

# Free-rider Episode Screening via Dual Partition Model

Xiang Ao<sup>1,2</sup>, Yang Liu<sup>1,2</sup>, Zhen Huang<sup>3\*</sup>, Luo Zuo<sup>1,2</sup>, and Qing He<sup>1,2</sup>

<sup>1</sup>Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),  
Institute of Computing Technology, CAS, Beijing 100190, China.

<sup>2</sup>University of Chinese Academy of Sciences, Beijing 100049, China.

<sup>3</sup>Tsinghua University, Beijing, China.

{aoxiang, liuyang17z, zuo1, heqing}@ict.ac.cn  
huangzh2.13@sem.tsinghua.edu.cn

**Abstract.** One of the drawbacks of frequent episode mining is that overwhelmingly many of the discovered patterns are redundant. Free-rider episode, as a typical example, consists of a real pattern doped with some additional noise events. Because of the possible high support of the inside noise events, such free-rider episodes may have abnormally high support that they cannot be filtered by frequency based framework. An effective technique for filtering free-rider episodes is using a partition model to divide an episode into two consecutive subepisodes and comparing the observed support of such episode with its expected support under the assumption that these two subepisodes occur independently. In this paper, we take more complex subepisodes into consideration and develop a novel partition model named EDP for free-rider episode filtering from a given set of episodes. It combines (1) a dual partition strategy which divides an episode to an underlying real pattern and potential noises; (2) a novel definition of the expected support of a free-rider episode based on the proposed partition strategy. We can deem the episode interesting if the observed support is substantially higher than the expected support estimated by our model. The experiments on synthetic and real-world datasets demonstrate EDP can effectively filter free-rider episodes compared with existing state-of-the-arts.

**Keywords:** Episode dual partition · Interesting pattern discovery · Episode mining · Sequence mining

## 1 Introduction

One of the defects in frequent pattern mining is that there are abundant redundant patterns in the very large number of output patterns [22]. As a result, how to effectively reduce redundancy of the output becomes an essential problem of current research [23,13,11,18,16,10,6,8]. Frequent episode mining [14] (FEM for short), as one of the sub-topics of frequent pattern mining, which aims at discovering frequently appeared ordered sets of events from a single symbol (event) sequence, is facing the similar problem as well. Moreover, FEM algorithms usually suffer from inefficient processing since checking whether a sequence covers a general episode is an **NP-hard**

\* This work was done when Zhen was visiting Institute of Computing Technology, CAS.

problem [20]. It makes redundancy reduction in FEM particularly crucial because users will be reluctant to obtain redundant and useless patterns after a long time waiting.

In this paper, our purpose is to reduce redundancy of frequent episodes. Specifically, we focus on screening *free-rider episodes* and identifying underlying patterns from a given set of frequent episodes. Here free-rider episode, as a prototypical example of redundancy in episodes, consists of a real pattern doped with additional noise events.

While filtering patterns to reduce redundancy is well-studied for itemsets, it is still underexplored for episodes. The most straightforward approach to solve such problem is to compare observations against expectations predicted by independence model [9,17,12]. Interestingly enough, episodes occurrences in a single sequence may be dependent to each other, and hence support is no longer a sum of independent variables [18]. A more effective manner is to utilize partition models in which an episode is divided into subepisodes and the observed support is compared with the expected support produced by the model under specialized assumptions [18,16]. However, constructing a good partition model has significant challenges that existing approaches present some limitations, and we summarize them as follows.

- Noise events could be doped *before*, *after*, or *inside* a real pattern. For example,  $a \rightarrow x \rightarrow b$  could be a free-rider episode, where  $a \rightarrow b$  is the real pattern and  $x$  is the noise. Here  $a \rightarrow b$  is a non-prefix subepisode of  $a \rightarrow x \rightarrow b$ . However, most existing partition models fail to take non-prefix subepisodes into account. For example, in [18],  $a \rightarrow x \rightarrow b$  can only be divided into two consecutive subepisodes, e.g.  $\{a \rightarrow x, b\}$  or  $\{a, x \rightarrow b\}$ .
- It is challenging to define the expected support of an episode in one long sequence. The major reason is that the support of an episode is defined as its number of occurrences in a sequence, and the events in such sequence may be dependent to each other. Hence we have to carefully use independence assumptions in this scenario. However, independence assumptions are widely used in redundancy reduction in database of itemsets or sequences [23,18]. In that problem, the support of pattern is usually defined as the number of the transactions/sequences covering the pattern, and each individual transaction/sequence is naturally regarded independent with each other. As a result, it is difficult to adapt traditional ideas to our problem.
- The search space of partitions suffers from “combination explosion” as it increases exponentially to the number of events in an episode. Hence pruning strategies are needed to perform an efficient filtering.

To address the challenges and overcome the limitations of the existing methods, we propose a novel partition model named EDP (**E**pisode **D**ual **P**artition).

In EDP, we partition the events of an episode into two distinct categories, called *informative* and *random* respectively, and we make the assumption that informative events form the real pattern while random events are noises. Here we do not restrict the position of either informative or random events and thus non-prefix subepisodes are taken into consideration. Then we build a *generative model* for random sequences which are considered producing free-rider episodes by the following ways. First, we fix the occurrences of informative events according to their positions in the original sequence. Second, we generate random events with an assumption that they are independent of

potential real patterns. Next, based on these random sequences, we define the expected support of a free-rider episode and adopt an automaton based algorithm to compute. Finally we deem an episode redundant if its observed support can be explained by the estimated expected support. Otherwise, it will be an informative pattern. The experiments on both synthetic and real-world dataset about finance and text demonstrate the effectiveness of our model.

## 2 Related Work

We categorize related papers in the literature as follows.

**Mining frequent episodes.** Mining frequent episodes from event sequence was first introduced by Mannila et al. [14] where they defined episodes as directed acyclic graphs and considered two kinds of methods counting support, i.e. sliding window and minimal occurrence. Mining general episodes can be intricate and computing-intensive because discovering whether a sequence covers a general episode is **NP**-hard [20]. As a result, efforts have been focused on mining subclasses of episodes, such as serial episodes [4,3,5], closed episodes [20,19], episodes with unique labels [2,15].

**Filtering episodes based on expectation.** Our EDP model belongs to such category. That is to define the expected support for an episode and compare with its observed support for filtering or ranking tasks. Some existing efforts compare the observed support against the independence model [9,17,12]. In such models, they use a null hypothesis that the data are generated by a stationary and independent stochastic process. Recent work estimated the expected support by considering different partitions of episodes. Tatti [18] divided an episode into two consecutive subepisodes and downplayed the importance of such episode if its observed support could be explained by the two subepisodes. Such model failed to take non-prefix subepisodes into account when perform partitioning. Therefore it may be difficult to detect free-rider episodes when noise events are interweaved with a real pattern. Some idea for filtering sequential patterns can be adapted to episodes as well. For example, Petitjean et al. proposed SkO-PUS [16] in which re-ordering candidates were made from two subpattern partitions, and the expected support was defined based on the mean of support of all possible candidates in the original sequences. However, such work may distort the expected support as some re-orderings may be surprisingly rare in the original sequence.

