algorithms

MDPI

Article

A Data-Driven Approach to Discovering Process Choreography

Jaciel David Hernandez-Resendiz [1], Edgar Tello-Leal [2,*] and Marcos Sepúlveda [3]

[1] Multidisciplinary Academic Unit Reynosa-Rodhe, Autonomous University of Tamaulipas, Reynosa 88779, Mexico
[2] Faculty of Engineering and Science, Autonomous University of Tamaulipas, Victoria 87000, Mexico
[3] Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, Santiago 8331150, Chile
* Correspondence: etello@docentes.uat.edu.mx

Abstract: Implementing approaches based on process mining in inter-organizational collaboration environments presents challenges related to the granularity of event logs, the privacy and autonomy of business processes, and the alignment of event data generated in inter-organizational business process (IOBP) execution. Therefore, this paper proposes a complete and modular data-driven approach that implements natural language processing techniques, text similarity, and process mining techniques (discovery and conformance checking) through a set of methods and formal rules that enable analysis of the data contained in the event logs and the intra-organizational process models of the participants in the collaboration, to identify patterns that allow the discovery of the process choreography. The approach enables merging the event logs of the inter-organizational collaboration participants from the identified message interactions, enabling the automatic construction of an IOBP model. The proposed approach was evaluated using four real-life and two artificial event logs. In discovering the choreography process, average values of 0.86, 0.89, and 0.86 were obtained for relationship precision, relation recall, and relationship F-score metrics. In evaluating the quality of the built IOBP models, values of 0.95 and 1.00 were achieved for the precision and recall metrics, respectively. The performance obtained in the different scenarios is encouraging, demonstrating the ability of the approach to discover the process choreography and the construction of business process models in inter-organizational environments.

Keywords: process choreography; IOBP; data-driven; process mining; discovery

check for updates

Citation: Hernandez-Resendiz, J.D.; Tello-Leal, E.; Sepúlveda, M. A Data-Driven Approach to Discovering Process Choreography. *Algorithms* **2024**, *17*, 188. https://doi.org/10.3390/a17050188

Academic Editors: Zixin Huang, Li Jin, Xiongbo Wan and Sheng Du

Received: 10 April 2024
Revised: 28 April 2024
Accepted: 28 April 2024
Published: 29 April 2024

1. Introduction

In collaborative networks, the partners work together to create competitive advantages by defining the activities to be carried out by each organization, the business processes to be executed, the roles to be played, the communication channels, and the definition of interoperability at both the process and system levels in order to achieve common business goals [1–3]; e.g., a supply chain process may involve several organizations [4]. Collaborative networks foster joint problem-solving through resource sharing and the fusion of complementary skills. This collaborative environment enhances organizations' potential to create and acquire knowledge, leading to the innovation of products or services [5]. In this context of collaborative innovation, members of the supply chain plan and implement actions for knowledge sharing and knowledge application to develop new products and services quickly and efficiently, enabling them to maintain and improve their performance in the long term [6]. Furthermore, in Industry 4.0, end-to-end digital integration is required in the supply chain, with a business process design logic that crosses organizational boundaries. These business processes can be defined using the business process model and notion (BPMN) language [7,8], a standard for graphically representing the logic of the business process and its subsequent automation [9], which not only makes the logic more understandable but also makes it easier to integrate the perspective of the control flow, the subprocesses, the data flows (internal or external), and the resources involved in the processes into a BPMN diagram [10].

Data-driven approaches are characterized by decision-making based on the analysis and interpretation of data, rather than observations, allowing decisions and solutions to be supported by facts [11,12]. Process mining techniques are implemented through data-driven approaches, making it possible to discover process patterns within event logs and detect and diagnose differences between observed and modeled behavior [10,13], which helps in decision-making to improve and optimize business processes [14,15]. In this way, approaches based on process mining techniques have been implemented to verify and enhance business processes. These techniques are characterized by supporting discovery, conformance checking, enhancement, and predictive analytic tasks [16–18]. In a discovery approach to the business process model, event data generated by the execution of business processes are analyzed, making it possible to identify the logic and behavior of the process from these event data, known as event logs. Process conformance checking consists of evaluating the alignment of the behavior of the actual business process model against the behavior discovered in the event log (generated by the business process itself), to detect any possible deviations. In the process improvement task, various analyses are carried out, considering all the attributes available in the event log and in the real business process model to detect possible bottlenecks, high time consumption in task execution, deviations, and duplication of the execution of tasks by different resources, among others; making it possible to identify opportunities for improvement in the business process.

In approaches based on process mining techniques that implement tasks such as predictive process monitoring or trace clustering, an event log preprocessing stage is included, in which the input data must be encoded to feed the prediction or inference algorithm. At this stage, an encoding method is typically implemented to transform complex event data into a numerical or representative feature space [19]. One of the most important methods for this purpose is Doc2Vec, based on representation learning, developed in natural language processing (NLP) [20]. This learning uses neural network architecture models to automatically learn distributed vector representations of a concept of interest (for example, an activity or a trace) with high quality. Doc2Vec is an architecture for computing continuous vector representations of words from large datasets with high dimensionality. In the process mining domain, several approaches based on representation learning techniques have been presented to significantly improve the performance of the inference algorithm. In [21], the authors proposed several activity-level models, traces, models, and logs to deal with the high dimensionality of real-life event logs and to generate a distributed representation that can be used in different process mining tasks. In [22], the authors presented a case-level solution that uses word embeddings for business process data to better encode process instances. For their part, ref. [23] expounded an approach for conformance verification based on vector representations of each activity/task present in the model and the event log. Therefore, the vectors generated by Doc2Vec can be used to find similarities between traces, allowing for the quick analysis of large event logs by expressing words in the vector-space model and considering the context when learning through the co-occurrence of activities.

