A Simple Characterization of Serially Constructible Episodes

Takashi Katoh[1] and Kouichi Hirata[2⋆]

[1] Graduate School of Computer Science and Systems Engineering
[2] Department of Artificial Intelligence
Kyushu Institute of Technology
Kawazu 680-4, Iizuka 820-8502, Japan
{f673024t,hirata}@ai.kyutech.ac.jp
Tel: +81-948-29-7622, Fax: +81-948-29-7601

Abstract. This paper gives a simple characterization of episodes in episode mining that are constructible from just information for occurrences of serial episodes, called *serially constructible* episodes. First, we formulate an episode as an *acyclic transitive labeled digraph* of which label is an event type in episode mining. Next, we introduce a *parallel-free* episode that always has an arc between vertices with the same label. Also we formulate a *serially constructible* episode as an episode embedded into every parallel-free episode containing all of the serial episodes occurring in it. Then, we show that an episode is parallel-free if and only if it is serially constructible.

1 Introduction

It is one of the important tasks for data mining to discover frequent patterns from time-related data. Agrawal and Srikant [1, 10] have introduced one method for such a task called *sequential pattern mining* to discover frequent *subsequences* as patterns in a sequential database. The sequential pattern mining has been developed by designing the efficient algorithms [9], in particular, with a non-redundant form called a frequent *closed* subsequence [2, 11, 12]. While the sequential pattern mining is efficient in general, the pattern represents just a totally ordered relationship in time-related data.

Mannila *et al.* [8] have introduced another method for such a task called *episode mining* to discover frequent *episodes* as patterns in an *event sequence* that are a collection of events occurring frequently together in event sequences. In episode mining, the frequency is formulated as the number of occurrences of episodes in every *window* that is a subsequence of event sequences under a fixed time span.

Then, Mannila *et al.* [8] have formulated the episodes as *acyclic labeled digraphs* of which label is an event type and of which edges specify the temporal

⋆ The author is partially supported by Grand-in-Aid for Scientific Research 17200011 and 19300046 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

precedent-subsequent relation, and designed an algorithm to construct episodes from a *parallel episode* as a set of events and a *serial episode* as a sequence of events. Hence, the pattern in episode mining is more rich to represent temporal relationship than one in sequential pattern mining.

On the other hand, Mannila's work [8] for episode mining is known to be general but inefficient. In order to avoid such inefficiency of episode mining, the episode mining has been developed by introducing the specific form of episodes for every target area together with efficient algorithms [3–6]. In particular, Katoh *et al.* have introduced a *sectorial episode* [5, 7], a *diamond episode* [6] and an *elliptic episode* [4]. Then, they have designed efficient algorithms to extract such episodes by scanning an event sequence just once.

The above efficiency follows from the construction of specific forms of episodes from just information for occurrences of serial episodes in an event sequence, which is obtained by scanning an event sequence just once. On the other hand, there has remained an open problem of how form of episodes is constructible from such information. Note that the solution of this problem gives one of the theoretical limitations on efficiently constructing episodes, because such information is essential for constructing episodes in episode mining [8].

We explain the above problem by using an example. Consider the event sequence W consisting of a pair (e, t) of an event type e and the occurrence time t of e described as Figure 1 (upper left). Also all of the serial episodes occurring in W are described as Figure 1 (upper right). Then, consider the episodes D_1, D_2 and D_3 described in Figure 1 (lower) as acyclic labeled digraphs. Note first that all of D_i are embedded into W.

On the other hand, D_2 is *not* embedded into W with distinguishing an event type a in D_2, so D_2 is *not* constructible from just information for occurrences of serial episodes in D_2. Furthermore, all of the serial episodes occurring in D_3 are equal to ones in a serial episode *bcab*, which are underlined in Figure 1 (upper right). Hence, D_3 is *not* constructible from just information for occurrences of serial episodes in D_3, because we cannot distinguish an episode D_3 and a serial episode *bcab* from just such information.

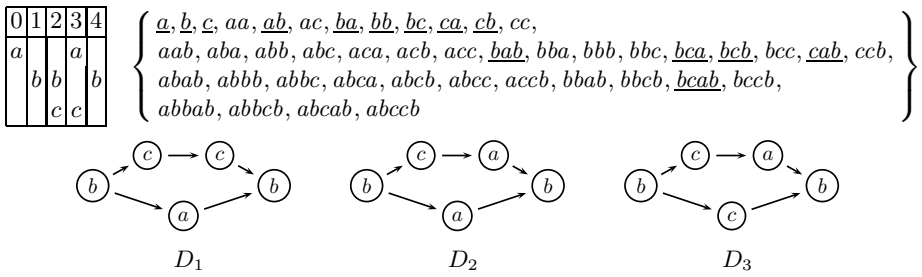


Fig. 1. An event sequence W (upper left), all of the serial episodes occurring in W (upper right), and episodes D_1, D_2 and D_3 (lower).

In order to solve the problem of *how form of episodes is constructible from just information for occurrences of serial episodes* with capturing the above situations, in this paper, we formulate both an episode and an event sequence as an acyclic *transitive* labeled digraph (*ATL-digraph*, for short) of which label is an event type. Here, we adopt the transitivity to represent all of the serial episodes contained in an episode or an event sequence explicitly. Then, we introduce the concept of a *parallel-free* a *serially constructible ATL-digraphs*.

