Efficient Temporal Pattern Mining in Big Time Series Using Mutual Information

Van Long Ho  
Aalborg University  
Aalborg, Denmark  
vlh@cs.aau.dk

Nguyen Ho  
Aalborg University  
Aalborg, Denmark  
nnth@cs.aau.dk

Torben Bach Pedersen  
Aalborg University  
Aalborg, Denmark  
tbp@cs.aau.dk

ABSTRACT

Very large time series are increasingly available from an ever wider range of IoT-enabled sensors deployed in different environments. Significant insights can be gained by mining temporal patterns from these time series. Unlike traditional pattern mining, temporal pattern mining (TPM) adds event time intervals into extracted patterns, making them more expressive at the expense of increased time and space complexities. Existing TPM methods either cannot scale to large datasets, or work only on pre-processed temporal events rather than on time series. This paper presents our Frequent Temporal Pattern Mining from Time Series (FTPMfTS) approach providing: (1) The end-to-end FTPMfTS process taking time series as input and producing frequent temporal patterns as output. (2) The efficient Hierarchical Temporal Pattern Graph Mining (HTPGM) algorithm that uses efficient data structures for fast support and confidence computation, and employs effective pruning techniques for significantly faster mining. (3) An approximate version of HTPGM that uses mutual information, a measure of data correlation, to prune unpromising time series from the search space. (4) An extensive experimental evaluation showing that HTPGM outperforms the baselines in runtime and memory consumption, and can scale to big datasets. The approximate HTPGM is up to two orders of magnitude faster and less memory consuming than the baselines, while retaining high accuracy.

1 INTRODUCTION

IoT-enabled sensors have enabled the collection of many big time series, e.g., from smart-meters, -plugs, and -appliances in households, weather stations, and GPS-enabled mobile devices. Extracting patterns from these time series can offer new domain insights for evidence-based decision making and optimization. As an example, consider Fig. 1 that shows the electricity usage of a water boiler with a hot water tank collected by a 20 euro wifi-enabled smart-plug, and accurate CO2 intensity (g/kWh) forecasts of local electricity, e.g., as supplied by the Danish Transmission System Operator [12]. From Fig. 1, we can identify several useful patterns. First, the water boiler switches On once a day, for one hour between 6 and 8AM. This indicates that the resident takes only one hot shower per day which starts between 5.30 and 6.30AM. Second, all water boiler On events are contained in CO2 High events, i.e., the periods when CO2 intensity is high. Third, between two consecutive On events of the boiler, there is a CO2 Low event lasting for one or more hours which occurs at most 4 hours before the hot shower (so water heated during that event will still be hot at 6AM). Pattern mining can be used to extract the relations between CO2 intensity and water boiler events. However, traditional sequential patterns only capture the sequential occurrence of events, e.g., that one boiler On event follows after another, but not that there is at least 23 hours between them; or that there is a CO2 Low event between the two boiler On events, but not when or for how long it lasts. In contrast, temporal pattern mining (TPM) adds temporal information into patterns, providing details on when certain relations between events happen, and for how long. For example, TPM expresses the above relations as: [(7:00 - 8:00, Day X) BoilerOn → [6:00 - 7:00, Day X+1] BoilerOn] (meaning BoilerOn is followed by BoilerOn), [(6:00 - 10:00, Day X) HighCO2 > [7:00 - 8:00, Day X] BoilerOn) (meaning HighCO2 contains BoilerOn), and [(7:00 - 8:00, Day X) BoilerOn → [0:00 - 2:00, Day X+1] LowCO2 → [6:00 - 7:00, Day X+1] BoilerOn). As the resident is very keen on reducing her CO2 footprint, we can rely on the above temporal patterns to automatically (using the smart-plug) delay turning on the boiler until the CO2 intensity is low again, saving CO2 without any loss of comfort for the resident.

Another example is in the smart city domain in which temporal patterns extracted from vehicle GPS data [41] can reveal spatio-temporal correlations between traffic jams. For example, if the pattern [(07:30, 08:00) SlowSpeedTunnel → [08:00, 08:30] SlowSpeedMainBoulevard] is found with high frequency and high confidence on weekdays, it can be used to advise drivers to take another route for their morning commute.

Although temporal patterns are useful, mining them is much more expensive than sequential patterns. Not only does the temporal information add extra computation to the mining process, the complex relations between events also add an additional exponential factor $O(3^h)$ to the complexity $O(m^h)$ of the search space ($m$ is the number of events and $h$ is the length of temporal patterns), yielding an overall complexity of $O(m^h3^h)$ (see Lemma 1 in Section 4.4). Existing TPM methods [8, 35, 36] do not scale on big datasets, i.e., many time series and many sequences, and/or do not work directly on time series but rather on pre-processed temporal events.

Contributions. In this paper, we present our comprehensive Frequent Temporal Pattern Mining from Time Series (FTPMfTS) approach which overcomes the above limitations. Our key contributions are: (1) We present the first end-to-end FTPMfTS process that

Figure 1: CO2 intensity and water boiler electricity usage
receives time series as input, and produces frequent temporal patterns as output. Within this process, a splitting strategy is proposed to convert time series into event sequences while ensuring the preservation of temporal patterns. (2) We propose the efficient Hierarchical Temporal Pattern Graph Mining (HTPGM) algorithm that employs: a) efficient data structures, Hierarchical Pattern Graph and bitmap, to enable fast support and confidence computation; and b) pruning techniques based on the Apriori principle and the transitivity property of temporal relations to enable faster mining. (3) Based on the concept of mutual information which measures the correlation among time series, we propose a novel approximate version of HTPGM that prunes unpromising time series to significantly reduce the search space and can scale on big datasets, i.e., many time series and many sequences. (4) We perform extensive experiments on synthetic and real-world datasets which show that HTPGM outperforms the baselines in both runtime and memory usage. The approximate HTPGM is up to two orders of magnitude faster and less memory consumption than the baselines while retaining high accuracy compared to the exact HTPGM.

2 RELATED WORK
Temporal pattern mining: Compared to sequential pattern mining, TPM is rather a new research area. One of the first papers in this area is [20] from Kam et al. that uses a hierarchical representation to manage temporal relations, and based on that mines temporal patterns. However, the approach in [20] suffers from ambiguity when presenting temporal relations. In [39], Wu et al. develop TPrefix to mine temporal patterns from non-ambiguous temporal relations. However, TPrefix has several inherent limitations: it scans the database repeatedly, and the algorithm does not employ any pruning strategies to reduce the search space. In [32], Moskovitch et al. design a TPM algorithm using the transitivity property of temporal relations. They use this property to generate candidates by inferring new relations between events. In comparison, our HTPGM uses the transitivity property for effective pruning. In [3], Iyad et al. propose a TPM framework to detect events in time series. However, their focus is to find irregularities in the data. In [38], Wang et al. propose a temporal pattern mining algorithm HUTPMiner to mine high-utility patterns. Different from our HTPGM which uses support and confidence to measure the frequency of patterns, HUTPMiner uses utility to measure the importance or profit of an event/pattern, thereby addressing an orthogonal problem. In [37], Amit et al. propose STIPA which uses a Hoppenner matrix representation to compress temporal patterns for memory savings. However, STIPA does not use any pruning/optimization strategies and thus, despite the efficient use of memory, it cannot scale to large datasets, unlike our HTPGM. Other work [4], [7] proposes TPM algorithms to classify health record data. However, these methods are very domain-specific, thus cannot generalize to other domains.

The state-of-the-art TPM methods that currently achieve the best performance are our baselines: H-DFS [35], TPMiner [8], IEMiner [36], and Z-Miner [28]. H-DFS is a hybrid algorithm that uses breadth-first and depth-first search strategies to mine frequent arrangements of temporal intervals. H-DFS uses a data structure called ID-List to transform event sequences into vertical representations, and temporal patterns are generated by merging the ID-Lists of different events. This means that H-DFS does not scale well when the number of time series increases. In [36], Patel et al. design a hierarchical lossless representation to model event relations, and propose IEMiner that uses Apriori-based optimizations to efficiently mine patterns from this new representation. In [8], Chen et al. propose TPMiner that uses endpoint and endtime representations to simplify the complex relations among events. Similar to [35], IEMiner and TPMiner do not scale to datasets with many time series. Z-Miner [28], proposed by Lee et al., is the most recent work addressing TPM. Z-Miner improves the mining efficiency over existing methods by employing two data structures: a hierarchical hash-based structure called Z-Table for time-efficient candidate generation and support count, and Z-Arrangement, a structure to efficiently store event intervals in temporal patterns for efficient memory consumption. Although using efficient data structures, Z-Miner neither employs the transitivity property of temporal relations nor mutual information for pruning. Thus, Z-Miner is less efficient than our exact and approximate HTPGM in both runtimes and memory usage, and does not scale to large datasets with many sequences and many time series (see Section 6).

Our HTPGM algorithm improves on these methods by: (1) using efficient data structures and applying pruning techniques based on the Apriori principle and the transitivity property of temporal relations to enable fast mining, (2) the approximate HTPGM can handle datasets with many time series and sequences, and (3), providing an end-to-end FTPMfTS process to mine temporal patterns directly from time series, a feature that is not supported by the baselines.

Using correlations in TPM: Different correlation measures such as expected support [1], all-confidence [27], and mutual information (MI) [6, 11, 15–18, 21–25, 40] have been used to optimize the pattern mining process. However, these only support sequential patterns. To the best of our knowledge, our proposed approximate HTPGM is the first that uses MI to optimize TPM.

3 PRELIMINARIES
In this section, we introduce the notations and the main concepts that will be used throughout the paper.

3.1 Temporal Event of Time Series
Definition 3.1 (Time series) A time series \( X = x_1, x_2, ..., x_n \) is a sequence of data values that measure the same phenomenon during an observation time period, and are chronologically ordered.

Definition 3.2 (Symbolic time series) A symbolic time series \( X_S \) of a time series \( X \) encodes the raw values of \( X \) into a sequence of symbols. The finite set of permitted symbols used to encode \( X \) is called the symbol alphabet of \( X \), denoted as \( \Sigma_X \).

The symbolic time series \( X_S \) is obtained using a mapping function \( f: X \rightarrow \Sigma_X \) that maps each value \( x_i \in X \) to a symbol \( \omega \in \Sigma_X \). For example, let \( X = 1.61, 1.21, 0.41, 0.0 \) be a time series representing the energy usage of an electrical device. Using the symbol alphabet \( \Sigma_X = \{\text{On}, \text{Off}\} \), \( \text{On} \) represents that the device is on and operating (e.g., \( x_i \geq 0.5 \)), and \( \text{Off} \) that the device is off (\( x_i < 0.5 \)). The symbolic representation of \( X \) is: \( X_S = \text{On, On, Off, Off} \). The mapping function \( f \) can be defined using existing time series representation techniques such as SAX [29] or MVQ [30].

Definition 3.3 (Symbolic database) Given a set of time series \( X = \{X_1, ..., X_n\} \), the set of symbolic representations of the time series in \( X \) forms a symbolic database \( D_{SYB} \).
An example of the symbolic database $D_{SYB}$ is shown in Table 1. There are 6 time series representing the energy usage of 6 electrical appliances: [Stove, Toaster, Microwave, Clothes Washer, Dryer, Iron]. For brevity, we name the appliances respectively as [S, T, M, W, D, I]. All appliances have the same alphabet $Σ = \{\text{On}, \text{Off}\}$.

**Definition 3.4** (Temporal event in a symbolic time series) A temporal event $E$ in a symbolic time series $X$ is a tuple $E = (i, T)$ where $i \in X$ is a symbol, and $T = \{[t_1, t_2], \ldots\}$ is the set of time intervals during which $X$ is associated with the symbol $i$.

Given a time series $X$, a temporal event is created by first converting $X$ into symbolic time series $X_σ$ and then combining identical consecutive symbols in $X_σ$ into one single time interval. For example, consider the symbolic representation of $S$ in Table 1. By combining its consecutive On symbols, we form the temporal event "Stove is On" as: (SOn, [[10:00, 10:15], [10:35, 10:40], [11:15, 11:25], [11:50, 12:00], [12:15, 12:20], [12:35, 12:45]])

**Definition 3.5** (Instance of a temporal event) Let $E = (i, T)$ be a temporal event, and $[t_1, t_2] \in T$ be a time interval. The tuple $e = (i, [t_1, t_2])$ is called an instance of the event $E$, representing a single occurrence of $E$ during $[t_1, t_2]$. We use the notation $E_{e\epsilon}$ to denote that event $E$ has an instance $e$.

### 3.2 Relations between Temporal Events

We adopt the popular Allen’s relations model [2] and define three basic temporal relations between events. Furthermore, to avoid the exact time mapping problem in Allen’s relations, we adopt the buffer idea from [35], adding a tolerance buffer $\epsilon$ to the relation’s endpoints. However, we change the way $\epsilon$ is used in [35] to ensure the relations are mutually exclusive (proof is in the full paper [19]).

Consider two temporal events $E_1$ and $E_2$ and their corresponding instances, $e_1 = (i, [t_1, t_2])$ and $e_2 = (j, [t_3, t_4])$, let $\epsilon$ be a non-negative number ($\epsilon \geq 0$) representing the buffer size. The following relations can be defined between $E_1$ and $E_2$ through $e_1$ and $e_2$.

**Definition 3.6** (Follows) $E_1$ and $E_2$ form a Follows relation through $e_1$ and $e_2$, denoted as $\text{Follows}(E_{i\epsilon}, E_{j\epsilon})$ or $E_{i\epsilon} \rightarrow E_{j\epsilon}$, iff $t_2 - \epsilon \leq t_1$.

**Definition 3.7** (Contains) $E_1$ and $E_2$ form a Contains relation through $e_1$ and $e_2$, denoted as $\text{Contains}(E_{i\epsilon}, E_{j\epsilon})$ or $E_{i\epsilon} \supset E_{j\epsilon}$, iff $t_1 \leq t_2$ and $(t_4 - \epsilon \geq t_3)$.

**Definition 3.8** (Overlaps) $E_1$ and $E_2$ form an Overlaps relation through $e_1$ and $e_2$, denoted as $\text{Overlaps}(E_{i\epsilon}, E_{j\epsilon})$ or $E_{i\epsilon} \cap E_{j\epsilon}$, iff $(t_1 < t_2) \land (t_4 - \epsilon \leq t_3)$ and $(t_3 - \epsilon \geq t_3)$, where $d_0$ is the minimal overlapping duration between two event instances, and $0 \leq \epsilon < d_0$.