Recent studies have proposed solutions for different process mining tasks applied in intra-organizational business processes [24–27]. However, when process mining solutions are implemented in inter-organizational business processes (IOBP), aspects such as the process's privacy and autonomy; data with different levels of granularity; and event data stored in other sources, formats, and distribution form must be considered. Therefore, managing independently generated event logs requires methodologies and algorithms to process, align, and merge the event logs generated by process-oriented information systems [28]. Importantly, events need to be correlated across organizational boundaries. Then, by implementing process mining techniques, the tasks of discovery, monitoring, compliance, and improvement of IOBPs, which have yet to be studied to date, can be carried out. Furthermore, the analysis can be extended to discover and verify the process choreography, which represents the formalization of interactions through messages from the participants in an inter-organizational collaboration [29].

In this sense, automatic analysis of the historical information recorded from the execution of the business processes of the participating organizations can help to find relationships within the IOBP. The above can be achieved through data-driven and process model-level analysis. At the structured data level, the organizations participating in the IOBP are responsible for selecting and structuring the data from their information systems and consequently choosing the appropriate level of abstraction and the point of view of the data. At the level of process models, the business process flow of the participating organizations is analyzed, in search of patterns that can complement the analysis of structured data, to obtain sufficient information for identifying collaboration patterns between organizations and discovering the IOBP model. Different approaches are available in the state of the art that partially address analyzing and discovering process choreography, focusing on the analysis of the information contained in event logs [30–32], in business process models [33–37], document electronics, and information related to the business process [38].

Therefore, this paper proposes a data-driven methodology supported by semi-automatic methods that enable the discovery of the IOBP model and the process choreography in a collaborative environment. The relationships between the organizations participating in the business process are identified, labeled, and defined using a method based on the Doc2Vec algorithm and by calculating the cosine similarity measure between events to identify possible message-type tasks and their task subtype (send/receive), for which a set of definitions are specified to formalize the relationships, as well as a group of rules for assigning the message task subtype. These criteria are formulated in terms of relationships at the trace level and the event level. Next, each collaboration participant's intra-organizational business process model is determined, marking in the model the tasks previously identified as message-type tasks and their subtype, and defining the relationship between the processes through flow message connectors, which allows building an IOBP, including its process choreography. Subsequently, an inter-organizational event log is generated from the intra-organizational event logs, applying a fusion of traces from the relationships identified by the message-type tasks and their subtype, containing the event data of both traces. Finally, the process choreography and intra- and inter-organizational models are evaluated using the metrics of precision, recall, F-score, and generalization. The proposed approach was evaluated using four event logs derived from real-life IOBPs and two artificial event logs. The results achieved are very acceptable, with an overall performance in the discovery of the process choreography of 0.86 for the *relationship precision* metric, a *relationship recall* of 0.89, and with a measurement *F-score of the relationship* of 0.86, with a performance over 89% in the message-type task identification task. On the other hand, for the average evaluation of the quality level of the IOBP discovered, a *precision* of 0.94 was achieved, with a *recall* value of 0.99, and *generalization* indicator of 0.63, which indicates that the model of the inter-organizational process discovered could reflect more than 94% of the behavior contained in the merged event log.

2. Related Work

In [39], a technique to discover collaboration models from intra-organizational event logs was proposed. The structure of the event log was extended to support interaction data between participants by adding attributes to contain the message name, message identifier, participant role, and type of communication between participants. Interactions between participants are identified through an event data analysis in the event log, determining the correlation between the messages exchanged. Subsequently, the intra-organizational models discovered with the information from the interaction of messages are combined, which enables the generation of an inter-organizational business process model aligned with the BPMN language. Intra-organizational business process models are discovered for each participant in the collaboration by applying algorithms available in the state of the art. Similarly, ref. [40] presented a process mining approach to discover inter-organizational business processes and process choreography from an extended event log. This log requires information about the participants and the messages exchanged between the participants, to discover a model of the inter-organizational process represented by the BPMN language. The extended event log includes information required for the inter-organizational process model and process choreography. For example, the participant attribute identifies the participant that executes the event, and an attribute contains the type of event; in the case of message-type events, the information of the participant who receives the message is required. A fundamental stage in this proposal is extracting all message-type events and the information related to the message: the participant who sends or receives it. With this extra event log, a model of the process with the message interactions between the participants involved in the collaboration is discovered. This model is used to build process choreography and inter-organizational models in conjunction with the intra-organizational process models discovered for each participant. Our proposal takes a different approach from the studies mentioned above. It does not require the extension of the event log or adding information about the messages and resources exchanged between the collaboration participants. Instead, our method is based on a unique set of methods and formal rules. These tools allow for the identification of potential message tasks and the determination of the task's subtype, which in turn defines the message's meaning (send/receive).

