

A Formal Query Language and Automata Model for Aggregation in Complex Event Recognition

Pierre Bourhis ✉ 

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISTAL, INRIA Lille, F-59000 Lille, France

Cristian Riveros ✉ 

Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Institute for Foundational Research on Data, Santiago, Chile

Amaranta Salas ✉ 

Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Institute for Foundational Research on Data, Santiago, Chile

Abstract

Complex Event Recognition (CER) systems are used to identify complex patterns in event streams, such as those found in stock markets, sensor networks, and other similar applications. An important task in such patterns is aggregation, which involves summarizing a set of values into a single value using an algebraic function, such as the maximum, sum, or average, among others. Despite the relevance of this task, query languages in CER typically support aggregation in a restricted syntactic form, and their semantics are generally undefined.

In this work, we present a first step toward formalizing a query language with aggregation for CER. We propose to extend Complex Event Logic (CEL), a formal query language for CER, with aggregation operations. This task requires revisiting the semantics of CEL, using a new semantics based on bags of tuples instead of sets of positions. Then, we present an extension of CEL, called Aggregation CEL (ACEL), which introduces an aggregation operator for any commutative monoid operation. The operator can be freely composed with previous CEL operators, allowing users to define complex queries and patterns. We showcase several queries in practice where ACEL proves to be natural for specifying them. From the computational side, we present a novel automata model, called Aggregation Complex Event Automata (ACEA), that extends the previous proposal of Complex Event Automata (CEA) with aggregation and filtering features. Moreover, we demonstrate that every query in ACEL can be expressed in ACEA, illustrating the effectiveness of our computational model. Finally, we study the expressiveness of ACEA through the lens of ACEL, showing that the automata model is more expressive than ACEL.

2012 ACM Subject Classification Information systems → Stream management; Theory of computation → Database theory; Theory of computation → Database query languages (principles)

Keywords and phrases Streams, complex event recognition, query language, aggregation

Digital Object Identifier 10.4230/LIPIcs.ICDT.2026.15

Related Version *Full Version:* <https://arxiv.org/abs/2601.00967> [8]

Funding The work of Bourhis was supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-23-PECL-0008. The work of Riveros and Salas was supported by ANID Fondecyt Regular project 1230935 and ANID – Millennium Science Initiative Program – Code ICN17_002.

1 Introduction

Complex Event Recognition (CER) systems are a group of data stream management systems for the detection of special events in real-time, called complex events, that satisfy a pattern, considering their position, the order, and other constraints between them [17, 12]. Some examples of its use are maritime monitoring [26], network intrusion detection [24], industrial



© Pierre Bourhis, Cristian Riveros, and Amaranta Salas;
licensed under Creative Commons License CC-BY 4.0

29th International Conference on Database Theory (ICDT 2026).

Editors: Balder ten Cate and Maurice Funk; Article No. 15; pp. 15:1–15:20

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

control systems [21], and real-time analytics [30]. In the literature, people have proposed multiple systems and query languages based on different formalisms for approaching complex events, such as automata-based, logic-based, tree-based, or a combination of them [17]. Examples of CER systems developed in academic and industrial contexts include SASE [34], EsperTech [1], and CORE [9, 7], among others.

A problem in CER systems is that their query languages, used to declare complex events, are, unfortunately, underspecified with respect to both their syntax and semantics. As observed in previous works [36, 15, 11, 5], CER query languages in systems typically lack a simple, compositional, and denotational semantics. In general, its semantics is defined indirectly through examples [2, 10], or by translation into evaluation models [25, 31, 33]. Recently, this issue in CER systems has been studied more thoroughly, and a query language that has successfully defined the semantics of several CER operators is *Complex Event Logic* (CEL) [19, 20], alongside a computational model called *Complex Event Automata* (CEA), which is based on the theory of finite state transducers and symbolic automata.

An open problem in formalizing CER query language is that several interesting queries in practice include *aggregation*, which previous proposals have not addressed. Aggregation refers to any subprocess in a query that combines and merges several (most often numerical) values into a single one [18], such as taking the average, the sum, or the maximum of a list of values. Examples of CER systems that used aggregation are SASE [14, 35], EsperTech [1], GLORIA [23], GRETA [28], and others [17, 12]. For illustrating a prototypical query with aggregation, consider the following (simplified) query from SASE [35, p. 3]:

```
Q1: PATTERN seq(JobStart a, Mapper+ b[ ], JobEnd c)
WHERE a.job_id = b[i].job_id and a.job_id = c.job_id
RETURN AVG(b[ ].period), MAX(b[ ].period)
```

Intuitively, the previous query aims to retrieve the average and maximum periods from a list of running times of mappers. For this, it looks for events from the stream, in the “PATTERN” clause that match the pattern: a JobStart typed event, followed by one or more Mapper events, and finally a JobEnd event. In turn, it uses the “WHERE” clause to ensure that each event matching the pattern has the same id attribute, and finally, it returns the average and maximum of the events that meet the conditions.

Previous proposals for formalizing CER query languages do not include such queries with aggregation, as they must not only detect and retrieve complex events but also produce new events and values. In particular, aggregation queries cannot be defined by logics like CEL or computational models like CEA, or any other formalization of CER as it is currently defined. These issues imply that CER query languages with aggregation are difficult to compare, unclear how to compose queries, and difficult to evaluate (i.e., without knowing the real meaning of a query). Furthermore, computational models for compiling queries with aggregation are not well understood, and systems usually rely on ad-hoc evaluation strategies suitable for specific queries and patterns.

In this work, we propose an extension of the logic CEL, and its corresponding computational model CEA, to express queries with aggregation, which we call *Aggregation Complex Event Logic* (ACEL) and *Aggregation Complex Event Automata* (ACEA), respectively. Our main goal is to design a logic and computational model, with a formal semantics that formalizes aggregation in CER and serves as a base for all CER languages.

For extending CEL with aggregation, we need to revisit its semantics. One of the first problems to arise with the current semantics of CEL is that a CEL formula retrieves the positions in the streams that fire the complex events, but it does not allow the creation of

new values or events. For this reason, we propose a new, equivalent semantics for CEL that returns events instead of positions. Additionally, since we also need to maintain duplicates for aggregation, we extend the semantics by using bags of events instead of sets. We then prove that the new semantics is equivalent to the previous one. Interestingly, the new semantics enable us to define new relevant operators for CER, such as attribute projection.

To formalize the aggregation in CER, we consider a general setting of aggregation based on *aggregate functions* [22, 18]; these are functions that go from a bag of values to single values, and they aim to summarize information (like count, sum, etc). By using this general framework of aggregation functions, they support our proposal in providing a general framework for aggregation in CER. Furthermore, we introduce an operator $\text{Agg}_Y^{\mathbf{b} \leftarrow \otimes X(\mathbf{a})}$ for variables named X and Y , attributes named \mathbf{a} and \mathbf{b} , and aggregation function \otimes , which takes a bag of events stored in X and \otimes -operates it corresponding attribute \mathbf{a} , storing the result in another attribute \mathbf{b} of an event in another variable Y . We formally define its syntax and semantics in Section 5. An advantage of this definition is that we can compose the Agg operator and every other operator in CEL. We show that most CER queries with aggregation from previous works are definable with ACEL.

An advantage of CEL is that one can characterize its expressive power with the so-called Complex Event Automata (CEA); specifically, that for every CEL formula, there exists an equivalent CEA, and vice versa. The practical relevance of this result is that CEL is useful for users to define queries, where CEA is useful for systems to evaluate them. In this work, we aim to achieve an equivalent result, so our next step was to find a machine model that can extend CEA and formally define ACEL. We introduce an automata model with aggregation for ACEL, which we call *Aggregation Complex Event Automata* (ACEA), an extension of CEA with registers to aggregate values. This extension employs the same concept of operating values in transitions and maintaining registers as cost register automata [4]. Specifically, in each transition, an ACEA takes an event and updates its register with those new values. Then, it performs an operation based on the assignments, checks if it satisfies a predicate, and finally, it creates a new tuple with the aggregated values. One of our main results is that we can compile every ACEL formula into an ACEA, namely, we can prove that the expressive power of ACEL is a subset of ACEA.

Outline. We present the preliminaries in Section 2. In Section 3, we discuss the necessary changes in CEL semantics and we show a new operator, projection by attribute. In Section 4, we discuss the setting of aggregate functions. In Section 5, we formally introduce ACEL, and we introduce ACEA in Section 6, to study the compilation of CEL formulas into ACEA and its equivalence with CEA. We conclude and discuss future work in Section 7. An extended version with all missing proofs and examples can be found in [8].

2 Preliminaries

Sets, intervals, and mappings. Given a set A , we denote by $\mathcal{P}(A)$ the set of all finite subsets of A . We denote by \mathbb{N} the natural numbers. Given $n, m \in \mathbb{N}$ with $n \leq m$, we denote by $[n]$ the set $\{1, \dots, n\}$ and by $[n..m]$ the interval $\{n, n+1, \dots, m\}$ over \mathbb{N} . As usual, we write $f : A \rightarrow B$ to denote a *function* f from the set A to B where every element in A has an image. A *mapping* M is a partial function that maps a finite number of elements from A to elements over B . We write $M : A \mapsto B$ to denote a mapping M from A to B . We denote by $\text{dom}(M)$ the domain of M (i.e., all $a \in A$ such that $M(a)$ is defined), and by $\text{img}(M)$ the image of M . We will usually use the notation $[a_1 \mapsto b_1, \dots, a_k \mapsto b_k]$ to define a mapping

15:4 A Formal Query Language for Aggregation in CER

M with $\text{dom}(M) = \{a_1, \dots, a_k\}$, $\text{img}(M) = \{b_1, \dots, b_k\}$ and $M(a_i) = b_i$ for every $i \in [k]$. Furthermore, for a map M and $a \notin \text{dom}(M)$ we write $[M, a \mapsto b]$ to specify a new map M' that extends M mapping a to b .

Bags. A *bag* or *multiset* (with own identity) B is a mapping $B : I \mapsto U$ where $I = \text{dom}(B)$ is a finite set of identifiers (or ids) and $U = \text{img}(B)$ is the underlying set of the bag. Given any bag B , we refer to these components as $I(B)$ and $U(B)$, respectively. For example, a bag $B = \{\{a, a, b\}\}$ (where a is repeated twice) can be represented with a mapping $B_0 = [1 \mapsto a, 2 \mapsto a, 3 \mapsto b]$ where $I(B_0) = \{1, 2, 3\}$ and $U(B_0) = \{a, b\}$. In general, we will use the standard notation for bags $\{\{a_1, \dots, a_n\}\}$ to denote the bag B whose identifiers are $I(B) = \{1, \dots, n\}$ and $B(i) = a_i$ for each $i \in I(B)$. We will use \uplus to refer to the union of bags: for every bags A and B , we define the bag $C := A \uplus B$ such that $\text{dom}(C) = (\text{dom}(A) \times \{0\}) \cup (\text{dom}(B) \times \{1\})$ and $C((i, b)) = X(i)$ where $X = A$ if $b = 0$ and $X = B$, otherwise. In other words, we take the disjoint union of the identifiers mapping every identifier to its corresponding element in A or B . Further, we define $\mathcal{P}_{\text{bags}}(A)$ as the set of all finite bags that one can form from a set A .