Let $D = (V, A)$ be an ATL-digraph. Then, we say that D is *parallel-free* if, for every pair $(u, v) \in V \times V$ of vertices such that u and v have the same label, D has an arc either $(u, v) \in A$ or $(v, u) \in A$. In the above example, D_1 is parallel-free, but D_2 and D_3 are not.

Also we say that D is *serially constructible* if D is embedded into every parallel-free ATL-digraph containing all of the serial episodes occurring in D. Intuitively, this definition requires that, for ATL-digraphs D and W such that W contains all of the serial episodes occurring in D, every serial episode in W is corresponding to exactly one serial episode in D without duplications. In the above example, neither D_2 nor D_3 is embedded into the event sequence W without duplications.

Then, we show that *an episode* (as *an ATL-digraph*) *is parallel-free if and only if it is serially constructible*. This equivalence result gives a simple characterization of serially constructible episodes, and also implies that a parallel-free episode is one of the theoretical limitations on efficiently constructing episodes.

2 Episodes as Acyclic Transitive Labeled Digraphs

As similar as Mannila's episode mining [8], we assume that an event has an associated time of occurrence as a natural number. Formally, let $\mathcal{E}$ be a set of *event types*. Then, a pair (e, t) is called an *event*, where $e \in \mathcal{E}$ and t is a natural number which is the (*occurrence*) *time* of the event. For a set $E \subseteq \mathcal{E}$ of event types, we denote $\{(e, t) \mid e \in \mathcal{E}\}$ by (E, t).

An *event sequence* $\mathcal{S}$ on $\mathcal{E}$ is a triple $\langle S, T_s, T_e \rangle$, where

$$S = \langle (E_1, t_1), \ldots, (E_n, t_n) \rangle$$

is an ordered sequence of events satisfying the following conditions.

1. $E_i \subseteq \mathcal{E}$ $(1 \leq i \leq n)$,
2. $T_s \leq t_1 < \cdots < t_n \leq T_e$.

Here, T_s and T_e are called the *starting* time and the *ending* time of $\mathcal{S}$. In particular, a *window* in an event sequence $\mathcal{S} = (S, T_s, T_e)$ is an event sequence $W = (w, t_s, t_e)$ such that $t_s < T_e$, $T_s < t_e$ and w consists of all of the events (e, t) in S where $t_s \leq t < t_e$.

Mannila *et al.* [8] have formulated an episode as an *acyclic labeled digraph*. On the other hand, in this paper, we formulate an episode as an acyclic *transitive* labeled digraph. Note that we adopt the transitivity to represent all of the serial

episodes contained in an episode explicitly. Then, we prepare some notions for digraphs necessary for discussion bellow according to [4].

A *digraph* (or a *directed graph*) $D = (V, A)$ consists of a finite, nonempty set V of *vertices* and a (possibly empty) set A of ordered pairs of distinct vertices. We sometimes denote V by $V(D)$ and A by $A(D)$. We denote $|V|$ by $|D|$. A digraph $(\emptyset, \emptyset)$ is called *empty* and denoted by $\emptyset$.

An element of A is called an *arc*. For an arc $(u, v) \in A$, u is said to be *adjacent to* v and v is *adjacent from* u. For a digraph $D = (V, A)$ and a vertex $v \in V$, the *outdegree* of v in D, denoted by $od_D(v)$, is the number of vertices adjacent from v in D and the *indegree* of v in D, denoted by $id_D(v)$, is the number of vertices adjacent to v in D. Then, we define $ini(D) = \{v \in V \mid id(v) = 0\}$ and $fin(D) = \{v \in V \mid od(v) = 0\}$.

For two digraphs $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$, we denote a digraph $(V_1 \cup V_2, A_1 \cup A_2)$ by $D_1 \cup D_2$. For a digraph $D = (V, A)$ and $W \subseteq V$, we denote a digraph $(V - W, A - \{(v, u) \in A \mid v \in W \text{ or } u \in W\})$ by $D - W$. Furthermore, we denote $in(D, W) = \{v \in V \mid (v, w) \in A, w \in W\}$ and $out(D, W) = \{v \in V \mid (w, v) \in A, w \in W\}$.

For digraphs $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$, D_1 is a *subgraph* of D_2 if $V_1 \subseteq V_2$ and $A_1 \subseteq A_2$. For a digraph $D = (V, A)$ and a non-empty set $S \subseteq V$, the *subgraph of D induced by S*, denoted by $\langle S \rangle_D$, is the maximal subgraph of D of which vertices is S, that is, $\langle S \rangle_D = (S, \{(u, v) \in A \mid u, v \in S\})$.

Let D be a digraph (V, A). Then, a *walk* in D is an alternating sequence $W = v_0 a_1 v_1 \cdots a_n v_n$ of vertices and arcs, beginning and ending with vertices, such that $a_i = (v_{i-1}, v_i)$ for $1 \leq i \leq n$, and refer to W as a v_0-v_n walk. For vertices u and v in V, u is *accessible to* v (in D) if there exists a u-v walk in D.