On the other hand, ref. [41] proposed a process mining technique to merge intra-organizational event logs and discover an inter-organizational process model represented by a directly-follows graph. This approach is characterized by only using the common elements of an event log: case ID, timestamps, and activity. Furthermore, it is based on the premise that two activities of different organizations occur consecutively with a very short time difference, for which several time thresholds are defined. Therefore, adjacent activities with the minimum time difference should be interconnected and extracted, since they belong to the same trace within an inter-organizational event log, forming the sequence of the activities of the merged event log, ordered by the timestamp value. Each extracted activity pair will be identified in this log by concatenation with the original case IDs. The rest of the events of the same trace (which were not extracted) of each participant are embedded in the trace according to the timestamp value, with which the trace is constructed with all its events. This procedure is executed until no adjacent activities are identified in the event logs of each collaboration participant. In our case, the relationship between message tasks is determined by a cosine similarity measure that ensures that two tasks (from different participants) are close and possibly related. Furthermore, the task's subtype is determined through a set of rules that allow analyzing the context of the message-type task, that is, the antecedent and consequent tasks for both parties of the collaboration.

Differently, ref. [42] presented an approach based on a Petri net extension that supports the management of message attributes and resources exchanged in workflows (called RM_WF_net) to formalize healthcare processes in hospitals, particularly inter-departmental processes. From the formalization, algorithms are applied to discover intra-departmental models and identify collaboration patterns in each intra-departmental model, with which a collaboration model is built. The first algorithm discovers a control-flow structure based on WF-net. Subsequently, the event log is processed to identify messages and resources, which generates a RM_WF_net for each department. On the other hand, ref. [43] presented a process mining approach in an inter-organizational environment for a cloud computing multi-tenancy architecture through declarative models. Through a set of business rules, information related to the processes of systems that run in the cloud is extracted, and distributed data are identified, enabling the building of an event log. This approach makes it possible to represent processes with high variability. The previous proposals differ from our approach since using Petri nets reduces the expressiveness of the discovered model notation and does not support high-level notations compared to a BPMN-based model. In addition, there may be some difficulties in representing complex behaviors in the process logic, for example, in event-based gateways, which does not happen in BPMN-based models.

3. Preliminary Formalization

This section introduces the main foundations of the proposed approach, which formalizes the methodology phases and enables the identification of message-type tasks from direct tasks (previous or subsequent) or non-direct tasks. The above facilitates the marking of message tasks by their subtype (send/receive), making it possible to merge event logs and discover the correlation of messages exchanged in a collaboration.

Definition 1 (**Mapping an event to a sentence).** *This refers to a sentence of words that represent each event E_j. The sentence of words is generated from the values of the attributes Et_k that makeup E_j, representing the sentence's words.*

Definition 2 (**Mapping a trace to a document).** *This refers to statements representing each case T_i in the event log L. This document is generated from the values of the Et attributes of each activity $E_j \in T_i$, which represent the document's words (see Definition 1).*

Definition 3 (**Incoming ($\bullet\theta$) and Outgoing edges ($\theta\bullet$) for the task** θ). *Given a BPMN model $M = (i, o, T, G, E_m)$ and a task $\theta \in T$, its incoming edges $\bullet\theta = \{(c, d) \in E_m \mid d = \theta\}$ and its outgoing edges $\theta\bullet = \{(c, d) \in E_m \mid c = \theta\}$ [24].*

Definition 4 (**Direct predecessors of task** m). *Given a BPMN model $M = (i, o, T, G, E_m)$ and a task $m \in T$, its set of t-predecessors of task m are all tasks $p \notin G$, such that there is a direct path between p and m; and this path is contained in its set of incoming edges $\bullet m$ (see Definition 3).*

Definition 5 (**Direct successors of task** m). *Given a BPMN model $M = (i, o, T, G, E_m)$ and an event $m \in T$, its set of t-successors of m are all tasks $s \notin G$ such that there is a direct path between m and s, and this path is contained in the set of outgoing edges $m\bullet$ (see Definition 3).*

Definition 6 (**Non-direct predecessors of task** m). *Given a BPMN model $M = (i, o, T, G, E_m)$ and a task $m \in T$, the set of its t-non-directpredecessors is the set of tasks $TSK \in T$ such that for each $tsk \in TSK$, there are one or more paths between the tasks i and m that visit the event $tsk \notin G$.*

4. Materials and Methods

This section describes a conceptual representation of the scheme for discovering process choreography in inter-organizational environments. Figure 1 shows, in general terms, the phases and methods that compose the proposed methodology.

4.1. Event Log Processing

Our approach assumes that the event logs do not have empty or missing attributes. Furthermore, our methodology requires at least three common attributes in event logs: case ID, activity name, and timestamp. If additional attributes are available, they can be integrated and processed. However, only the value in the activity name attribute is used when marking and labeling the message-type task.

4.1.1. Method 1: Construction of the Vector Representation Matrix (VRM) of the Cases

For each event log L and L', a VRM and VRM' matrix of dimensions of $m \times n$ is generated, respectively. Where one row (a trace in the event log) of the matrix is a VRM representation of $T \in L$ and VRM' of $T' \in L'$, and this representation is mathematically implemented using the Doc2Vec word embedding technique, storing contextual information, in a low-dimensional vector, of all the attributes of a case (within the event log) that describe each of the events of the event logs L and L'. Next, the value of each attribute is identified to separate it into the words contained. The set of identified words form the vocabulary of the word embedding model, enabling the construction of a Doc2Vec representation, where a trace is treated as a document. The size of each word document is equal to the number of different values within the attributes that describe the event contained in a trace. The VRM and VRM' matrices are generated using the following rules:

- Consider any case $T_i \in L$.
- The mapping of a trace T_i to a D_i document is defined, considering all the values of the attributes et_k of each task $e_j \in T_i$, according to Definition 2.
- Stop-words in each D_i document are identified and removed.
- The remaining words in the D_i document form the T_i document. This procedure is performed for L and L', generating a corpus of D and D' documents.
- The cDoc2Vec model is built using the Doc2Vec method.
- The D and D' corpus create a general vocabulary of words.
- The cDoc2Vec model is trained with the corpus of D and D' documents.
- Then, the VRM and VRM' matrices are built from the representation $VRM_i \leftarrow cDoc2Vec.infer(D_i)$ and $VRM'_j \leftarrow cDoc2Vec.infer(D'_j)$, where the function $cDoc2Vec.$-$infer(D_i)$ and $cDoc2Vec.infer(D'_j)$ allows extracting a mathematical representation of each document D_i and D'_j through $cDoc2Vec$ model inference.