The Follows relation represents sequential occurrences of one event after another. For example, $E_{S\epsilon}$ is followed by $E_{T\epsilon}$ if the end time $t_2$ of $e_1$ occurs before the start time $t_1$ of $e_2$. Here, the buffer $\epsilon$ is used as a tolerance, i.e., the Follows relation between $E_{S\epsilon}$ and $E_{T\epsilon}$ holds if $(t_2 + \epsilon) < (t_1 - \epsilon)$.

The timespan of another event. Finally, in an Overlaps relation, the timespans of the two occurrences overlap each other. In Table 2, the followings illustrate the three temporal relations and their conditions.

### 3.3 Temporal Pattern

**Definition 3.9** (Temporal sequence) A list of $n$ event instances $S = <e_1, \ldots, e_n>$ forms a temporal sequence if the instances are chronologically ordered by their start times. Moreover, $S$ has size $n$, denoted as $|S| = n$.

**Definition 3.10** (Temporal sequence database) A set of temporal sequences forms a temporal sequence database $D_{SEQ}$ where each row $i$ contains a temporal sequence $S_i$.

Table 3 shows the symbolic temporal database $D_{SEQ}$, created from the symbolic database $D_{SYB}$ in Table 1.

**Definition 3.11** (Temporal pattern) Let $R = \{\text{Contains, Overlaps, Follows}\}$ be the set of temporal relations. A temporal pattern $P = <(E_1, E_2), \ldots, (r_{ij}, E_1, E_2)\rangle$ or a list of triples $(r_{ij}, E_1, E_2)$, each representing a relation $r_{ij} \in R$ between two events $E_1$ and $E_2$.

Note that the relation $r_{ij}$ in each triple is formed by using the specific instances of $E_1$ and $E_2$. A temporal pattern that has $n$ events is called an $n$-event pattern. We use $E_i \in P$ to denote that the event $E_i$ occurs in $P$, and $P_i \subseteq P$ to say that a pattern $P_i$ is a sub-pattern of $P$.

**Definition 3.12** (Temporal sequence supports a pattern) Let $S = <e_1, \ldots, e_n>$ be a temporal sequence. We say that $S$ supports a temporal pattern $P$, denoted as $P \in S$, if $|S| \geq n \land \forall (r_{ij}, E_1, E_2) \in P$, there exists $E_1, E_2 \in S$ such that $r_{ij}$ holds between $E_{i\epsilon}$ and $E_{j\epsilon}$.

If $P$ is supported by $S$, then $P$ can be written as $P = <(r_{ij}, E_1, E_2)\rangle$ or $P = <(r_{ij}, E_1, E_2)\rangle$, where the relation between two events in each triple is expressed using the event instances.

In Fig. 1, consider the sequence $<\text{Dryer On}, [6:00, 7:00]>$, $\text{BoilerOn, [7:00, 8:00]}$, $\text{LowCO2, [13:00, 15:00]}>$ representing the order of CO2 intensity and boiler events. Here, $S$ supports a 3-event pattern $P = <(\text{Contains, HighCO2seq}), (\text{BoilerOn seq}), (\text{LowCO2seq})>$, (Follows, $\text{HighCO2seq}$, $\text{LowCO2seq}$), (Follows, $\text{BoilerOn seq}$, $\text{LowCO2seq}$).

**Maximal duration constraint:** Let $P \in S$ be a temporal pattern supported by the sequence $S$. The duration between the start time of the instance $e_1$, and the end time of the instance $e_n$ in $S$ must not exceed the predefined maximal time duration $t_{max} = t_n - t_s \leq t_{max}$.

The maximal duration constraint guarantees that the relation between any two events is temporally valid. This enables the pruning of invalid patterns. For example, under this constraint, a Follows relation between a “Washer On” event and a “Dryer On” event in Table 3 happening one year apart should be considered invalid.

### 3.4 Frequent Temporal Pattern

Given a temporal sequence database $D_{SEQ}$, we want to find patterns that occur frequently in $D_{SEQ}$. We use support and confidence [34] to measure the frequency and the likelihood of a pattern.

**Definition 3.13** (Support of a temporal pattern) The support of a temporal event $E$ in $D_{SEQ}$ is the number of sequences $S \in D_{SEQ}$...
Temporal sequences
(SOn, [10:00, 10:15]), (TOn, [10:00, 10:05]), (MOt, [10:00, 10:20]), (WOt, [10:00, 10:20]),
(DOt, [10:00, 10:40]), (IOt, [10:00, 10:35]), (TOm, [10:15, 10:35]),
(TOn, [10:15, 10:35]), (MOm, [10:20, 10:30]), (WOm, [10:20, 10:30]), (DOm, [10:30, 10:40]),
(MOmt, [10:35, 10:40]), (SOn, [10:35, 10:40]), (TOm, [10:35, 10:40]), (MOt, [10:35, 10:40]).

The confidence of a temporal pattern is defined as

\[ \text{conf}(P) = \frac{\text{supp}(P)}{\text{max}\{\text{supp}(E_1), \text{supp}(E_j)\}} \]

which contain at least one instance \( e \) of \( E \).

\[ \text{supp}(E) = |\{ S \in \mathcal{D}_{\text{SEQ}} : S \in E \}| \]

The support of a pattern \( P \) is the fraction of sequences \( S \in \mathcal{D}_{\text{SEQ}} \) that support \( P \).

\[ \text{supp}(P) = |\{ S \in \mathcal{D}_{\text{SEQ}} : P \in S \}| \]

The confidence of an event pair \((E_i, E_j)\) is the fraction between \( \text{supp}(E_i, E_j) \) and the support of its most frequent event:

\[ \text{conf}(E_i, E_j) = \frac{\text{supp}(E_i, E_j)}{\max\{\text{supp}(E_i), \text{supp}(E_j)\}} \]

Confidence of a temporal pattern is the fraction between \( \text{supp}(P) \) and the support of its most frequent event:

\[ \text{conf}(P) = \frac{\text{supp}(P)}{\text{max}\{\text{supp}(E_k)\}} \]

where \( E_k \in P \) is a temporal event. Since the denominator in Eq. (6) is the maximum support of the events in \( P \), the confidence computed in Eq. (6) is the minimum confidence of a pattern \( P \) in \( \mathcal{D}_{\text{SEQ}} \), which is also called the all-confidence as in [34].

Note that unlike association rules, temporal patterns do not have antecedents and consequents. Instead, they represent pairwise temporal relations between events based on their temporal occurrences. Thus, while the support and relative support of event(s)/pattern(s) defined in Eqs. (1) – (4) follow the same intuition as the traditional support concept, indicating how frequently an event/pattern occurs in a given database, the confidence computed in Eqs. (5) – (6) instead represents the minimum likelihood of an event pair/pattern, knowing the likelihood of its most frequent event.

Frequent Temporal Pattern Mining from Time Series (FTP MTS). Given a set of univariate time series \( X = \{X_1, ..., X_n\} \), let \( \mathcal{D}_{\text{SEQ}} \) be the temporal sequence database obtained from \( X \), and \( \sigma \) and \( \delta \) be the support and confidence thresholds, respectively. The FTPMfTS problem aims to find all temporal patterns \( P \) that have high enough support and confidence in \( \mathcal{D}_{\text{SEQ}} \) : \( \text{supp}(P) \geq \sigma \land \text{conf}(P) \geq \delta \).

4 FREQUENT TEMPORAL PATTERN MINING

Fig. 2 gives an overview of the FTPMfTS process which consists of 2 phases. The first phase, Data Transformation, converts a set of time series \( X \) into a symbolic database \( \mathcal{D}_{\text{SYB}} \), and then converts \( \mathcal{D}_{\text{SYB}} \) into a temporal sequence database \( \mathcal{D}_{\text{SEQ}} \). The second phase, Frequent Temporal Pattern Mining, mines frequent patterns which includes 3 steps: (1) Frequent Single Event Mining, (2) Frequent 2-Event Pattern Mining, and (3) Frequent k-Event Pattern Mining (k>2).

The final output is a set of all frequent patterns in \( \mathcal{D}_{\text{SEQ}} \).

4.1 Data Transformation

4.1.1 Symbolic Time Series Representation. Given a set of time series \( X \), the symbolic representation of each time series \( X \in X \) is obtained by using a mapping function as in Def. 3.2.

4.1.2 Temporal Sequence Database Conversion. To convert \( \mathcal{D}_{\text{SYB}} \) to \( \mathcal{D}_{\text{SEQ}} \), a straightforward approach is to split the symbolic series in \( \mathcal{D}_{\text{SYB}} \) into equal-length sequences, each belongs to a row in \( \mathcal{D}_{\text{SEQ}} \). For example, if each symbolic series in Table 1 is split into 4 sequences, then each sequence will last for 40 minutes. The first sequence \( S_1 \) of \( \mathcal{D}_{\text{SEQ}} \) therefore contains temporal events of \( S, T, M, W, D \), and \( I \) from 10:00 to 10:40. The second sequence \( S_2 \) contains events from 10:45 to 11:25, and similarly for \( S_3 \) and \( S_4 \).

However, the splitting can lead to a potential loss of temporal patterns. The loss happens when a splitting point accidentally divides a temporal pattern into different sub-patterns, and places these into separate sequences. We explain this situation in Fig. 3a. Consider 2 sequences \( S_1 \) and \( S_2 \), each of length \( t \). Here, the splitting point divides a pattern of 4 events, \( [\text{SON}, \text{TON}, \text{MOn}, \text{WOn}] \), into two sub-patterns, in which \( \text{SON} \) and \( \text{TON} \) are placed in \( S_1 \), and \( \text{MOn} \) and \( \text{WOn} \) in \( S_2 \). This results in the loss of this 4-event pattern which can be identified only when all 4 events are in the same sequence.

To prevent such a loss, we propose a splitting strategy using overlapping sequences. Specifically, two consecutive sequences are overlapped by a duration \( t_{ov} = t_{ov} \neq t_{max} \), where \( t_{max} \) is the...
maximal duration of a temporal pattern. The value of $t_{ov}$ decides how large the overlap between $S_i$ and $S_{i+1}$ is: $t_{ov} = 0$ results in no overlap, i.e., no redundancy, but with a potential loss of patterns, while $t_{ov} = t_{max}$ creates large overlaps between sequences, i.e., high redundancy, but all patterns are preserved. As illustrated in Fig. 3b, the overlapping between $S_1$ and $S_2$ keeps the 4 events together in the same sequence $S_2$, and thus helps preserve the pattern.

4.2 Frequent Temporal Patterns Mining

We now present our method, called Hierarchical Temporal Pattern Graph Mining (HTPGM), to mine frequent patterns from $D_{SEQ}$. The main novelties of HTPGM are: a) the use of efficient data structures, i.e., the proposed Hierarchical Pattern Graph and bitmap indexing, to fast computations of support and confidence, and b) the proposal of two groups of pruning techniques based on the Apriori principle and the temporal transitivity property of temporal events. In Section 5, we introduce an approximate version of HTPGM based on mutual information to further optimize the mining process. We first discuss the data structures used in HTPGM.

Hierarchical Pattern Graph (HPG): We use a hierarchical graph structure, called the Hierarchical Pattern Graph, to keep track of the frequent events and patterns found in each mining step. The HPG allows HTPGM to mine iteratively (e.g., 2-event patterns are mined on frequent single events, 3-event patterns are mined on 2-event patterns, and so on) and perform effective pruning. Fig. 4 shows the HPG built from $D_{SEQ}$ in Table 3: the root is the empty set $\emptyset$, and each level $L_k$ maintains frequent $k$-event patterns. As HTPGM constructs, HPG is constructed gradually. We explain this process for each mining step.

Efficient bitmap indexing: We use bitmaps to index the occurrences of events and patterns in $D_{SEQ}$, enabling fast computations of support and confidence. Specifically, each event $E$ or pattern $P$ found in $D_{SEQ}$ is associated with a bitmap indicating where $E$ or $P$ occurs. Each bitmap $b$ has length $|D_{SEQ}|$ (i.e., the number of sequences), and has value $b[i] = 1$ if $E$ or $P$ is present in sequence $i$ of $D_{SEQ}$, or $b[i] = 0$ otherwise. An example bitmap can be seen at $L_1$ in Fig. 4. The event TOn has the bitmap $b_{TOn} = [1,0,1,1]$, indicating that TOn occurs in all but the second sequence of $D_{SEQ}$.

Constructing the bitmap is also done step by step. For single events in $D_{SEQ}$, bitmaps are built by scanning $D_{SEQ}$ only once. Algorithm 1 provides the pseudo-code of HTPGM. The details are explained in each mining step.

4.3 Mining Frequent Single Events

The first step in HTPGM is to find frequent single events (Alg. 1, lines 1-4) which is easily done using the bitmap. For each event $E_i$ in $D_{SEQ}$, the support $sup(E_i)$ is computed by counting the number of set bits in bitmap $b_{E_i}$, and comparing against $\sigma$. Note that for single events, confidence is not considered since it is always 1.

After this step, the set $1Freq$ contains frequent single events is created to build $L_1$ of HPG. We illustrate this process using Table 3, with $\sigma = 0.7$ and $\delta = 0.7$. Here, $1Freq$ contains 11 frequent events, each belongs to one node in $L_1$. The event DOn is not frequent (only appears in sequences 2 and 4), and is thus omitted. Each $L_1$ node has a unique event name, a bitmap, and a list of instances corresponding to that event (see SON at $L_1$).

Complexity: The complexity of finding frequent single events is $O(m \cdot |D_{SEQ}|)$, where $m$ is the number of distinct events.

Proof: Detailed proofs of all complexities, lemmas and theorems in this article can be found in the Appendix of the full paper [19].