**Episode set mining.** There is also another kind of efforts to score a set of episodes which best explain the data. Most of them are Minimum Description Length (MDL) based approaches for scoring a set of patterns relative to a dataset as well as a heuristic search method to construct the set [21,11,10,8,7]. Our work is different with such episode set mining since episode set mining methods select episodes based on how well we can describe the data using the episodes while our purpose is to filter free-rider episodes from real patterns based on how well their support can be explained by random sequences.

## 3 Definitions and Problem Statement

In this section, we introduce preliminaries, definitions and problem statement. Let  $\Omega = \{e_1, e_2, \dots, e_m\}$  be a finite event alphabet. We will use  $\Omega = \{a, b, c, d\}$  for all examples

in this paper. An event sequence  $S = \langle (E_1, t_1), (E_2, t_2), \dots, (E_n, t_n) \rangle$ , is an ordered sequence of events, where each  $E_i \subseteq \Omega$  consists of all events associated with time stamp  $t_i$ ,  $1 \leq i \leq n$ , and  $t_j < t_k$  for any  $1 \leq j < k \leq n$ . In addition,  $n$  is the *length* of event sequence  $S$ , denoted by  $\text{len}(S) = n$ .

For example, Fig. 1 shows an event sequence  $S$  with  $\text{len}(S) = 10$ .  $S$  will be used as the running example in this paper.

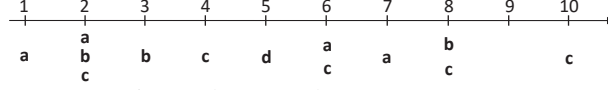


Fig. 1: The example event sequence.

An episode (refer to serial episode in this paper)  $\alpha$  is a totally ordered events in the form of  $e_{\alpha_1} \rightarrow e_{\alpha_2} \rightarrow \dots \rightarrow e_{\alpha_k}$  where  $\alpha_i \in [1, m]$  and thus  $e_{\alpha_i} \in \Omega$  for all  $1 \leq i \leq k$ . We can abbreviate  $\alpha = e_{\alpha_1} \rightarrow e_{\alpha_2} \rightarrow \dots \rightarrow e_{\alpha_k}$  as  $e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k}$ . We will use such abbreviation to represent episode in this paper unless otherwise specified. The *length* of an episode is defined as the number of events in the episode. An episode of length  $k$  is called a  $k$ -episode. We call an episode with length 0 as an *empty episode*, and denote it by  $\emptyset$ . A minimal occurrence window of  $\alpha$  is a time-window  $[t_s, t_e]$  which contains an occurrence of  $\alpha$ , such that no proper sub-window of it contains another occurrence of  $\alpha$ .  $t_s$  and  $t_e$  are called *start time* and *end time*, respectively. Usually there is an additional threshold of maximal window size  $\delta$  such that  $t_e - t_s < \delta$ . The set of all distinct minimal occurrence window of an episode  $\alpha$  is denoted by  $\text{moSet}(\alpha)$ .

For example,  $\text{moSet}(\text{abc}) = \{[2, 4], [7, 10]\}$  if  $\delta = 4$  in Fig. 1. Though  $[1, 4]$  contains an occurrence of episode  $\text{abc}$ , it is not a minimal occurrence window of  $\text{abc}$  because it subsumes the time window  $[2, 4]$  which contains another occurrence of  $\text{abc}$ .

The support of an episode  $\alpha$ , denoted by  $\text{sp}(\alpha)$ , is defined as the number of distinct minimal occurrence windows of  $\alpha$  in sequence  $S$ , i.e.  $\text{sp}(\alpha) = |\text{moSet}(\alpha)|$ . We call an episode  $\alpha$  *frequent* if its support passes a user-specific *minimum support threshold*  $\text{min\_sup} > 0$ , i.e.  $\text{sp}(\alpha) \geq \text{min\_sup}$ . Otherwise, it is infrequent. Consider two episodes  $\alpha = e_{\alpha_1} \dots e_{\alpha_k}$  and  $\beta = e'_{\alpha_1} \dots e'_{\alpha_m}$  where  $m \leq k$ , the episode  $\beta$  is a *subepisode* of  $\alpha$ , denoted by  $\beta \preceq \alpha$ , if and only if there exist  $m$  integers  $1 \leq i_1 \leq i_2 \leq \dots \leq i_m \leq k$  such that  $e'_{\alpha_j} = e_{\alpha_{i_j}}$  where  $1 \leq j \leq m$ . We call an episode  $H_j^\alpha = e_{\alpha_1} \dots e_{\alpha_j}$  a  $j$ -*prefix* episode of  $\alpha$  if  $0 \leq j \leq k$ . Specifically, 0-prefix episode of any non-empty episode  $\alpha$  is the empty episode  $\emptyset$ .

For example, all possible subepisodes of  $\alpha = \text{abc}$  include  $\emptyset$ ,  $\text{a}$ ,  $\text{b}$ ,  $\text{c}$ ,  $\text{ab}$ ,  $\text{ac}$ ,  $\text{bc}$  and  $\text{abc}$ . All prefix episodes of  $\alpha = \text{abc}$  are  $H_0^{\text{abc}} = \emptyset$ ,  $H_1^{\text{abc}} = \text{a}$ ,  $H_2^{\text{abc}} = \text{ab}$ ,  $H_3^{\text{abc}} = \text{abc}$ .

Next, we define the concept of free-rider episode and give a formal statement of the problem. We assume the knowledge of generative rules behind real patterns in advance and give the following assumption: A *generative rule*  $\alpha \Rightarrow e$  states that an event  $e$  will be more likely to occur after an occurrences of an episode  $\alpha$ .

For example,  $\text{ab} \Rightarrow \text{c}$  is the generative rule indicates that the probability of occurrence of event  $\text{c}$  will increase in some time stamps after episode  $\text{ab}$  occurs. With the defined generative rule, free-rider events in an episode can be seen as those events that no generative rule is associated with.

**Free-rider Event.** A free-rider event  $e_{\alpha_i}$  in an episode  $\alpha = e_{\alpha_1} \cdots e_{\alpha_k}$ , where  $1 \leq i \leq k$ , is an event that there is no generative rule associated with in the form as follows.

1.  $\beta \Rightarrow e_{\alpha_i}$ , where  $\beta$  is a non-empty subepisode of  $(i-1)$ -prefix episode of  $\alpha$ , i.e.,  $\beta \preceq H_{i-1}^\alpha$  or
2.  $\beta' \Rightarrow e_{\alpha_j}$ , where  $j > i$  and  $\beta'$  containing  $e_{\alpha_i}$  is a non-empty subepisode of  $(j-1)$ -prefix episode of  $\alpha$ , i.e.,  $e_{\alpha_i} \in \beta' \preceq H_{j-1}^\alpha$ .

For instance, consider episode  $\alpha = abc$ ,  $b$  will not be a free-rider event if  $a \Rightarrow b$  holds or if  $ab \Rightarrow c$  holds. However, generative rule  $ac \Rightarrow b$  has no impact on whether  $b$  is a free-rider event of the episode  $abc$  or not. In other words, the event order in  $\alpha$  should be respected. Then we can describe free-rider episodes as follows.