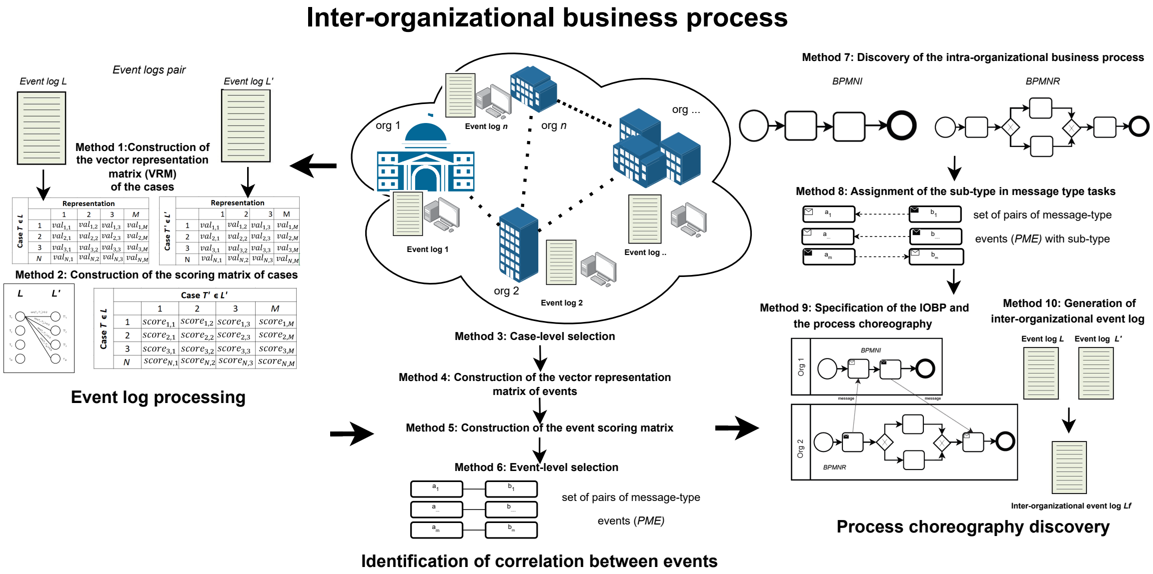

Inter-organizational business process
Event logs pair
Event log L
Event log L'
Method 1:Construction of the vector representation matrix (VRM) of the cases
Representation
Case T ∈ L
Case T' ∈ L'
Method 2: Construction of the scoring matrix of cases
L
L'
Case T' ∈ L'
Case T ∈ L
Event log processing
org 1
Event log n
org n
org ...
Event log 1
Event log ..
org 2
Event log 2
Method 3: Case-level selection
Method 4: Construction of the vector representation matrix of events
Method 5: Construction of the event scoring matrix
Method 6: Event-level selection
set of pairs of message-type events (PME)
Identification of correlation between events
Method 7: Discovery of the intra-organizational business process
BPMNI
BPMNR
Method 8: Assignment of the sub-type in message type tasks
set of pairs of message-type events (PME) with sub-type
Method 9: Specification of the IOBP and the process choreography
Org 1
Org 2
message
message
Method 10: Generation of inter-organizational event log
Event log L
Event log L'
Inter-organizational event log Lf
Process choreography discovery

Figure 1. Overview of the proposed data-driven methodology.

4.1.2. Method 2: Construction of the Scoring Matrix of Cases

In this method, for each representation VRM_i contained in VRM, the cosine similarity measure is calculated with all representations contained in VRM'. The value obtained in the measure calculation allows us to know the similarity between vectors representing the traces in an internal product space. Then, a score matrix (SM) is generated with the similarity values between the VRM and VRM' vectors, constructing a matrix of size $|L| * |L'|$.

4.2. Identification of the Correlation between Events

4.2.1. Method 3: Case-Level Selection

The scoring matrix SM is filtered by applying the condition that, for each value $SM_{i,j}$ that exceeds threshold U_t, it must be extracted, generating a set of document pairs D_i and D'_j. For these documents $D_i \leftarrow T_i \in L$ y $D'_j \leftarrow T'_j \in L'$, there is a relationship at the case level, since they share information of the IOBP. The pair of traces of these selected documents are stored in the set $SCP \leftarrow (T_i, T'_j)$.