Computational model. We assume the model of *random access machines (RAM)* with uniform cost measure, and addition and subtraction as basic operations [3]. This implies, for example, that the access to a lookup table (i.e., a table indexed by a key) takes constant time. These are common assumptions in the literature of the area [9, 32].

3 Revisiting the semantics of Complex Event Logic

In this section, we revisit the semantics of CEL [20] and present a new semantics based on tuples instead of positions. We then prove the equivalence between the two versions. This new semantics allows, for example, the definition of a new operator for CEL, called *attribute-projection*, which cannot be defined with the old semantics. Furthermore, the new semantics is crucial to introduce aggregation in CEL in the next section.

Events and streams. We fix a countably infinite set of *attribute names* \mathbf{A} and a countably infinite of *data values* \mathbf{D} (e.g. integers, strings). An (untyped) *event* e is a pair (M, i) such that $M : \mathbf{A} \mapsto \mathbf{D}$ maps attribute names from \mathbf{A} to data values in \mathbf{D} , and $i \in \mathbb{N}$ is the time of the event [16] (we prefer to use *discrete time*, which is enough for our purposes). Intuitively, M defines the data of the event (i.e., as a tuple). We denote by $e(\mathbf{a}) \in \mathbf{D}$ the value of the attribute $\mathbf{a} \in \mathbf{A}$ assigned by M (i.e., $e(\mathbf{a}) = M(\mathbf{a})$). If e is not defined on attribute \mathbf{a} , then we write $e(\mathbf{a}) = \text{NULL}$. Furthermore, for the sake of simplification we also denote $e(\text{time}) = i$ (note, however, that *time* is not an attribute). We define by $\text{Att}(e)$ the set of attributes of e , namely, $\text{Att}(e) = \text{dom}(M)$. We write \mathbf{E} to denote the *set of all events* over attributes names \mathbf{A} and data values \mathbf{D} . We will usually use bold letters \mathbf{a} , \mathbf{b} , and \mathbf{c} to denote attribute names in \mathbf{A} and (normal) letters a , b , and c to denote data values in \mathbf{D} .

Fix now a finite set of *event types* \mathbf{T} and assume that $\mathbf{T} \subseteq \mathbf{D}$ and $\text{NULL} \in \mathbf{D}$. In this work, we assume the existence of a distinguished attribute *type* that defines the *type* of an event. Specifically, let *type* be an attribute such that *type* $\in \mathbf{A}$. For every event e , we assume that *type* $\in \text{Att}(e)$ and $e(\text{type}) \in \mathbf{T} \cup \{\text{NULL}\}$ is the type of e . Notice that e could be *typed* (i.e., $e(\text{type}) \in \mathbf{T}$) or *untyped* in which case we have $e(\text{type}) = \text{NULL}$. A *schema* Σ is a function $\Sigma : \mathbf{T} \rightarrow \mathcal{P}(\mathbf{A}_\Sigma)$ where $\mathbf{A}_\Sigma \subseteq \mathbf{A}$ is a finite set of attributes. We say that an event e satisfies the schema Σ if, and only if, e is a typed event and $\text{Att}(e) = \Sigma(e(\text{type})) \cup \{\text{type}\}$. In particular, untyped events do not satisfy a schema by definition.

Let $\Sigma : \mathbf{T} \rightarrow \mathcal{P}(\mathbf{A}_\Sigma)$ be a schema. A *stream* over a schema Σ is an (arbitrary long) sequence $\mathcal{S} = e_1 e_2 \dots e_n$ of typed events such that, for every $i \in [n]$, it holds that e satisfies Σ and $e_i(\text{time}) = i$. In other words, a stream consists of typed events according to Σ and every time of an event is the position in the stream. Note that we defined the type and time of an event for its later use in the semantics of CEL. The first will allow us to know the attributes of each event when we compile CEL into an automata model, and the second will allow us to differentiate between tuples by adding its *origin* in the stream [6].

► **Example 1.** As a running example, consider that we have a stream $\mathcal{S}_{\text{Stocks}}$ that is emitting buy and sell events of particular stocks [9]. Here, we assume a schema Σ_{Stocks} with attributes **name** and **price** that represents the name of the stock (e.g., INTL for intel) and its price (e.g., US\$80), respectively. We have two types, called **BUY** and **SELL**, and $\Sigma_{\text{Stocks}}(\text{BUY}) = \Sigma_{\text{Stocks}}(\text{SELL}) = \{\text{name}, \text{price}\}$. A possible stream $\mathcal{S}_{\text{Stocks}}$ could be the following:

$\mathcal{S}_{\text{Stocks}}$:	0	1	2	3	4	5	6	7	8	9
	[SELL MSFT 101]	[SELL MSFT 102]	[SELL INTL 80]	[BUY INTL 80]	[SELL AMZN 1900]	[SELL INTL 81]	[BUY AMZN 1920]	[BUY MSFT 101]	[BUY INTL 79]	[SELL INTL 80]

Note that each event contains a type (i.e., **BUY** or **SELL**), its attributes values (i.e., **name** and **price**) and its time (i.e., the position above the event). Further, each event satisfies Σ_{Stocks} .

The notion of a *renaming of a event* will be useful in this work (e.g., see Section 6). Formally, we define a *renaming* r as a mapping $r : \mathbf{A} \mapsto \mathbf{A}$. We say that an event e is consistent with a renaming r iff $\text{dom}(r) \subseteq \text{dom}(e)$ and, if $r(\mathbf{a}) = r(\mathbf{b})$, then $e(\mathbf{a}) = e(\mathbf{b})$ for every $\mathbf{a}, \mathbf{b} \in \text{dom}(r)$. Given an event e consistent with r , we define the *renamed* event $r(e)$ such that $[r(e)](\text{time}) = e(\text{time})$ and $[r(e)](r(\mathbf{a})) = e(\mathbf{a})$ for every attribute $\mathbf{a} \in \text{dom}(r)$. In other words, r renames some attributes \mathbf{a} of e to $r(\mathbf{a})$. We define by Ren the set of all renamings over \mathbf{A} .

Predicates of events. A *predicate* is a possibly infinite set P of events. For instance, P could be the set of all events e such that $e(\mathbf{a}) \leq 20$. In our examples, we will use the notation $\mathbf{a} \sim a$ where $\mathbf{a} \in \mathbf{A}$, $a \in \mathbf{D}$, and \sim is a binary relation over \mathbf{D} to denote the predicate $P = \{e \mid e(\mathbf{a}) \sim a\}$. We say that an event e satisfies predicate P , denoted $e \models P$, if, and only if, $e \in P$. We generalize this notation from events to a bag of events E such that $E \models P$ if, and only if, $e \models P$ for every $e \in E$.

In this work, we assume a fix *set of predicates* \mathbf{P} that is close under intersection, negation, and renaming, namely, $P_1 \cap P_2 \in \mathbf{P}$, $\mathbf{E} \setminus P \in \mathbf{P}$, and $r(P) \in \mathbf{P}$ for every $P, P_1, P_2 \in \mathbf{P}$ and $r \in \text{Ren}$ where $r(P) = \{r(e) \mid e \in P \wedge e \text{ is consistent with } r\}$ and \mathbf{E} , the set of all events, is a predicate in \mathbf{P} that we usually denote by **TRUE**.

Complex events. In this work, we will use a slightly different definition of complex event: we will store events inside valuations, instead of storing positions like in [9]. Formally, fix a finite set \mathbf{X} of *variables*, which includes all event types (i.e. $\mathbf{T} \subseteq \mathbf{X}$). Let \mathcal{S} be a stream of length n . A *complex event* of \mathcal{S} is a triple (i, j, μ) where $i, j \in [n]$, $i \leq j$, and $\mu : \mathbf{X} \rightarrow \mathcal{P}_{\text{bags}}(\mathbf{E})$ is a function from variables to finite bags of events. Intuitively, i and j marks the beginning and end of the interval where the complex event happens, and μ stores the events in the interval $[i..j]$ that fired the complex event. In the following, we will usually denote C to denote a complex event (i, j, μ) of \mathcal{S} and omit \mathcal{S} if the stream is clear from the context. We will use $\text{time}(C)$, $\text{start}(C)$, and $\text{end}(C)$ to denote the interval $[i..j]$, the start i , and the end j of C , respectively. Further, by some abuse of notation we will also use $C(X)$ for $X \in \mathbf{X}$ to denote the bag $\mu(X)$ of C .

$$\begin{aligned}
 \llbracket R \rrbracket(\mathcal{S}) &= \{(i, i, \mu) \mid i \in [k] \wedge e_i(\text{type}) = R \wedge \mu(R) = \{\{e_i\}\} \wedge \forall Y \neq X. \mu(Y) = \emptyset\} \\
 \llbracket \varphi \text{ AS } X \rrbracket(\mathcal{S}) &= \{C \mid \exists C' \in \llbracket \varphi \rrbracket(\mathcal{S}). \text{time}(C) = \text{time}(C') \wedge C(X) = \biguplus_Y C'(Y) \\
 &\quad \wedge \forall Z \neq X. C(Z) = C'(Z)\} \\
 \llbracket \varphi \text{ FILTER } X[P] \rrbracket(\mathcal{S}) &= \{C \mid C \in \llbracket \varphi \rrbracket(\mathcal{S}) \wedge C(X) \models P\} \\
 \llbracket \pi_L(\varphi) \rrbracket(\mathcal{S}) &= \{\pi_L(C) \mid C \in \llbracket \varphi \rrbracket(\mathcal{S})\} \\
 \llbracket \varphi_1 \text{ OR } \varphi_2 \rrbracket(\mathcal{S}) &= \llbracket \varphi_1 \rrbracket(\mathcal{S}) \cup \llbracket \varphi_2 \rrbracket(\mathcal{S}) \\
 \llbracket \varphi_1 \text{ AND } \varphi_2 \rrbracket(\mathcal{S}) &= \llbracket \varphi_1 \rrbracket(\mathcal{S}) \cap \llbracket \varphi_2 \rrbracket(\mathcal{S}) \\
 \llbracket \varphi_1 : \varphi_2 \rrbracket(\mathcal{S}) &= \{C_1 \uplus C_2 \mid C_1 \in \llbracket \varphi_1 \rrbracket(\mathcal{S}) \wedge C_2 \in \llbracket \varphi_2 \rrbracket(\mathcal{S}) \wedge \text{end}(C_1) + 1 = \text{start}(C_2)\} \\
 \llbracket \varphi_1 ; \varphi_2 \rrbracket(\mathcal{S}) &= \{C_1 \uplus C_2 \mid C_1 \in \llbracket \varphi_1 \rrbracket(\mathcal{S}) \wedge C_2 \in \llbracket \varphi_2 \rrbracket(\mathcal{S}) \wedge \text{end}(C_1) < \text{start}(C_2)\} \\
 \llbracket \varphi \oplus \rrbracket(\mathcal{S}) &= \llbracket \varphi \rrbracket(\mathcal{S}) \cup \llbracket \varphi : \varphi \oplus \rrbracket(\mathcal{S}) \\
 \llbracket \varphi + \rrbracket(\mathcal{S}) &= \llbracket \varphi \rrbracket(\mathcal{S}) \cup \llbracket \varphi ; \varphi + \rrbracket(\mathcal{S})
 \end{aligned}$$

■ **Figure 1** The semantics of CEL defined over a stream $\mathcal{S} = e_1 e_2 \dots e_n$ where each e_i is an event.