4.4 Mining Frequent 2-event Patterns

4.4.1 Search space of HTPGM. The next step in HTPGM is to mine frequent 2-event patterns. A straightforward approach would be to enumerate all possible event pairs, and check whether each pair can form frequent patterns. However, this naive approach is very expensive. Not only does it need to repeatedly scan $D_{SEQ}$ to check each combination of events, the complex relations between
Algorithm 1: Hierarchical Temporal Pattern Graph Mining

| Line | Description |
|------|-------------|
| 1. | **Input:** Temporal sequence database \( D_{SEQ} \), a support threshold \( \sigma \), a confidence threshold \( \delta \) |
| 2. | **Output:** The set of frequent temporal patterns \( P \) |
| 3. | // Mining frequent single events |
| 4. | **foreach** event \( E_i \in D_{SEQ} \) **do** |
| 5. | \( supp(E_i) \leftarrow \text{countBitmap}(b_{E_i}) \); |
| 6. | **if** \( supp(E_i) \geq \sigma \) **then** |
| 7. | Insert \( E_i \) to \( 1Freq \); |
| 8. | // Mining frequent \( k \)-event patterns |
| 9. | **foreach** \( kEventCombinations \in \text{FrequentPairs} \) **do** |
| 10. | **Check** frequent relations; |
| 11. | //Mining frequent \( k \)-event patterns |
| 12. | EventPairs \( \leftarrow \text{Cartesian}(1Freq, 1Freq) \); |
| 13. | FrequentPairs \( \leftarrow \emptyset \); |
| 14. | **foreach** \( (E_i, E_j) \) in EventPairs **do** |
| 15. | \( b_{ij} \leftarrow \text{AND}(b_{E_i}, b_{E_j}) \); |
| 16. | \( supp(E_i, E_j) \leftarrow \text{countBitmap}(b_{ij}) \); |
| 17. | **if** \( supp(E_i, E_j) \geq \sigma \) **then** |
| 18. | FrequentPairs \( \leftarrow \text{ApplyLemma3}(E_i, E_j) \); |
| 19. | **foreach** \( (E_i, E_j) \) in FrequentPairs **do** |
| 20. | Retrieve event instances; |
| 21. | Check frequent relations; |
| 22. | //Mining frequent \( k \)-event patterns |
| 23. | Filtered1Freq \( \leftarrow \text{TransitivityFiltering}(1Freq) \); |
| 24. | //Lemmas 4, 5 |
| 25. | \( kEventCombinations \leftarrow \text{Cartesian}(\text{Filtered1Freq}, \text{kEventCombinations}) \); |
| 26. | \( \text{FreqentEvents} \leftarrow \text{AprioriFiltering}(kEventCombinations) \); |
| 27. | **foreach** \( kEventCombinations \) in \( \text{FreqentEvents} \) **do** |
| 28. | Retrieve relations; |
| 29. | Iteratively check frequent relations; |
| 30. | //Lemmas 4, 6, 7 |

From Lemma 3, the confidence of a pattern \( P \) is always at most the confidence of its events. Thus, a low-confidence event pair cannot form any high-confidence patterns and therefore, can be safely pruned. We note that the Apriori principle has already been used in other work, e.g., [8, 35], for mining optimization. However, they only apply this principle to the support (Lemma 2), while we further extend it to the confidence (Lemma 3). Applying Lemmas 2 and 3 to the first filtering step will remove infrequent or low-confidence event pairs, reducing the candidate patterns of HTPGM. We detail this filtering below.

Step 2.1. Mining frequent event pairs: This step finds frequent event pairs in \( D_{SEQ} \) using the set \( 1Freq \) found in \( L_1 \) of HPG (Alg. 1, lines 5-11). First, HTPGM generates all possible event pairs by calculating the Cartesian product \( 1Freq \times 1Freq \). Next, for each pair \( (E_i, E_j) \), the joint bitmap \( b_{ij} \) (representing the set of sequences where both events occur) is computed by \( \text{ANDing} \) the two individual bitmaps \( b_{ij} = \text{AND}(b_{E_i}, b_{E_j}) \). Finally, HTPGM computes the support \( supp(E_i, E_j) \) by counting the set bits in \( b_{ij} \), and comparing against \( \sigma \). If \( supp(E_i, E_j) \geq \sigma \), then \( (E_i, E_j) \) has high enough support. Next, \( (E_i, E_j) \) is further filtered using Lemma 3: \( (E_i, E_j) \) is selected only if its confidence is at least \( \delta \). After this step, only frequent and high-confidence event pairs remain and form the nodes in \( L_2 \).

Step 2.2. Mining frequent \( 2 \)-event patterns: This step finds frequent \( 2 \)-event patterns from the nodes in \( L_2 \) (Alg. 1, lines 12-14). For each node \( (E_i, E_j) \in L_2 \), we use the bitmap \( b_{ij} \) to retrieve the set of sequences \( S \) where both events are present. Next, for each sequence \( S \in S \), the pairs of event instances \( (e_i, e_j) \) are extracted, and the relations between them are verified. The support and confidence of each relation \( r(E_{i, e_i}, E_{j, e_j}) \) are computed and compared against the thresholds, after which only frequent relations are selected and stored in the corresponding node in \( L_2 \). Examples of the relations in \( L_2 \) can be seen in Fig. 4, e.g., node (SOn, TOn).

Step 2.2 results in two different sets of nodes in \( L_2 \). The first set contains nodes that have frequent events but do not have any frequent patterns. These nodes (colored in brown in Fig. 4) are removed from \( L_2 \). The second set contains nodes that have both frequent events and frequent patterns (colored in green), which remain in \( L_2 \) and are used in the subsequent mining steps.

**Complexity:** Let \( m \) be the number of frequent single events in \( L_1 \), and \( i \) be the average number of event instances of each frequent event. The complexity of frequent \( 2 \)-event pattern mining is \( O(m^2 i^2 |D_{SEQ}|^2) \).

4.5 Mining Frequent \( k \)-event Patterns

Mining frequent \( k \)-event patterns \((k \geq 3)\) follows a similar process as \( 2 \)-event patterns, with additional prunings based on the transitivity property of temporal relations.

Step 3.1. Mining frequent \( k \)-event combinations: This step finds frequent \( k \)-event combinations in \( L_4 \) (Alg. 1, lines 15-17). Let \((k-1)Freq\) be the set of frequent \((k-1)\)-event combinations found in \( L_{k-1} \), and \( 1Freq \) be the set of frequent single events in \( L_1 \). To generate all \( k \)-event combinations, the typical process is to compute the Cartesian product: \((k-1)Freq \times 1Freq\). However, we observe that using \( 1Freq \) to generate \( k \)-event combinations at \( L_4 \) can create redundancy, since \( 1Freq \) might contain events that when combined with nodes in \( L_{k-1} \), result in combinations that clearly cannot form any frequent patterns. To illustrate this observation,
consider node IOn at L₁ in Fig. 4. Here, IOn is a frequent event, and thus, can be combined with frequent nodes in L₂ such as (SOn, TOn) to create a 3-event combination (SOn, TOn, IOn). However, (SOn, TOn, IOn) cannot form any frequent 3-event patterns, since IOn is not present in any frequent 2-event patterns in L₂. To reduce the redundancy, the combination (SOn, TOn, IOn) should not be created in the first place. We rely on the transitivity property of temporal relations to identify such event combinations.

Lemma 4. Let \( S = \langle e₁, ..., e_{n-1} \rangle \) be a temporal sequence that supports an \((n-1)\)-event pattern \( P = \langle (r_{12}, E_{e₁e₂}, E_{e₂e₃}), ..., (r_{(n-2)-(n-1)}, E_{e_{n-2}e_{n-1}}, E_{e_{n-1}e₁}) \rangle \). Let \( e_n \) be a new event instance added to \( S \) to create the temporal sequence \( S = \langle e₁, ..., e_n \rangle \).

The set of temporal relations \( R \) is transitive on \( S : \forall e₁ + i, x < n, \exists r \in R \) s.t. \( r(E_{eᵢeᵢ₊₁}) \) holds.

Lemma 4 says that given a temporal sequence \( S \), a new event instance added to \( S \) will always form at least one temporal relation with existing instances in \( S \). This is due to the transitivity property, which can be used to prove the following lemma.

Lemma 5. Let \( N_{k-1} = \langle E₁, ..., E_{k-1} \rangle \) be a frequent \((k-1)\)-event combination, and \( E_k \) be a frequent single event. The combination \( N_k = N_{k-1} \cup E_k \) can form frequent \( k \)-event temporal patterns if \( \forall E_l \in N_{k-1}, \exists r \in R \) s.t. \( r(E_l, E_k) \) is a frequent temporal relation.

From Lemma 4, only single events in L₁ that occur in L₁ could be used to create k-event combinations. Using this result, a filtering on \( IFreq \) is performed before calculating the Cartesian product. Specifically, from the nodes in L₁, we extract the distinct single events D₁, and intersect them with \( IFreq \) to remove redundant single events: \( FilteredIFreq = D₁ \cap IFreq \). Next, the Cartesian product \((k-1)Freq \times FilteredIFreq \) is calculated to generate k-event combinations. Finally, we apply Lemmas 2 and 3 to select frequent and high-confidence k-event combinations \( kFreq \) to form L₂.

Step 3.2 Mining frequent \( k \)-event patterns: This step finds frequent k-event patterns from the nodes in L₁ (Alg. 1, lines 18-20). Unlike 2-event patterns, determining the relations in a k-event combination \( \geq 3 \) is much more expensive, as it requires to verify the frequency of \( \frac{k(k-1)}{2} \) triples. To reduce the cost of relation checking, we propose an iterative verification method that relies on the transitivity property and the Apriori principle.

Lemma 6. Let \( P \) and \( P' \) be two temporal patterns. If \( P' \subseteq P \), then \( \text{conf}(P') \geq \text{conf}(P) \).

Lemma 7. Let \( P \) and \( P' \) be two temporal patterns. If \( P' \subseteq P \) and \( \frac{\max_{\text{supp}(E_k \subseteq P \cup \text{supp}(E_k))} \text{supp}(E_k)}{\text{supp}(P')} \leq \delta \), then \( \text{conf}(P') \leq \delta \).

Lemma 6 says that, the confidence of a pattern \( P \) is always at most the confidence of its sub-patterns. Consequently, from Lemma 7, a temporal pattern \( P \) cannot be high-confidence if any of its sub-patterns are low-confidence.

Let \( N_{k-1} = \langle E₁, ..., E_{k-1} \rangle \) be a node in L₁, \( N₁ = \langle E₁ \rangle \) be a node in L₁, and \( N₁ = N_{k-1} \cup N₁ = \langle E₁, ..., E_{k-1} \rangle \) be a node in L₂. To find k-event patterns for \( L₂ \), we first retrieve the set \( P_{k-1} \) containing frequent (k-1)-event patterns in node \( N_{k-1} \). Each \( p_{k-1} \in P_{k-1} \) is a list of \( \frac{k(k-1)}{2} \) triples: \( (r_{12}, E_{e₁e₂}, E_{e₂e₃}), ..., (r_{(k-2)-(k-1)}, E_{e_{k-2}e_{k-1}}, E_{e_{k-1}e₁}) \). We iteratively verify the possibility of \( p_{k-1} \) forming a frequent k-event pattern with \( E_k \) as follows:

We first check whether the triple \( (r_{(k-1)k}, E_{e₁e₂e₃}, E_{e₃e₁}) \) is frequent and high-confidence by accessing the node \( (E_{e₁e₂e₃}, E_{e₃e₁}) \) in L₂. If the triple is not frequent (using Lemmas 4 and 5) or high-confidence (using Lemmas 4, 6, and 7), the verifying process stops immediately for \( p_{k-1} \). Otherwise, it continues on the triple \( (r_{(k-2)k}, E_{e₂e₃e₁}, E_{e₁e₂}) \), until it reaches \( (r_{kk}, E_{e₁e₂e₃}, E_{e₃e₁}) \). We note that the transitivity property of temporal relations has been exploited in [32] to generate new relations. Instead, we use this property to prune unpromising candidates (Lemmas 4, 5, 6, 7).

Complexity: Let \( r \) be the average number of frequent \((k-1)\)-event patterns in \( L_{k-1} \). The complexity of frequent k-event pattern mining is \( O(|IFreq| \cdot |L_{k-1}| \cdot r \cdot k^2 \cdot |DFreq|) \).

HTPGM overall complexity: Throughout this section, we have seen that HTPGM complexity depends on the size of the search space \((O(m^h n^k))\) and the complexity of the mining process itself, i.e., \( O(m |DFreq|) + O(m^2 n^2 |DFreq|^2) + O(|IFreq| \cdot |L_{k-1}| \cdot r \cdot k^2 \cdot |DFreq|) \). While the parameters \( m, h, i, r \) and \( k \) depend on the number of time series, other factors such as \( |IFreq| \), \( |L_{k-1}| \) and \( |DFreq| \) also depend on the number of temporal sequences. Thus, given a dataset, HTPGM complexity is driven by two main factors: the number of time series and the number of temporal sequences.

5 APPROXIMATE HTPGM

5.1 Correlated Symbolic Time Series

Let \( X_S \) and \( Y_S \) be the symbolic series representing the time series \( X \) and \( Y \), respectively, and \( S_X, S_Y \) be their symbolic alphabets.

Definition 5.1 (Entropy) The entropy of \( X_S \), denoted as \( H(X_S) \), is defined as

\[
H(X_S) = - \sum_{x \in S_X} p(x) \cdot \log p(x)
\] (7)

Intuitively, the entropy measures the amount of information or the inherent uncertainty in the possible outcomes of a random variable. The higher the \( H(X_S) \), the more uncertain the outcome of \( X_S \).

The conditional entropy \( H(X_S|Y_S) \) quantifies the amount of information needed to describe the outcome of \( X_S \), given the value of \( Y_S \), and is defined as

\[
H(X_S|Y_S) = - \sum_{x \in S_X} \sum_{y \in S_Y} p(x,y) \cdot \log \frac{p(x,y)}{p(y)}
\] (8)

Definition 5.2 (Mutual information) The mutual information of two symbolic series \( X_S \) and \( Y_S \), denoted as \( I(X_S; Y_S) \), is defined as

\[
I(X_S; Y_S) = \sum_{x \in S_X} \sum_{y \in S_Y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}
\] (9)

The MI represents the reduction of uncertainty of one variable (e.g., \( X_S \)), given the knowledge of another variable (e.g., \( Y_S \)). The larger \( I(X_S; Y_S) \), the more information is shared between \( X_S \) and \( Y_S \), and thus, the less uncertainty about one variable given the other.

We demonstrate how to compute the MI between the symbolic series \( S \) and \( T \) in Table 1. We have: \( p(SOn)=\frac{17}{35}, p(SOff)=\frac{19}{35}, p(TOn)=\frac{14}{35}, \) and \( p(TOff)=\frac{21}{35}. \) We also have the joint probabilities: \( p(SOn,TOn)=\frac{12}{35}, p(SOn,TOff)=\frac{16}{35}, p(SOff,TOn)=\frac{3}{35}, \) and \( p(SOff,TOff)=\frac{3}{35}. \) Applying Eq. 9, we have \( I(S; T) = 0.29 \).