4.2.2. Method 4: Construction of the Vector Representation Matrix of Events

From the set of case pairs (SCP) selected in method 3, the event attributes that can provide information on the exchange of messages and business documents of the collaborative process are identified. Then, for each event $E_j \in T_i$ and $E'_l \in T'_k$ of each pair of cases $T_i \in L$ and $T'_k \in L'$ contained in SCP, this is mapped as an event-level statement RE_j and RE'_l, according to Definition 1. Their event representation vector is generated through the RE_j and RE'_l statements by inferring the previously trained *eDoc2Vec* model. This model is trained similarly to the *cDoc2Vec* model, but using event-level input data. The representations RE_j and RE'_l for each of the pairs of cases $T, T' \in SCP$ are generated by the following rules:

- We consider a pair of cases $T, T' \in SCP$.
- The events contained in $E_j \in T$ and $E'_l \in T'$ are mapped to a statement document $SD_j \leftarrow E_j$ and $SD'_l \leftarrow E_l$, considering all the values of their attributes $Et_k \in E_j$ and $E't_k \in E'_l$, according to Definition 1.
- Stop-words are identified and removed from the SD_j and SD'_l statement documents. The remaining words in SD_j and SD'_l form the final version of these statement documents.
- Finally, the vectors RE_j and RE'_l are generated from the representation $RE_j \leftarrow eDoc2Vec.infer(SD_j)$ and $RE'_l \leftarrow eDoc2Vec.infer(SD'_l)$, where the inference function allows you to extract a mathematical representation of each document SDj and SD'_l, using the previously trained *eDoc2Vec* model.

4.2.3. Method 5: Construction of the Event Scoring Matrix

For each of the representations RE_j, their similarity with all representations of RE' is measured using the cosine distance metric. The similarity measure values allow the construction of an event-level scoring matrix (ESM), where each row of $ESM_{j,l}$ contains the similarity measure of the representations RE_j and $RE'_{1,2,\dots,l}$.

4.2.4. Method 6: Event-Level Selection

The event score matrix (ESM) is traversed row-wise for each position within ESM_j to identify the highest similarity value in $ESM_{j,l}$. This similarity value will be selected if it meets the condition of exceeding the threshold U_a. The event pairs E_j and E'_l that satisfy the condition are considered possible message-type events, allowing a vector to be generated that stores all pairs of message-type events (PME). Subsequently, filtering is applied to select the pair of events with the highest similarity value, since the case may arise that an event E_j or E'_l may have more than one relationship with another event that contains similar information and exceeded the threshold U_a. Then, each of the events E_j that is related to the event E'_l with the highest similarity value is added to PME.

4.3. Process Choreography Discovery

4.3.1. Method 7: Discovery of the Intra-Organizational Business Process

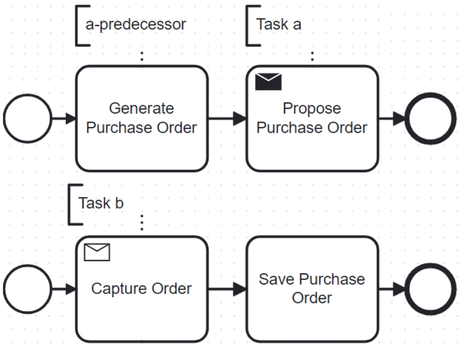
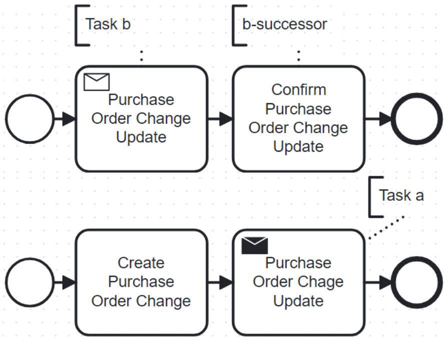
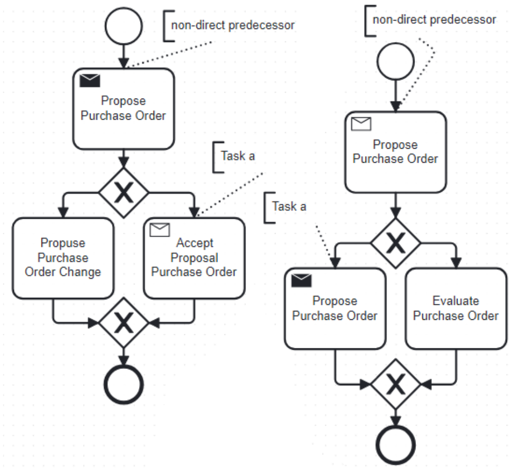
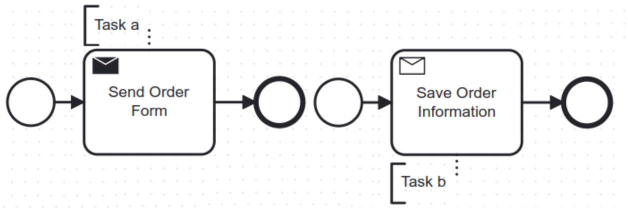
In our experiment, considering the event logs L and L' (for the minimum number of participants required in a collaborative process), the BPMN process models of the initiating participant and the receiving participant are generated, denoted as $BPMNI \leftarrow L$ and $BPMNR \leftarrow L'$. This method uses the split-miner algorithm [24] to perform business process discovery. The essential operation of the algorithm is as follows:

- Consider an event log L.
- From the event log L, a directly-follows graph (DFG) is generated. This graph is a component $g = N, E$, where N represents the set of events (nodes) identified in the event log, and the set E represents the edges or paths that connect the set of nodes N.
- With the resulting DFG graph, a process model is generated based on the syntax of the BPMN language.