The following operations on complex events will be useful throughout the paper. We define the *union* of complex events C_1 and C_2 , denoted by $C_1 \uplus C_2$, as the complex event C' such that $\text{start}(C') = \min\{\text{start}(C_1), \text{start}(C_2)\}$, $\text{end}(C') = \max\{\text{end}(C_1), \text{end}(C_2)\}$, and $C'(X) = C_1(X) \uplus C_2(X)$ for every $X \in \mathbf{X}$. Further, we define the *projection over* $L \subseteq \mathbf{X}$ of a complex event C , denoted by $\pi_L(C)$, as the complex event C' such that $\text{time}(C') = \text{time}(C)$ and $C'(X) = C(X)$ whenever $X \in L$, and $C'(X) = \emptyset$, otherwise. Finally, we denote by (i, j, μ_\emptyset) the complex event with the trivial function μ_\emptyset such that $\mu_\emptyset(X) = \emptyset$ for every $X \in \mathbf{X}$.

A new semantics for CEL. In this work, we use the *Complex Event Logic* (CEL) introduced in [20] and implemented in CORE [9] as our basic query language for CER. However, we revisit its semantics in order to extend it with aggregation. In particular, we use the same CEL syntax as in [20] which is given by the following grammar:

φ	:=	R	(event type selection)		$\varphi \text{ AS } X$	(variable binding)	
			$\varphi \text{ FILTER } X[P]$	(predicate filtering)		$\pi_L(\varphi)$	(variable projection)
			$\varphi \text{ OR } \varphi$	(disjunction)		$\varphi \text{ AND } \varphi$	(conjunction)
			$\varphi : \varphi$	(contiguous sequencing)		$\varphi ; \varphi$	(non-cont. sequencing)
			$\varphi \oplus$	(contiguous iteration)		$\varphi +$	(non-cont. iteration)

where R is an event type, $X \in \mathbf{X}$ is a variable, $P \in \mathbf{P}$ is a predicate, and $L \subseteq \mathbf{X}$ is a finite set of variables. Similar to [20, 9], we define the semantics of a CEL formula φ over a stream $\mathcal{S} = e_1 e_2 \dots e_n$, recursively, as a set of complex events over \mathcal{S} . The main difference is the notion of complex events, that now contains events instead of positions. In Figure 1, we define the semantics of each CEL operator like in [9, 20]. Given a formula φ , the semantics $\llbracket \varphi \rrbracket(\mathcal{S})$ defines a set of complex events. Notice that $\llbracket \varphi \rrbracket(\mathcal{S})$ has a *set-semantics* and, instead, complex events store bags of events.

Next, we present an example for showing how to use the syntax and semantics of CEL to extract complex events from streams (see also Section 5). In this example, we use conjunction and disjunction in filtering that one can read them as:

$$\begin{aligned}
 \varphi \text{ FILTER } (X[P_1] \wedge Y[P_2]) &\equiv (\varphi \text{ FILTER } X[P_1]) \text{ FILTER } Y[P_2] \\
 \varphi \text{ FILTER } (X[P_1] \vee Y[P_2]) &\equiv (\varphi \text{ FILTER } X[P_1]) \text{ OR } (\varphi \text{ FILTER } Y[P_2])
 \end{aligned}$$

for every CEL formula φ , variables $X, Y \in \mathbf{X}$, and predicates P_1, P_2 .

► **Example 2** (from [9]). Consider the stream $\mathcal{S}_{\text{Stocks}}$ from Example 1. Suppose that we are interested in all triples of **SELL** events where the first is a sale of Microsoft over US\$100, the second is a sale of Intel (of any price), and the third is a sale of Amazon below US\$2000. Then, we can specify this pattern by the following CEL formula:

$$\begin{aligned} \varphi_2 = & (\text{SELL AS msft} ; \text{SELL AS intel} ; \text{SELL AS amzn}) \\ & \text{FILTER} (\text{msft}[\text{name} = \text{"MSFT"}] \wedge \text{msft}[\text{price} > 100] \wedge \text{intel}[\text{name} = \text{"INTC"}] \\ & \wedge \text{amzn}[\text{name} = \text{"AMZN"}] \wedge \text{amzn}[\text{price} < 2000]). \end{aligned}$$

Intuitively, the expression $(\text{SELL AS msft} ; \text{SELL AS intel} ; \text{SELL AS amzn})$ specifies that we want to see three **SELL** events that we named by the variables *msft*, *intel* and *amzn*, respectively. The semicolon operator ($;$) indicates non-contiguous sequencing among them, namely, there could be more events between them. Finally, the **FILTER** clause requires the data of the events to satisfy the necessary restrictions.

As we already mentioned, in this work we change the semantics used in [20, 9] to use events instead of positions, called it here *event-based* semantics. The old semantics of CEL, called *position-based* semantics, was obtained by outputting complex events of the form $C = (i, j, \mu_{\text{index}})$ where μ_{index} a mapping such that $\mu_{\text{index}} : \mathbf{X} \mapsto \mathcal{P}(\mathbb{N})$. Namely, μ_{index} contains the positions of the events that participates in C . One can easily see that the event-based semantics of CEL is equivalent to the position-based semantics where i -th position must be replaced by the i -th event of the stream. In other words, we have the following equivalence.

► **Theorem 3.** *The (old) position-based semantics of CEL is equivalent to the (new) event-based semantics of CEL.*

A formalization of the previous statement can be found in [8].

A new operator for projecting attributes. An advantage of providing a new event-based semantics is that one can extend CEL with new operators, such as aggregation, which we will discuss in the next chapters. More interestingly, we can introduce new natural operators for managing complex events that cannot be defined using the old semantics in [20]. In this work, we use events instead of positions, which makes it possible to extend the CEL syntax with the *attribute-projection operator*, an operator for projecting tuples within complex events. Formally, we extend the syntax of CEL formulas with the following operator:

$$\varphi := \pi_{X(\mathbf{a}_1, \dots, \mathbf{a}_k)}(\varphi) \quad (\text{tuple projection})$$

where φ is an arbitrary CEL formula, X is a variable in \mathbf{X} , and $\mathbf{a}_1, \dots, \mathbf{a}_k$ is a list of attributes in \mathbf{A} . Intuitively, it means that it will only consider the attributes in $\mathbf{a}_1, \dots, \mathbf{a}_k$ in the events that are in the variable X .

We define the formal semantics of the attribute-projection operator $\pi_{X(\mathbf{a})}$ recursively as follows. For a list of attributes $\mathbf{a}_1, \dots, \mathbf{a}_k$ and an event e , we define $\pi_{\mathbf{a}_1, \dots, \mathbf{a}_k}(e)$ as the new event e' such that $\text{Att}(e') = \text{Att}(e) \cap \{\mathbf{a}_1, \dots, \mathbf{a}_k\}$, $e'(\text{time}) = e(\text{time})$, and $e'(\mathbf{a}_i) = e(\mathbf{a}_i)$ whenever $\mathbf{a}_i \in \text{Att}(e')$. Let $\mathcal{S} = e_1 e_2 \dots e_n$ where each e_i is an event. Then:

$$\begin{aligned} \llbracket \pi_{X(\mathbf{a}_1, \dots, \mathbf{a}_k)}(\varphi) \rrbracket(\mathcal{S}) = & \{ C \mid \exists C' \in \llbracket \varphi \rrbracket(\mathcal{S}). \text{time}(C) = \text{time}(C') \wedge \forall Y \neq X. C(Y) = C'(Y) \\ & \wedge C(X) = \{ \{ \pi_{\mathbf{a}_1, \dots, \mathbf{a}_k}(e) \mid e \in C'(X) \} \} \} \end{aligned}$$

Intuitively, given a complex event $C' \in \llbracket \varphi \rrbracket(\mathcal{S})$ with $C'(X) = \{ \{ e_1, \dots, e_l \} \}$, the projection formula above creates a new complex event C which has the same interval than C' and events in variables $Y \neq X$, but it redefines events in X as $C(X) = \{ \{ \pi_{\mathbf{a}_1, \dots, \mathbf{a}_k}(e_1), \dots, \pi_{\mathbf{a}_1, \dots, \mathbf{a}_k}(e_l) \} \}$.

► **Example 4.** We consider again the setting as in Examples 1 and 2. Now we are interested in getting the price of the sale of Intel, subject to the same constraints. Then, we can write this query by using tuple projection as follows:

$$\begin{aligned} \varphi_4 = & \pi_{\text{intel}(\text{price})}(\text{SELL AS msft}; \text{SELL AS intel}; \text{SELL AS amzn}) \\ & \text{FILTER}(\text{msft}[\text{name} = \text{"MSFT"}] \wedge \text{msft}[\text{price} > 100] \wedge \text{intel}[\text{name} = \text{"INTC"}] \\ & \wedge \text{amzn}[\text{name} = \text{"AMZN"}] \wedge \text{amzn}[\text{price} < 2000]). \end{aligned}$$

4 Modelling aggregate functions in CER

Before introducing our logic for aggregation, we present a framework to model aggregate functions in CER based on monoids. Our goal is to present a logic that is as general as possible, encompassing most of the aggregations queries used in practice, such as `sum`, `max`, `count`, or `range`. In the following, we recall the definitions of monoids and aggregate functions. We end by stating our main assumptions regarding aggregation in CER.

Monoids. A *monoid* is an algebraic structure (M, \oplus, \mathbb{O}) where (M, \oplus) forms a semigroup and $\mathbb{O} \in M$ is an identity element over \oplus . Similar to semigroups, we will further assume that \oplus is commutative. For example, the natural numbers with addition $(\mathbb{N}, +, 0)$ or with product $(\mathbb{N}, \times, 1)$ form commutative monoids. Other examples are $(\mathbb{N} \cup \{\infty\}, \min, \infty)$ with `min` and $(\mathbb{N}, \max, 0)$ with `max`. Given a commutative monoid (M, \oplus, \mathbb{O}) , a finite bag $A = \{\{a_1, \dots, a_n\}\} \subseteq M$ and a function $f : M \rightarrow M$ we define the operator: $\bigoplus_{a \in A} f(a) = f(a_1) \oplus \dots \oplus f(a_n)$, namely, the generalization of \oplus from a binary operator to a set of elements. In particular, if $A = \emptyset$, we define $\bigoplus_{a \in A} f(a) = \mathbb{O}$. In the sequel, we will use (M, \oplus, \mathbb{O}) or $(M, \otimes, \mathbb{1})$ for denoting arbitrary commutative monoid over some set M .