A digraph D is *acyclic* if there exists no v-v walk in D. Also a digraph D is *transitive* if, for $u, v, w \in V$, it holds that $(u, w) \in A$ whenever it holds that $(u, v) \in A$ and $(v, w) \in A$. Furthermore, for a set L of labels, a digraph D is *labeled* (by L) if every vertex $v \in V$ has a label $l(v) \in L$. We call an acyclic transitive labeled digraph an *ATL-digraph*. For an ATL-digraph $D = (V, A)$, D^- denotes an acyclic labeled digraph obtained by removing an arc $(u, v) \in A$ such that there exist arcs $(u, w) \in A$ and $(w, v) \in A$ from A as possible. Note that D^- is uniquely determined for every ATL-digraph D.

Two digraphs $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$ are *isomorphic as labeled digraphs*, denoted by $D_1 \cong D_2$, if there exists a bijection φ from V_1 to V_2 such that $(u, v) \in A_1$ if and only if $(\varphi(u), \varphi(v)) \in A_2$, and $l(v) = l(\varphi(v))$ for every $v \in V_1$. A digraph $D_1 = (V_1, A_1)$ is *embedded into* a digraph $D_2 = (V_2, A_2)$ *as labeled digraphs*, denoted by $D_1 \sqsubseteq D_2$, if there exists an injection from V_1 to V_2 such that $(\varphi(u), \varphi(v)) \in A_2$ whenever $(u, v) \in A_1$, and $l(v) = l(\varphi(v))$ for every $v \in V_1$.

In this paper, we formulate an *episode* as an ATL-digraph of which label is an event type. Also Mannila *et al.* [8] have introduced a *serial episode* as a sequence of events. In this paper, we formulate a serial episode $a_1 \cdots a_n$ as an ATL-digraph $S = (\{v_1, \ldots, v_n\}, \{(v_i, v_j) \mid 1 \leq i < j \leq n\})$ such that $l(v_i) = a_i$. We sometimes identify a serial episode S with a label sequence $a_1 \cdots a_n$ of S.

For an episode D, we denote the set of all serial episodes embedded into D by $se(D)$, that is, $se(D) = \{S \mid S \sqsubseteq D \text{ and } S : \text{serial episode}\}$.

Furthermore, we also formulate an event sequence $\mathcal{S}$ as an ATL-digraph $d(\mathcal{S}) = (V, A)$ satisfying the following conditions.

1. For every event $(e, t) \in \mathcal{S}$, there exists a vertex $v_{e,t} \in V$ such that $l(v_{e,t}) = e$.
2. For every pair $((e, t), (e', t')) \in \mathcal{S} \times \mathcal{S}$ of events, $(v_{e,t}, v_{e',t'}) \in A$ iff $t < t'$.

It is obvious that, for an event sequence $\mathcal{S}$, $d(\mathcal{S})$ is determined uniquely.

3 Equivalence between Parallel-Free and Serially Constructible Episodes

In this section, we newly introduce a *parallel-free* and a *serially constructible* ATL-digraphs. Then, we show the main result of this paper that an episode as an ATL-digraph is parallel-free if and only if it is serially constructible.

Definition 1 (Katoh & Hirata [4]). For ATL-digraphs $W = (V_1, A_1)$ and $D = (V_2, A_2)$, we say that D is *parallel-free in* W if for every pair $(u, v) \in V_2 \times V_2$ such that $u \neq v$ and $l(u) = l(v)$, it holds that either $(u, v) \in A_1$ or $(v, u) \in A_1$.

Also we say that D is *parallel in* W if there exists a pair $(u, v) \in V_2 \times V_2$ such that $u \neq v$, $l(u) = l(v)$ and $(u, v), (v, u) \notin A_1$.

Furthermore, we say that D is *parallel-free* (*resp.*, *parallel*) if D is parallel-free (*resp.*, parallel) in D itself.

Note that the notion of "a parallel episode" in this paper is different from one in Mannila *et al.* [8]. For example, every serial episode is parallel-free. Also $d(\mathcal{S})$ for an event sequence $\mathcal{S}$ is parallel-free. Furthermore, if an ATL-digraph D is parallel-free, then D is parallel-free in an ATL-digraph W such that $D \sqsubseteq W$.

Definition 2 (Katoh & Hirata [4]). An ATL-digraph D is *serially constructible* if it holds that $D \sqsubseteq W$ for every parallel-free ATL-digraph W such that $se(D) \subseteq se(W)$.

Definition 2 requires that, for ATL-digraphs D and W, every serial episode in W is corresponding to exactly one serial episode in D. Hence, by regarding D as an episode and W as a window, Definition 2 claims that a window W contains the information of occurrences of serial episodes in D.

Example 1 (Katoh & Hirata [4]). Let D be a serial episode. Since $D \sqsubseteq D$ and by the definition of $se(D)$, it holds that $D \in se(D)$. Then, for every parallel-free ATL-digraph W such that $se(D) \subseteq se(W)$, it holds that $D \in se(W)$. By the definition of $se(W)$, it holds that $D \sqsubseteq W$, which implies that D is serially constructible. Hence, every serial episode is serially constructible.