**Free-rider Episode.** An episode  $\alpha = e_{\alpha_1} \cdots e_{\alpha_k}$  is a *free-rider episode* if it contains at least one free-rider event  $e_{\alpha_i}$  where  $1 \leq i \leq k$ .

We argue that filtering free-rider episodes by definition is intricate since it may be hard to identify free-rider events by checking these implicit generative rules. To close the gap, however, we devise a novel expected support of a free-rider episode and take *Lift*, which is the ratio of observed support to the expectation, as the measure to determine whether an episode is a free-rider episode or not. We can then filter the episodes whose Lift values fail to exceed the minimum lift threshold *min\_lift* and rank the rest by Lift with a descending order as their interestingness scores. Hence, the problem of free-rider episode screening in this paper is formulated as follows.

**Problem Statement of Free-rider Episode Screening.** Given an event sequence  $S$  and a set of frequent episodes on  $S$ , the free-rider episode screening problem is to rank the episodes with the Lift produced by the EDP model and filter the episodes whose Lift values are less than *min\_lift*.

## 4 The EDP Model

The key ingredient to our problem is to compute the expected support of a free-rider episode, and we thus propose the EDP model. It involves an episode partition strategy, a generative model for random sequences based on episode partitions, and an expected support definition of a free-rider episode based on the generated random sequences. The inherent idea behind our model is that if the support of an episode can be simulated by that from a group of random sequences, it cannot be a real pattern.

### 4.1 Episode Partition Strategy

We first detail the episode partition strategy in EDP. Since pattern and noise events in an episode may interweave in complex ways we first make the following assumption to simplify the problem. Specifically, we assume that an event of an episode will not simultaneously be a part of real pattern and a noise event. Such hard assignment could simplify the processes of sequence generation and expectation calculation that will be introduced in the following parts. Formally, given an episode  $\alpha = e_{\alpha_1} \cdots e_{\alpha_k}$ , there is no  $i, j \in [1, k]$  and  $i \neq j$  such that  $e_{\alpha_i} = e_{\alpha_j}$  while  $e_{\alpha_i}$  is a free-rider event and  $e_{\alpha_j}$  is not.

Take the episode  $abb$  as an example. There are two event  $b$  in such episode. We assume it is not possible that the first  $b$  is a noise event while the second  $b$  is not. The practical usefulness of our model will not be limited by this assumption, as shown in our experiments. Under such assumption, we have the following partition strategy.

**Definition 1. (Dual Partition)** Let  $\Omega_\alpha$  be the set of event alphabet of an episode  $\alpha$ . We divide  $\Omega_\alpha$  into two distinct parts called informative and random events, which are denoted by  $\mathcal{I}_\alpha$  and  $\overline{\mathcal{I}_\alpha}$ , respectively. We assume the events in  $\mathcal{I}_\alpha$  form the real pattern, and the events in  $\overline{\mathcal{I}_\alpha}$  are regarded as noises. Then the episode  $\alpha$  is partitioned as a subepisode of  $\alpha$  which consists of events in  $\mathcal{I}_\alpha$ , and noise events in  $\overline{\mathcal{I}_\alpha}$ .

Such partition indeed takes non-prefix subepisodes of  $\alpha$  into account when performing partitioning. For example, consider an episode  $\alpha = abcd$ , we have  $\Omega_\alpha = \{a, b, c, d\}$ .  $\Omega_\alpha$  can be partitioned as  $\mathcal{I}_\alpha = \{a, c\}$  and  $\overline{\mathcal{I}_\alpha} = \{b, d\}$  or  $\mathcal{I}_\alpha = \{a, b, d\}$  and  $\overline{\mathcal{I}_\alpha} = \{c\}$ . Under these two partitions,  $\alpha$  can be divided into a subepisode  $ac$  and a noise event set  $\{b, d\}$  or  $abd$  and  $\{c\}$ . Both  $ac$  and  $abd$  are non-prefix subepisodes of  $abcd$ .

## 4.2 Generate Random Sequences

Next we devise a generative model (denoted by  $\mathcal{M}_{\mathcal{I}_\alpha}$ ), for a given episode  $\alpha$ , generating random sequences based on the aforementioned episode dual partition  $\mathcal{I}_\alpha$  and  $\overline{\mathcal{I}_\alpha}$ .

The basic idea for  $\mathcal{M}_{\mathcal{I}_\alpha}$  is that we consider the potential real pattern, which consists of informative events, follows implicit generative rules. Their occurrences in the original event sequence are thus meaningful. While the random events, which are independent of the real patterns, could occur at any time stamp. Hence the model generates random sequences as follows. First, the length of every random sequence is the same as the original event sequence. Second, the model produces informative events according to their occurrences in the original sequence. Third, the random events are generated independently over all possible time stamps.

Formally, we first define the probability of generating an event  $e$  at a specific time stamps  $t_j$  under  $\mathcal{M}_{\mathcal{I}_\alpha}$  as follows.

$$P(e|\mathcal{M}_{\mathcal{I}_\alpha}, t_j) = \begin{cases} 1 & e \in \mathcal{I}_\alpha \ \& \ e \in E_j \\ 0 & e \in \mathcal{I}_\alpha \ \& \ e \notin E_j \\ p_{ind}(e) & e \in \overline{\mathcal{I}_\alpha} \end{cases} \quad (1)$$

In Formula 1,  $E_j$  is the event set associated with time stamp  $t_j$  in the original event sequence  $\mathbf{S}$ , and  $p_{ind}(e)$  is the occurring probability<sup>1</sup> of event  $e$  in  $\mathbf{S}$ .

Next we define the probability for generating an event set  $E'_j$  based on  $\mathcal{M}_{\mathcal{I}_\alpha}$  at a specific time stamp  $t_j$  as follows.

$$P(E'_j|\mathcal{M}_{\mathcal{I}_\alpha}, t_j) = \prod_{e \in E'_j} P(e|\mathcal{M}_{\mathcal{I}_\alpha}, t_j) \prod_{e \in \Omega_\alpha \setminus E'_j} (1 - P(e|\mathcal{M}_{\mathcal{I}_\alpha}, t_j)) \quad (2)$$

We can generate event set for every time stamp with Formula 2 and eventually obtain a random sequence by assuming every event set is independent to each other.

<sup>1</sup>  $p_{ind}(e)$  can be calculated by  $\frac{sp(e)}{len(\mathbf{S})}$ , where  $sp(e)$  is the support of event  $e$  in  $\mathbf{S}$ .