Since \( 0 \leq I(X_S; Y_S) \leq \min(H(X_S), H(Y_S)) \) [10], the MI value has no upper bound. To scale the MI into the range \([0, 1]\), we use normalized mutual information as defined below.

Definition 5.3 (Normalized mutual information) The normalized mutual information (NMI) of two symbolic time series \( X_S \) and \( Y_S \), denoted as \( \tilde{I}(X_S; Y_S) \), is defined as

\[
\tilde{I}(X_S; Y_S) = \frac{I(X_S; Y_S)}{H(X_S)} = 1 - \frac{H(X_S|Y_S)}{H(X_S)}
\] (10)
We is reduced by \( (X) \) shows that NMI is not symmetric, i.e., \( \tilde{T}(X; Y) \neq \tilde{T}(Y; X) \).

Using Table 1, we have \( I(S; T) = 0.29 \). However, we do not know what the 0.29 reduction means in practice. Applying Eq. (10), we can compute NMI \( I(S; T) = 0.43 \), which says that the uncertainty of \( S \) is reduced by 43% given \( T \). Moreover, we also have \( I(T; S) = 0.42 \), showing that \( I(S; T) \neq I(T; S) \).

**Definition 5.4** (Correlated symbolic time series) Let \( 0 < \mu \leq 1 \) be the mutual information threshold. We say that the two symbolic series \( X_S \) and \( Y_S \) are correlated iff \( \tilde{T}(X; Y) \geq \mu \vee \tilde{T}(Y; X) \geq \mu \), and uncorrelated otherwise.

**5.2 Lower Bound of the Confidence**

5.2.1 *Derivation of the lower bound.* Consider 2 symbolic series \( X_S \) and \( Y_S \). Let \( X_i \) be a temporal event in \( X_S \), and \( Y_i \) be a temporal event in \( Y_S \), and \( D_{SYB} \) and \( D_{SEQ} \) be the symbolic and the sequence databases created from \( X_S \) and \( Y_S \), respectively. We first study the relationship between the support of \( (X_i, Y_i) \) in \( D_{SYB} \) and \( D_{SEQ} \).

**Lemma 8.** Let \( supp(X_i, Y_i) \) \( D_{SYB} \) and \( supp(X_i, Y_i) \) \( D_{SEQ} \) be the support of \( (X_i, Y_i) \) in \( D_{SYB} \) and \( D_{SEQ} \), respectively. We have the following relation: \( supp(X_i, Y_i) \) \( D_{SYB} \leq supp(X_i, Y_i) \) \( D_{SEQ} \).

From Lemma 8, if an event pair is frequent in \( D_{SYB} \), it is also frequent in \( D_{SEQ} \). We now investigate the connection between \( \tilde{T}(X_S; Y_S) \) in \( D_{SYB} \), and the confidence of \( (X_i, Y_i) \) in \( D_{SEQ} \).

**Theorem 1.** (Lower bound of the confidence) Let \( \sigma \) and \( \mu \) be the minimum support and mutual information thresholds, respectively. Assume that \( (X_i, Y_i) \) is frequent in \( D_{SEQ} \), i.e., \( supp(X_i, Y_i) \) \( D_{SEQ} \geq \sigma \).

If the NMI \( \tilde{T}(X_S; Y_S) \geq \mu \), then the confidence of \( (X_i, Y_i) \) in \( D_{SEQ} \) has a lower bound:

\[
\text{conf}(X_i, Y_i) \geq \sigma \cdot \frac{\mu}{\lambda_1} \cdot \left( \frac{n_x - 1}{1 - \sigma} \right)^{\frac{1}{\nu_x}}
\]

\[ (11) \]

where \( n_x \) is the number of symbols in \( \Sigma_X \), \( \lambda_1 \) is the minimum support of \( X_i \) in \( X_S \), and \( \lambda_2 \) is the support of \( (X_i, Y_j) \) \( (X_S, Y_S) \) such that \( p(X_i|Y_j) \) is minimal, \( \forall (i \neq 1 \& j \neq 1) \).

**Proof.** (Sketch - Detailed proof in [19]). From Eq. (10), we have:

\[
H(X_S|Y_S) = 1 - H(X|Y) \geq \mu
\]

\[ (12) \]

\[
\frac{H(X_S|Y_S)}{H(X_S)} = - \frac{1}{\lambda_2} \sum p(X_i) \log p(X_i)
\]

\[
+ \sum_{i \neq 1 \& j \neq 1} p(X_i, Y_j) \log \frac{p(X_i, Y_j)}{p(Y_j)} \leq 1 - \mu
\]

\[ (13) \]

Let \( \lambda_1 = p(X_k) \) such that \( p(X_k) = \min (p(X_i)) \forall i \), and \( \lambda_2 = p(Y_m, Y_n) \) such that \( p(Y_m|Y_n) = \min (p(Y_i|Y_j)) \forall (i \neq 1 \& j \neq 1) \). Then, by applying the min-max inequality theorem for the sum of ratio \( [5] \) to the numerator of Eq. (13), we obtain:

\[
\begin{align*}
H(X_S|Y_S) & \geq \frac{p(X_i, Y_i) \cdot \log p(X_i|Y_i) + \lambda_2 \cdot \log \frac{1-p(X_i, Y_i)}{n_x - p(Y_i)}}{\log \lambda_1} \\
& \geq \sigma \cdot \frac{p(X_i, Y_i)}{p(Y_i)} + \lambda_2 \cdot \log \frac{1-\sigma}{n_x - 1} \\
& \geq \frac{\sigma \cdot \frac{1-\sigma}{n_x - 1}}{\log \lambda_1}
\end{align*}
\]

\[ (14) \]

Next, assume that \( supp(Y_i) \) \( D_{SYB} \geq supp(X_i) \) \( D_{SYB} \). From Eqs. (13), (14), the confidence lower bound of \( (X_i, Y_i) \) in \( D_{SYB} \) is derived as:

\[
\text{conf}(X_i, Y_i) \geq \frac{\sigma \cdot \frac{1-\sigma}{n_x - 1}}{\log \lambda_1}
\]

\[ (15) \]

Since:

\[
\text{conf}(X_i, Y_i) \geq \frac{\sigma \cdot \text{conf}(X_i, Y_i)}{\log \lambda_1}
\]

\[ (16) \]

It follows that:

\[
\text{conf}(X_i, Y_i) \geq \frac{\sigma \cdot \frac{1-\sigma}{n_x - 1}}{\log \lambda_1}
\]

\[ (17) \]

**Interpretation of the confidence lower bound:** Theorem 1 says that, given an MI threshold \( \mu \), if the two symbolic series \( X_S \) and \( Y_S \) are correlated, then the confidence of a frequent event pair in \( (X_S, Y_S) \) is at least the lower bound in Eq. (11). Combining Theorem 1 and Lemma 3, we can conclude that given \( (X_S, Y_S) \), if its event pair has a confidence less than the lower bound, then any pattern \( P \) formed by that event pair also has a confidence less than that lower bound. This allows to approximate HTPGM (discussed in Section 5.3).

5.2.2 *Shape of the confidence lower bound.* To understand how the confidence changes w.r.t. the support \( \sigma \) and the MI \( \mu \), we analyze its shape, shown in Fig. 5 (\( \sigma \) and \( \mu \) vary between 0 and 1). First, it can be seen that the confidence lower bound has a direct relationship with \( \sigma \) and \( \mu \) (one increases if the other increases and vice versa).

While the direct relationship between the confidence and \( \sigma \) can be explained using Eq. (5), it is interesting to observe the connection between \( \mu \) and the confidence. As the MI represents the correlation between two symbolic series, the larger the value of \( \mu \), the more correlated the two series. Thus, when the confidence increases together with \( \mu \), it implies that patterns with high confidence are more likely to be found in highly correlated series, and vice versa.

**Algorithm 2:** Approximate HTPGM using MI

**Input:** A set of time series \( X \), an MI threshold \( \mu \), support threshold \( \sigma \), confidence threshold \( \delta \)

**Output:** The set of frequent temporal patterns \( P \)

1. convert \( X \) to \( D_{SYB} \) and \( D_{SEQ} \);
2. scan \( D_{SYB} \) to compute the probability of each event and event pair;
3. foreach pair of symbolic time series \( (X_S, Y_S) \) \( D_{SYB} \) do
4. compute \( \tilde{T}(X_S; Y_S) \) and \( \tilde{T}(Y_S; X_S) \);
5. compute \( \mu \);
6. if \( \tilde{T}(X_S; Y_S) \geq \mu \vee \tilde{T}(Y_S; X_S) \geq \mu \) then
7. insert \( X_S \) and \( Y_S \) into \( C_G \); 8. create an edge between \( X_S \) and \( Y_S \) in \( G_C \);
9. \( X_S \) \( X_C \) do
10. mine frequent single events from \( X_S \);
11. foreach event pair \( (E_i, E_j) \) in \( L_1 \) do
12. if there is an edge between \( X_S \) and \( Y_S \) in \( G_C \) then
13. mine frequent patterns for \( (E_i, E_j) \);
14. if \( k \geq 3 \) then
15. perform HTPGM using \( L_1 \) and \( L_2 \).
Fig. 5 also shows that, when \( \sigma \) is low, e.g., \( \sigma < 0.1 \), we obtain a very low value of the confidence lower bound regardless of \( \mu \) value. This implies that the confidence is less sensitive to \( \mu \) when the support is low. The opposite is obtained when the support is high, e.g., \( \sigma > 0.1 \), where we see a visible increase of the confidence lower bound as \( \mu \) increases. This indicates that the "insensitive" area of the lower bound (when \( \sigma \leq 0.1 \)) is less accurate than the "sensitive" area (\( \sigma > 0.1 \)) when performing the approximate mining, as we will discuss in Section 6.

5.3 Using the Bound to Approximate HTPGM

5.3.1 Correlation graph. Using Theorem 1, we propose to approximate HTPGM by performing the mining only on the set of correlated symbolic series \( X_C \subseteq X \). We first define the correlation graph.

Definition 5.5 (Correlation graph) A correlation graph is an undirected graph \( G_C = (V, E) \) where \( V \) is the set of vertices, and \( E \) is the set of edges. Each vertex \( v \in V \) represents one symbolic series \( X_v \in X_C \). There is an edge \( e_{uv} \) between a vertex \( u \) containing \( X_u \) and a vertex \( v \) containing \( X_v \) if \( \overline{I}(X_u; X_v) \geq \mu \lor \overline{I}(Y_u; Y_v) \geq \mu \).

Fig. 6 shows an example of the correlation graph \( G_C \) built from \( D_{SYB} \) in Table 1. Here, each node corresponds to one electrical appliance. There is an edge between two nodes if their NMI is at least \( \mu \). The number on each edge is the NMI between two nodes.

Constructing the correlation graph: Given a symbolic database \( D_{SYB} \), the correlation graph \( G_C \) can easily be constructed by computing the NMI for each symbolic series pair, and comparing their NMI against the threshold \( \mu \). A symbolic series pair is included in \( G_C \) if their NMI is at least \( \mu \), and vice versa.

Setting the value of \( \mu \): While NMI can easily be computed using Eq. (10), it is not trivial how to set the value for \( \mu \). Here, we propose a method to determine \( \mu \) using the lower bound in Eq. (11).

Recall that HTPGM relies on two user-defined parameters, the support threshold \( \sigma \) and the confidence threshold \( \delta \), to look for frequent temporal patterns. Based on the confidence lower bound in Theorem 1, we can derive \( \mu \) using \( \sigma \) and \( \delta \) as the following:

**Corollary 1.1.** The confidence of an event pair \((X_1, Y_1) \in (X_S, Y_S)\) in \( D_{SEQ} \) is at least \( \mu \) if \( \overline{I}(X_1; Y_1) \geq \mu \), where:

\[
\mu \geq 1 - \sigma \cdot \log_3 \left( \frac{\delta}{\sigma} \cdot \frac{1 - \sigma}{\sigma_1 - 1} \right) \quad (18)
\]

Note that \( \mu \) in Eq. (18) only ensures that the event pair \((X_1, Y_1)\) has a minimum confidence of \( \delta \). Thus, given \((X_S, Y_S)\), \( \mu \) has to be computed for each event pair in \((X_S, Y_S)\). The final chosen \( \mu \) value to be compared against \( \overline{I}(X_1; Y_1) \) is the minimum \( \mu \) value among all the event pairs in \((X_S, Y_S)\).

5.3.2 Approximate HTPGM using the correlation graph. Using the correlation graph \( G_C \), the approximate HTPGM is described in Algorithm 2. First, \( D_{SYB} \) is scanned once to compute the probability of each single event and pair of events (line 2). Next, NMI and \( \mu \) are computed for each pair of symbolic series \((X_v, Y_v)\) in \( D_{SYB}\) (lines 4-5). Then, only pairs whose \( \overline{I}(X_v; Y_v) \) or \( \overline{I}(Y_v; X_v) \) is at least \( \mu \) are inserted into \( X_C \), and an edge between \( X_v \) and \( Y_v \) is created (lines 6-8). Next, at \( L_1 \) of HPG, only the correlated symbolic series in \( X_C \) are used to mine frequent single events (lines 9-10). At \( L_2 \), \( G_C \) is used to filter 2-event combinations: for each event pair \((E_i, E_j)\), we check whether there is an edge between the corresponding

### Table 4: Parameters and values

| Params       | Values                                      |
|--------------|---------------------------------------------|
| Support \( \sigma \) | User-defined: \( \sigma = 0.5\%, 1\%, 10\%, 20\%, \ldots \) |
| Confidence \( \delta \) | User-defined: \( \delta = 0.5\%, 1\%, 10\%, 20\%, \ldots \) |
| Overlapping duration \( t_{ov} \) | \( t_{ov} \) (hours) = 0, 1, 2, 3 (NIST, UKDALE, DataPort, and Smart City) |
| Tolerance buffer \( \epsilon \) | \( \epsilon \) (mins) = 0, 1, 2, 3 (NIST, UKDALE, DataPort), \( \epsilon \) (mins) = 0, 5, 10, 15 (Smart City), \( \epsilon \) (frames) = 0, 30, 45, 60 (ASL) |

6 EXPERIMENTAL EVALUATION

We evaluate HTPGM (both exact and approximate), using real-world datasets from three application domains: smart energy, smart city, and sign language. Due to space limitations, we only present here the most important results, and discuss other findings in [19].