Aggregate functions. In this work, we consider the most general definition of an aggregate function that can be defined through a monoid (see [22]). Specifically, an *aggregate function* is a function from a bag of values to values, formally, $f : \mathcal{P}_{bags}(\mathbf{D}) \rightarrow \mathbf{D}$ for some set of values M . An aggregate function f is *self-decomposable* if there exists a commutative monoid¹ $(\mathbf{D}, \oplus, \mathbb{O})$ such that $f(X \uplus Y) = f(X) \oplus f(Y)$ for every disjoint bags $X, Y \subseteq \mathbf{D}$. Examples of functions that are self-decomposable are `sum`, `max`, `min`, and `count`. For instance, `sum`($X \uplus Y$) is equal to 0 if $X \uplus Y = \emptyset$, to x if $X \uplus Y = \{\{x\}\}$, and to `sum`(X) + `sum`(Y), otherwise. Similarly, `count`($X \uplus Y$) is equal to 0 if $X \uplus Y = \emptyset$, to 1 if $X \uplus Y = \{\{x\}\}$, and to `count`(X) + `count`(Y) otherwise. Finally, `min`($X \uplus Y$) is equal to ∞ if $X \uplus Y = \emptyset$, to x if $X \uplus Y = \{\{x\}\}$, and `min`(`min`(X), `min`(Y)) otherwise.

Unfortunately, in practice not all aggregate function are self-decomposable; however, most of them can still be decomposed before we apply a simple operation. Formally, an aggregate function f is called *decomposable* if there exist a function $g : \mathbf{D} \rightarrow \mathbf{D}$ and a self-decomposable aggregate function h such that $f = g \circ h$. Furthermore, we assume that g can be computed in constant time (i.e., in the RAM model). This last condition is necessary, as we want g to perform a simple operation (i.e., constant time), as the final step after h has completed the aggregation, and not to be powerful enough to perform the aggregation itself.

Every self-decomposable functions is also decomposable (i.e., where g is the identity function). Other examples of aggregate functions that are decomposable (but not self-decomposable) are `avg` and `range`. For instance, one can define `avg` as `avg`(X) = $g(h(X))$

¹ In [22], the definition of self-decomposable is not given in terms of a monoid. However, one can easily see that the definition in [22] implies the existence of a monoid.

where $h(\{\!\{x\}\!\}) = (x, 1)$ and $h(X \uplus Y) = h(X) + h(Y)$ where $+$ is the standard pointwise sum of pairs, and $g((s, c)) = s/c$. Another example is the *range* which can be defined as $\text{range}(X) = g(h(X))$ such that $h(x) = (x, x)$, $h(X \uplus Y) = (\max(X \uplus Y), \min(X \uplus Y))$, and $g((s, c)) = s - c$. In both cases, one can check that h is a self-decomposable function and g can be computed in constant time in the RAM model.

Notice that, although self-decomposable functions can be decomposed through a monoid, they are not entirely specified by it (e.g., *count*). Nevertheless, as the following lemma shows, we can restrict to monoids by first mapping the values to the underlying monoid.

► **Lemma 5.** *f is self-decomposable if, and only if, there exist a commutative monoid (M, \oplus, \mathbb{O}) and a function $f' : \mathbf{D} \rightarrow M$ such that $f(X) = \bigoplus_{a \in X} f'(a)$ for every bag X .*

Given the previous lemma, we say that f is *strong* self-decomposable if there exists a pair (M, f') such that M is a commutative monoid and f' is the identity function. In other words, f can be directly defined by a commutative monoid. The functions that are strong self-decomposable are *sum*, *min*, and *max*. On the other hand, *count* needs to map each value to 1 before adding them.

For the sake of simplification, in the following we assume that all aggregate functions are *strong self-decomposable*. In other words, we can directly define the semantics of the aggregate functions through a commutative monoids. We can make this assumption since the functions f' and g (i.e., of decomposable aggregate functions) can be computed in constant time when the each data item is read or after the aggregation is done, like, for example, the function f' to map a single element to 1 (e.g., *count*), and the final function g to calculate the difference between two elements (e.g., *range*), or divide one by the other (e.g., *avg*). This assumption considerably simplifies our setting, allowing us to focus on the most relevant details of aggregation without discarding relevant aggregate functions from practice.

5 Aggregation complex event logic

In this section, we present our proposal to extend CEL with aggregation. Specifically, we demonstrate how to extend CEL with an operation for aggregations, building upon previous work experience. We provide examples of how this new operator is sufficient to model most queries used in earlier works. We start by introducing the algebraic structure for modelling aggregate functions, which we then use to define the aggregation operator in CEL.

The algebraic structure for aggregation. Recall that \mathbf{A} and \mathbf{D} are our fix sets of attributes names and data values, respectively. We fix an algebraic structure:

$$\mathcal{D} = (\mathbf{D}, \oplus_1, \dots, \oplus_k, \mathbb{O}_1, \dots, \mathbb{O}_k) \quad (\dagger)$$

over \mathbf{D} such that each $(\mathbf{D}, \oplus_i, \mathbb{O}_i)$ forms a commutative monoid for every $i \in [k]$. For example, $(\mathbb{N} \cup \{\infty\}, +, \min, \max, \mathbb{O}, \infty, \mathbb{O})$ forms such an algebraic structure where we assume that $n + \infty = \infty$ for every $n \in \mathbb{N}$. Without loss of generality, we assume that $\text{NULL} \in \mathbf{D}$ and $a \oplus_i \text{NULL} = \text{NULL}$ for every $a \in \mathbf{D}$ and $i \in [k]$ (if this is not the case, one can extend \mathcal{D} with a fresh value NULL). The purpose of NULL is to define the aggregation operator over events e where an attribute is not defined (i.e., $e(\mathbf{a}) = \text{NULL}$ for some $\mathbf{a} \in \mathbf{A}$).

The single-attribute aggregation operator. Our goal is to extend the syntax of CEL with an *aggregation operator* that aggregates values in a single event. For the sake of presentation, we will first introduce the operation for a single attribute to then show how to extend it to multiple attributes.

15:10 A Formal Query Language for Aggregation in CER

Specifically, we extend the CEL syntax with the *single-attribute aggregation operator*, called *Aggregation CEL* (ACEL), as follows:

$$\varphi := \text{Agg}_{Y[\mathbf{b} \leftarrow \otimes X(\mathbf{a})]}(\varphi)$$

where φ is an arbitrary CEL formula, X and Y are variables in \mathbf{X} , \mathbf{a} and \mathbf{b} are attributes names in \mathbf{A} , and \otimes is a binary operator from \mathcal{D} where $(\mathbf{D}, \otimes, \mathbf{1})$ forms a monoid. Intuitively, the syntax $Y[\mathbf{b} \leftarrow \otimes X(\mathbf{a})]$ means that the aggregation will create a new event e that will be stored at the variable Y , such that e will have a single attribute \mathbf{b} that stores the \otimes -aggregation of the \mathbf{a} -attribute of events in X . We define the formal semantics of the single aggregation operator Agg recursively as follows. Let $\mathcal{S} = e_1 e_2 \dots e_n$ be a stream. Then:

$$\begin{aligned} \llbracket \text{Agg}_{Y[\mathbf{b} \leftarrow \otimes X(\mathbf{a})]}(\varphi) \rrbracket(\mathcal{S}) = \\ \{ C \mid \exists C' \in \llbracket \varphi \rrbracket(\mathcal{S}). \text{time}(C) = \text{time}(C') \wedge \forall Z \neq Y. C(Z) = C'(Z) \\ \wedge C(Y) = C'(Y) \uplus \{ e \mid e = [\mathbf{b} \mapsto \bigotimes_{e' \in C'(X)} e'(\mathbf{a})] \wedge e(\text{time}) = \text{end}(C') \} \} \end{aligned}$$

Intuitively, given a complex event $C' \in \llbracket \varphi \rrbracket(\mathcal{S})$ with $C'(X) = \{e_1, \dots, e_\ell\}$, the aggregation formula above creates a new complex event C which has the same interval and events than C' except that Y has an additional event e (i.e., $C(Y) = C'(Y) \uplus \{e\}$) and $e(\mathbf{b}) = e_1(\mathbf{a}) \otimes \dots \otimes e_\ell(\mathbf{a})$. In case that $C'(X) = \emptyset$, it will return the identity $\mathbf{1}$ of \otimes . Further, the new event e has $e(\text{time}) = \text{end}(C')$, namely, the last time inside C' . Notice that the event e is always well-defined, since we assume that, if \mathbf{a} is not defined for some e_i , it holds that $e_i(\mathbf{a}) = \text{NULL}$ and then $e(\mathbf{a}) = \text{NULL}$.

In the following, we use some special notation for useful functions like sum , max , min instead of \oplus . For example, if we use the sum function, we write $\text{Agg}_{Y[\mathbf{b} \leftarrow \text{sum}(X(\mathbf{a}))]}(\varphi)$. Further, recall that, although we use commutative monoids to define the semantics, this semantics can easily be generalized to *decomposable aggregate functions* like count , avg , or range . Therefore, without loss of generality, we also write, for example, $\text{Agg}_{Y[\mathbf{b} \leftarrow \text{count}(X(\mathbf{a}))]}(\varphi)$ although strictly speaking count is not a strong self-decomposable aggregate function.

► **Example 6.** Consider again the setting as in Examples 1 and 2. Now we are interested in getting the maximum price in a sequence of Intel sales between a Microsoft and an Amazon sale under the same constrains and store it in an attribute MAX in a variable M . Then, we can specify this query by using the aggregation operator as:

$$\begin{aligned} \varphi_6 = \text{Agg}_{M[\text{MAX} \leftarrow \text{max}(\text{intel}(\text{price}))]}(\\ \text{[SELL AS msft; (SELL AS intel)+; SELL AS amzn]} \\ \text{FILTER [msft[name = "MSFT"]} \wedge \text{msft[price} > 100] \wedge \text{intel[name = "INTC"]} \\ \wedge \text{amzn[name = "AMZN"]} \wedge \text{amzn[price} < 2000]\text{]}) \end{aligned}$$

As the reader can check from the semantics of Agg , the max value in the intel sequence will be stored in a new event at the variable M .