Here we can safely use the independent assumption here since we attempt to generate random sequences without specialized rules. Remember that the length of a random sequence is the same as the length of  $\mathcal{S}$ . Therefore, the probability to generate a specific random sequence  $\hat{\mathcal{S}}$  can be given as Eq. 3, where  $n$  denotes the length of  $\mathcal{S}$ .

$$P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n) = \prod_{j=1}^n P(E'_j|\mathcal{M}_{\mathcal{I}_\alpha}, t_j) \quad (3)$$

For example, consider the event sequence shown in Fig. 1 and the episode  $\alpha = abc$ . Suppose  $\mathcal{I}_\alpha = \{a, b\}$  and  $\overline{\mathcal{I}_\alpha} = \{c\}$ . The generative model  $\mathcal{M}_{\{a,b\}}$  fixes all the occurrences of event  $a$  and  $b$  as in the sequence shown in Fig. 1 and generates event  $c$  by  $p_{ind}(c) = 0.5$  ( $sp(c) = 5$  and  $len(\mathcal{S}) = 10$ ). The illustration of  $\mathcal{M}_{\{a,b\}}$  is shown as Fig. 2 in which the fixed informative events are shown under the time stamps while the random event  $c$  is shown above with the grey font. For probability of the candidate event set at each time stamp, it is the multiplication of all possible events in it, e.g.  $P(E'_7 = \{a, c\}|\mathcal{M}_{\{a,b\}}, 7) = P(E'_7 = \{a\}|\mathcal{M}_{\{a,b\}}, 7) = 0.5$  and  $P(E'_9 = \{c\}|\mathcal{M}_{\{a,b\}}, 9) = P(E'_9 = \{\emptyset\}|\mathcal{M}_{\{a,b\}}, 9) = 0.5$ . Though  $E_9$  has no event in the original sequence, the model may generate events based on the dual partition.

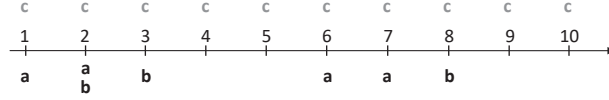


Fig. 2: The generative model  $\mathcal{M}_{\{a,b\}}$  for episode  $abc$ . Occurrences of event  $a$  and  $b$  are fixed according to their occurrences in the original sequence. Event  $c$  is generated with its occurring probability in the original sequence.

### 4.3 Expected Support Definition

Next we define the expected support of a free-rider episode on these random sequences. Recall that given an episode  $\alpha$ , EDP generates random sequences with events in  $\Omega_\alpha$  based on the model  $\mathcal{M}_{\mathcal{I}_\alpha}$ , and we assume, in  $\mathcal{M}_{\mathcal{I}_\alpha}$ , events belonging to  $\mathcal{I}_\alpha$  have implicit generative rules but events in  $\overline{\mathcal{I}_\alpha}$  do not. Hence based on our definition every event  $e'$  in  $\overline{\mathcal{I}_\alpha}$  will be a free-rider event to  $\alpha$  in each generated sequence and  $\alpha$  in these sequences will become a free-rider since  $e' \in \alpha$ . As a consequence, denoted the set of random sequences generated by a given generative model  $\mathcal{M}_{\mathcal{I}_\alpha}$  as  $\text{Seq}(\mathcal{M}_{\mathcal{I}_\alpha})$ , we match the minimal occurrences of  $\alpha$  on each generated sequence  $\hat{\mathcal{S}} \in \text{Seq}(\mathcal{M}_{\mathcal{I}_\alpha})$  and define the expected support of  $\alpha$ , which is regarded as a free-rider episode, as follows.

$$\mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}[sp(\alpha|\hat{\mathcal{S}})] = \sum_{\hat{\mathcal{S}} \in \text{Seq}(\mathcal{M}_{\mathcal{I}_\alpha})} sp(\alpha|\hat{\mathcal{S}}) \cdot P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n) \quad (4)$$

where  $sp(\alpha|\hat{\mathcal{S}})$  denotes the support of the episode  $\alpha$  in a generated sequence  $\hat{\mathcal{S}}$ .

It is no doubt different generative models will derive different expected support for an episode  $\alpha$ . To minimize the false positive, we define the formal expected support as the maximum expected support value produced by a specific generative model. Hence we define the expected support of an episode  $\alpha$ , denoted as  $\text{ExpSup}(\alpha)$ , as follows.

$$\text{ExpSup}(\alpha) = \max_{\mathcal{M}_{\mathcal{I}_\alpha}} \mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}[sp(\alpha|\hat{\mathcal{S}})] \quad (5)$$

With such definition, we adopt the following  $Lift(\alpha)$  to measure the deviation between the observed support of an episode  $\alpha$  and its expectation and regard it as a free-rider episode if  $Lift(\alpha)$  does not exceed the minimum lift threshold  $min\_lift$ .

$$Lift(\alpha) = \frac{sp(\alpha)}{ExpSup(\alpha)} \quad (6)$$

Under this formula, we actually make the following assumption: For an episode  $\alpha$  in an event sequence  $S$ , if we could find a generative model producing free-rider episodes in the same form of  $\alpha$  by EDP which can generate  $\alpha$  more frequently or close to the observed support of  $\alpha$  in  $S$ , then  $\alpha$  in  $S$  will be redundant.

#### 4.4 Time Complexity of EDP

In this subsection, we analyze the time complexity of the proposed EDP. Though the search space of possible partitions of an episode  $\alpha$  is  $2^{|\Omega_\alpha|}$ , where  $\Omega_\alpha$  is the alphabet of  $\alpha$ , we could stop calculation and screen  $\alpha$  if we find one  $\mathcal{M}_{\mathcal{I}_\alpha}$  such that the value of  $\frac{sp(\alpha)}{\mathbb{E}_{S \sim P(S|\mathcal{M}_{\mathcal{I}_\alpha}, n)}[sp(\alpha|S)]}$  is clearly less than  $min\_lift$ . It indicates there exists one group of random sequences that provides free-rider episodes of  $\alpha$  can explain  $\alpha$ 's observed support in the input sequence. As a result, we can explore pruning techniques based on such property.

For a real pattern, on the other hand, we need to enumerate all partitions to ascertain its interestingness. Hence the time complexity will be  $O(2^{|\Omega_\alpha|})$ . However, in our EDP each partition has no impact on others, the process can thus be highly parallelized without hurting the performance. The experimental results will show EDP can achieve significant speedup in running time if we distribute the checking into multi-processes.

### 5 Expected Support Calculation

So far our problem is how to compute the expected support of an episode  $\alpha$  given a generative model  $\mathcal{M}_{\mathcal{I}_\alpha}$ . In this section, we devise an algorithm computing the expected support defined in Eq. 4 with the help of automaton.

#### 5.1 Tracking Minimal Occurrences with Automaton

Automaton is an effective tool to track minimal occurrence of episode in sequences. Given an episode  $\alpha = e_{\alpha_1}e_{\alpha_2} \cdots e_{\alpha_k}$ , we can obtain a unique automaton  $\mathcal{A}_\alpha$  in which every node (also known as state) denotes a prefix subepisode of  $\alpha$ , namely  $H_i^\alpha$  where  $0 \leq i \leq k$ . The label on each edge denotes the event that can trigger state transition. We denote the transition function by  $T(H, E)$  where  $H$  is the current state of an automaton and  $E$  is an event set. The first state and the last state of an automaton are called the source state and the sink state, respectively. Since minimal occurrences may overlap to each other on the event sequence, we need to simultaneously manage multiple automata which do not reach the sink state during scanning the sequence. We call them active automata.



## 5.2 The Algorithm

The pseudo code of expected support calculation is shown as Algorithm 1.