4.3.2. Method 8: Assignment of the Sub-Type in Message Type Tasks

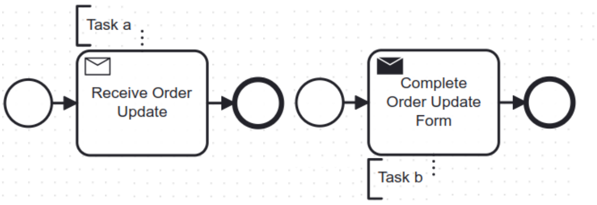
In the discovered BPMN business process models (intra-organizational level), the message type tasks and the subtype of these tasks are identified. For each pair of events $(a, b) \in PME_i$, where $a \in L$ and $b \in L'$, the message subtype is defined, which can have the value of (*send* or *receive*). This subtype specifies the flow and direction of the message in the interaction between the participants in the collaboration. The assignment of the task subtype for each pair of events $(a, b) \in PME_i$ is performed based on the rules described in Table 1, complying with at least one condition defined in the rules.

Table 1. Rules for marking the subtype of the message-type task.

Description	Graphic Representation
Rule 1: Given the *BPMNI* and *BPMNR* models, if the *name* of one of the events in its set of direct *a-predecessors* of the event *a* (see Definition 4) is a compound of one of the elements included in the set $A = \{$*Get, Preparation, Notice, Need, Generate, Evaluate, Information Request, Approval, Process, Analyze, Make a decision, Acceptance, Validate, Communicate, Transport, Calculate, Demand, Order*$\}$, then the event *a* will be a message-type task and subtype *send* ($a.subtype = send$), and task *b* will be a message-type task and subtype *receive* ($b.subtype = receive$).	
Rule 2: Given the *BPMNI* and *BPMNR* models, if the *name* of one of the events in the set of direct *b-successors* of the event *b* (see Definition 5) is a compound of one of the elements included in the set $C = \{$*Question, Notification, Decision Making, Request, Application, Transfer, Order, Manage, Confirm, Delivery, Safety, Evaluate, Approval*$\}$, then event *b* will be a message-type task and subtype *receive* ($b.subtype = receive$), and task *a* will be a message-type task and subtype *send* ($a.subtype = send$).	
Rule 3: Given the *BPMNI* model, if a non-direct predecessor event to event *a* is a message-type task and the subtype is to *send* (see Definition 6), then event *a* will be assigned as a message-type task and subtype *receive* ($a.subtype = receive$). On the other hand, if the non-direct predecessor event to event *a* is a task of type message and subtype *receive* (see Definition 6), then event *a* will be assigned as a message task with subtype *send* ($a.subtype = send$). In the case of event *b*, the same rules are applied using the *BPMNR* model and event *b* to assign its task type and subtype.	
Rule 4: Given the models *BPMNI*, *BPMNR*; the events *a* and *b*, if the *name* of the event *a* is a compound of one of the elements included in the set $S = \{$*Generate, Send, Communicate, Make, Confirm, Set Up, Turn, Order, Report, Transportation*$\}$, then event *a* will be a message-type task and subtype *send* ($a.subtype = send$), and the task *b* will be message-type task and subtype *receive* ($b.subtype = receive$).	

Task b
b-successor
Purchase Order Change Update
Confirm Purchase Order Change Update
Task a
Create Purchase Order Change
Purchase Order Chage Update

Table 1. *Cont.*

Description	Graphic Representation
Rule 5: Given the models *BPMNI*, *BPMNR*; the events *a* and *b*, if the *name* of the event *a* is compound of one of the elements included in the set $R = \{$*Receive, Accept, Status, Arrival, Notice, Admit*$\}$, then event *a* will be a message-type task and subtype *receive* ($a.subtype =$ *receive*), and the task *b* will be message-type task and subtype *send* ($b.subtype = send$)	 Task a Receive Order Update Complete Order Update Form Task b

4.3.3. Method 9: Specification of the IOBP and the Process Choreography

The construction of the IOBP model and the process choreography representation are defined using the results (data and models) generated by implementing methods 7 and 8. Therefore, the exchange of messages is specified on the IOBP model based on the data discovered in the previous methods, enabling the process choreography to be visualized through the formalization of the interaction between the participants of the collaboration. The IOBP model and process choreography are built through the following steps:

- Let us consider a pair of BPMN process models from participants $BPMNI = (i, o, T, G, E_m)$ and $BPMNR = (i, o, T, G, E_m)$.
- The set of nodes $T \in BPMNI$ and $T' \in BPMNR$ refers to the set of events contained in the event registers L and L', considering that each node in T and T' has a label with the name of the events in L and L', respectively. This means that a $\subseteq T$ is represented by the events $a \in PME$ and $\subseteq T'$ is represented by the events $b \in PME$.
- The inter-organizational business process (IOBP) is generated by adding the process models $BPMNI$ and $BPMNR$, each model within a *pool* element identified with the *name* of the participant.
- With the set of task event pairs of type message and its subtype (PME), the message flow connection (Mfc) between the subsets of nodes $\subseteq T$ and $\subseteq T'$ is specified. Mfc represents the links that relate message-type tasks (events) between the pools of the *IOBP* model. Connectors Mfc are added to the *IOBPmodel* $\leftarrow Mfc$, representing the process choreography that supports the message exchange logic between the process models $BPMNI$ and $BPMNR$.
- The direction of each flow connector Mfc is given by the following conditions (1):

$$Mfc \leftarrow \begin{cases} (a \in PME)x(b \in PME) \mid |a \rightarrow b| = 1, & \text{if } a.subtype = send \\ & \textbf{and } b.subtype = receive, \\ (a \in PME)x(b \in PME) \mid |b \rightarrow a| = 1, & \text{if } a.subtype = receive \\ & \textbf{and } b.subtype = send \end{cases} \quad (1)$$

4.3.4. Method 10: Generation of Inter-Organizational Event Log

The inter-organizational event log is constructed by merging the event logs L and L' and the output data from the previous methods, applying the following procedure.

