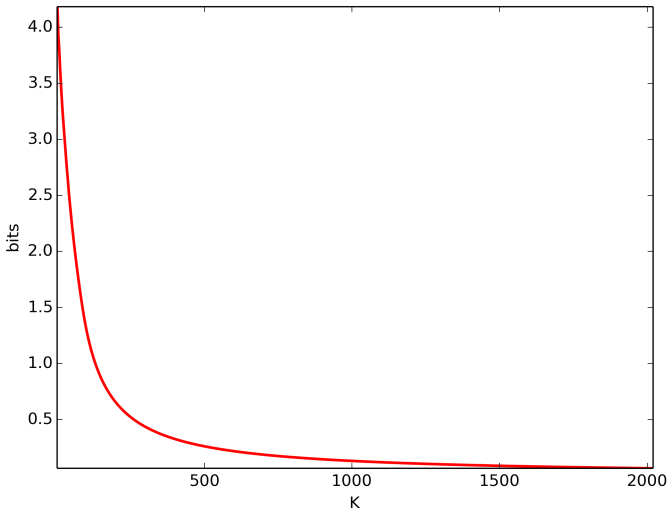

Fig. 17: $\Delta(\Phi)$ vs. K for *C10D50K-A2B2-N3* and $|P| = 5215$

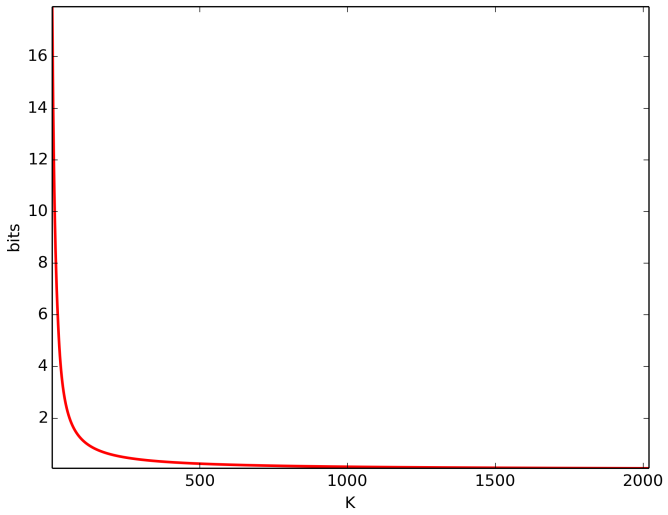

Fig. 18: $\Delta(\Phi)$ vs. K for the Greenland dataset and $|P| = 5223$

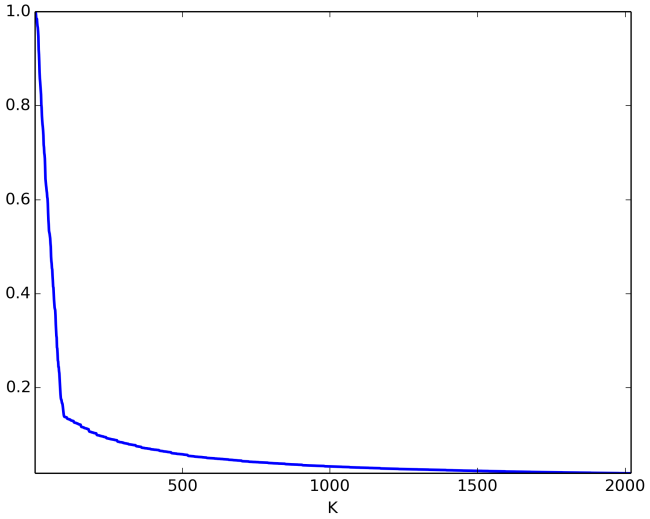

Fig. 19: $T(\Phi)$ vs. K for *C10D50K-A2B2-N3* and $|P| = 5215$

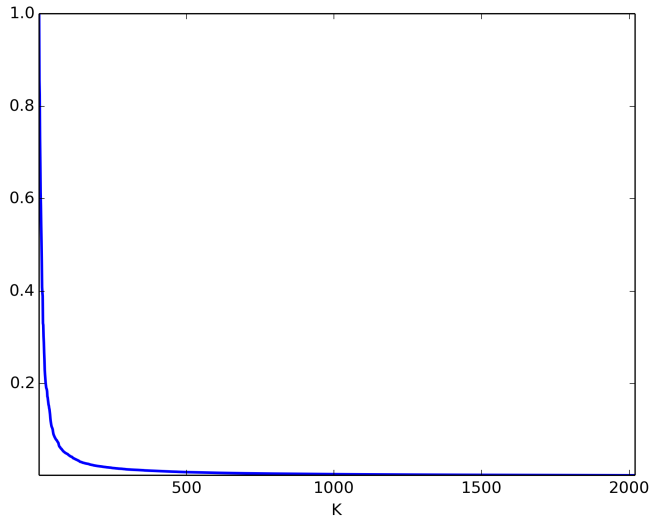