Example 2 (Katoh & Hirata [4]). Let D and W be ATL-digraphs such that D^- and W^- are described as Figure 2. Then, it holds that W is parallel-free and $se(D) = se(W) = \{a, b, c, ab, bb, bc, abc, bbc, abbc\}$. However, there exists no injection from $V(D)$ to $V(W)$, so $D \not\sqsubseteq W$. Hence, D is *not* serially constructible.

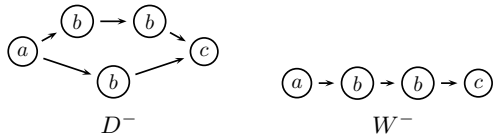


Fig. 2. D^- and W^- in Example 2.

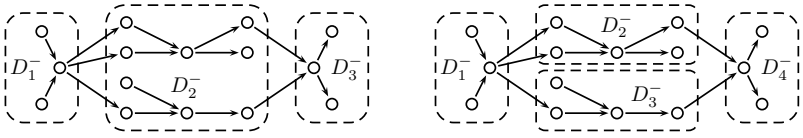
In order to show the equivalence between parallel-free and serially constructible episodes, it is necessary to decompose a digraph into a certain collection of digraphs. Then, we introduce a *decomposition* of digraphs.

For three digraphs $D_i = (V_i, A_i)$ $(i = 1, 2, 3)$ (that are possibly empty) and two sets $B_1 = \{(u, v) \mid u \in V_1, v \in V_2\}$ and $B_2 = \{(u, v) \mid u \in V_2, v \in V_3\}$ of arcs, $D_1 \oplus_{B_1} D_2 \oplus_{B_2} D_3$ denotes a digraph (V, A) such that $V = V_1 \cup V_2 \cup V_3$, $V_i \cap V_j = \emptyset$ $(1 \leq i < j \leq 3)$, and $A = A_1 \cup A_2 \cup A_3 \cup B_1 \cup B_2$.

Definition 3. Let $D = (V, A)$ be an ATL-digraph.

1. We say that D has a *serial decomposition* if there exist ATL-digraphs $D_i = (V_i, A_i)$ $(i = 1, 2, 3)$ and sets B_i $(i = 1, 2)$ of arcs such that $D^- = D_1^- \oplus_{B_1} D_2^- \oplus_{B_2} D_3^-$. In this case, we denote D by $[D_1, D_2, D_3]$.
2. We say that D has a *diamond decomposition* if there exist ATL-digraphs $D_i = (V_i, A_i)$ $(i = 1, 2, 3, 4)$ and sets B_i $(i = 1, 2, 3, 4)$ of arcs such that $D^- = (D_1^- \oplus_{B_1} D_2^- \oplus_{B_2} D_4^-) \cup (D_1^- \oplus_{B_3} D_3^- \oplus_{B_4} D_4^-)$ and $D_2 \cup D_3$ is parallel-free in D. In this case, we denote D by $[D_1, \langle D_2, D_3 \rangle, D_4]$.

Example 3. Consider the ATL-digraph D such that D^- is described as Figure 3, where every label in D is assumed to be distinct. Then, D has a serial decomposition $[D_1, D_2, D_3]$ and a diamond decomposition $[D_1, \langle D_2, D_3 \rangle, D_4]$, where D_i^- is described as the dashed boxes in Figure 3.



serial decomposition diamond decomposition

Fig. 3. D^- in Example 3

Lemma 1. *Let W be an ATL-digraph. Also let D and E be ATL-digraphs embedded into W having serial decompositions $[D_1, D_2, D_3]$ and $[E_1, E_2, E_3]$, respectively. If $D_1 \cong E_1$, $D_3 \cong E_3$ and $D_2 \cup E_2$ are parallel-free in W, then there*

exists an ATL-digraph F that is parallel-free in W, embedded into W and has a diamond decomposition $[F_1, \langle F_D, F_E\rangle, F_3]$ satisfying the following conditions. See Figure 4.

1. *$[F_1, F_D, F_3] \cong [D_1, D_2, D_3]$ and $[F_1, F_E, E_3] \cong [E_1, E_2, E_3]$.*
2. *$F_1 \cong D_1 (\cong E_1)$ and $F_3 \cong D_3 (\cong E_3)$.*
3. *$F_D \cong D_2$ and $F_E \cong E_2$.*
4. *For every $v \in V(F_1)$, it holds that either $v \in V(D_1)$ or $v \in V(E_1)$.*
5. *For every $v \in V(F_3)$, it holds that either $v \in V(D_3)$ or $v \in V(E_3)$.*

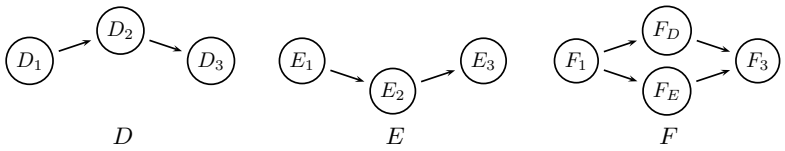


Fig. 4. Intuitive figures of D, E and F in Lemma 1.

Proof. We show the statement by induction on $|D_1|$ and $|D_3|$.