► **Example 7.** For a second example, suppose that now we are interested in getting the length of a sequence (BUY OR SELL) in a trend between prices 100 and 2000 and store it in an attribute QNT in a variable Q . Further, we want to also check that this length is greater than 5. Then, we can express the query as:

$$\begin{aligned} \varphi_7 = \left[\text{Agg}_{M[\text{QNT} \leftarrow \text{count}(m(\text{price}))]}(\\ \text{[(BUY OR SELL) AS } l; \text{(BUY OR SELL) + AS } m; \text{(BUY OR SELL) AS } h] \\ \text{FILTER [l[price} < 100] \wedge m[\text{price} \geq 100] \\ \wedge m[\text{price} \leq 2000] \wedge h[\text{price} > 2000]\text{]}) \right] \text{FILTER } M[\text{QNT} > 5] \end{aligned}$$

Notice that, although in the previous example the aggregation was applied over a simple CEL formula (i.e., at the topmost level), in ACEL all operators, including the aggregation operator, can be freely composed. In particular, we can apply a filter (e.g. `FILTER M.QNT > 5`) over an aggregation that was computed.

The multi-attribute aggregation operator. We present now the generalization of the aggregation operator to multiple attributes. Although this generalized version is more verbose, it is needed in practice for aggregating different sets simultaneously in different attributes. We extend the syntax of CEL with the *(multi-attribute) aggregation operator*:

$$\varphi := \text{Agg}_{Y[\mathbf{b}_1 \leftarrow \otimes_1 X_1(\mathbf{a}_1), \dots, \mathbf{b}_\ell \leftarrow \otimes_\ell X_\ell(\mathbf{a}_\ell)]}(\varphi)$$

where φ is an arbitrary CEL formula, X_1, \dots, X_ℓ and Y are variables in \mathbf{X} , $\mathbf{a}_1, \dots, \mathbf{a}_\ell$ and $\mathbf{b}_1, \dots, \mathbf{b}_\ell$ are attributes names in \mathbf{A} , and $\otimes_1, \dots, \otimes_\ell$ are binary operators from \mathcal{D} . Intuitively, the syntax $Y[\mathbf{b}_1 \leftarrow \otimes_1 X_1(\mathbf{a}_1), \dots, \mathbf{b}_\ell \leftarrow \otimes_\ell X_\ell(\mathbf{a}_\ell)]$ states that the aggregation will create a new event e with attributes $\mathbf{b}_1, \dots, \mathbf{b}_\ell$ that will be stored at the variable Y , such that each attribute \mathbf{b}_i will store the \otimes_i -aggregation of the \mathbf{a}_i -attribute of events in X_i .

Given a stream \mathcal{S} , the formal semantics of the generalization of **Agg** is given as follows:

$$\begin{aligned} \llbracket \text{Agg}_{Y[\mathbf{b}_1 \leftarrow \otimes_1 X_1(\mathbf{a}_1), \dots, \mathbf{b}_\ell \leftarrow \otimes_\ell X_\ell(\mathbf{a}_\ell)]}(\varphi) \rrbracket(\mathcal{S}) = \\ \{C \mid \exists C' \in \llbracket \varphi \rrbracket(\mathcal{S}). \text{time}(C) = \text{time}(C') \wedge \forall Z \neq Y. C(Z) = C'(Z) \wedge C(Y) = C'(Y) \uplus \\ \{e \mid e = [\mathbf{b}_1 \mapsto \bigotimes_1^{e' \in C'(X_1)} e'(\mathbf{a}_1), \dots, \mathbf{b}_\ell \mapsto \bigotimes_\ell^{e' \in C'(X_\ell)} e'(\mathbf{a}_\ell)] \wedge e(\text{time}) = \text{end}(C')\}\} \end{aligned}$$

Intuitively, the general version of **Agg** allows to define several attributes $\mathbf{b}_1, \dots, \mathbf{b}_\ell$ by performing aggregation over the attributes $\mathbf{a}_1, \dots, \mathbf{a}_\ell$, respectively. The idea is similar to the single-attribute aggregation operator but with several attributes $\mathbf{b}_1, \dots, \mathbf{b}_\ell$ at once.

In Appendix A (see also [8]), we show how to use ACEL to specify some examples from previous academic proposals and real-life systems. In particular, we present examples from the literature where the multi-attribute aggregation operator is required. Similar to the simple-attribute aggregation operator, in ACEL, one can freely compose all operators, including this new aggregation operator. We conclude this section by discussing several relevant design decisions we made in defining the aggregation operator in CEL.

Why this semantics for aggregation in CEL? There are multiple ways to define a semantics for aggregation in CEL; however, our proposal for CEL and ACEL has some crucial design decisions that need to be justified. Specifically, we propose a semantics that (1) outputs a set of complex events (i.e., no repetitions), (2) each complex event contains bags of events, and (3) each event has a timestamp that defines the time when it arrives or was created. Indeed, we could consider other alternatives, such as a semantics that outputs bags of complex events, sets inside a complex event, or events without a timestamp, or any combination of these alternatives. In the following, we discuss why we proposed a semantics based on (1), (2), and (3), and what the consequences are of taking other alternatives.

For (1), if we choose a semantics based on bags of complex events, independent of the other choices, we will get a semantics that outputs duplicated results depending on how we specify the query. For example, assume a bag-based semantics and a user writes the query:

$$\varphi_1 = \pi_X(A \text{ AS } X; B+; A \text{ AS } X)$$

15:12 A Formal Query Language for Aggregation in CER

over a stream $\mathcal{S}_1 = A_1 B_2 B_3 A_4$ where A and B are the types of the events (i.e., the data in the attributes is not relevant). If we evaluate φ_1 over \mathcal{S}_1 with a bag-based semantics, we will have the same result (1, 4) multiple times (potentially exponentially many times) depending on how many B were captured for each result. Instead, a set-based semantics ensures that each complex event appears only once, no matter how the query is specified.

For (2), if we choose that each variable inside a complex event maps to a set of events, instead of a bag of events, we could get some answers that do not consider some results as they will be taken as repeated elements. For example, if we consider a query:

$$\varphi_2 = \text{Agg}_{Y(\mathbf{b} \leftarrow \text{sum}(X(\mathbf{a})))} [(\text{Agg}_{X(\mathbf{a} \leftarrow \text{sum}(A(\text{value})))} [B : A \oplus]) \oplus]$$

and a stream $\mathcal{S}_2 = B_1 A_2[\mathbf{a} : 3] A_3[\mathbf{a} : 5] B_4 A_5[\mathbf{a} : 2] A_6[\mathbf{a} : 4] A_7[\mathbf{a} : 2]$, first we will get two matches (one from $B_1 A_2 A_3$ and the other from $B_4 A_5 A_6 A_7$), then we will make the aggregation in each of them, but the result of each aggregation is the same (i.e., 8), they come from different values and they will not be saved as two different values, so finally the outer aggregation will be applied over one element and not two.

Finally, for (3), if we consider that each event that arrives or is created does not have a timestamp (i.e., a mark of origin), then we can still lose some information during aggregation (even if we used bags inside complex events to store events). For example, consider the query

$$\varphi_3 = (\text{Agg}_{Y(\mathbf{b} \leftarrow \text{sum}(X(\mathbf{a})))} (X \oplus); \psi_1) \text{ OR } (\psi_2; \text{Agg}_{Y(\mathbf{b} \leftarrow \text{sum}(W(\mathbf{a})))} (W \oplus))$$

for some subformulas ψ_1 and ψ_2 . For formula φ_3 over some stream, the results of the aggregation in the left and right parts of the disjunction (i.e., OR) could be equal, and it will be impossible to differentiate which part the aggregation is coming from when we apply the OR operator. Instead, by assuming that each event has a timestamp (even those created through aggregation), for φ_3 , there will be at least two outputs, and we can differentiate the position where the aggregation was performed.

It is important to note that another semantics for CEL and the aggregation operator is possible, and our argument above does not invalidate them. However, there could be consequences for the query language with unintuitive behavior for the users. In this work, we have chosen to focus on a semantics based on (1), (2), and (3), studying its properties, and reserve the study of other variants of the logic for future work.

Expressive power of ACEL. When introducing a new operator, such as aggregation, one wants it to model only what it is meant to; however, combining it with other operators can lead to unexpected properties that can be expressed. In particular, combining the aggregate operator with filters is very powerful, as it allows one to check equivalence between events. For example, consider the following query with aggregation and filtering:

$$\pi_{X,Y}(\text{Agg}_{Z(\mathbf{b}_1 \leftarrow \text{sum}(X(\mathbf{a})), \mathbf{b}_2 \leftarrow \text{sum}(Y(\mathbf{a})))} (R \text{ AS } X; T \text{ AS } Y) \text{ FILTER } [Z(\mathbf{b}_1 = \mathbf{b}_2)])$$

Intuitively, the previous query checks that an event of type T (naming it Y) happens after an event of type R (naming it X) and sum the values of attribute \mathbf{a} in both events separately, saving those values in attributes \mathbf{b}_1 and \mathbf{b}_2 of variable Z in one event. Then, the query filters it by checking if the values of attributes \mathbf{b}_1 and \mathbf{b}_2 are equal, as they correspond to the same event. Finally, it projects variables X and Y . One can see that the query correlates events from different types only using aggregation and filter operators over single events, as it generates two new variables in an event with the aggregation, that come from two different events, and then it uses filter which only is applied to a single event.

This unexpected behaviour of combining aggregation with filtering is an interesting side effect that could lead to a better understanding of aggregation in CER. Note that observing this interaction between aggregates, filters, and other operators will not be possible without having a concrete and formal semantics of the query language.

6 Automata model for aggregation in CER

Here we present an automata model for aggregation that extends complex event automata with registers similar to the model of cost register automata [4]. We start by recalling the model of complex event automata (CEA) to provide then the necessary definitions for introducing our new automata model for aggregation.

CEA. A *Complex Event Automaton (CEA)* [20, 9] is a tuple $\mathcal{A} = (Q, \Delta, q_0, F)$ where Q is a finite set of states, $\Delta \subseteq Q \times \mathbf{P} \times \mathcal{P}(\mathbf{X}) \times Q$ is a finite transition relation, $q_0 \in Q$ is the initial state, and $F \subseteq Q$ is the set of final states. A *run of \mathcal{A} over stream $\mathcal{S} = e_1 \dots e_n$ from positions i to j* is a sequence of transition:

$$\rho := q_i \xrightarrow{P_i/L_i} q_{i+1} \xrightarrow{P_{i+1}/L_{i+1}} \dots \xrightarrow{P_j/L_j} q_{j+1}$$

such that $q_i = q_0$ is the initial state of \mathcal{A} and for every $k \in [i..j]$ it holds that $(q_k, P_k, L_k, q_{k+1}) \in \Delta$ and $e_k \models P_k$. A run ρ is *accepting* if $q_{j+1} \in F$. An accepting run ρ of \mathcal{A} over \mathcal{S} from i to j naturally defines the complex event $C_\rho := (i, j, \mu_\rho)$ such that $\mu_\rho(X) = \{e_k \mid i \leq k \leq j \wedge X \in L_k\}$ for every $X \in \mathbf{X}$. If position i and j are clear from the context, we say that ρ is a run of \mathcal{A} over \mathcal{S} . Finally, we define the semantics of \mathcal{A} over a stream \mathcal{S} as: $\llbracket \mathcal{A} \rrbracket(\mathcal{S}) := \{C_\rho \mid \rho \text{ is an accepting run of } \mathcal{A} \text{ over } \mathcal{S}\}$.