Fig. 20: $T(\Phi)$ vs. K for the Greenland dataset and $|P| = 5223$

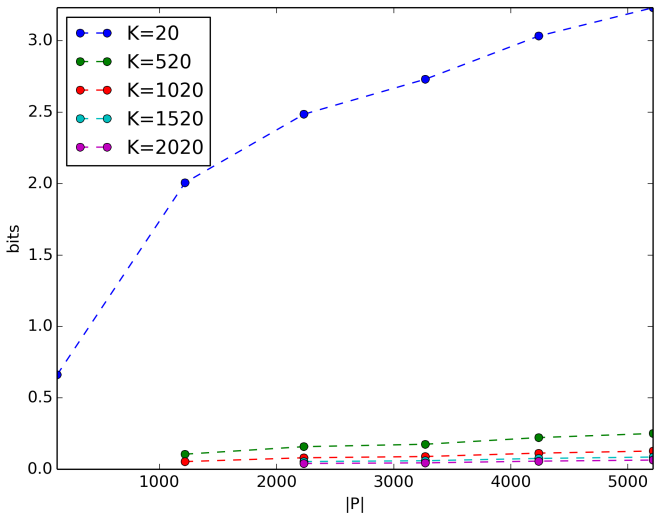

Fig. 21: $\Delta(\Phi)$ vs. $|P|$ for *C10D50K-A2B2-N3*

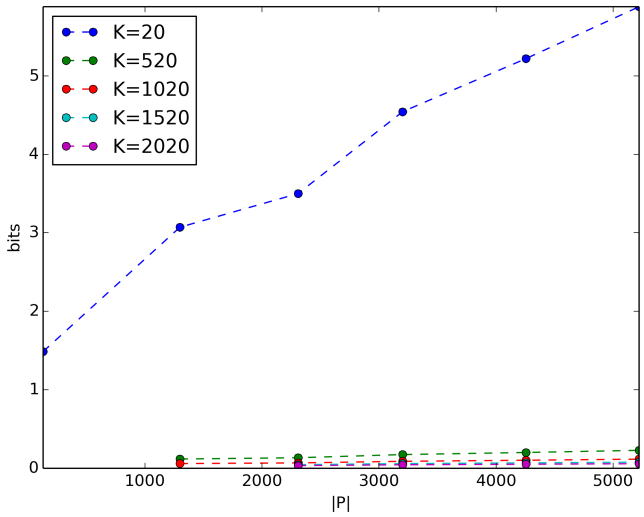

Fig. 22: $\Delta(\Phi)$ vs. $|P|$ for the Greenland dataset

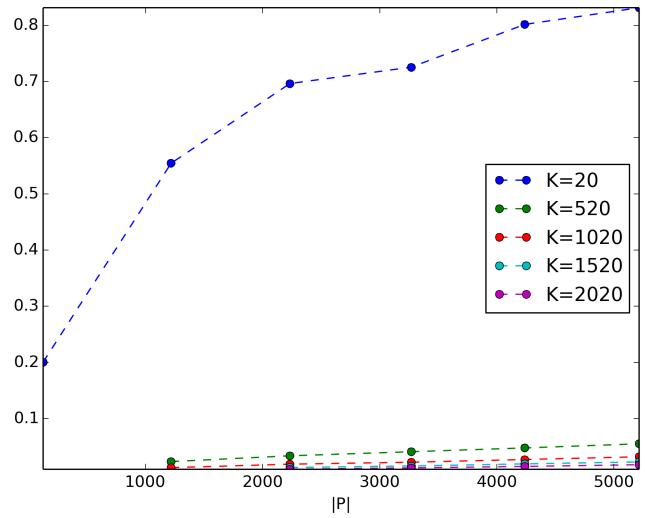

Fig. 23: $T(\Phi)$ vs. $|P|$ for *C10D50K-A2B2-N3*