Consider the case that $|D_1| = |D_3| = 0$, that is, D_1 and D_3 are empty. Since $D_1 \cong E_1$ and $D_3 \cong E_3$, E_1 and E_3 are empty. Let F_1 and F_3 be empty digraphs. Then, the condition 1 obviously holds. Furthermore, since $D_2 \cup E_2$ is parallel-free, it holds that $[\emptyset, \langle D_2, E_2\rangle, \emptyset]$ is parallel-free.

Suppose that the statement holds for $|D_1| = |E_1| < n$ and $|D_3| = |E_3| < m$, and consider the case that $|D_3| = |E_3| = m$. Let φ be a bijection on $D_1 \cong E_1$ and $D_3 \cong E_3$. Also let v_1 be a vertex in $\mathit{fin}(D_3)$ and v_2 be $\varphi(v_1) \in E_3$. It is obvious that $v_2 \in \mathit{fin}(E_3)$. Furthermore, let $D'_1 = D_1$, $D'_2 = D_2$, $D'_3 = D_3 - \{v_1\}$ and $D'_4 = \langle\{v_1\}\rangle_{D_3}$. Then, we can write D as Figure 5. Here, for a set $A_{i,j} = \{(u, v) \in A(D) \mid u \in D_i, v \in D_j\}$ $(1 \leq i < j \leq 3)$ of arcs in D, every set $A'_{i,j}$ of arcs from $V(D'_i)$ to $V(D'_j)$ in D $(1 \leq i < j \leq 4)$ satisfies the following statements.

$A'_{1,4} = A_{1,3} \cap \{(v, v_1) \in A(D) \mid v \in V(D_1)\}$, $A'_{1,3} = A_{1,3} - A'_{1,4}$, $A'_{1,2} = A_{1,2}$,
$A'_{2,4} = A_{2,3} \cap \{(v, v_1) \in A(D) \mid v \in V(D_2)\}$, $A'_{2,3} = A_{2,3} - A'_{2,4}$,
$A'_{3,4} = \{(v, v_1) \in A(D) \mid v \in V(D_3)\}$.

Also let $E'_1 = E_1$, $E'_2 = E_2$, $E'_3 = E_3 - \{v_2\}$ and $E'_4 = \langle\{v_2\}\rangle_{E_3}$. Then, we can write E as Figure 5. Here, for a set $B_{i,j} = \{(u, v) \in A(E) \mid u \in E_i, v \in E_j\}$ $(1 \leq i < j \leq 3)$ of arcs in E, every set $B'_{i,j}$ of arcs from $V(E'_i)$ to $V(E'_j)$ in E $(1 \leq i < j \leq 4)$ satisfies the following statements.

$B'_{1,4} = B_{1,3} \cap \{(v, v_2) \in A(E) \mid v \in V(E_1)\}$, $B'_{1,3} = B_{1,3} - B'_{1,4}$, $B'_{1,2} = B_{1,2}$,
$B'_{2,4} = B_{2,3} \cap \{(v, v_2) \in A(E) \mid v \in V(E_2)\}$, $B'_{2,3} = B_{2,3} - B'_{3,4}$,
$B'_{3,4} = \{(v, v_2) \in A(E) \mid v \in V(E_3)\}$.

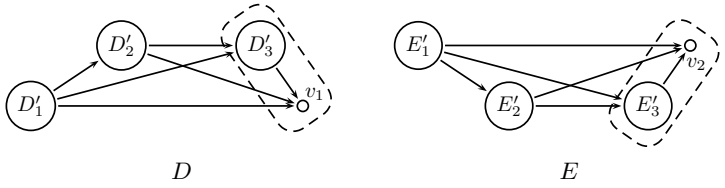


Fig. 5. Intuitive figures of D and E, where the dashed boxes mean D_3 and E_3, respectively.

Let $D' = D - \{v_1\}$ and $E' = E - \{v_2\}$. Then, it holds that $D' = [D'_1, D'_2, D'_3]$ and $E' = [E'_1, E'_2, E'_3]$. Since $|D'_3| = |E'_3| < n$ and by induction hypothesis, there exist ATL-digraphs F'_1 and F'_3 satisfying the following conditions. See Figure 6.

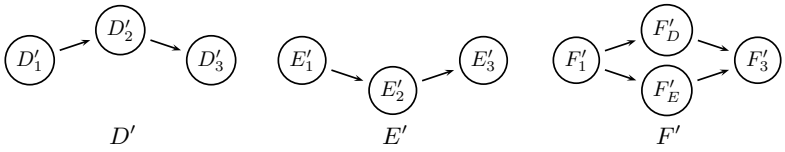


Fig. 6. Intuitive figures of D', E' and F'.

1. $F' = [F'_1, \langle F'_D, F'_E \rangle, F'_3]$ is parallel-free and embedded into D.
2. $[F'_1, F'_D, F'_3] \cong [D'_1, D'_2, D'_3]$ and $[F'_1, F'_E, F'_3] \cong [E'_1, E'_2, E'_3]$.
3. $F'_1 \cong D'_1 (\cong E'_1)$ and $F'_3 \cong D'_3 (\cong E'_3)$.
4. $F'_D \cong D'_2$ and $F'_E \cong E'_2$.
5. For every $v \in V(F'_1)$, either $v \in V(D'_1)$ or $v \in V(E'_1)$.
6. For every $v \in V(F'_3)$, either $v \in V(D'_3)$ or $v \in V(E'_3)$.