In our algorithm, we do not need to generate any specific random sequence. Instead, we adopt the probability of event set generation and the probability of the appearance of active automatons to estimate the expected support defined in Eq. 4. Since  $\mathcal{M}_{\mathcal{I}_\alpha}$  may generate multiple different event sets for each time stamp, we need to simultaneously manage more than one possible active automaton lists and consider all possible event sets for every time stamp. In more detail, we use a variable  $L_i$  to store active automatons at time stamp  $t_i$  and first initialize  $L_0 = \{H_0^\alpha\}$ . Since  $L_0$  is determined, its appearance probability is one, and the list of active automatons, denoted as  $\mathcal{L}_0$ , contains the only element  $L_0$  (Line 1–2).

---

### Algorithm 1: Expected support calculation given $\mathcal{M}_{\mathcal{I}_\alpha}$

---

**Input:**  $\alpha$ : a  $k$ -episode  
 $\mathcal{M}_{\mathcal{I}_\alpha}$ : a generative model constructed by episode partition  $\mathcal{I}_\alpha$   
 $\mathcal{S}$ : the original event sequence  
**Output:**  $\mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}(\text{sp}(\alpha|\hat{\mathcal{S}}))$ : the expected support of an episode  $\alpha$  given a generative model  $\mathcal{M}_{\mathcal{I}_\alpha}$

```

1 initialize  $L_0 = \{H_0^\alpha\}$ ,  $\mathcal{L}_0 = \emptyset$ ,  $P(L_0) = 1.0$ ,  $\mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}(\text{sp}(\alpha|\hat{\mathcal{S}})) = 0$ 
2  $\mathcal{L}_0 \leftarrow \mathcal{L}_0 \cup \{L_0\}$ 
3 for  $i = 1$  to  $\text{len}(\mathcal{S})$  do
4    $\mathcal{L}_i = \emptyset$ 
5   foreach possible active automaton list  $L_{i-1} \in \mathcal{L}_{i-1}$  do
6      $P(L_i) = 0$ 
7     foreach possible event set  $E'_i$  produced by  $\mathcal{M}_{\mathcal{I}_\alpha}$  do
8       foreach active  $\mathcal{A}(\alpha) \in L_{i-1}$  do
9          $H_c \leftarrow$  current state of  $\mathcal{A}(\alpha)$ 
10         $H_t = T(H_c, E'_i)$ 
11        if TRANSIT i.e.  $H_t \neq H_c$  then
12          if SINK i.e.  $H_t = H_k^\alpha$  then
13            increase  $\mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}(\text{sp}(\alpha|\hat{\mathcal{S}}))$  by  $P(L_{i-1}) \cdot P(E'_i|\mathcal{M}_{\mathcal{I}_\alpha}, t_i)$ 
14          else
15            add a copy of  $\mathcal{A}(\alpha)$  whose current state is  $H_t$  to  $L_i$ 
16          if SOURCE i.e.  $H_c = H_0^\alpha$  then
17            add an initialized  $\mathcal{A}(\alpha)$  whose current state is  $H_0^\alpha$  to  $L_i$ 
18        forall  $\mathcal{A}(\alpha) \in L_{i-1}$  not TRANSIT do
19           $H_c \leftarrow$  current state of  $\mathcal{A}(\alpha)$ 
20          if there is no  $\mathcal{A}(\alpha)' \in L_i$  whose current state is  $H_c$  then
21            add a copy of  $\mathcal{A}(\alpha)$  to  $L_i$ 
22      increase  $P(L_i)$  by  $P(L_{i-1}) \cdot P(E'_i|\mathcal{M}_{\mathcal{I}_\alpha}, t_i)$ , update  $\mathcal{L}_i$ 
23 return  $\mathbb{E}_{\hat{\mathcal{S}} \sim P(\hat{\mathcal{S}}|\mathcal{M}_{\mathcal{I}_\alpha}, n)}(\text{sp}(\alpha|\hat{\mathcal{S}}))$ 

```

---

Then we sequentially consider every pair of possible active finite automaton list  $L_{i-1}$  and possible event set  $E'_i$  produced by the generative model  $\mathcal{M}_{\mathcal{I}_\alpha}$  for every time stamp  $t_i$  where  $1 \leq i \leq \text{len}(\mathcal{S})$  (Line 3–7). For each active  $\mathcal{A}(\alpha)$  in  $L_{i-1}$ , we invoke the state transitive function by consuming its current state  $H_c$  and the event set  $E'_i$  and acquire the output state  $H_t$  (Line 8–10). If a transition occurs, we check both  $H_c$  and  $H_t$  for automaton management. In particular, if  $H_t$  reaches the sink state  $H_k^\alpha$ , we obtain a minimal occurrence of  $\alpha$  and increase the expected support of  $\alpha$  by  $P(L_{i-1}) \cdot$

$P(E'_i|\mathcal{M}_{\mathcal{I}_\alpha}, t_i)$  (Line 11–13). Otherwise, we add the updated  $\mathcal{A}(\alpha)$  to  $L_i$ . We meanwhile check whether  $H_c$  is a source state and add an initialized automaton to  $L_i$  if it is (Line 16–17). For the rest automata without state transitions, we will determine whether they can be added to  $L_i$  since we are interested in minimal occurrence of episode (Line 18–21). Finally we update the probability of the appearance of active automata  $L_i$  for further computations (Line 22).

For example, consider the episode  $abc$  and the generative model  $\mathcal{M}_{ab}$  as shown in Fig. 2. When  $i = 3$ , the only possible active automaton list  $L_2 \in \mathcal{L}_2$ , having three active automata, contains  $\{H_0^{abc}, H_1^{abc}, H_2^{abc}\}$ . Here  $H_i^{abc}$  denotes the current state of the corresponding automaton. For possible event sets, there are two possible event sets, i.e.  $E'_3 = \{b\}$  and  $E''_3 = \{b, c\}$ , with the probability of 0.5 of each, respectively. Considering every pair of the active automaton and the event set, we detect  $T(H_2^{abc}, E''_3)$  derives the SINK state. We thus add the expected support of the episode  $abc$  under the generative model  $\mathcal{M}_{ab}$  by  $P(L_2) \cdot P(E''_3|\mathcal{M}_{ab}, t_3) = 0.5$ . Then we update the probability of active automata and derive  $\mathcal{L}_3 = \{L_3\} = \{H_0^{abc}, H_2^{abc}\}$  with  $P(L_3) = 1$ .

### 5.3 Time Complexity Analysis

Here we analyze the time complexity of the algorithm, i.e., Algorithm 1, for expected support calculation given a specific generative model. We only discuss the worst case for simplicity. Given a specific generative model  $\mathcal{M}_{\mathcal{I}_\alpha}$  of an episode  $\alpha$ , for each time stamp  $t_i$ , there are at most  $|\alpha|$  active automata in each active automata list  $L_i$ ,  $2^{|\mathcal{I}_\alpha|}$  possible event set  $E_j$  and  $2^{|\alpha|}$  possible lists of active automata, respectively. Hence the time complexity for precessing a time stamp  $t_i$  in the worst case is  $O(|\alpha| \cdot 2^{|\mathcal{I}_\alpha|+|\alpha|})$ . Since there are all together  $n$  time stamps on the event sequence  $\mathcal{S}$ , the overall time complexity of Algorithm 1 will be  $O(n \cdot |\alpha| \cdot 2^{|\mathcal{I}_\alpha|+|\alpha|})$  where  $n$  denotes the length of the event sequence. Though it is exponential in  $|\mathcal{I}_\alpha|$  and  $|\alpha|$ , this number might not be very large as usually the episode to be checked has limited length and so is  $|\mathcal{I}_\alpha|$ .