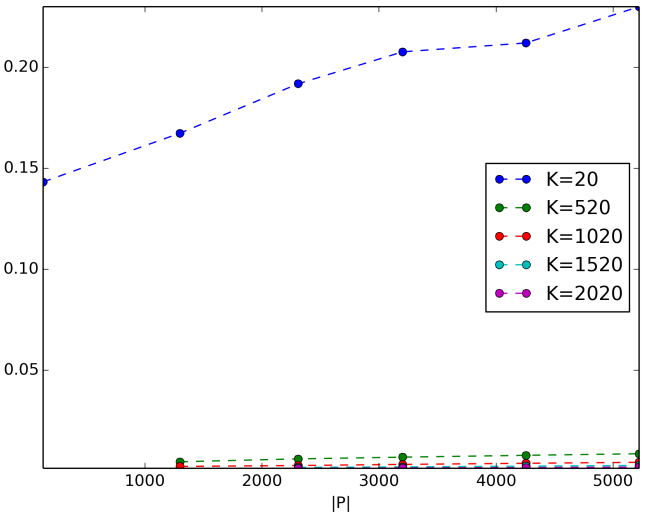

Fig. 24: $T(\Phi)$ vs. $|P|$ for the Greenland dataset

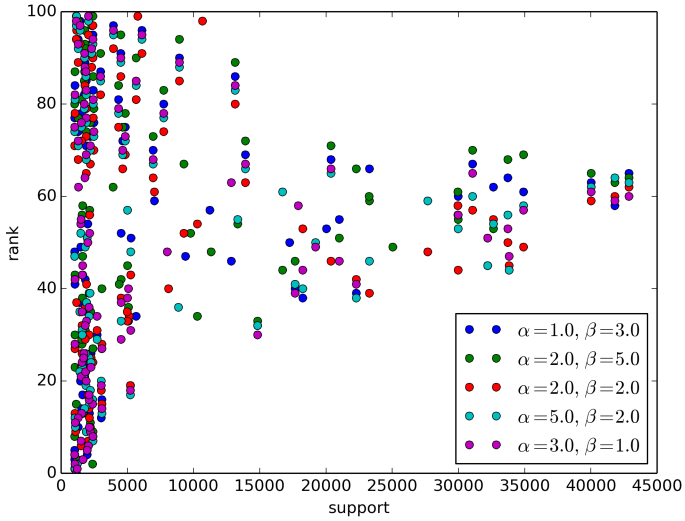

Fig. 25: Rank of the selected patterns against their support for synthetic datasets

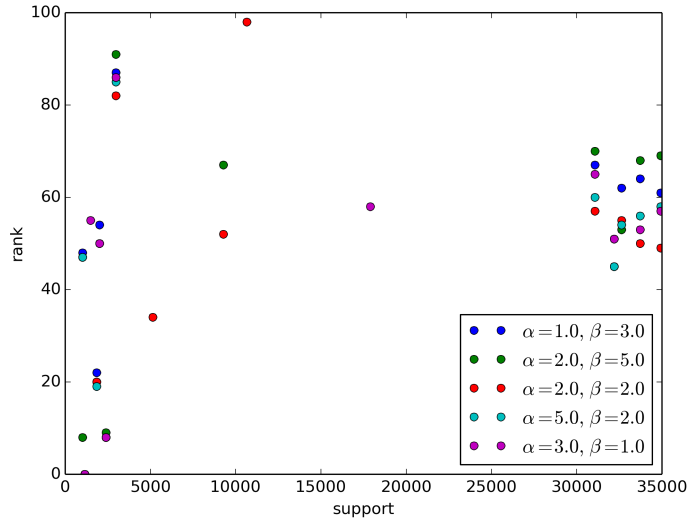

Fig. 26: Rank of the selected Quest patterns against their support for synthetic datasets