6.1 Experimental Setup

Datasets: We use 3 smart energy datasets, NIST [14], UKDALE [26], and DataPort [13], all of which measure the energy/power consumption of electrical appliances in residential households. For the smart city, we use weather and vehicle collision data obtained from NYC Open Data Portal [9]. For sign language, we use the American Sign Language (ASL) datasets [33] containing annotated video sequences of different ASL signs and gestures. Table 5 summarizes their characteristics.

Baseline methods: Our exact method is referred to as E-HTPGM, and the approximate one as A-HTPGM. We use 4 baselines (described in Section 2): Z-Miner [28], TPMiner [8], IEMiner [36], and H-DFS [35]. Since E-HTPGM and the baselines provide the same exact solutions, we use the baselines only for the quantitative evaluation, and compare only E-HTPGM and A-HTPGM qualitatively.

Infrastructure: The experiments are run on virtual machines (VM) with AMD EPYC Processor 32 cores (2GHz) CPU, 256 GB main memory, and 1 TB storage. For scalability evaluation, we use VMs with 512 GB main memory.

Parameters: Table 4 lists the parameters and their values used in our experiments.

6.2 Qualitative Evaluation

Our goal is to make sense and learn insights from extracted patterns. Table 7 lists some interesting patterns found in the datasets.

Patterns P1 - P9 are extracted from the energy datasets, showing how the residents interact with electrical devices in their houses. Patterns P10 - P15 extracted from the smart city datasets, while patterns P16 - P19 are from the ASL dataset.
6.3 Quantitative Evaluation

6.3.1 Baselines comparison on real world datasets. We compare E-HTPGM and A-HTPGM with the baselines in terms of the runtime and memory usage. Tables 8 and 9 show the experimental results on the energy and the smart city datasets. The quantitative results of other datasets are reported in the full paper [19].

As shown in Table 8, A-HTPGM achieves the best runtime among all methods, and E-HTPGM has better runtime than the baselines. On the tested datasets, the range and average speedups of A-HTPGM compared to other methods are: [1.21-4.82] and 2.31 (E-HTPGM), [2.52-25.86] and 7.85 (Z-Miner), [7.43-69.68] and 21.65 (TPMiner), [8.61-188.16] and 40.75 (IE Miner), and [14.50-332.98] and 61.36 (H-DFS). The speedups of E-HTPGM compared to the baselines are: [1.47-5.64] and 3.19 on average (Z-Miner), [3.59-30.97] and 9.08 on avg. (TPMiner), [4.63-78.41] and 15.86 on avg. (IE Miner), and [5.54-118.21] and 23.37 on avg. (H-DFS). Note that the time to compute MI and μ for the NIST and the smart city datasets in Table 8 are 28.01 and 20.82 seconds, respectively.

Moreover, A-HTPGM is most efficient, i.e., achieves highest speedup and memory saving, when the support threshold is low, e.g., σ = 20%. This is because typical datasets often contain many patterns with very low support and confidence. Thus, using A-HTPGM to prune uncorrelated series early helps save computational time and resources. However, the speedup comes at the cost of a small loss in accuracy (discussed in Sections 6.3.2 and 6.3.4).

In terms of memory consumption, as shown in Table 9, A-HTPGM is the most efficient method, while E-HTPGM is more efficient than the baselines. The range and the average memory consumption of A-HTPGM compared to other methods are: [1.1-3.2] and 1.6 (E-HTPGM), [3.7-105.1] and 19.1 (Z-Miner), [1.3-7.9] and 3.4 (TPMiner), [1.4-10.4] and 4.5 (IE Miner), and [2.1-13.9] and 6.7 (H-DFS). The memory usage of E-HTPGM compared to the baselines are: [2.9-52.5] and 11.4 on avg. (Z-Miner), [1.2-4.7] and 2.1 on average (TPMiner), [1.3-6.2] and 2.7 on avg. (IE Miner), and [1.9-7.5] and 4.1 on avg. (H-DFS).

Finally, in Table 11, we provide the pre-processing times to convert the raw time series to $D_{SYB}$, $D_{SEYB}$ to $D_{SEQ}$. We also report the sizes of $D_{SYB}$ and $D_{SEQ}$ stored on disk. We see that while the storage costs for $D_{SYB}$ and $D_{SEQ}$ are small, the pre-processing times are 10-25 seconds. This is a one-time cost which can be reused for many mining runs, making it negligible in all non-trivial cases.

6.3.2 Scalability evaluation on synthetic datasets. As discussed in Section 4, the complexity of HTPGM is driven by two main factors: (1) the number of temporal sequences, and (2) the number of time series. The evaluation on real-world datasets has shown that E-HTPGM and A-HTPGM outperform the baselines significantly in both runtimes and memory usage. However, to further assess the scalability, we scale these two factors on synthetic datasets. Specifically, starting from the real-world datasets, we generate 10 times more sequences, and create up to 1000 synthetic time series. We evaluate the scalability using two configurations: varying the number of sequences, and varying the number of time series.

Figs. 7 and 8 show the runtimes of A-HTPGM, E-HTPGM and the baselines when the number of sequences changes (y-axis is in log scale). The range and average speedups of A-HTPGM w.r.t. other methods are: [1.5-3.7] and 2.5 (E-HTPGM), [3.1-13.6] and 8.1 (Z-Miner), [5.1-31.2] and 16.8 (TPMiner), [6.4-45.8] and 24.9 (IE Miner), and [9.4-59.1] and 31.8 (H-DFS). In particular, A-HTPGM obtains even higher speedup for more sequences. Similarly, the range and average speedups of E-HTPGM are: [1.6-5.3] and 3.2 (Z-Miner), [2.2-12.1] and 6.7 (TPMiner), [3.5-17.4] and 10.1 (IE Miner), and [4.9-22.8] and 12.9 (H-DFS).

Figs. 9 and 10 compare the runtimes of A-HTPGM with other methods when changing the number of time series (y-axis is in log scale). It is seen that, A-HTPGM achieves even higher speedup with more time series. The range and average speedups of A-HTPGM are: [2.1-4.9] and 2.9 (E-HTPGM), [2.9-10.4] and 6.8 (Z-Miner), [3.6-21.5] and 12.8 (TPMiner), [4.7-30.2] and 18.1 (IE Miner), and [6.1-39.6] and 23.2 (H-DFS), and of E-HTPGM are: [1.4-4.1] and 2.4 (Z-Miner), [1.7-8.1] and 4.4 (TPMiner), [2.3-11.3] and 6.2 (IE Miner), and [2.7-16.3] and 8.1 (H-DFS).

In Figs. 9 and 10, to illustrate the computation time of MI and $\mu$, we add an additional bar chart for A-HTPGM. Each bar represents the runtime of A-HTPGM with two separate components: the time to compute MI and $\mu$ (top red), and the mining time (bottom blue). However, note that for each dataset, we only need to compute MI and $\mu$ once (the computed values are used across the mining process with different support and confidence thresholds). Thus, the times to compute MI and $\mu$, for example, in Figs. 9a, 9b, and 9c, are added and not actually used.

Moreover, most baselines fail for the larger configurations in the scalability study, e.g., Z-Miner on the NIST dataset when $\sigma=20\%$ (Fig. 7a), and Z-Miner, TPMiner, IE Miner and H-DFS when the number of time series grows to 1000 (Fig. 9a). The scalability test shows that A-HTPGM and E-HTPGM can scale well on big datasets, both vertically (many sequences) and horizontally (many time series), unlike the baselines.

Furthermore, the number of time series and events pruned by A-HTPGM in the scalability test are provided in Table 10. Here, we can see that high confidence threshold leads to more time series (events) to be pruned. This is because confidence has a direct relationship with MI, therefore, high confidence results in higher $\mu$, and thus, more pruned time series.

6.3.3 Evaluation of the pruning techniques in E-HTPGM. We compare different versions of E-HTPGM to understand how effective the pruning techniques are: (1) NoPrune: E-HTPGM with no pruning, (2) Apriori: E-HTPGM with Apriori-based pruning (Lemmas 2, 3), (3) Trans: E-HTPGM with transitivity-based pruning (Lemmas 4, 5, 6, 7), and (4) All: E-HTPGM applied both pruning techniques.

We use 3 different configurations that vary: the number of sequences, the confidence, and the support. Figs. 11, 12 show the results (the y-axis is in log scale). It can be seen that (All)-E-HTPGM achieves the best performance among all versions. Its speedup w.r.t. (NoPrune)-E-HTPGM ranges from 5 up to 60 depending on the configurations, showing that the proposed prunings are very effective in improving E-HTPGM performance. Furthermore, (Trans)-E-HTPGM delivers larger speedup than (Apriori)-E-HTPGM. The average speedup is from 8 to 20 for (Trans)-E-HTPGM, and from 3 to 12 for (Apriori)-E-HTPGM. However, applying both always yields better speedup than applying either of them.

6.3.4 Evaluation of A-HTPGM. We proceed to evaluate the accuracy of A-HTPGM and the quality of patterns pruned by A-HTPGM.

To evaluate the accuracy, we compare the patterns extracted by A-HTPGM and E-HTPGM. Table 6 shows the accuracies of
Table 5: Characteristics of the Datasets

| Method   | NIST | UKDALE | DataPort | Smart City | ASL |
|----------|------|--------|----------|------------|-----|
| # sequences | 1460 | 1520 | 1460 | 1216 | 1908 |
| # variables | 49   | 24   | 21  | 26  | 25  |
| # distinct events | 98   | 48   | 82  | 130 | 174 |
| # instances/seq | 55   | 190  | 49  | 162 | 20  |

Table 6: The Accuracy of A-HTPGM (%)

| Method   | NIST | UKDALE | DataPort | Smart City | ASL |
|----------|------|--------|----------|------------|-----|
| Supp. (%) | 10   | 87     | 89       | 91         | 94  |
| Conf. (%) | 50   | 50     | 50       | 50          | 50  |

Table 7: Summary of Interesting Patterns

| Pattern | NIST | Conf. (%) | UKDALE | Conf. (%) |
|---------|------|-----------|--------|-----------|
| (P1) [[05:58, 08:24]] First Floor Lights | 10 | 30 |
| (P2) [[06:00, 07:31]] Upstairs Bathroom Lights | 20 | 20 |
| (P3) [[18:01, 20:16]] Children Room Plugs | 25 | 25 |
| (P4) [[18:59, 21:00]] Kitchen Lights | 25 | 25 |
| (P5) [[06:02, 09:17]] Kitchen Lights | 25 | 25 |
| (P6) [[18:15, 19:00]] Kitchen Plugs | 25 | 25 |
| (P7) [[16:15, 17:30]] Washer | 25 | 25 |
| (P8) [[18:00, 18:25]] Kitchen Lights | 25 | 25 |
| (P9) [[18:05, 18:10]] Microwave | 25 | 25 |

Table 8: Runtime Comparison (seconds)

| Methods   | NIST | Conf. (%) | UKDALE | Conf. (%) |
|-----------|------|-----------|--------|-----------|
| H-DFS     | 591.1 | 461.3 | 494.1 | 461.3 |
| EMiner    | 547.7 | 481.9 | 461.3 | 461.3 |
| Z-Miner   | 538.6 | 461.3 | 481.9 | 481.9 |
| E-HTPGM   | 529.6 | 481.9 | 481.9 | 481.9 |
| A-HTPGM   | 522.6 | 478.6 | 481.9 | 481.9 |

Table 9: Memory Usage Comparison (MB)

| Methods   | NIST | Conf. (%) | UKDALE | Conf. (%) |
|-----------|------|-----------|--------|-----------|
| H-DFS     | 617.7 | 434.6 | 434.6 | 434.6 |
| EMiner    | 575.9 | 434.6 | 434.6 | 434.6 |
| Z-Miner   | 568.7 | 434.6 | 434.6 | 434.6 |
| E-HTPGM   | 561.7 | 434.6 | 434.6 | 434.6 |
| A-HTPGM   | 554.7 | 434.6 | 434.6 | 434.6 |

Table 10: Pruned Time Series and Events from A-HTPGM

| Method   | NIST | Conf. (%) | UKDALE | Conf. (%) |
|-----------|------|-----------|--------|-----------|
| Pruned Time Series | 20   | 20       | 20     | 20        |
| Pruned Events     | 20   | 20       | 20     | 20        |

Table 11: Building DSBY and DSEQ

| Dataset | NIST | Conf. (%) | UKDALE | Conf. (%) |
|---------|------|-----------|--------|-----------|
| Time (sec) | 249.2 | 103       | 210.8 | 42        |
| Storage (MB) | 138.8 | 241       | 8.9   | 114       |

A-HTPGM for different supports and confidences. It is seen that, A-HTPGM obtains high accuracy (≥ 71%) when σ and δ are low, e.g., σ = δ = 10%, and very high accuracy (≥ 95%) when σ and δ are high, e.g., σ = δ = 50%. Next, we analyze the quality of patterns pruned by A-HTPGM. These patterns are extracted from the uncorrelated time series. Fig. 13 shows the cumulative distribution of the confidences of the pruned patterns. It is seen that most of these patterns have low confidences, and can thus safely be pruned. For NIST and Smart City, A-HTPGM obtains high accuracy (≥ 71%) when σ and δ are low, e.g., σ = δ = 10%, and very high accuracy (≥ 95%) when σ and δ are high, e.g., σ = δ = 50%. Next, we analyze the quality of patterns pruned by A-HTPGM.
This paper presents our comprehensive Frequent Temporal Pattern Graph Mining (E-HTPGM) algorithm that employs efficient data structures and multiple pruning techniques to achieve fast mining, and (3) an approximate A-HTPGM that uses mutual information to prune unpromising time series, allows HTPGM to scale on big datasets. Extensive experiments conducted on real-world and synthetic datasets show that both A-HTPGM and E-HTPGM out-perform the baselines, consume less memory, and scale well to big datasets. Compared to the baselines, the approximate A-HTPGM delivers an order of magnitude speedup on large synthetic datasets and up to 2 orders of magnitude speedup on real-world datasets. In future work, we plan to extend HTPGM to prune at the event level to further improve its performance.

City datasets, 80% of pruned patterns have confidences less than 20% when the support is 10% and 20%, and 70% of pruned patterns have confidences less than 30% when the support is 30%. For the ASL dataset, 80% of pruned patterns have confidences less than 5%.

Other experiments: We analyze the effects of the tolerance buffer $\epsilon$, and the overlapping duration $t_{ov}$, to the quality of extracted patterns. The analysis can be seen in the full paper [19].