- Consider the event logs L and L', the set of pairs of the cases $(T, T') \in SCP$, and the set of pairs of the events considered message-type tasks $(a, b) \in PME$.
- A new Lf event log is created, which will contain the values of the merged event log.
- For each pair of cases $(T, T') \in SCP$:
  - In the case of T, the message-type tasks found in case T' are added. Therefore, $T \leftarrow E \in T \cup E' \in T'$ as long as $(a, b) \in PME | E' \in b$.
  - In the case of T', the message-type tasks found in case T are added. Therefore, $T' \leftarrow E' \in T' \cup E \in T$ as long as $(a, b) \in PME | E \in a$.
- Cases T and T' are added to the event log $Lf \leftarrow T$ and $Lf \leftarrow T'$.
- Cases $Tl \in L$ not found in $SCP(Tl \notin SCP)$ are added to the event log $Lf \leftarrow Tl$.

- Cases $Tl' \in L'$ not found in $SCP(Tl' \notin SCP)$ are added to the event log $Lf \leftarrow Tl'$.

The resulting event log Lf (merged) allows you to visualize the IOBP model, including all participants and their interactions. In this log, the event attributes that compound the cases contain information about the interaction through messages from the participants involved in the inter-organizational collaboration.

4.4. Evaluation Metrics

The discovered process choreography is evaluated in a supervised manner, for which a reference inter-organizational process model ($RIOBP$) is required. This model contains the process logic, behavior, and real interaction of the participants of the IOBP. Then, process choreography data are extracted from the $RIOBP$ model, including the message-type tasks, the subtype of each message-type task, and the message exchange sequence (noted as *relevant relationships* previously). Furthermore, in the set PME (see Method 6), the relationships selected by our proposal are searched and counted because they met the condition of exceeding the threshold U_a and with the highest value in the measure of cosine similarity, which we call $found\ relationships$. Moreover, relations with a cosine similarity value greater than or equal to the threshold U_a are recovered. These relationships are identified as $recovered\ relationships$. Then, $found\ relationships$ are the subset of $relevant\ relationships$ found by our proposal in the set $recovered\ relationships$; that is,

$$found\ relationships \subset recovered\ relationships | found\ relationships \in relevant\ relationships \tag{2}$$

Considering the above, we propose the following metrics to evaluate the discovered process choreography:

- **Relationship Precision (RP)** is the proportion of $relevant\ relationships$ encountered out of all $recovered\ relationships$. This metric evaluates whether the complete process choreography has been discovered ($found\ relationships = |relevant\ relationships \cap recovered\ relationships|$) without adding relations not found in the process choreography of the reference model ($relevant\ relationships$).

$$RP : \frac{|relevant\ relationships \cap recovered\ relationships|}{recovered\ relationships} \tag{3}$$

- **Relationship Recall (RR)** is the proportion of $found\ relationships$ matching $relevant\ relationships$. This metric indicates the percentage of the process choreography discovered versus the process choreography of the reference IOBP model $relevant\ relationships$.

$$RR : \frac{|relevant\ relationships \cap recovered\ relationships|}{relevant\ relationships} \tag{4}$$

- **F-score of the Relationship (FsR)** is the harmonic mean between the RP and RR, which allows us to determine the performance of the proposed approach. This metric indicates the ability of the method to discriminate between relevant and non-relevant relationships.

$$FsR = 2 \times \frac{RP\ \times RR}{RP + RR} \tag{5}$$

Furthermore, the quality of intra-organizational business process models are evaluated through the metrics *precision*, *recall*, and *generalization*, as defined in [44,45]. These metrics have as input an event log and an intra-organizational business process model, comparing the information available in the event log and the discovered business process model. In our experimentation, it is also required to evaluate the inter-organizational business process model ($IOBP$) built from the merged event log (Lf), for which it is required to implement the following modifications in the process logic of the $IOBP$ model, to represent it as an intra-organizational model, making it possible to determine the quality of the model discovered through the application of the *precision*, *recall*, and *generalization* metrics.