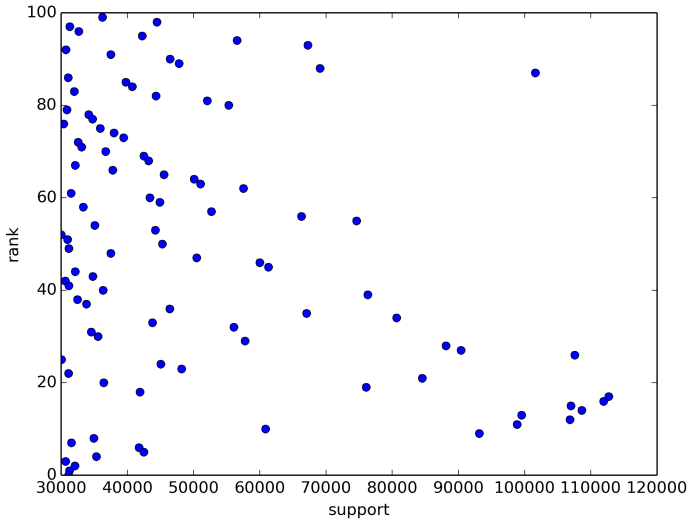

Fig. 27: Rank of the selected patterns against their support for the Greenland dataset

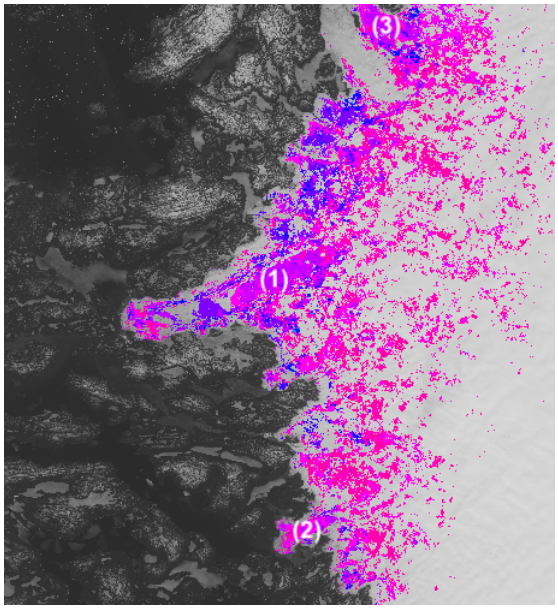

Fig. 28: Map of pattern $p1$: $3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1$. (1): Nordenskjöld glacier, (2): Polonia glacier, (3): Alangordliup Sermia glacier

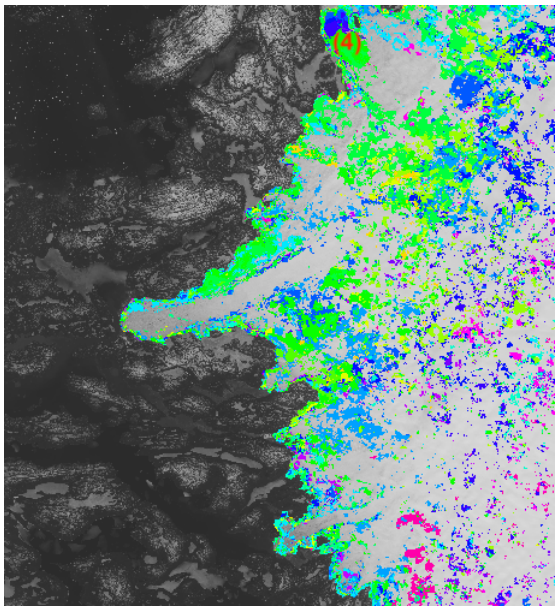

Fig. 29: Map of pattern $p2$: $3 \to 1 \to 1 \to 3$. (4): Sarqardliup Sermia glacier

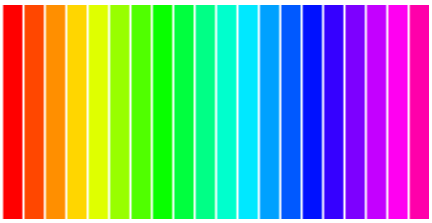

Fig. 30: color scale: from 1985 in red to 2013 in magenta decomposed in 20 timestamps