7 CONCLUSION AND FUTURE WORK

This paper presents our comprehensive Frequent Temporal Pattern Mining from Time Series (FTPMfTS) solution that offers: (1) an end-to-end FTPMfTS process to mine frequent temporal patterns from time series, (2) an efficient and exact Hierarchical Temporal Pattern Graph Mining (E-HTPGM) algorithm that employs efficient data structures and multiple pruning techniques to achieve fast mining, and (3) an approximate A-HTPGM that uses mutual information to prune unpromising time series, allows HTPGM to scale on big datasets. Extensive experiments conducted on real-world and synthetic datasets show that both A-HTPGM and E-HTPGM out-perform the baselines, consume less memory, and scale well to big datasets. Compared to the baselines, the approximate A-HTPGM
REFERENCES

[1] Akiz Uddin Ahmed, Chowdhury Farhan Ahmed, Md Samiullah, Nahim Adnan, and Carson Kai-Sang Leung. 2016. Mining interesting patterns from uncertain databases. Information Sciences 354 (2016).

[2] James F Allen. 1983. Maintaining knowledge about temporal intervals. Commun. ACM 26 (1983).

[3] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. 2012. Mining recent temporal patterns for event detection in multi-variate time series data. In SIGKDD.

[4] Iyad Batal, Hamed Valizadegan, Gregory F Cooper, and Milos Hauskrecht. 2013. A temporal pattern mining approach for classifying electronic health record data. TIST 4 (2013).

[5] Edwin F Beckenbach, Richard Bellman, and Richard Ernest Bellman. 1964. An introduction to inequalities. Technical Report. Mathematical Association of America Washington, DC.

[6] Julien Blanchard, Fabrice Guillet, Regis Gras, and Henri Briand. 2005. Using information-theoretic measures to assess association rule interestingness. In ICMD’05.

[7] Elizabeth A Campbell, Ellen J Bass, and Aaron J Masino. 2020. Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma. JAMIA 27 (2020).

[8] Yi-Cheng Chen, Wen-Chih Peng, and Wen-Yueh Shih, Lin Yang, Yu-Shao Liu, and Chen-Chang Hsieh. 2018. A temporal condition pattern mining approach for classifying electronic health record data. In: Kutsche RD., Zimányi E. (eds) Big Data Management and Analytics. Cham, 2014.

[9] Jack Kelly and William Knottenbelt. 2015. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Scientific Data (2015).

[10] Vasileios Megalogiannakos, Liang Wang, Quoc Li, and Christos Faloutsos. 2005. A multiresolution symbolic representation of time series. In 21st International Conference on Data Engineering (ICDE’05). IEEE, 668–679.

[11] Robert Moskovitch and Yuval Shahar. 2015. Fast time intervals mining using the transitivity of temporal relations. KAIS 42 (2015).

[12] Edward R Omiecinski. 2003. Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15, 1 (2003), 57–69.

[13] Panagiotis Papapetrou, George Kollias, Stan Sclaroff, and Dimitrios Gunopulos. 2009. Mining frequent arrangements of temporal intervals. KAIS 21 (2009).

[14] Dhaavl Patel, Wynne Hsu, and Mong Li Lee. 2008. Mining relationships among interval-based events for classification. In SIGMOD.

[15] Amir Kumar Sharma and Dhaval Patel. 2018. Stipa: A memory efficient technique for interval pattern discovery. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, 1767–1776.

[16] Jun-Zhe Wang, Yi-Cheng Chen, Wen-Yueh Shih, Lin Yang, Yu-Shao Liu, and Chen-Chang Hsieh. 2018. Mining high-utility temporal patterns on interval-based data. ACM Transactions on Intelligent Systems and Technology (TIST) 11, 4 (2020), 1–31.

[17] Shu-Yi Wu and Yen-Liang Chen. 2007. Mining nonambiguous temporal patterns for interval-based events. TKDE 19 (2007).

[18] YY Yao. 2003. Information-theoretic measures for knowledge discovery and data mining. In Entropy measures, maximum entropy principle and emerging applications. 115–136.

[19] Torp K., Andersen O., Thomsen C. (2020) Travel-Time Computation Based on GPS Data. In: Kutsche RD., Zimányi E. (eds) Big Data Management and Analytics. eBIS 2019. Lecture Notes in Business Information Processing, vol 390. Springer, Cham. https://doi.org/10.1007/978-3-030-61627-4_4.
A DETAILED PROOFS OF COMPLEXITIES, LEMMAS AND THEOREMS

A.1 Mutual exclusive property of temporal relations

Property 1. (Mutual exclusive) Consider the set of temporal relations $\mathcal{R} = \{\text{Follows}, \text{Contains}, \text{Overlaps}\}$. Let $E_i$ and $E_j$ be two temporal events, and $e_i$ occurring during $[t_i, \tau_i]$, $e_j$ occurring during $[t_j, \tau_j]$ be their corresponding event instances, and $\epsilon$ be the tolerance buffer. The relations in $\mathcal{R}$ are mutually exclusive on $E_i$ and $E_j$.

Proof. * Case 1: Assume the relation $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$. Thus, we have:

$$t_{e_i} \pm \epsilon \leq t_{e_j}$$

and:

$$t_{e_j} < t_{e_i} \Rightarrow t_{e_i} \pm \epsilon < t_{e_j}$$

Hence, $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$, since $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $(t_{e_i} \leq t_{e_j}) \land (t_{e_i} \pm \epsilon \geq t_{e_j})$ (contradict Eq. (20)).

Similarly, $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$ since $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $(t_{e_j} < t_{e_i}) \land (t_{e_i} \pm \epsilon < t_{e_j}) \land (t_{e_j} - t_{e_j} \geq \epsilon \pm \epsilon)$ (contradict Eq. (21)).

In conclusion, if $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$, then the two remaining relations cannot exist between $E_i$ and $E_j$.

* Case 2: Assume the relation $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$. Thus, we have:

$$t_{e_i} \leq t_{e_j}$$

$$t_{e_j} \pm \epsilon \geq t_{e_i}$$

Hence, $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$ since $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $t_{e_i} \pm \epsilon < t_{e_j}$ (contradict Eq. (22)).

Similarly, $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$ since $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $(t_{e_j} < t_{e_i}) \land (t_{e_i} \pm \epsilon < t_{e_j}) \land (t_{e_i} - t_{e_i} \geq \epsilon \pm \epsilon)$ (contradict Eq. (23)).

In conclusion, if $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$, then the two remaining relations cannot exist between $E_i$ and $E_j$.

* Case 3: Assume the relation $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$. Thus, we have:

$$t_{e_i} < t_{e_j}$$

$$t_{e_i} \pm \epsilon < t_{e_j}$$

$$t_{e_j} - t_{e_i} \geq \epsilon \pm \epsilon \Rightarrow t_{e_j} \leq t_{e_i} - \epsilon \pm \epsilon$$

Hence, $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$, since $\text{Follows}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $t_{e_i} \pm \epsilon < t_{e_j}$ (contradict Eq. (24)).

Similarly, $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ cannot exist between $E_i$ and $E_j$, since $\text{Contains}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds iff $t_{e_i} \pm \epsilon \geq t_{e_j}$ (contradict Eq. (25)).

In conclusion, if $\text{Overlaps}(E_{i\epsilon_i}, E_{j\epsilon_j})$ holds between $E_i$ and $E_j$, then the two remaining relations cannot exist between $E_i$ and $E_j$.

A.2 Mining frequent single event

Complexity: The complexity of finding frequent single events is $O(m |\mathcal{D}_{SEQ}|)$, where $m$ is the number of distinct events.

Proof. Computing the support for each event $E_i$ takes $O(|\mathcal{D}_{SEQ}|)$ (to count the set bits of the bitmap $b_{E_i}$ of length $|\mathcal{D}_{SEQ}|$). Thus, counting the support for $m$ events takes $O(m |\mathcal{D}_{SEQ}|)$.  

A.3 Lemma 1

Lemma 1. Let $m$ be the number of events in $\mathcal{D}_{SEQ}$, and $h$ be the longest length of a temporal pattern. The total number of temporal patterns in HPG from $L_1$ to $L_h$ is $O(m^{h^3} h^3)$.  

Lemma 1 shows that the exponential search space of HTPGM is driving by: the number of distinct events ($m$) which is correlated to the number of time series in the datasets, the max pattern length ($h$), and the number of temporal relations ($3$). A dataset of just a few hundred events can create a very large search space with billions of candidate patterns.

Proof. At $L_1$, the number of nodes is: $N_1 = m \sim O(m)$. At $L_2$, the number of permutations of $m$ distinct events taken 2 at a time is: $P(m, 2)$. However, since the same event can form a pair of events with itself, e.g., $(S|O|S|O|)$, the total number of nodes at $L_2$ is: $N_2 = P(m, 2) + m - O(m^2)$. Each node in $N_2$ can form 3 different temporal relations, and thus, the total number of 2-event patterns in $L_2$ is: $N_2 \times 3^3 \sim O(m^23^3)$. Similarly, the number of 3-event nodes at $L_3$ is: $N_3 = P(m, 3) + P(m, 2) + m \sim O(m^3)$, and the number of 3-event patterns is: $N_3 \times 3^3 \sim O(m^33^3)$. At level $L_h$, the number of nodes is $O(m^h)$, while the number of $h$-event patterns is $O(m^h \times 3^{h(h-1)}) \sim O(m^h3^{h^2})$. Therefore, the total number of temporal patterns from $L_1$ to $L_h$ in HPG is $O(m) + O(m^23^3) + O(m^33^3) + \ldots + O(m^h3^{h^2}) \sim O(m^h3^{h^2})$.  

A.4 Lemma 2

Lemma 2. Let $P$ be a 2-event pattern formed by an event pair $(E_i, E_j)$. Then, $\text{supp}(P) \leq \text{supp}(E_i, E_j)$.

From Lemma 2, the support of a pattern is at most the support of its events. Thus, infrequent event pairs cannot form frequent patterns and thereby, can be safely pruned.

Proof. Derived directly from Defs. 3.11, 3.12, 3.13 and 3.14.  

A.5 Lemma 3

Lemma 3. Let $(E_i, E_j)$ be a pair of events occurs in a 2-event pattern $P$. Then $\text{conf}(P) \leq \text{conf}(E_i, E_j)$.

From Lemma 3, the confidence of a pattern $P$ is always at most the confidence of its events. Thus, a low-confidence event pair cannot form any high-confidence patterns and therefore, can be safely pruned.

Proof. Can derived directly from Defs. 3.15 and 3.16.  

A.6 Mining frequent 2-event pattern

Complexity: Let $m$ be the number of frequent single events in $L_1$, and $i$ be the average number of event instances of each frequent event. The complexity of frequent 2-event pattern mining is $O(m^2i^2 |\mathcal{D}_{SEQ}|^2)$.  

Proof. The Cartesian product of \( m \) events in \( L_1 \) generates \( m^2 \) event pairs. To compute the support of \( m^2 \) event pairs, we count the set bits of the bitmap that takes \( O(m^2|D_{SEQ}) \).

For each node in \( L_2 \), we need to compute the support and confidence of their temporal relations, which takes \( O(|I^2|D_{SEQ}) \). We have potentially \( m^2 \) nodes. And thus, it takes \( O(m^2|I^2|D_{SEQ}) \).

The overall complexity is: \( O(m^2|D_{SEQ}| + m^2|I^2|D_{SEQ}) \sim O(m^2|D_{SEQ}|^2) \).

\[ \Box \]

A.8 Lemma 5

Lemma 5. Let \( N_{k-1} = (E_1, ..., E_{k-1}) \) be a frequent \((k-1)\)-event combination, and \( E_k \) be a frequent single event. The combination \( N_k = N_{k-1} \cup E_k \) can form frequent \( k \)-event temporal patterns if \( \forall E_i \in N_{k-1}, \exists r(E_i, E_k) \) is a frequent temporal relation.

From Lemma 5, only single events in \( L_1 \) that occur in \( L_{k-1} \) should be used to create \( k \)-event combinations, as those that do not occur in \( L_{k-1} \) cannot form frequent temporal patterns and thus, should not be considered.

Proof. Let \( p_k \) be any \( k \)-event pattern formed by \( N_k \). Then \( p_k \) is a list of \( \frac{1}{2}k(k-1) \) triples \( (E_i, r(E_i, E_j)) \) where each represents a relation \( r(E_i, E_j) \) between two events. In order for \( p_k \) to be frequent, each of the relations in \( p_k \) must be frequent (Defs. 3.11 and 3.15, and Lemma 4). However, since \( \forall E_i \in N_{k-1}, r(E_i, E_k) \) is frequent, \( p_k \) is not frequent.

\[ \Box \]

A.9 Lemma 6

Lemma 6. Let \( P \) and \( P' \) be two temporal patterns. If \( P' \subset P \), then \( \text{conf}(P') \geq \text{conf}(P) \).

Lemma 6 says that the confidence of a pattern \( P \) is always at most the confidence of its sub-patterns. Thus, a temporal pattern at level \( L_{k-1} \) that is infrequent and/or low-confidence cannot be part of frequent and high-confidence patterns at \( L_k \).

Proof. Can be derived directly from Def. 3.16. \[ \Box \]

A.10 Lemma 7

Lemma 7. Let \( P \) and \( P' \) be two temporal patterns. If \( P' \subset P \) and \( \max_{1 \leq k \leq |P|} \left\{ \text{supp}(E_k) \right\} \leq \delta \), then \( \text{conf}(P') \leq \delta \).

From Lemma 7, a temporal pattern \( P \) cannot be high-confidence if any of its sub-patterns are low-confidence.

Proof. We have:

\[
\text{conf}(P) = \frac{\text{supp}(P)}{\max_{1 \leq k \leq |P|} \left\{ \text{supp}(E_k) \right\}} \leq \frac{\text{supp}(P')}{\max_{1 \leq k \leq |P'|} \left\{ \text{supp}(E_k) \right\}} \leq \delta
\]

\[ \Box \]

A.11 Mining frequent \( k \)-event pattern

Complexity: Let \( r \) be the average number of frequent \((k-1)\)-event patterns in \( L_{k-1} \). The complexity of frequent \( k \)-event pattern mining is \( O(|1\text{Freq} \cdot |L_{k-1}| \cdot r \cdot k^2|D_{SEQ}|) \).