In the following, we explain how to construct F_3 satisfying the statements. Since W is parallel-free and by the definition of v_1 and v_2, v_1 and v_2 satisfy one of the conditions (a) $v_1 = v_2$, (b) $(v_1, v_2) \in A(W)$ or (c) $(v_2, v_1) \in A(W)$. We denote $A'_{1,4} \cup A'_{2,4} \cup A'_{3,4}$ and $B'_{1,4} \cup B'_{2,4} \cup B'_{3,4}$ by $A'_{*,4}$ and $B'_{*,4}$, respectively.

(a) In the case that $v_1 = v_2$, construct the following ATL-digraph F_3, see Figure 7 (a).

$$F_3 = (V(F'_3) \cup \{v_1\}, A(F'_3) \cup \{(v, v_1) \in A(W) \mid v \in V(F'_3)\}).$$

We denote an ATL-digraph by adding arcs $A'_{*,4} \cup B'_{*,4}$ to $[F'_1, \langle F'_D, F'_E \rangle, F_3]$ by F. Since F' is parallel-free, so is F. Furthermore, we check the conditions in the statement as follows.

Since $[F'_1, F'_D, F'_3] \cong D'$ and by the definition of D'_3 and D'_4, it holds that $[F'_1, F'_D, F_3] \cong D$. Also, since $[F'_1, F'_E, F'_3] \cong E'$ and by the definition of E'_3 and

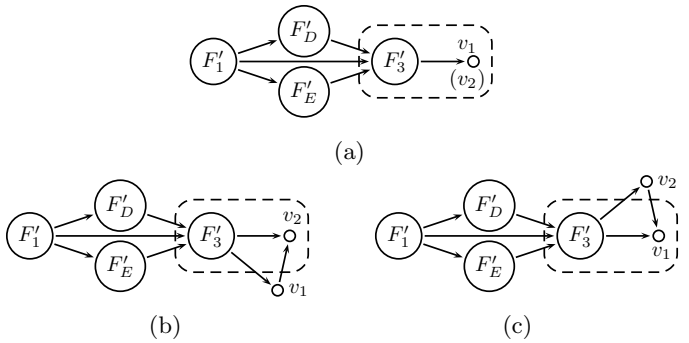


Fig. 7. Intuitive figures of F, where the dashed box is F_3 and we omit the arcs from F'_1, F'_D and F'_E to v_1 and v_2.

E'_4, it holds that $[F'_1, F'_E, F_3] \cong E$. Hence, the condition 1 holds. By induction hypothesis, the conditions 2, 3 and 4 also hold. Finally, we show the condition 5. For every $v \in V(F_3)$, if $v \neq v_1$, then either $v \in V(D'_3)$ or $v \in V(E'_3)$. Since $D'_3 = D_3 - \{v_1\}$ and $E'_3 = E_3 - \{v_1\}$, it holds that either $v \in V(D_3)$ or $v \in V(E_3)$. If $v = v_1$, then it holds that $v_1 \in V(D_3)$, since $v_1 \in \mathit{fin}(D_3)$. Hence, for every $v \in F_3$, either $v \in V(D_3)$ or $v \in V(E_3)$.

(b) Consider the case that $(v_1, v_2) \in A(W)$. Since $F'_3 \cong D'_3 \cong E'_3$ and $D_3 \cong E_3$, for a given bijection φ, $(v, v_1) \in A(D_3)$ if and only if $(\varphi(v), v_2) \in A(E_3)$, where $v_2 = \varphi(v_1)$. Since either $v \in V(D'_3)$ or $v \in V(E'_3)$ for every $v \in V(F'_3)$, there exists a vertex $v \in V(F'_3)$ such that either $(v, v_1) \in A(D_3)$ or $(v, v_2) \in A(E_3)$. For the former case, if $(v, v_1) \in A(D_3)$, then there exists an arc $(v, v_2) \in A(W)$, since W is transitive and by the supposition that $(v_1, v_2) \in A(W)$. Then, in both cases, there exists a vertex $v \in V(F'_3)$ such that $(v, v_2) \in A(W)$.

Hence, construct the following ATL-digraph F_3, see Figure 7 (b).

$$F_3 = (V(F'_3) \cup \{v_2\}, A(F'_3) \cup \{(v, v_2) \in A(W) \mid v \in V(F'_3)\}).$$

We denote an ATL-digraph by adding arcs $A'_{*,4} \cup B'_{*,4}$ to $[F'_1, \langle F'_D, F'_E \rangle, F_3]$ by F. Then, we can check that F satisfies the conditions as similar as the case (a).

(c) Consider the case that $(v_2, v_1) \in A(W)$. By the same reason of the case (b), there exists a vertex $v \in V(F_3)$ such that either $(v, v_1) \in A(D_3)$ or $(v, v_2) \in A(E_3)$. For the latter case, if $(v, v_2) \in A(E_3)$, then there exists an arc $(v, v_1) \in A(W)$, since W is transitive and by the supposition that $(v_2, v_1) \in A(W)$. Then, in both cases, there exists a vertex $v \in V(F_3)$ such that $(v, v_1) \in A(W)$.