## 6 Experiments

In this section, we present the results of our experimental studies.

### 6.1 Datasets and Experiment Settings

We conducted the experiments on both synthetic and real-world datasets, and Table 1 illustrates their statistics.

The **SYN** dataset refers to a synthetic dataset which is generated as follows. We use an alphabet of 52 events, i.e.,  $a \cdots z$  and  $A \cdots Z$  for generating a sequence whose length is set to 10,000. In such sequence, we planted two episodes and a high frequency noise event. The first episode  $abc$  with no gaps was embedded 300 times into randomly selected time stamps. The second episode, a 4-episode  $defg$ , can have a gap between any two consecutive events. The gap is a positive integer drawn from a truncated normal distribution  $N(2, 2)$ . The episode  $defg$  was randomly planted 300 times in the sequence as well. These two patterns might overlap but there are no generative rules between them. Third, we planted a high frequency event  $X$  with  $p_{ind}(X) = 0.3$  into randomly selected

time stamps of the sequence. Such event might become a free-rider event that doped with the embedded patterns. Finally, we filled the sequence using the rest events which were generated from a uniform distribution. They are considered as random noises. As a consequence, all the episodes rather than the non-empty subepisodes of *abc* and *defg* are considered as free-riders in such dataset.

Table 1: Statistical information of datasets

Dataset	Sequence Len.	Events in alphabet	Avg. events per timestamp	Avg. occurrences per event
SYN	10,000	52	1.4	267.3
STK	9,334	50	5.55	1,037
JMLR	75,645	3,846	1.0	19.67

The **STK** dataset consists of daily prices of stocks from Chinese stock market. It includes 50 blue chip stocks with their price information over 26 years from December 19th, 1990 to July 14th, 2016. The events for a stock are generated by the increase ratio of price on each trading day. If the ratio is positive, we generate an increase event for such stock on the corresponding day. For example, if the stock of Bank of China increases on December 19th, 1990, we will generate an event as “Bank of China+” for that day. Otherwise, we will not produce event for the stock of Bank of China on that day. Similar operations were performed for all the 50 stocks in such dataset.

The **JMLR** dataset consists of the abstracts of the papers published on the Journal of Machine Learning Research which was adopted in [18]. We treat each word as an event and built an event sequence by connecting every sentence together. Such dataset was preprocessed by stemming the words and removing the stop words. The details can be found in [18].

We compare our EDP model against the following state-of-the-art methods:

1. **PRT** [18]: A partition model that divides an episode into two consecutive sub-episodes and detects whether its frequency can be explained by the two sub-episodes.
2. **SkOPUS** [16]: A partition model that compares the frequency of a pattern with its re-ordering candidates consisting of two sub-pattern partitions.
3. **EGH** [12]: An independence baseline that connects the significance of an episode with a stationary hidden markov model.
4. **IND**: The degraded version of the proposed EDP model which considers every event is a random event.

Since all the compared methods require a set of candidate patterns, we start from a set of frequent episodes, denoted as  $\mathcal{F}$ , in which we require  $|\mathcal{Q}_\alpha| > 1$  for every  $\alpha \in \mathcal{F}$ .<sup>2</sup> PRT and SkOPUS are designed for multiple sequences rather than a single event sequence, we thus transfer our event sequence into database of short sequences by segmenting it with fixed length sliding windows to fit the methods. For mining frequent episodes and deriving  $\mathcal{F}$ , we adopted the DFS algorithm [1], which is a state-of-the-art minimal occurrence based frequent episode mining algorithm.

<sup>2</sup> Here we do not take any frequent event or the episode consisting of single event into account.

## 6.2 Results on data with known patterns

We first demonstrate our results on the SYN dataset having known patterns. In particular, we mined  $\mathcal{F}$  by setting  $min\_sup = 200$  and  $\delta = 12$  and finally obtained  $|\mathcal{F}| = 325$ . Such setting ensures the embedded real patterns are contained in  $\mathcal{F}$ . Our purpose is to check whether the compared methods can unearth these planted patterns and rank them as high as possible.

Table 2: Precision@ $k$  on SYN dataset.

Top $k$	Precision@ $k$				
	IND	EGH	SkOPUS	PRT	EDP
1	100.0%	0.0%	100.0%	100.0%	100.0%
2	100.0%	50.0%	100.0%	100.0%	100.0%
3	100.0%	66.7%	100.0%	100.0%	100.0%
4	100.0%	75.0%	100.0%	100.0%	100.0%
5	80.0%	60.0%	100.0%	100.0%	100.0%
6	66.7%	50.0%	100.0%	83.3%	100.0%
7	71.4%	57.1%	85.7%	71.4%	100.0%
8	75.0%	62.5%	75.0%	62.5%	100.0%
9	66.7%	66.7%	66.7%	66.7%	100.0%
10	60.0%	60.0%	60.0%	60.0%	100.0%
11	54.5%	54.5%	54.5%	63.6%	100.0%
12	50.0%	58.3%	50.0%	66.7%	100.0%
13	46.2%	53.8%	46.2%	69.2%	100.0%
14	42.9%	57.1%	50.0%	71.4%	100.0%
15	40.0%	60.0%	53.3%	73.3%	100.0%

Table 3: Details of top 15 episodes.

Top $k$	SkOPUS	PRT	EDP
1	$d \rightarrow e \rightarrow f \rightarrow g$	$d \rightarrow e \rightarrow f \rightarrow g$	$e \rightarrow f \rightarrow g$
2	$a \rightarrow b \rightarrow c$	$d \rightarrow e \rightarrow g$	$f \rightarrow g$
3	$d \rightarrow e \rightarrow f$	$d \rightarrow f \rightarrow g$	$e \rightarrow f$
4	$e \rightarrow f \rightarrow g$	$e \rightarrow f \rightarrow g$	$e \rightarrow g$
5	$d \rightarrow e \rightarrow g$	$d \rightarrow e \rightarrow f$	$d \rightarrow e \rightarrow f$
6	$c \rightarrow X \rightarrow d$	$a \rightarrow c \rightarrow d$	$d \rightarrow f$
7	$b \rightarrow c \rightarrow X \rightarrow d$	$a \rightarrow b \rightarrow c \rightarrow d$	$d \rightarrow e$
8	$b \rightarrow c \rightarrow d$	$b \rightarrow c \rightarrow d$	$a \rightarrow b$
9	$b \rightarrow X \rightarrow d$	$a \rightarrow b \rightarrow c$	$b \rightarrow c$
10	$a \rightarrow b \rightarrow c \rightarrow d$	$a \rightarrow b \rightarrow d$	$a \rightarrow c$
11	$a \rightarrow b \rightarrow X \rightarrow d$	$f \rightarrow g$	$a \rightarrow b \rightarrow c$
12	$a \rightarrow c \rightarrow X \rightarrow d$	$e \rightarrow f$	$d \rightarrow e \rightarrow f \rightarrow g$
13	$f \rightarrow g$	$e \rightarrow g$	$d \rightarrow f \rightarrow g$
14	$e \rightarrow f$	$d \rightarrow e$	$d \rightarrow e \rightarrow g$
15	$e \rightarrow g$	$d \rightarrow f$	$d \rightarrow g$