CEA was crucial to capture the expressiveness of CEL, compile queries from CEL into CEA, and efficiently evaluate them. Unfortunately, one can easily notice that CEA are not useful for our ACEL semantics, since there is no way to *remember* the values of the attributes that we have seen to do the aggregation. In other words, there is no mechanism for aggregating values and producing new events as in the new semantics of ACEL. We will show how to overcome these shortcomings in the next definitions.

Expressions. Recall that \mathbf{A} is a fixed set of attributes and \mathbf{D} a fix set of data values. Further, recall that in Section 5 we fix an algebraic structure of the form (\dagger) over \mathbf{D} such that each $(\mathbf{D}, \oplus_i, \mathcal{O}_i)$ forms a commutative monoid for every $i \in [k]$. We define an $(\mathcal{D}, \mathbf{A})$ -*expression e* (or just *expression*) as a syntactical formula over \mathcal{D} and \mathbf{A} generated by the grammar:

$$\alpha := d \mid \mathbf{a} \mid \alpha \oplus_i \alpha \quad i \in [k]$$

where $d \in \mathbf{D}$ and $\mathbf{a} \in \mathbf{A}$. We define the *set of all $(\mathcal{D}, \mathbf{A})$ -expressions* by $\text{Expr}(\mathcal{D}, \mathbf{A})$. For an expression $\alpha \in \text{Expr}(\mathcal{D}, \mathbf{A})$ we denote by $\text{Att}(\alpha)$ the set of all attributes in α . Given an event $e : \mathbf{A} \mapsto \mathbf{D}$ and an expression α such that $\text{Att}(\alpha) \subseteq \text{Att}(e)$, we define the semantics of α over e , denoted by $\llbracket \alpha \rrbracket(e)$, as the value in \mathbf{D} of evaluating α by replacing every $\mathbf{a} \in \mathbf{A}$ by $e(\mathbf{a})$.

► **Example 8.** Let $\mathcal{D} = (\mathbf{D}, \min, \max, +, \infty, 0, 0)$, the expressions $\alpha = \mathbf{a} + \mathbf{b}$ and $\beta = \min(\mathbf{a}, \mathbf{b}) + \max(\mathbf{b}, \mathbf{c})$, and an event e where $e(\mathbf{a}) = 4$, $e(\mathbf{b}) = 2$ and $e(\mathbf{c}) = 5$. Then the result of each expression α and β over the event e are $\llbracket \alpha \rrbracket(e) = 6$ and $\llbracket \beta \rrbracket(e) = 7$, respectively.

Assignments. An $(\mathcal{D}, \mathbf{A})$ -assignment (or just *assignment* when \mathcal{D} and \mathbf{A} are clear from the context) is a program that assigns attributes in \mathbf{A} to expressions in $\text{Expr}(\mathcal{D}, \mathbf{A})$. Formally, an *assignment* is defined as a mapping $\sigma : \mathbf{A} \mapsto \text{Expr}(\mathcal{D}, \mathbf{A})$. Similar to expressions, we define $\text{Att}_{\text{in}}(\sigma) = \bigcup_{\mathbf{a} \in \text{dom}(\sigma)} \text{Att}(\sigma(\mathbf{a}))$ to be all the attributes used in expressions of σ , and $\text{Att}_{\text{out}}(\sigma) = \text{dom}(\sigma)$ all the attributes that are assigned. Given an event $e : \mathbf{A} \rightarrow \mathbf{D}$ and an $(\mathcal{D}, \mathbf{A})$ -assignment σ such that $\text{Att}_{\text{in}}(\sigma) \subseteq \text{Att}(e)$, the semantics of an assignment σ over e is an event $e' := \llbracket \sigma \rrbracket(e) : \mathbf{A} \mapsto \mathbf{D}$ such that $\text{Att}(e') = \text{Att}_{\text{out}}(\sigma)$ and $e'(\mathbf{a}) = \llbracket \sigma(\mathbf{a}) \rrbracket(e)$ for every $\mathbf{a} \in \text{Att}_{\text{out}}(\sigma)$. In other words, $\llbracket \sigma \rrbracket(e)$ is the result of applying the assignment σ with the values in the event e . We denote the set of all $(\mathcal{D}, \mathbf{A})$ -assignments by $\text{Asg}(\mathcal{D}, \mathbf{A})$.

► **Example 9.** Consider again the setting of Example 8 and the assignment σ defined as: $\sigma : \mathbf{a} \leftarrow \max(\mathbf{a} + \mathbf{b}, \mathbf{c})$. Here, we think σ as a program where the left side of \leftarrow is updated with the right side, namely, $\sigma(\mathbf{a}) = \max(\mathbf{a} + \mathbf{b}, \mathbf{c})$. Then, $\llbracket \sigma \rrbracket(e)(\mathbf{a}) = 6$.

Finally, we recall the notion of renamings (Section 3) and define updates of events that will be useful for our automata model. So, remember that a *renaming* r is defined as $r : \mathbf{A} \mapsto \mathbf{A}$, which maps each attribute to a new attribute. We can note that a renaming is also a particular case of an assignment $r : \mathbf{A} \mapsto \text{Expr}(\mathcal{D}, \mathbf{A})$ such that $r(\mathbf{a}) \in \mathbf{A}$. Also, recall that we define by Ren the set of all tuple renaming over \mathbf{A} . Given events e and e' , we define the *update of e' by e* , denoted by $e \gg e'$, as a new event such that $\text{Att}(e \gg e') = \text{Att}(e) \cup \text{Att}(e')$ and $[e \gg e'](\mathbf{a}) = e(\mathbf{a})$ if $\mathbf{a} \in \text{Att}(e)$, and $[e \gg e'](\mathbf{a}) = e'(\mathbf{a})$ otherwise.

Aggregation Complex Event Automata. We are ready to define the model of CEA with aggregation. An *Aggregation Complex Event Automaton (ACEA)* is a tuple $\mathcal{A} = (Q, \Delta, q_0, F)$ where Q is a finite set of states, $q_0 \in Q$ is the initial state, $F \subseteq Q$ are the final states, and:

$$\Delta \subseteq Q \times \text{Asg}(\mathcal{D}, \mathbf{A}) \times \mathbf{P} \times \{\lambda : \mathbf{X} \mapsto \mathcal{P}_{\text{bags}}(\text{Ren})\} \times Q$$

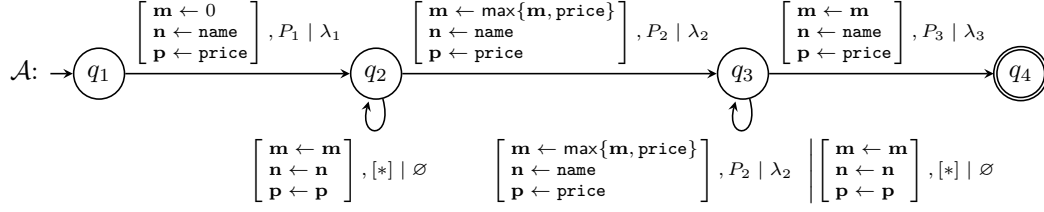
is a finite transition relation where $\{\lambda : \mathbf{X} \mapsto \mathcal{P}_{\text{bags}}(\text{Ren})\}$ is the set of all mappings λ that maps a variable X to a finite bag of renamings $\{\{r_1, \dots, r_k\}\}$. A transition $(p, \sigma, P, \lambda, q) \in \Delta$ specifies that \mathcal{A} can move from state p to state q after reading an event, by updating some internal registers with σ and checking a condition (over the registers) with P . Similar to CEA, λ will be in charge of creating the outputs of the complex event where the renamings $\lambda(X) = \{\{r_1, \dots, r_k\}\}$ will create k new tuples in the variable X coming from the values stored in the internal registers. We assume that the renamings in λ for a transition of the form $(p, \sigma, P, \lambda, q) \in \Delta$ are consistent with σ , namely, $\text{Att}_{\text{in}}(\lambda(X)) \subseteq \text{dom}(\sigma)$ for every $X \in \text{dom}(\lambda)$.

A pair (q, ν) is a configuration of \mathcal{A} where $q \in Q$ and $\nu : \mathbf{A} \mapsto \mathbf{D}$ is an event which represents the current values of the attributes. For the sake of simplification, in ACEA, we use attributes as “registers” for storing temporary values. For this reason, the configuration (q, ν) represents that the automata is in the state q and the registers $\text{dom}(\nu)$ (i.e., a subset of attributes) store the current computed values.

Let $\mathcal{S} = e_1 \dots e_n$ be a stream. A *run of \mathcal{A} over stream \mathcal{S} from positions i to j* is a sequence of configurations and transitions:

$$\rho := (q_i, \nu_i) \xrightarrow{\sigma_i, P_i / \lambda_i} (q_{i+1}, \nu_{i+1}) \xrightarrow{\sigma_{i+1}, P_{i+1} / \lambda_{i+1}} \dots \xrightarrow{\sigma_j, P_j / \lambda_j} (q_{j+1}, \nu_{j+1}) \quad (\ddagger)$$

such that q_i is the initial state q_0 , ν_i is the empty event (i.e., $\text{dom}(\nu_i) = \emptyset$), and for every $k \in [i..j]$, $(q_k, \sigma_k, P_k, \lambda_k, q_{k+1}) \in \Delta$, (q_{k+1}, ν_{k+1}) is a configuration of \mathcal{A} with $\nu_{k+1} := \llbracket \sigma_k \rrbracket(e_k \gg \nu_k)$, and $\nu_{k+1} \models P_k$. Also, it must hold that $\text{Att}_{\text{in}}(\sigma_k) \subseteq \text{dom}(e_k \gg \nu_k)$. Intuitively, the new values ν_{k+1} are produced by first updating ν_k by the new event e_k (i.e., $e_k \gg \nu_k$) and then operate $e_k \gg \nu_k$ by the assignment σ_k . After the new values ν_{k+1} are computed, we check if they satisfy the predicate P_k of the transition.



■ **Figure 2** An ACEA \mathcal{A} representing the given query in Example 6 where $P_1 := \mathbf{n} = \text{"MSFT"} \wedge \mathbf{p} > 100$, $P_2 := \mathbf{n} = \text{"INTC"}$ and $P_3 := \mathbf{n} = \text{"AMZN"} \wedge \mathbf{p} < 2000$. Further, $\lambda_1(\text{msft}) = \lambda_2(\text{intel}) = \lambda_3(\text{amzn}) = \{\{[\text{name} \mapsto \mathbf{n}, \text{price} \mapsto \mathbf{p}]]\}$, and $\lambda_3(M) = \{\{[\text{MAX} \mapsto \mathbf{m}]\}\}$.