Proof. For each frequent \((k-1)\)-event pattern of a node in \( L_4 \), we need to compute the support and confidence of \( \frac{1}{k}k(k-1) \) \((k-2)\) triples, which takes \( O\left(\frac{1}{k}k(k-1)(k-2)|D_{SEQ}| \right) \sim O(k^2|D_{SEQ}|) \).

We have \(|1\text{Freq}| \times |L_{k-1}| \) nodes in \( L_k \), each has \( r \) frequent \((k-1)\)-event patterns. Thus, the total complexity is: \( O(|1\text{Freq}| \cdot |L_{k-1}| \cdot r \cdot k^2|D_{SEQ}|) \).

\[ \Box \]

A.12 Lemma 8

Lemma 8. Let \( \text{supp}(X_1, Y_1)_{D_{SYB}} \) and \( \text{supp}(X_1, Y_1)_{D_{SEQ}} \) be the supports of \((X_1, Y_1)\) in \( D_{SYB} \) and \( D_{SEQ} \), respectively. We have the following relation: \( \text{supp}(X_1, Y_1)_{D_{SYB}} \leq \text{supp}(X_1, Y_1)_{D_{SEQ}} \).

From Lemma 8, if an event pair is frequent in \( D_{SYB} \), it is also frequent in \( D_{SEQ} \).

Proof. Recall that when converting \( D_{SYB} \) to \( D_{SEQ} \), we divide the symbolic time series in \( D_{SYB} \) into equal length temporal sequences. Let \( n \) be the length of each symbolic time series in \( D_{SYB} \), and \( m \) be the length of each temporal sequence. The number of temporal sequences obtained in \( D_{SEQ} \) is: \( \left\lfloor \frac{n}{m} \right\rfloor \).

The support of \((X_1, Y_1)\) in \( D_{SYB} \) is computed as:

\[
\text{supp}(X_1, Y_1)_{D_{SYB}} = \sum_{i=1}^{\left\lfloor \frac{n}{m} \right\rfloor} \sum_{j=1}^{m} s_{ij} / n
\]

where

\[
s_{ij} = \begin{cases} 1, & \text{if } (X_1, Y_1) \text{ occurs in row } j \text{ of the sequence } s_i \text{ in } D_{SYB} \\ 0, & \text{otherwise} \end{cases}
\]

Intuitively, Eq. 26 computes the relative support of \((X_1, Y_1)\) in \( D_{SYB} \) by counting the number of times \((X_1, Y_1)\) occurs in \( D_{SYB} \), and then dividing to the size of \( D_{SYB} \).

On the other hand, the support of \((X_1, Y_1)\) in \( D_{SEQ} \) is computed by counting the number of sequences in \( D_{SEQ} \) that \((X_1, Y_1)\) occurs. Note that if \((X_1, Y_1)\) occurs more than one in the same sequence,
we will count only 1 for that sequence. Thus, we have:

\[ \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} = \frac{\sum_{i=1}^{\frac{m}{n}} g_i}{n/m} = \frac{m \cdot \sum_{i=1}^{\frac{m}{n}} g_i}{n} \]  

(27)

where

\[ g_i = \begin{cases} 1, & \text{if } (X_i, Y_j) \text{ occurs in the sequence } g_i \text{ in } \mathcal{D}_{\text{SEQ}} \\ 0, & \text{otherwise} \end{cases} \]

Compare Eqs. 26 and 27, we have:

\[ \sum_{i=1}^{\frac{m}{n}} s_{ij} \leq m \cdot \sum_{i=1}^{\frac{m}{n}} g_i \]  

(28)

Hence:

\[ \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SYB}}} \leq \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} \]  

(29)

A.13 Theorem 1

Theorem 1. (Lower bound of the confidence) Let \( \sigma \) and \( \mu \) be the minimum support and mutual information thresholds, respectively. Assume that \((X_i, Y_j)\) is frequent in \(\mathcal{D}_{\text{SEQ}}\), i.e., \(\text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} \geq \sigma\).

If the NMI \( I(X; Y) \geq \mu \), then the confidence of \((X_i, Y_j)\) in \(\mathcal{D}_{\text{SEQ}}\) has a lower bound:

\[ \text{conf}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} \geq \sigma \cdot \lambda_1 \cdot \left( \frac{n_x - 1}{1 - \sigma} \right)^{\frac{1}{\lambda_2}} \]  

(30)

where \( n_x \) is the number of symbols in \( \Sigma_X \), \( \lambda_1 \) is the minimum support of \( X_i \in \Sigma_X \), \( \forall i \), and \( \lambda_2 \) is the support of \((X_i, Y_j) \in \Sigma_X \cdot \Sigma_Y \) such that \( p(X_i, Y_j) = \min \{ p(X_i), Y_j \} \).

Proof. From Eq. (10), we have:

\[ I(X_i; Y_j) = 1 - \frac{H(X_i|Y_j)}{H(X_i)} \geq \mu \]  

(31)

Hence:

\[ \frac{H(X_i|Y_j)}{H(X_i)} \leq 1 - \mu \]  

(32)

First, we derive a lower bound for \( \frac{H(X_i, Y_j)}{H(X_i)} \). We have:

\[ \frac{H(X_i, Y_j)}{H(X_i)} = \frac{p(X_i, Y_j) \cdot \log p(X_i|Y_j)}{\sum_i p(X_i) \cdot \log p(X_i)} + \frac{\sum_{i \neq j} p(X_i, Y_j) \cdot \log p(X_i|Y_j)}{\sum_i p(X_i)} \]  

(33)

We first consider the numerator in Eq. (33). Suppose that:

\[ \frac{p(X_m, Y_n)}{p(Y_n)} = \min \{ \frac{p(X_i, Y_j)}{p(Y_j)} \}, \forall (i \neq 1 & j \neq 1) \]  

(34)

Then, applying the min-max inequality theorem [5], we have:

\[ \frac{p(X_m, Y_n)}{p(Y_n)} \leq \frac{1 - p(X_m, Y_n)}{n_x - p(Y_n)} = \log \frac{p(X_m, Y_n)}{p(Y_n)} \leq \log \frac{1 - p(X_m, Y_n)}{n_x - p(Y_n)} \]  

(35)

\[ \Rightarrow \frac{p(X_m, Y_n)}{p(Y_n)} \leq \log \frac{1 - p(X_m, Y_n)}{n_x - p(Y_n)} \]  

(36)

\[ \Rightarrow p(X_m, Y_n) \cdot \log \frac{p(X_m, Y_n)}{p(Y_n)} \leq p(X_m, Y_n) \cdot \log \frac{1 - p(X_m, Y_n)}{n_x - p(Y_n)} \]  

(37)

Now, consider the second term of the numerator in Eq. (33). Since we have \( p(X_i, Y_j) < 0 \), hence:

\[ \sum_{i \neq 1 \& j \neq 1} p(X_i, Y_j) \cdot \log \frac{p(X_i, Y_j)}{p(Y_j)} \leq p(X_m, Y_n) \cdot \log \frac{p(X_m, Y_n)}{p(Y_n)} \]  

(38)

From Eqs. (37), (38), it follows that:

\[ \sum_{i \neq 1 \& j \neq 1} p(X_i, Y_j) \cdot \log \frac{p(X_i, Y_j)}{p(Y_j)} \leq p(X_m, Y_n) \cdot \log \frac{1 - p(X_m, Y_n)}{n_x - p(Y_n)} \]  

(39)

where \( \lambda_2 = p(X_m, Y_n) \).

Next, we consider the denominator in Eq. (33). Suppose that:

\[ p(X_i) = \min \{ p(X_i) \}, \forall i \]  

(40)

Then we have:

\[ p(X_i) \geq p(X_k), \forall i \]  

\[ \Rightarrow \log p(X_i) \geq \log p(X_k) \]  

\[ \Rightarrow \sum_i p(X_i) \log p(X_i) \geq \sum_i p(X_i) \log p(X_k) \]  

\[ = \log p(X_k) \sum_i p(X_i) \]  

\[ = \log p(X_k) \]  

\[ = \log \lambda_1 \]  

(41)

where \( \lambda_1 = p(X_k) \).

Replace Eq. (39) and (41) into Eq. (33), we get:

\[ \frac{H(X_i|Y_j)}{H(X_i)} \geq \frac{p(X_i, Y_j) \cdot \log p(X_i|Y_j) + \lambda_2 \cdot \log \frac{1 - p(X_i, Y_j)}{n_x - p(Y_i)}}{\log \lambda_1} \]  

(42)

Next, we derive the confidence lower bound of \((X_i, Y_j)\) in \(\mathcal{D}_{\text{SYB}}\).

Assuming that \( \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SYB}}} \geq \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} \geq \sigma \). We consider two cases.

- **Case 1:** Assuming that \( \text{supp}(Y_j)_{\mathcal{D}_{\text{SYB}}} \geq \text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SEQ}}} \). Thus, the confidence of \((X_i, Y_j)\) in \(\mathcal{D}_{\text{SYB}}\) is computed as

\[ \text{conf}(X_i, Y_j)_{\mathcal{D}_{\text{SYB}}} = \frac{\text{supp}(X_i, Y_j)_{\mathcal{D}_{\text{SYB}}}}{\text{supp}(Y_j)_{\mathcal{D}_{\text{SYB}}}} = \frac{p(X_i, Y_j)}{p(Y_j)} \]  

(43)

Since we have: \( \log \frac{p(X_i, Y_j)}{p(Y_j)} < 0 \), and \( \sigma \leq p(X_i, Y_j) \), we can deduce:

\[ p(X_i, Y_j) \cdot \log \frac{p(X_i, Y_j)}{p(Y_j)} \leq \sigma \cdot \log \frac{p(X_i, Y_j)}{p(Y_j)} \]  

(44)

We also have: \( 1 - p(X_i, Y_j) \leq 1 - \sigma \), and \( n_x - 1 \leq n_x - p(Y_j) \), hence:

\[ 1 - p(X_i, Y_j) \leq 1 - \sigma \]  

\[ n_x - p(Y_j) \leq n_x - 1 \]  

(45)
From Eqs. (32) and (47), it follows that:

\[
p(X_1, Y_1) \cdot \log \frac{p(X_1, Y_1)}{p(Y_1)} + \lambda_2 \cdot \log \frac{1 - p(X_1, Y_1)}{\pi_x - p(Y_1)} \leq \sigma \cdot \log \frac{p(X_1, Y_1)}{p(Y_1)} + \lambda_2 \cdot \log \frac{1 - \sigma}{\pi_x - 1} \leq 0 \quad (46)
\]

Replace Eq. (46) into the numerator of Eq. (42), we get:

\[
\frac{H(X_S | Y_S)}{H(X_S)} \geq \frac{\sigma \cdot \log \frac{p(X_1, Y_1)}{p(Y_1)} + \lambda_2 \cdot \log \frac{1 - \sigma}{\pi_x - 1}}{\log \lambda_1} \quad (47)
\]

From Eqs. (32) and (47), it follows that:

\[
(1 - \mu) \geq \frac{\sigma \cdot \log \frac{p(X_1, Y_1)}{p(Y_1)} + \lambda_2 \cdot \log \frac{1 - \sigma}{\pi_x - 1}}{\log \lambda_1} \quad (48)
\]

Hence:

\[
\text{conf}(X_1, Y_1)_{D_{SBY}} = \frac{p(X_1, Y_1)}{p(Y_1)} \geq \frac{\lambda_1^{\frac{1 - \mu}{\sigma}} \cdot \left(\frac{n_x - 1}{1 - \sigma}\right)^{\frac{\lambda_2}{\sigma}}}{\left(\frac{n_x - 1}{1 - \sigma}\right)^{\frac{\lambda_2}{\sigma}}} \quad (49)
\]

**Case 2:** Assuming that \(supp(Y_1)_{D_{SBY}} < supp(X_1)_{D_{SBY}}\). Thus, the confidence of \((X_1, Y_1)\) in \(D_{SBY}\) is computed as

\[
\text{conf}(X_1, Y_1)_{D_{SBY}} = \frac{supp(X_1, Y_1)_{D_{SBY}}}{supp(X_1)_{D_{SBY}}} = \frac{p(X_1, Y_1)}{p(X_1)} \quad (50)
\]

From Eq. (42), we have:

\[
\frac{H(X_S | Y_S)}{H(X_S)} \geq \frac{p(X_1, Y_1) \cdot \log \left(\frac{p(X_1, Y_1)}{p(X_1)} \cdot \frac{p(Y_1)}{p(Y_1)}\right) + \lambda_2 \cdot \log \frac{1 - p(X_1, Y_1)}{\pi_x - p(Y_1)}}{\log \lambda_1} \quad (51)
\]

\[
\geq \frac{\sigma \cdot \log \left(\frac{p(X_1, Y_1)}{p(X_1)} \cdot \frac{1}{\sigma}\right) + \lambda_2 \cdot \log \frac{1 - \sigma}{\pi_x - 1}}{\log \lambda_1} \quad (52)
\]

From Eqs. (32), (52), it follows that:

\[
(1 - \mu) \geq \frac{\sigma \cdot \log \left(\frac{p(X_1, Y_1)}{p(X_1)} \cdot \frac{1}{\sigma}\right) + \lambda_2 \cdot \log \frac{1 - \sigma}{\pi_x - 1}}{\log \lambda_1} \quad (53)
\]

Hence:

\[
\text{conf}(X_1, Y_1)_{D_{SBY}} = \frac{p(X_1, Y_1)}{p(X_1)} \geq \lambda_1^{\frac{1 - \mu}{\sigma}} \cdot \left(\frac{n_x - 1}{1 - \sigma}\right)^{\frac{\lambda_2}{\sigma}} \cdot \sigma \quad (54)
\]

In Eq. (54), we have \(\sigma < 1\). Thus, from Eqs. (49), (54), the confidence lower bound of \((X_1, Y_1)\) in \(D_{SBY}\) in both cases is:

\[
\text{conf}(X_1, Y_1)_{D_{SBY}} \geq \lambda_1^{\frac{1 - \mu}{\sigma}} \cdot \left(\frac{n_x - 1}{1 - \sigma}\right)^{\frac{\lambda_2}{\sigma}} \quad (55)
\]

Next, we derive the confidence of \((X_1, Y_1)\) in the temporal sequence database \(D_{SEQ}\). From Lemma 8, we have:

\[
supp(X_1)_{D_{SEQ}} \leq supp(X_1)_{D_{SEQ}} \quad (56)
\]

\[
supp(Y_1)_{D_{SEQ}} \leq supp(Y_1)_{D_{SEQ}} \quad (57)
\]

\[
supp(X_1, Y_1)_{D_{SEQ}} \leq supp(X_1, Y_1)_{D_{SEQ}} \quad (58)
\]