Hence, construct the following ATL-digraph F_3, see Figure 7 (c).

$$F_3 = (V(F'_3) \cup \{v_1\}, A(F'_3) \cup \{(v, v_1) \in A(W) \mid v \in V(F'_3)\}).$$

We denote an ATL-digraph by adding arcs $A'_{*,4} \cup B'_{*,4}$ to $[F'_1, \langle F'_D, F'_E \rangle, F_3]$ by F. Then, we can check that F satisfies the conditions as similar as the case (a).

Furthermore, we can give the similar proof of the case that $|D_1| = |E_1| = n$. Hence, the statement holds. □

Theorem 1. *Every parallel-free ATL-digraph is serially constructible.*

Proof. For a parallel-free ATL-digraph F, we show the statement by induction on $|F|$. If $|F| \leq 1$, then F is a serial episode, so the statement holds.

Suppose that the statement holds for $|F| < n$ and consider the case that $|F| = n$. If F is a serial episode, then the statement obviously holds, so suppose that F is not a serial episode. Then, there exist vertices u and v in F such that (1) $(u, v) \notin A(F)$ and (2) $(v, u) \notin A(F)$. Let $acc(F, v) = \{v\} \cup in(F, \{v\}) \cup out(F, \{v\})$. Since F is transitive, $u \in acc(F, v)$ implies that either u is accessible to v in F or v is accessible to u in F. Then, for this v, we construct the following ATL-digraphs F_1, F_2, F_3 and F_4:

$$
\begin{aligned}
F_2 &= \langle V(F) - acc(F, v) \rangle_F, \\
F_1 &= \langle in(F, \{v\}) \cap in(F_2, \{v\}) \rangle_F, \\
F_4 &= \langle out(F, \{v\}) \cap out(F_2, \{v\}) \rangle_F, \\
F_3 &= \langle V(F) - (V(F_1) \cup V(F_2) \cup V(F_4)) \rangle_F.
\end{aligned}
$$

For example, consider an ATL-digraph F such that F^- is described in Figure 8 (left). Suppose that $v \in V(F)$ in Figure 8 (left) satisfies the above condition. Then, we obtain F_i^- $(i = 1, 2, 3, 4)$ as the dashed boxes in Figure 8 (right).

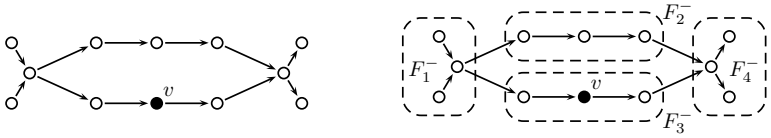


Fig. 8. ATL-digraphs F^- (left) and F_i^- $(i = 1, 2, 3, 4)$ (right).

Then, from F, we can construct the ATL-digraphs $X = [F_1, F_2, F_4]$ and $Y = [F_1, F_3, F_4]$. By the construction of F_i, it is obvious that $v \in V(F_3)$ and $u \in V(F_2)$, so it holds that $|X| < n$ and $|Y| < n$. Also it holds that $se(X) \subseteq se(F)$ and $se(Y) \subseteq se(F)$. Since F is parallel-free, $F_2 \cup F_3$ is also parallel-free

Let W be a parallel-free ATL-digraph such that $se(F) \subseteq se(W)$. Then, it holds that $se(X) \subseteq se(W)$ and $se(Y) \subseteq se(W)$. Since X and Y are serially constructible and by induction hypothesis, it holds that $X \sqsubseteq W$ and $Y \sqsubseteq W$.

By regarding X and Y as D and E in Lemma 1, there exists an ATL-digraph $Z = [F_1, \langle F_D, F_E \rangle, F_4]$ that is parallel-free in W and embedded into W. In particular, it holds that $F \cong Z$, so it holds that $F \sqsubseteq W$. Hence, F is serially constructible. □

Theorem 2. *Every serially constructible ATL-digraph is parallel-free.*

Proof. By contraposition, it is sufficient to show that, for every parallel ATL-digraph $D = (V, A)$, there exists an ATL-digraph W such that $se(D) \subseteq se(W)$ but $D \not\sqsubseteq W$.

Since D is parallel, there exists a pair $(u, v) \in V \times V$ such that $u \neq v$, $l(u) = l(v)$ and $(u, v), (v, u) \notin A$. For such u and v, let A_u and $A_{u \to v}$ be the following sets of arcs.

$$A_u = \{(w, u) \in A \mid w \in in(D, \{u\})\} \cup \{(u, w) \in A \mid w \in out(D, \{u\})\},$$
$$A_{u \to v} = \{(w, v) \mid w \in in(D, \{u\})\} \cup \{(v, w) \mid w \in out(D, \{u\})\}.$$

Furthermore, let W be an ATL-digraph $W = (V - \{u\}, (A - A_u) \cup A_{u \to v})$. Then, $D - \{u\} (= (V - \{u\}, A - A_u))$ is a subgraph of W.