Similar to CEA, a run ρ is *accepting* if $q_{j+1} \in F$. An accepting run ρ like (‡) of \mathcal{A} over S from i to j defines the complex event $C_\rho := (i, j, \mu_\rho)$ such that:

$$\mu_\rho(X) = \{\{e \mid k \in [i..j] \wedge r \in \lambda_k(X) \wedge e = \llbracket r \rrbracket(\nu_{k+1}) \wedge e(\text{time}) = k\}\}$$

for every $X \in \mathbf{X}$. Finally, we define the semantics of \mathcal{A} over a stream S as:

$$\llbracket \mathcal{A} \rrbracket(S) := \{C_\rho \mid \rho \text{ is an accepting run of } \mathcal{A} \text{ over } S\}.$$

► **Example 10.** Consider the ACEL query from Example 6. We can obtain the same result with the ACEA \mathcal{A} in Figure 2 where $P_1 := \mathbf{n} = \text{"MSFT"} \wedge \mathbf{p} > 100$, $P_2 := \mathbf{n} = \text{"INTC"}$ and $P_3 := \mathbf{n} = \text{"AMZN"} \wedge \mathbf{p} < 2000$. Further, $\lambda_1(\text{msft}) = \lambda_2(\text{intel}) = \lambda_3(\text{amzn}) = \{\{[\text{name} \mapsto \mathbf{n}, \text{price} \mapsto \mathbf{p}]]\}$, and $\lambda_3(M) = \{\{[\text{MAX} \mapsto \mathbf{m}]\}\}$. Intuitively, in the first transition, \mathcal{A} initializes a register \mathbf{m} (i.e., an attribute) with 0 and checks that the price and name attributes satisfy the predicate P_1 , by storing the name in \mathbf{n} and the price in \mathbf{p} . Then, it waits in q_2 maintaining the registers. After, in the next transition it updates the maximum value in \mathbf{m} with the new price and again checks that the name satisfies P_2 . Then, in the loop of q_3 , it repeats the same as the previous transition or it maintains the registers. Finally, in the last transition, it maintains the maximum value in \mathbf{m} and verifies that the attributes name and price satisfy P_3 . The mappings λ_1 , λ_2 , and λ_3 are in charge of outputting the events in variables msft , intel , and amzn , respectively. Further, λ_3 is in charge of producing the final event in variable M that contains the max-aggregate of Intel's prices.

A first natural question to answer is whether the expressive power of the new model ACEA includes queries defined by CEA or not. Similar to the question of ACEL versus CEL, CEA outputs complex events with positions, where our new model outputs complex events with events among other new features. Below, we show that a ACEA can define every CEA by mapping the positions of the stream to the events.

► **Theorem 11.** *ACEA can define the same as CEA over streams over a schema Σ , namely, for every CEA \mathcal{A} there exists an ACEA \mathcal{A}' such that $\llbracket \mathcal{A} \rrbracket(S) = \llbracket \mathcal{A}' \rrbracket(S)$ for every S over Σ .*

Equivalence with ACEL. The first main goal of this paper is to provide a query language with a formal and denotational semantics for performing aggregation in CER. The second main goal is to provide a computational model to compile queries from this language. In the following result, we show that ACEA is a computational model to fulfill this goal. Specifically, we show that every formula φ in ACEL can be compiled into a ACEA, proving that the model has all the feature to perform complex event extraction and aggregation.

► **Theorem 12.** *Let Σ be a schema. For every ACEL formula φ , there exists an ACEA \mathcal{A}_φ such that $\llbracket \varphi \rrbracket(\mathcal{S}) = \llbracket \mathcal{A}_\varphi \rrbracket(\mathcal{S})$ for every stream \mathcal{S} over Σ .*

We present the proof in [8]. It goes by induction over the formula showing how to compile each operator into an ACEA. The standard CEL operators follow a similar construction to that in [20] (except the AND operator), but here we also have to make sure that the registers are correctly maintained to produce the output.

It is important to remark that ACEA is a *hybrid automata model* that needs to perform computation (i.e., for the aggregation), check filters (i.e., for the predicates), and produce outputs (i.e., events). Therefore, in designing the model, we seek an equilibrium that fulfills all these goals and, simultaneously, is as simple as possible. This simplicity could be helpful for understanding its expressiveness and designing efficient evaluation algorithms.

Despite its simplicity, ACEA has more expressive power than ACEL, namely, there are queries that can be defined with ACEA but not with ACEL. For example, consider the monoid of natural numbers $(\mathbb{N}, +, 0)$ (i.e., *sum*). Given a stream $R[\mathbf{a} : 1]$ n -times $R[\mathbf{a} : 1]$, one can define an ACEA with one register that always doubles the current value and outputs its content in an event $[\mathbf{b} : 2^n]$. Intuitively, ACEL with $(\mathbb{N}, +, 0)$ cannot specify this query since it can only produce values that grow linearly with respect to the sum of all values in the stream. Even if we restrict the use of registers in a copyless manner (see *copyless cost register automaton* in [4]), one can design ACEA that cannot be specified by ACEL. For instance, given the previous stream, one can code an ACEA that produces a complex event with the sequence of events: $[\mathbf{b} : 1][\mathbf{b} : 2] \dots [\mathbf{b} : n]$ (i.e., by adding in a register the input values and outputting its content in each transition). Given that in ACEL, each value of an event can contribute to a finite number of new events, one cannot specify this in ACEL. Therefore, ACEA is more expressive than ACEL, and it is an interesting open problem to characterize ACEL in terms of restrictions over ACEA. We leave this problem for future work.

7 Future Work

This paper provides logical foundations for aggregation in CER but leaves several open problems for future work. One relevant open problem is to better understand the equivalence between ACEL and ACEA, namely, which ACEA can be written in ACEL. Another interesting question is to understand the expressive power of aggregation combined with filters and other operators (see Section 5). Finally, a crucial line of research for making ACEL work in practice is to study how to evaluate ACEL queries efficiently, finding enumeration algorithms that, given an ACEA and a stream, run with *constant update time and constant delay enumeration*.

References

- 1 Esper Enterprise Edition Website. <https://www.espertech.com/>, 2025. [Accessed 23-06-2025].
- 2 Asaf Adi and Opher Etzion. Amit-the situation manager. *The VLDB journal*, 13:177–203, 2004.
- 3 Alfred V Aho and John E Hopcroft. *The design and analysis of computer algorithms*. Pearson Education India, 1974.
- 4 Rajeev Alur, Loris D’Antoni, Jyotirmoy V. Deshmukh, Mukund Raghothaman, and Yifei Yuan. Regular functions and cost register automata. In *LICS*, pages 13–22, 2013. doi:10.1109/LICS.2013.65.
- 5 Alexander Artikis, Alessandro Margara, Martín Ugarte, Stijn Vansummeren, and Matthias Weidlich. Complex event recognition languages: Tutorial. In *DEBS*, pages 7–10. ACM, 2017. doi:10.1145/3093742.3095106.

- 6 Mikołaj Bojańczyk. Transducers with origin information. In *ICALP*, pages 26–37, 2014.
- 7 Kyle Bossonney, Nicolás Buzeta, Vicente Calisto, Juan-Eduardo López, Cristian Riveros, and Stijn Vansummeren. CORE+: A complex event recognition engine in C++. In *SIGMOD demo*, pages 47–50, 2025. doi:10.1145/3722212.3725090.
- 8 Pierre Bourhis, Cristian Riveros, and Amaranta Salas. A formal query language and automata model for aggregation in complex event recognition. *CoRR*, abs/2601.00967, 2026.
- 9 Marco Bucci, Alejandro Grez, Andrés Quintana, Cristian Riveros, and Stijn Vansummeren. CORE: a complex event recognition engine. *VLDB*, 15(9):1951–1964, 2022. doi:10.14778/3538598.3538615.
- 10 Gianpaolo Cugola and Alessandro Margara. Raced: an adaptive middleware for complex event detection. In *ARM*, pages 1–6, 2009.
- 11 Gianpaolo Cugola and Alessandro Margara. TESLA: a formally defined event specification language. In *DEBS*, pages 50–61. ACM, 2010. doi:10.1145/1827418.1827427.
- 12 Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)*, 44(3):1–62, 2012. doi:10.1145/2187671.2187677.
- 13 Alan J. Demers, Johannes Gehrke, Biswanath Panda, Mirek Riedewald, Varun Sharma, and Walker M. White. Cayuga: A general purpose event monitoring system. In *CIDR*, pages 412–422, 2007. URL: <http://cidrdb.org/cidr2007/papers/cidr07p47.pdf>.
- 14 Yanlei Diao, Neil Immerman, and Daniel Gyllstrom. SASE+: An agile language for kleene closure over event streams. *UMass Technical Report*, 2007.
- 15 Antony Galton and Juan Carlos Augusto. Two approaches to event definition. In *DEXA*, pages 547–556, 2002. doi:10.1007/3-540-46146-9_54.
- 16 Julián García and Cristian Riveros. Complex event recognition under time constraints: Towards a formal framework for efficient query evaluation. *Proc. ACM Manag. Data*, 3(2):94:1–94:17, 2025. doi:10.1145/3725231.
- 17 Nikos Giatrakos, Elias Alevizos, Alexander Artikis, Antonios Deligiannakis, and Minos Garofalakis. Complex event recognition in the big data era: a survey. *The VLDB Journal*, 29:313–352, 2020. doi:10.1007/S00778-019-00557-W.
- 18 Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap. *Aggregation functions*, volume 127. Cambridge University Press, 2009.
- 19 Alejandro Grez, Cristian Riveros, and Martín Ugarte. A formal framework for complex event processing. In *ICDT*, volume 127, pages 5:1–5:18, 2019. doi:10.4230/LIPIcs.ICDT.2019.5.
- 20 Alejandro Grez, Cristian Riveros, Martín Ugarte, and Stijn Vansummeren. A formal framework for complex event recognition. *ACM TODS*, 46(4):16:1–16:49, 2021. doi:10.1145/3485463.
- 21 Mikell P Groover. *Automation, production systems, and computer-integrated manufacturing*. Pearson Education India, 2016.
- 22 Paulo Jesus, Carlos Baquero, and Paulo Sérgio Almeida. A survey of distributed data aggregation algorithms. *IEEE Communications Surveys & Tutorials*, 17(1):381–404, 2014. doi:10.1109/COMST.2014.2354398.
- 23 Lei Ma, Chuan Lei, Olga Poppe, and Elke A Rundensteiner. Gloria: Graph-based sharing optimizer for event trend aggregation. In *SIGMOD*, pages 1122–1135, 2022. doi:10.1145/3514221.3526145.
- 24 Biswanath Mukherjee, L Todd Heberlein, and Karl N Levitt. Network intrusion detection. *IEEE network*, 8(3):26–41, 1994. doi:10.1109/65.283931.
- 25 Peter R Pietzuch, Brian Shand, and Jean Bacon. A framework for event composition in distributed systems. In *Middleware*, pages 62–82. Springer, 2003. doi:10.1007/3-540-44892-6_4.
- 26 Manolis Pitsikalis, Alexander Artikis, Richard Dreo, Cyril Ray, Elena Camossi, and Anne-Laure Jousselme. Composite event recognition for maritime monitoring. In *DEBS*, pages 163–174, 2019. doi:10.1145/3328905.3329762.