Without loss of generality, we assume \(supp(X_1)_{D_{SEQ}} \geq supp(Y_1)_{D_{SEQ}}\). Hence, the confidence of \((X_1, Y_1)\) in \(D_{SEQ}\) is computed as

\[
\text{conf}(X_1, Y_1)_{D_{SEQ}} = \frac{supp(X_1, Y_1)_{D_{SEQ}}}{supp(X_1)_{D_{SEQ}}} \geq \frac{supp(X_1, Y_1)_{D_{SBY}}}{supp(X_1)_{D_{SBY}}} = \frac{supp(X_1)_{D_{SBY}}}{supp(X_1)_{D_{SEQ}}} \quad (59)
\]

Since we have \(\sigma \leq supp(X_1)_{D_{SBY}}\) and \(supp(X_1)_{D_{SEQ}} \leq 1\), it follows that:

\[
\frac{supp(X_1)_{D_{SBY}}}{supp(X_1)_{D_{SEQ}}} \geq \sigma \quad (60)
\]

Finally, from Eqs. (55), (59) and (60), we can derive the confidence lower bound of \((X_1, Y_1)\) in \(D_{SEQ}\):

\[
\text{conf}(X_1, Y_1)_{D_{SEQ}} \geq \lambda_1^{\frac{1 - \mu}{\sigma}} \cdot \left(\frac{n_x - 1}{1 - \sigma}\right)^{\frac{\lambda_2}{\sigma}} \quad (61)
\]

## B. ADDITIONAL EXPERIMENTAL RESULTS

### B.1 Baselines comparison

Tables 12 and 13 show the comparison results between A-HTPGM, E-HTPGM and the baselines on the UKDALE, DataPort, and ASL datasets. A-HTPGM achieves the best performance (both runtime and memory usage) among all methods, and E-HTPGM has better performance than the baselines. The range and average speedups of E-HTPGM compared to the baselines are: [1.1 – 5.4] and 2.2 (Z-Miner), [3.1 – 12.7] and 6.1 (TPMiner), [3.7 – 30.1] and 10.8 (IEMiner), and [4.3 – 35.1] and 12.8 (H-DFS). Instead, the speedups of A-HTPGM compared to E-HTPGM and the baselines respectively are: [1.5 – 8.8] and 4.1 (E-HTPGM), [2.3 – 14.2] and 7.9 (Z-Miner), [11.7 – 38.1] and 21.2 (TPMiner), [15.3 – 116.3] and 40.6 (IEMiner), and [17.2 – 135.4] and 48.2 (H-DFS). Note that the time to compute MI and \(\mu\) for the UKDALE, DataPort, and ASL datasets in Table 12 are 17.82, 8.37, 11.27 seconds, respectively.

In average, on the tested datasets, E-HTPGM consumes ~3.8 times less memory than the baselines due to the applied pruning techniques, while A-HTPGM uses ~4.7 times less memory than E-HTPGM and the baselines by pruning uncorrelated series.

### B.2 Evaluation of the pruning techniques in E-HTPGM

In this section, we report the evaluation results of E-HTPGM on UKDALE, DataPort and ASL datasets. We use 3 different configurations that vary: the number of sequences, the confidence, and the support. Figs. 14, 15, 16 show the results (the y-axis is in log scale). It can be seen that All-E-HTPGM achieves the best performance among all versions. Its speedup w.r.t. NoPrune-E-HTPGM ranges from 2.3 up to 9.8 depending on the configurations, showing that...
Table 12: Runtime Comparison (seconds)

| Supp. (%) | Methods | Conf. (%) |
|-----------|---------|-----------|
|           |         | UKDALE    | DataPort |
| 20        | H-DFS   | 3207.59   | 1802.76  |
|           | EIMiner | 3159.95   | 1506.34  |
|           | E-HTPGM | 353.68    | 306.19   |
|           | A-HTPGM | 93.94     | 73.16    |
|           |         | 3.25      | 12.42    |
|           | 50      | 45.35     | 91.24    |
|           | H-DFS   | 140.92    | 106.46   |
|           | EIMiner | 93.74     | 49.78    |
|           | E-HTPGM | 397.92    | 394.42   |
|           | A-HTPGM | 12.42     | 8.23     |
|           | 80      | 20.09     | 20.42    |
|           | H-DFS   | 93.94     | 12.42    |
|           | EIMiner | 397.92    | 394.42   |
|           | E-HTPGM | 12.42     | 8.23     |
|           | A-HTPGM | 93.94     | 12.42    |

Table 13: Memory Usage Comparison (MB)

| Supp. (%) | Methods | Conf. (%) |
|-----------|---------|-----------|
|           |         | UKDALE    | DataPort |
| 20        | H-DFS   | 1358.56   | 1014.96  |
|           | EIMiner | 1397.56   | 1084.34  |
|           | E-HTPGM | 93.74     | 12.42    |
|           | A-HTPGM | 12.42     | 8.23     |
|           |         | 80        | 10.51    |
|           | H-DFS   | 1358.56   | 1014.96  |
|           | EIMiner | 1397.56   | 1084.34  |
|           | E-HTPGM | 93.74     | 12.42    |
|           | A-HTPGM | 12.42     | 8.23     |
|           | 50      | 80        | 15.87    |
|           | H-DFS   | 1358.56   | 1014.96  |
|           | EIMiner | 1397.56   | 1084.34  |
|           | E-HTPGM | 93.74     | 12.42    |
|           | A-HTPGM | 12.42     | 8.23     |
|           | 80      | 20        | 20.42    |
|           | H-DFS   | 93.94     | 12.42    |
|           | EIMiner | 93.74     | 12.42    |
|           | E-HTPGM | 93.74     | 12.42    |
|           | A-HTPGM | 93.74     | 12.42    |

Table 14: The Accuracy and the Number of Extracted Patterns from A-HTPGM on UKDALE

| Supp. (%) | Methods | Accuracy (%) | # Patterns |
|-----------|---------|--------------|------------|
| 10        | H-DFS   | 75.92        | 100        |
|           | EIMiner | 75.92        | 100        |
|           | E-HTPGM | 75.92        | 100        |
|           | A-HTPGM | 75.92        | 100        |
|           | 50      | 75.92        | 100        |
|           | H-DFS   | 75.92        | 100        |
|           | EIMiner | 75.92        | 100        |
|           | E-HTPGM | 75.92        | 100        |
|           | A-HTPGM | 75.92        | 100        |
|           | 80      | 75.92        | 100        |

Table 15: The Accuracy and the Number of Extracted Patterns from A-HTPGM on DataPort

| Supp. (%) | Methods | Accuracy (%) | # Patterns |
|-----------|---------|--------------|------------|
| 10        | H-DFS   | 75.92        | 100        |
|           | EIMiner | 75.92        | 100        |
|           | E-HTPGM | 75.92        | 100        |
|           | A-HTPGM | 75.92        | 100        |
|           | 50      | 75.92        | 100        |
|           | H-DFS   | 75.92        | 100        |
|           | EIMiner | 75.92        | 100        |
|           | E-HTPGM | 75.92        | 100        |
|           | A-HTPGM | 75.92        | 100        |
|           | 80      | 75.92        | 100        |
the proposed prunings are very effective in improving E-HTPGM performance. Furthermore, Trans-E-HTPGM brings larger speedup than Apriori-E-HTPGM. The speedup range is from 1.5 to 5.2 for Trans-E-HTPGM, and from 1.3 to 3.6 for Apriori-E-HTPGM. However, applying both always yields better speedup than applying either of them.

Fig. 17 shows the cumulative distribution of the confidences for the pruned patterns. It is seen that most of these patterns have low confidences. For UKDALE and DataPort, 85% of patterns have confidences less than 20% when the support is 10% and 20%, and 75% of patterns have confidences less than 30% when the support is 30%. For ASL, 80% of patterns have confidences less than 5%. Such patterns are likely not interesting to explore, and thus, are pruned by A-HTPGM.

### B.4 Evaluation of the tolerance buffer $\epsilon$

We evaluate the impact of the buffer $\epsilon$ on extracted patterns. Tables 17 and 18 report the number of extracted patterns for different $\epsilon$ values, and the corresponding percentages of pattern loss compared to $\epsilon = 0$. For NIST and DataPort datasets, the percentage of pattern loss is the lowest with $\epsilon = 1$ minute. For UKDALE, the percentage of pattern loss is the lowest with $\epsilon = 1$ and 2 minutes. And for Smart City, the percentage of pattern loss is very low, and the losses among different $\epsilon$ values are not very much different since there is low noise level in the dataset. For ASL, the percentage of pattern loss is the lowest when $\epsilon = 30$ frames.

Next, we analyze the quality of extracted patterns using different $\epsilon$ values. Specifically, for each specific $\epsilon$ value, we filter the patterns that are not in the result set of $\epsilon = 0$, and calculate the cumulative distribution of the confidences for them. Figs. 18 and 19 show the corresponding results on different datasets. For NIST and DataPort, 65% of patterns have confidences greater than 50% when $\epsilon$ is 1 minute, and 40% of patterns have confidences greater than 50% when $\epsilon$ is 2 and 3 minutes. For UKDALE, 60% of patterns have confidences greater than 50% when $\epsilon$ is 2 and 3 minutes. For Smart City, 62% of patterns have confidences greater than 50% when $\epsilon$ is 10 minutes, and 49% of patterns have confidences greater than 50% when $\epsilon$ is 5 and 15 minutes. For ASL, 70% of patterns have confidences greater than 10% when epsilon is 30 frames, and 50% of patterns have confidences greater than 50% when epsilon is 45 and 60 frames.
Table 18: Number of Extracted Patterns and Percentages of Pattern Loss on Smart City and ASL

| 𝜖 value | Smart City | | ASL |
|---|---|---|---|
| | # Patterns | Patterns (%) | # Patterns | Patterns (%) |
| 5 minutes | 1284207 | 0.06 | 30 frames | 104627 |
| 10 minutes | 1263902 | 0.08 | 45 frames | 101683 |
| 15 minutes | 1263787 | 0.09 | 60 frames | 96475 |

Table 19: Number of Extracted Patterns with Different Overlapping Durations

| Overlapping Duration | NIST | UKDALE | DataPort | Smart City | Overlapping Duration | ASL |
|---|---|---|---|---|---|---|
| 0 hour | 1946265 | 1090165 | 221948 | 1264971 | 0 frames | 105196 |
| 1 hour | 1947491 | 1084565 | 222052 | 1265193 | 150 frames | 105362 |
| 2 hours | 1947564 | 1085119 | 222052 | 1265232 | 300 frames | 105362 |
| 3 hours | 1947564 | 1085719 | 222052 | 1265232 | 450 frames | 105362 |

B.5 Evaluation of the Splitting Strategy using Overlapping Sequences

We evaluate the impact of overlapping duration used in the sequences splitting strategy on extracted patterns. Table 19 reports the number of extracted patterns with different overlapping durations. For NIST, UKDALE, and Smart City, the number of extracted patterns increases when overlapping duration increases, and the number of extracted patterns becomes stable when the overlapping duration is equal to 2 hours or more. The same trend also applies for DataPort and ASL, where the number of extracted patterns becomes constant when overlapping duration is greater than 1 hour and 150 frames, respectively, for the two datasets.

B.6 Evaluation between E-HTPGM and Z-Miner using seven temporal relations

As shown in Table 20, E-HTPGM has better runtime than Z-Miner. On the tested datasets, the range and average speedups of E-HTPGM compared to Z-Miner is: [1.30-3.84] and 2.28. In terms of memory consumption, as shown in Table 21, E-HTPGM is more efficient than Z-Miner. The range and the average memory consumption of E-HTPGM compared to Z-Miner is: [1.4-130.7] and 16.2.
### Table 20: Runtime Comparison using Seven Relations Model (seconds)

| Supp. (%) | Methods | Conf. (%) | NIST       | Smart City  |
|-----------|---------|-----------|------------|-------------|
|           |         | 20        | 50         | 80          | 20 | 50 | 80 |
| 20        | Z-Miner | 7996.05   | 689.89     | 172.23      | 141.78 | 12.97 | 2.10 |
|           | E-HTPGM | 2135.12   | 193.45     | 67.97       | 27.89  | 5.42  | 1.07 |
| 50        | Z-Miner | 564.41    | 559.17     | 154.61      | 11.25  | 10.84 | 1.90 |
|           | E-HTPGM | 149.96    | 145.53     | 52.22       | 2.82   | 2.06  | 0.88 |
| 80        | Z-Miner | 163.46    | 154.62     | 153.51      | 1.39   | 1.33  | 1.33 |
|           | E-HTPGM | 52.60     | 51.70      | 50.01       | 0.51   | 0.50  | 0.48 |

### Table 21: Memory Usage Comparison using Seven Relations Model (MB)

| Supp. (%) | Methods | Conf. (%) | NIST       | Smart City  |
|-----------|---------|-----------|------------|-------------|
|           |         | 20        | 50         | 80          | 20 | 50 | 80 |
| 20        | Z-Miner | 74988.73  | 6469.43    | 1970.82     | 571.89 | 101.11 | 50.34 |
|           | E-HTPGM | 13530.18  | 1937.75    | 666.31      | 64.39  | 74.87  | 32.58 |
| 50        | Z-Miner | 6123.63   | 6028.29    | 1920.11     | 94.88  | 94.19  | 47.53 |
|           | E-HTPGM | 1882.01   | 1786.34    | 664.64      | 71.39  | 70.99  | 32.61 |
| 80        | Z-Miner | 1878.24   | 1859.17    | 1849.66     | 34.16  | 33.38  | 33.30 |
|           | E-HTPGM | 660.86    | 660.60     | 593.95      | 31.36  | 31.36  | 29.85 |

### Supp. (%) Methods

| Supp. (%) | Methods | Conf. (%) | ASL | 0.5  | 1   | 10  |
|-----------|---------|-----------|-----|------|-----|-----|
| 0.5       | Z-Miner | 194.02    | 107.62 | 19.41 |
|           | E-HTPGM | 120.75    | 86.73  | 5.79  |
| 1         | Z-Miner | 83.05     | 75.02  | 12.39 |
|           | E-HTPGM | 39.72     | 34.27  | 5.42  |
| 10        | Z-Miner | 3.09      | 2.98   | 2.94  |
|           | E-HTPGM | 1.78      | 1.66   | 1.64  |