Let $S_k^n(u) = v_1 \cdots v_{k-1} u v_{k+1} \cdots v_n$ and $S_k^n(v) = v_1 \cdots v_{k-1} v v_{k+1} \cdots v_n$ be serial episodes containing u and v at k with length n ($n \geq 1, 1 \leq k \leq n$). By the definition of W, for every serial episode $S_k^n(u)$ embedded into D, there exists a serial episode $S_k^n(v)$ embedded into W. Since $l(u) = l(v)$, it holds that $S_k^n(u) \cong S_k^n(v)$. Furthermore, since $D - \{u\}$ is a subgraph of W, every serial episode not containing u and embedded into D is also embedded into W. Hence, it holds that $se(D) \subseteq se(W)$.

On the other hand, since $|W| = |D| - 1$, there exists no injection from $V(D)$ to $V(W)$. Hence, it holds that $D \not\sqsubseteq W$. □

By summarizing Theorem 1 and 2 and by regarding an episode as an ATL-digraph, we obtain the following main result of this paper.

Theorem 3. *An episode is parallel-free if and only if it is serially constructible.*

4 Discussion

In this paper, by formulating both an episode and an event sequence as an ATL-digraph, we have introduced the concept of a *parallel-free* and a *serially constructible episodes*. Then, we have shown that *an episode is parallel-free if and only if it is serially constructible*. Since a serially constructible episode is an episode that is constructible from just information for occurrences of serial episodes, this equivalence result gives one of the theoretical limitations on efficiently constructing episodes.

Since the concept of parallel-free is very simple, it is a future work to design an algorithm to extract parallel-free episodes from an event sequence. In particular, it is a future work to restrict the forms of episodes such as sectorial, diamond and elliptic episodes [4–6] as target episodes, and then to design an efficient algorithm to extract such episodes.

Finally, we discuss the extension of our result to sequential pattern mining [1, 2, 9–12]. Since a serial episode is the special form of a sequence in sequential pattern mining without occurring two or more event types, we can regard a sequence as a serial episode consisting of a set of event types. Then, by extending the labels of an ATL-digraph from a single event type to a set of event types, we can formulate an *extended ATL-digraph* for sequences. We call the representation by an extended ATL-digraph an *extended episode*.

Note that we can give two definitions of a parallel-free extended episode $D = (V, A)$ if, for every pair $(u, v) \in V \times V$ of vertices that (A) u and v has

the same label or (B) u and v contain the same event type, D has an arc either $(u, v) \in A$ or $(v, u) \in A$. It is obvious that Definition (B) implies Definition (A). Also, in order to formulate a serially constructible extended episode, it is necessary to introduce the embedding relation $\sqsubseteq$ on extended ATL-digraphs with introducing the embedding relation on the labels of vertices.

Hence, it is an important future work to solve the problem of whether or not we can formulate a parallel-free and serially constructible extended episodes preserving the equivalence between them. Furthermore, if we can solve the problem positively, then it is also a future work to design an algorithm to extract extended episodes efficiently from just information for occurrences of sequences, by restricting the forms of extended episodes as similar as [4–6] and by using various methods [2, 9, 11, 12].

References

1. R. Agrawal, R. Srikant: *Mining sequential patterns*, Proc. 11th ICDE, 3–14, 1995.
2. H. Arimura, T. Uno: *An efficient polynomial space and polynomial delay algorithm for enumeration of maximal motif in a sequence*, J. Comb. Optim. **13**, 243–262, 2007.
3. C. Bettini, S. Wang, S. Jajodia, J.-L. Lin: *Discovering frequent event patterns with multiple granularities in time sequences*, IEEE Trans. Knowledge and Data Engineering **10**, 222–237, 1998.
4. T. Katoh, K. Hirata: *Mining frequent elliptic episodes from event sequences*, Proc. 5th LLLL, 46–52, 2007.
5. T. Katoh, K. Hirata, M. Harao: *Mining sectorial episodes from event sequences*, Proc. 10th DS, LNAI **4265**, 137–145, 2006.
6. T. Katoh, K. Hirata, M. Harao: *Mining frequent diamond episodes from event sequences*, Proc. 4th MDAI, LNAI **4617**, 477–488, 2007.
7. T. Katoh, K. Hirata, M. Harao, S. Yokoyama, K. Matsuoka: *Extraction of sectorial episodes representing changes for drug resistant and replacements of bacteria*, Proc. CME'07, 304–309, 2007.
8. H. Mannila, H. Toivonen, A. I. Verkamo: *Discovery of frequent episodes in event sequences*, Data Mining and Knowledge Discovery **1**, 259–289, 1997.
9. J. Pei, J. Han, B. Mortazavi-Asi, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu: *Mining sequential patterns by pattern-growth: The PrefixSpan approach*, IEEE Trans. Knowledge and Data Engineering **16**, 1–17, 2004.
10. R. Srikant, R. Agrawal: *Mining sequential patterns: Generalizations and performance improvements*, Proc. 5th EDBT, 3–17, 1996.
11. J. Wang, J. Han: *BIDE: Efficient mining of frequent closed sequences*, Proc. 20th ICDE, 2004.
12. X. Yan, J. Han, R. Afshar: *CloSpan: Mining closed sequential patterns in large datasets*, Proc. 3rd SDM, 2003.