- 27 Olga Poppe, Chuan Lei, Lei Ma, Allison Rozet, and Elke A Rundensteiner. To share, or not to share online event trend aggregation over bursty event streams. In *SIGMOD*, pages 1452–1464, 2021. doi:10.1145/3448016.3452785.
- 28 Olga Poppe, Chuan Lei, Elke A. Rundensteiner, and David Maier. GRETA: graph-based real-time event trend aggregation. *VLDB*, 11(1):80–92, 2017. doi:10.14778/3151113.3151120.
- 29 Olga Poppe, Chuan Lei, Elke A Rundensteiner, and David Maier. Event trend aggregation under rich event matching semantics. In *SIGMOD*, pages 555–572, 2019. doi:10.1145/3299869.3319862.
- 30 BS Sahay and Jayanthi Ranjan. Real time business intelligence in supply chain analytics. *Information Management & Computer Security*, 16(1):28–48, 2008. doi:10.1108/09685220810862733.
- 31 Nicholas Poul Schultz-Møller, Matteo Migliavacca, and Peter Pietzuch. Distributed complex event processing with query rewriting. In *DEBS*, pages 1–12, 2009.
- 32 Luc Segoufin. Enumerating with constant delay the answers to a query. In *ICDT*, pages 10–20, 2013. doi:10.1145/2448496.2448498.
- 33 Walker White, Mirek Riedewald, Johannes Gehrke, and Alan Demers. What is “next” in event processing? In *PODS*, pages 263–272, 2007. doi:10.1145/1265530.1265567.
- 34 Eugene Wu, Yanlei Diao, and Shariq Rizvi. High-performance complex event processing over streams. In *SIGMOD*, pages 407–418, 2006. doi:10.1145/1142473.1142520.
- 35 Haopeng Zhang, Yanlei Diao, and Neil Immerman. On complexity and optimization of expensive queries in complex event processing. In *SIGMOD*, 2014.
- 36 Detlef Zimmer and Rainer Unland. On the semantics of complex events in active database management systems. In *ICDE*, pages 392–399, 1999. doi:10.1109/ICDE.1999.754955.

A Examples from practice

In the following, we present several queries obtained from the literature and show how to model them with ACEL. Given the operators introduced in the previous sections, recall that we define ACEL as any formula φ that uses the standard operators of CEL (Section 2), the aggregation operator **Agg** (Section 5), or a combination of them. We start this section by introducing some new operators and predicates that work as syntax sugar for ACEL to define practical queries. Then we present three queries with aggregation obtained from three different CER proposals and specify them by using ACEL.

A.1 Useful operators in ACEL for specifying real-life queries

To specify CER queries in practice, the following operator will be useful. Let R be any event type. We define the ACEL formula $\text{NEXT}(R)$ with CEL such that:

$$\text{NEXT}(R) \equiv \pi_R((X+ : R) \text{ FILTER } X[\text{type} \neq R]) \text{ OR } R$$

where we assume that R can also be used as a variable name (i.e., $R \in \mathbf{X}$). In other words, $\text{NEXT}(R)$ finds the first event of R -type in an interval and discard all other events in between. As an example, one can use this operator to succinctly define the CEL query $S : \text{NEXT}(R)$ that finds all S -events directly followed by an R event. We define this operator for its later use in examples appearing from other systems.

Let \mathbf{a} be an attribute. We also define some auxiliary predicates that will be use with the **FILTER** operator to correlate two or more events. Recall that we define a predicate P as a possibly infinite subset of events and we generalize P from events to a multiset of events $E \subseteq \mathbf{E}$ such that $E \models P$ if, and only if, $e \models P$ for every $e \in E$. For specifying queries in practice, we need some special multisets predicates that cannot be defined directly as a

generalization of normal predicates. Formally, a *multiset predicate* P is a subset of multisets of events, namely, $P \subseteq \mathcal{P}_{bags}(\mathbf{E})$. For example, the generalization of an (event) predicate to multisets of events is a multiset predicate. These multiset predicates defined below will allow us to (1) ensure that the pattern has the same value at attribute \mathbf{a} , (2) the value of an attribute is increasing, or (3) the value of an attribute is decreasing, respectively. More specifically, we define the following three multiset predicates. For a multiset of events E and two events $e_1, e_2 \in E$, let $\text{Succ}_E(e_1, e_2)$ be the logical formula that checks if e_2 is the *successor* of e_1 in E , formally, $e_1(\text{time}) < e_2(\text{time})$ and there does not exist $e_3 \in E$ such that $e_1(\text{time}) < e_3(\text{time})$ and $e_3(\text{time}) < e_2(\text{time})$.

1. We define the auxiliary predicate $[\mathbf{a}]$ as:

$$[\mathbf{a}] := \{E \in \mathcal{P}_{bags}(\mathbf{E}) \mid \forall e_1, e_2 \in E. e_1(\mathbf{a}) = e_2(\mathbf{a})\}$$

2. We define the auxiliary predicate $[\text{increasing}(\mathbf{a})]$ as:

$$[\text{increasing}(\mathbf{a})] := \{E \in \mathcal{P}_{bags}(\mathbf{E}) \mid \forall e_1, e_2 \in E. \text{Succ}_E(e_1, e_2) \rightarrow e_1(\mathbf{a}) < e_2(\mathbf{a})\}$$

3. Finally, we define the auxiliary predicate $[\text{decreasing}(\mathbf{a})]$ as:

$$[\text{decreasing}(\mathbf{a})] := \{E \in \mathcal{P}_{bags}(\mathbf{E}) \mid \forall e_1, e_2 \in E. \text{Succ}_E(e_1, e_2) \rightarrow e_1(\mathbf{a}) > e_2(\mathbf{a})\}$$

Following ACEL syntax and semantics, we use the above predicates with variables in \mathbf{X} . For example, we write $\varphi \text{ FILTER } X[\mathbf{a}]$ to define that all events in variable X must satisfy $[\mathbf{a}]$. Similar as for standard predicates, we use conjunction and disjunction in FILTER as a syntax sugar for composing filters or using OR , respectively.

In the following we provide several examples from previous literature and how we can specify them by using ACEL.

► **Example 13.** We use as an example an adaptation of a query extracted from ESPER's documentation [1], which says:

“This example statement demonstrates the idea by selecting a total price per customer over pairs of events (ServiceOrder followed by a ProductOrder event for the same customer id within 1 minute), occurring in the last 2 hours, in which the sum of price is greater than 100, and using a where clause to filter on name.”

The difference between this example and the original one is that we do not consider group-by, window and slide operators and a slightly different filter. The query in ESPER's query language is the following:

```
Q13: SELECT a.custId, sum(a.price + b.price)
      FROM PATTERN [every a=ServiceOrder ->
                   b=ProductOrder(custId = a.custId)]
      WHERE a.name in (b.name)
      HAVING sum(a.price + b.price) > 100
```

We can define Q2 by using ACEL as follows:

$$\varphi_{13} = \left[\text{Agg}_{Y(e \leftarrow \text{sum}(X(\text{price})))} \left(\begin{aligned} &[(\text{ServiceOrder AS } a; \text{ProductOrder AS } b) + \text{AS } X] \\ &\text{FILTER } (X[\text{name}] \wedge X[\text{custId}]) \end{aligned} \right) \right. \\ \left. \text{FILTER } Y[e > 100] \right]$$

15:20 A Formal Query Language for Aggregation in CER

► **Example 14.** We use as an example an adaptation of query “ q_3 ” extracted from HAMLET’s paper [27, p. 1], which says

“All events in a trip must have the same driver and rider identifiers as required by the predicate [driver, rider] (...). Query q_3 tracks riders who cancel their accepted requests while the drivers were stuck in slow-moving traffic. All three queries contain the expensive Kleene sub-pattern $T+$ that matches arbitrarily long event trends.”

The difference between this example and the original one is that this does not considers within, slide and group-by operators.

```
Q14: RETURN T.district, COUNT(*), SUM(T.duration)
      PATTERN SEQ(Request R, Travel T+, Cancel C)
      WHERE [driver, rider]
```

A formula equivalent to the previous query in CEL with aggregation could be the following:

$$\varphi_{14} = \text{Agg}_{Y[e \leftarrow \text{sum}(T(\text{duration})), f \leftarrow \text{count}(C(\text{driver}))]} [((\text{Request AS } R; \text{Travel AS } T+; \text{Cancel AS } C) \text{ AS } X)+] \\ \text{FILTER } X[\text{driver}] \wedge X[\text{rider}]$$

► **Example 15.** We use as an example an adaptation of the query “ q_1 ” extracted from COGRA’s paper [29, p. 1], which says:

“Query q_1 detects minimal and maximal heartbeat during passive physical activities (e.g., reading, watching TV). Query q_1 consumes a stream of heart rate measurements of intensive care patients. Each event carries a time stamp in seconds, a patient identifier, an activity identifier, and a heart rate. For each patient, q_1 detects contiguously increasing heart rate measurements during a time window of 10 minutes that slides every 30 seconds. No measurements may be skipped in between matched events per patient, as expressed by the contiguous semantics.”

The difference between this example and the original one is that this version does not considers within, group-by, and slide operators. The query is:

```
Q15: RETURN patient, MIN(M.rate), MAX(M.rate)
      PATTERN Measurement M+
      SEMANTICS contiguous
      WHERE [patient] AND M.rate < NEXT(M).rate
      AND M.activity = passive
```

An ACEL formula equivalent to the previous query could be the following:

$$\varphi_{15} = \text{Agg}_{Y[e \leftarrow \text{min}(M(\text{rate})), f \leftarrow \text{max}(M(\text{rate}))]} [(\text{Measurement AS } M) \oplus \\ \text{FILTER } M[\text{patient}] \wedge M[\text{increasing}(\text{rate})] \wedge M.\text{activity} = \text{“passive”}]$$

One can check that formula φ_3 specifies the same query as Q3 with the difference is that φ_3 has a formal and denotational semantics.

More examples can be found in [8]. Finally, it is important to note that we also consider examples of other proposals (e.g., CAYUGA [13]) that use aggregation; however, their query languages are procedural, and they do not adapt to the concept of declarative aggregation that we use in this work.