We explored every compared method on the  $\mathcal{F}$  and ranked the episodes by the deviation between the observed support and the expectation. For our EDP and IND model, we set  $min\_lift = 1$  and ranked the episodes whose Lift exceeds  $min\_lift$  with descending order. For other methods we directly ranked by the deviation used in their papers. We report precision at  $k$  in Table 2, that is, proportions of embedded patterns that are included in the top- $k$  patterns returned by each approach. Remember that non-empty subepisodes of  $abc$  and  $defg$  except the single events are real patterns based on our definition as there exists underlying generative rules to explain it. As a result, there could be 15 embedded patterns in the dataset. What we concern about is the ability of each approach to recognize these embedded patterns.

From the table, we observe the independence models, i.e. EGH and IND, do not perform well as expected, while the partition based models are more effective. Our EDP model, with the highest precision, ranks all the embedded patterns at top 15 rankings and significantly outperforms other methods. PRT and SkOUPS begin to report false positive patterns after 5th and 6th, respectively. In addition, we observe EDP provides 20 outputs whose Lift is greater than the threshold which was 1 in our setting. However we observe a large gap between the lift measure of the 15th and 16th in the results of EDP, which doubles the gap between the 1st and 15th. Though 16th-20th are false positives, we argue that we can filter the additional noises by finer tuning the threshold based on the gaps in the lift and our method already returns the targets on top 15.

Next we look closer to the results returned by the three partition models. Table 3 demonstrates the top 15 episodes produced by the three partition based models. Among them, we mark the false positive patterns in blue font. From the table we observe that SkOPUS cannot handle the high frequency noise event, i.e. X, very well, and X could be

doped inside patterns. On the other hand, SkOPUS may underestimate the expectation of patterns with longer length as their re-ordering candidates may be rare in the original sequence. Thus, the top 5 of SkOPUS are 3-episode or longer, and it can return defg and abc very early. The ineffectiveness of PRT comes from its incorrectness in identifying the event from other patterns. For example, it fails to recognize d is a free-rider event to the pattern abc. However, our EDP can correctly filter such episodes. As the fact that EDP can recognize more complicated redundant patterns, we conjecture the reason is that our partition model takes non-prefix episodes into account during the partition.

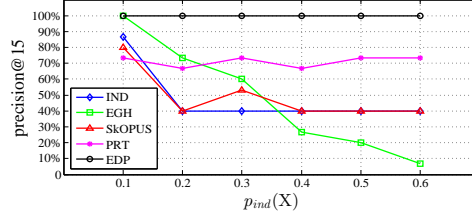


Fig. 3: The effect of  $p_{ind}(X)$ .

We next tune the probability of the embedded high frequency noise event, i.e.  $p_{ind}(X)$ , from 0.1 to 0.6 and visualize the Precision@15 of all compared methods as shown in Fig.3. From the figure, we observe that the probability of  $X$  has few effects on EDP. It always achieves the highest precisions with 100%. Second, the measure of IND, SkOPUS and PRT may vary around 40% to 75% as  $p_{ind}(X)$  varies. The only special case happens when  $p_{ind}(X) = 0.1$ , both IND and SkOPUS achieve more than 80% in precision. EGH, on the other hand, is much more sensitive to the  $p_{ind}(X)$ , and its performance drops significantly as the  $p_{ind}(X)$  increases. However, it performs well when the appearance probability of  $X$  is low.

As the synthetic experiments clearly show that IND and EGH overall are less effective than the others, we do not include them in the following experiments.

### 6.3 Results on real-world dataset

**Results on STK dataset.** We start from the most frequent 1,000 episodes mined from the STK dataset by setting  $min\_sup = 200$  and  $\delta = 5$ . We also ranked each episode by the deviation between the observed support and the expectation, and meanwhile we set  $min\_lift$  to 1 to screen redundant episodes for our method. We were surprised to find that each approach returned very different results on such dataset. Since the episodes related to stocks are not easy to understand, we design the following measure to demonstrate the results.

For the top  $k$  episodes given by each approach, we exhibit the percentage of the top 10 frequent events are included in. The reason we choose top 10 frequent events is that they have relatively high occurrence probabilities ranging from 0.2 to 0.26, which is twice to the average occurrence probability of events in such dataset (see Table 1). From the results on the synthetic dataset, we have known that real patterns may dope with some highly frequent but not related events, which makes the episode become a

Table 4: The percentage of the most frequent 10 events in the top  $k$  episodes in STK.

Top $k$	SkOPUS	PRT	EDP
1	33.3%	100.0%	0.0%
2	50.0%	100.0%	0.0%
3	44.4%	100.0%	0.0%
4	33.3%	100.0%	0.0%
5	33.3%	100.0%	20.0%
6	33.3%	100.0%	33.3%
7	33.3%	100.0%	28.6%
8	33.3%	100.0%	37.5%
9	33.3%	100.0%	33.3%
10	33.3%	100.0%	30.0%

free-rider. As a result, we argue that such ratio is higher may indicate that there might be more possibility to have free-rider episodes in the top  $k$  episodes.

Table 4 demonstrates the results, and we observe PRT achieves surprisingly high percentages. EDP outperforms SkOPUS especially when  $k$  is less than 5. To validate further, we check the top episodes produced by every method. We find that, as an example, the highest ranking episode returned by PRT indicates that the increase of a software company (SH.600100+) is followed by the increase of a security corporation (SH.600109+) and then leads to the increase of a metallurgical enterprise (SH.600111+). All the three events belong to different industry sectors and are included in top ten frequent events. In fact, from our common sense, we can hardly believe these three stocks have convincing correlations with each other. The results of SkOPUS is similar with that in the SYN dataset that the underlying patterns are always doped with the most frequent event, while the results provided by EDP are clearly different. For example, the episode “**Industrial and Commrc Bank of China+**”→“**Bank of China+**” and “**Minsheng Bank+**”→“**China Merchants Bank+**” rank at 1<sup>st</sup> and 2<sup>rd</sup>, respectively. Such patterns are more convincing and easy to understand from the stock names. The stocks in the same episode may have implicit correlations as they come from the same industry sector. We also find more examples from the results returned by EDP model but further discussions on possible related stocks are out of the scope of this paper.

Table 5: Top 10 output episodes on JMLR dataset.

Top $k$	SkOPUS	PRT	EDP
1	support→vector→machin (138)	support→vector (168)	support→vector (168)
2	support→vector (168)	support→vector→machin (138)	real→world (78)
3	vector→machin (151)	support→machin (142)	support→vector→machin (138)
4	support→machin (142)	vector→machin (151)	support→machin (142)
5	data→set→model (79)	real→world (78)	vector→machin (151)
6	base→algorithm→base (78)	featur→select (101)	bayesian→network (95)
7	data→set→data→set (76)	bayesian→network (95)	solv→problem (84)
8	paper→learn→algorithm (75)	problem→solv (79)	data→set (298)*
9	set→learn→set (75)	bayesian→model (92)	machin→learn (77)*
10	real→data (97)	solv→problem (84)	real→data (97)

**Results on JMLR dataset.** For the JMLR dataset, we are given 1,023 frequent episodes in  $\mathcal{F}$  by setting  $min\_sup = 50$  and  $\delta = 12$ . Then we explored the three partition models on such dataset. We visualize the top 10 episodes ranked by each method in Table 5. In such table, the number in the bracket is the support of corresponding episode. Since there are no ground truth patterns in such data, we can hardly compare which method is better in a quantitative manner. However, we find the following

interesting observations. First, confirm the ones in the previous comparisons, SkOPUS outputs free-rider episodes like “**data**→**set**→**model**” and “**base**→**algorithm**→**base**” at very early. In addition, SkOPUS seems to have negative correlations with support since it prefers longer patterns. We know the JMLR dataset is a sequence without simultaneous events, hence support based on minimal occurrence holds anti-monotonicity. PRT performs well on such data. Every pattern demonstrated in the table is meaningful, and all of them have lower frequency which is less than 200. Our EDP also provides meaningful results and has few correlations with pattern support. It can rank both high frequency as well as low frequency episode high, e.g. “**data**”→“**set**” (298) and “**machin**”→“**learn**” (77). We mark them by asterisk in the table. Since the basic idea of EDP is checking whether the observed support can be simulated by the random sequences generated by specific partitions. Episodes with high or low support may have possibility to pass such test. We conjecture it may be the reason why EDP can return both general and specific patterns.

#### 6.4 Scalability

Finally we evaluate the efficiency of EDP. Since EDP can be highly distributed, we mainly focus on the scalability of EDP by varying the number of processes. To this end, we implemented EDP with Python 3.5, distributed each partition model of an episode into different processes and verified its scalability on a shared memory computing system. The system equipped with two Intel Xeon E5-2620 CPUs and 128GB RAM running on Linux CentOS 6.5. Fig. 4 demonstrates the results. From the figure we can see EDP can achieve significant speedup as the number of processes increases on all datasets we investigated. Among them, EDP holds the most advanced speedup on the STK dataset. It is because we adopt small  $\delta$  when performing mining such that the episodes on STK have less length and smaller alphabet.

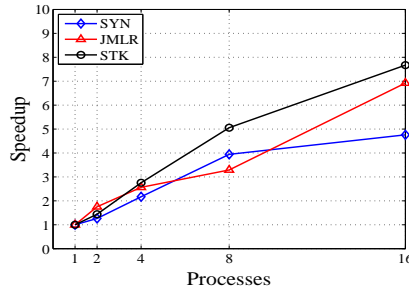


Fig. 4: The scalability of EDP.

## 7 Conclusion

In this paper, we proposed EDP model for reducing redundancy of frequent episodes in long single sequence. EDP first separates an episode into informative and random events, then builds a generative model for random sequences based on the partitions. Next we define the expected support of an episode if it is considered as a free-rider with these random sequences and devise an efficient algorithm to compute. Finally, the redundancy of an episode is determined by the deviation between the observation and the

estimated expectation. Experimental results on synthetic and real-world datasets clearly exhibit the effectiveness of the proposed method.

**Acknowledgements.** This research is supported by National key R&D program of China (No. 2017YFB1002104), National Natural Science Foundation of China (No.61602438, 91546122, 61573335), Guangdong provincial science and technology plan projects (No. 2015B010109005).

## References

1. Avinash Achar, A Ibrahim, and PS Sastry. Pattern-growth based frequent serial episode discovery. *DKE*, 2013.
2. Avinash Achar, Srivatsan Laxman, Raajay Viswanathan, and PS Sastry. Discovering injective episodes with general partial orders. *DMKD*, 2012.
3. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, and Qing He. Online frequent episode mining. In *ICDE*, 2015.
4. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, Qing He, and Zhongzhi Shi. Discovering and learning sensational episodes of news events. In *WWW*, 2014.
5. Xiang Ao, Ping Luo, Jin Wang, Fuzhen Zhuang, and Qing He. Mining precise-positioning episode rules from event sequences. *IEEE TKDE*, 2018.
6. Roel Bertens, Jilles Vreeken, and Arno Siebes. Keeping it short and simple: Summarising complex event sequences with multivariate patterns. In *KDD*, 2016.
7. Apratim Bhattacharyya and Jilles Vreeken. Efficiently summarising event sequences with rich interleaving patterns. In *SDM*, 2017.
8. Jaroslav Fowkes and Charles Sutton. A subsequence interleaving model for sequential pattern mining. In *KDD*, 2016.
9. Robert Gwadera, Mikhail J Atallah, and Wojciech Szpankowski. Reliable detection of episodes in event sequences. *KAIS*, 2005.
10. A Ibrahim, Shivakumar Sastry, and PS Sastry. Discovering compressing serial episodes from event sequences. *KAIS*, 2016.
11. Hoang Thanh Lam, Fabian Mörchen, Dmitriy Fradkin, and Toon Calders. Mining compressing sequential patterns. *Statistical Analysis and Data Mining*, 2014.
12. Srivatsan Laxman, PS Sastry, and KP Unnikrishnan. Discovering frequent episodes and learning hidden markov models: A formal connection. *IEEE TKDE*, 2005.
13. Michael Mampaey, Jilles Vreeken, and Nikolaj Tatti. Summarizing data succinctly with the most informative itemsets. *ACM TKDD*, 2012.
14. Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *DMKD*, 1997.
15. Jian Pei, Haixun Wang, Jian Liu, Ke Wang, Jianyong Wang, and Philip S Yu. Discovering frequent closed partial orders from strings. *IEEE TKDE*, 2006.
16. François Petitjean, Tao Li, Nikolaj Tatti, and Geoffrey I Webb. Skopus: Mining top-k sequential patterns under leverage. *DMKD*, 2016.
17. Nikolaj Tatti. Discovering episodes with compact minimal windows. *DMKD*, 2014.
18. Nikolaj Tatti. Ranking episodes using a partition model. *DMKD*, 2015.
19. Nikolaj Tatti and Boris Cule. Mining closed strict episodes. In *ICDM*, 2010.
20. Nikolaj Tatti and Boris Cule. Mining closed episodes with simultaneous events. In *KDD*, 2011.
21. Nikolaj Tatti and Jilles Vreeken. The long and the short of it: summarising event sequences with serial episodes. In *KDD*, 2012.
22. Jilles Vreeken and Nikolaj Tatti. Interesting patterns. In *Frequent pattern mining*. 2014.
23. Geoffrey I Webb. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. *ACM TKDD*, 2010.