- Let us consider the *IOBP* model and the collaborative *Lf* model previously generated.
- In the *IOBP* model, the following elements contained in the *pools* BPMNI = (i, o, T, G, E_m) and BPMNR = (i', o', T', G', E'_m) are updated:
    1. the initial activity i and i' of the BPMNI and BPMNR models, respectively, are eliminated, and
    2. a unique initial activity I is added, from which the intra-organizational models start, formed as BPMNI = (I, o, T, G, E_m) and BPMNR = (I, o', T', G', E'_m).
- Next, the *IOBP* models BPMNI = (I, o, T, G, E_m) and BPMNR $= (I, o', T', G', E'_m)$ are updated in the following way:
    1. their final activities o and o' are eliminated, respectively, and
    2. a unique final activity O is added, where the two *IOBP* models end with the following structure BPMNI = (I, O, T, G, E_m) and BPMNR = (I, O, T', G', E'_m).
- A set of edges or paths $Exor$ is created between the activities/nodes of message type; for each pair of activities/nodes $(a, b) \in PME$, the virtual edges are created according to the following conditions:
    1. if $a.subtype = send$, then a gateway of the XOR ($Gxor \leftarrow gxor$) type is created, and the paths connecting the nodes $Exor \leftarrow (a \rightarrow gxor)$, $Exor \leftarrow (for \rightarrow b)$ and $Exor \leftarrow (for \rightarrow a - successor)$ are added to $Exor$; removing the message flow connector $(a, b) \in IOBP$.
    2. if $b.subtype = send$, then a gateway XOR ($Gxor \leftarrow gxor$) type is created and the paths that connect the nodes $Exor \leftarrow (b \rightarrow gxor)$, $Exor \leftarrow (gxor \rightarrow a)$ and $Exor \leftarrow (gxor \rightarrow b - successor)$ are added to $Exor$, removing the message flow connector $(a, b) \in IOBP$.
- The resulting model is made up of the following components: $IOBPf = (I, O, (T \cup T'), (G \cup G' \cup Gxor), (E_m \cup E'_m \cup Exor))$.

5. Experimentation

5.1. Inter-Organizational Event Logs

The proposed approach was evaluated using the event logs of 4 real-life scenarios and 2 artificial scenarios of IOBPs. Table 2 summarizes the characteristics of each event log corresponding to the scenarios used in our experimentation.

1. **Air quality system**. This scenario was derived from an autonomous air quality monitoring system based on IoT technology. The collaborative process includes the interaction between 3 participants; consult the description at [46] for more details. The first participant, the *IoT Air quality monitor*, includes activities regarding the validation of the monitoring system's sensors, requests for access to the system, and assigning a valid network address for the monitoring system. Furthermore, it manages all activities for data census through sensors, validation, and sending of air pollution data and meteorological factors through a web service and system shutdown information. The second participant, the *System Access Service*, manages system access requests (accepted and rejected), assigns network addresses for the operation to the IoT air quality monitoring system, and registers active clients. This participant establishes communication with each instance of the *IoT Air quality monitor* participant. The third participant's *Repository Management Service* manages each request for data storage in a database located in a cloud service. It also manages the validation activities of the data sent by the participant *IoT Air quality monitor* (with acceptance or rejection options) and inserts these data into the database.
2. **Healthcare**. This scenario is made up of the activities of 4 participants (*Patient*, *Gynecologist*, *Laboratory*, and *Hospital*) involved in an IOBP within a healthcare scenario (e-healthcare) [31]. The process begins when the *Patient* participant provides information regarding her health status and waits for a response about her treatment or, if applicable, a request for hospitalization. The participant *Gynecologist* coordinates laboratory blood studies and hospitalization activities with the participant *Laboratory*

and the participant *Hospital*. The collaboration begins when a *Patient* sends information about her illness to the *Gynecologist*. The *Gynecologist* examines the *Patient* and, in parallel, takes blood samples from the *Patient* and sends them to the *Laboratory* for analysis. The *Laboratory* studies the blood samples, generating a report with the study results for the participant *Gynecologist*. Subsequently, the *Gynecologist* decides whether the *Patient* should be prescribed medicine or needs hospitalization, informing the *Patient*. When hospitalization is required, the *Gynecologist* communicates with the hospital to request the patient's admission and sends the clinical analysis results. When the *Hospital* begins its process, the patient's clinical history is created. Then, it decides whether to consider the blood test results sent by the *Gynecologist* or request a new analysis; in either case, the login information is sent to the *Patient*.

3. **Travel Agency**. This collaborative process involves the participants *Customer* and *Travel Agency* [31]. The process begins with the *Travel Agency* proposing a travel offer to the *Customer*. The *Customer* reviews the offer and can request a reservation for the trip. Next, the negotiation is executed to confirm the reservation, the payment of the services, and the generation of a reservation confirmation by the *Travel Agency*. Finally, the *Travel Agency* confirms the reservation number to the *Customer* and sends the electronic tickets.
4. **Purchase order**. The event log contains instances of the execution of the purchase order management process from 2017 to 2018 involving two organizations in the telecommunications industry [47]. The organization *M-Repair* plays the role of customer, and the organization *M-Parts* plays the role of the supplier of electronic components. The collaborative business process has the business goal of reducing component acquisition management time and accelerating the purchasing process in *M-Repair*, electronically automating confirmation decisions by the supplier.
5. **Transfer of goods**. This case study explores the management of multimodal transportation business processes, artificially generated [48]. The global business process involves the processes of the participants *Sender and Buyer* (owner of the goods), *Consigner* (responsible for carrying out the procedure for transporting the goods through two types of transport), *Carrier* (first means of transport), and *Shipper* (second means of transport). The business process begins with the *Sender and Buyer* organization, which generates and sends a merchandise order request to the *Consigner* organization. The request is processed, and a transport contract is generated, which both participants sign. Subsequently, the participant *Consigner* requests the reservation from the participant's *Carrier* and *Shipper*. The participant, *Carrier* and *Shipper*, evaluate the reservation request and return a response to the participant's *Consigner*. When acceptance of reservation requests is received, the *Consigner* receives a notification to prepare the packaging of the merchandise and transmits the merchandise to the *Carrier*. The participant *Carrier* loads the received goods onto a means of transport and generates an invoice for the *Consigner*. Then, payment for the service is processed by the participant *Consigner*. At the end of the merchandise transfer journey, the participant *Carrier* transfers the merchandise to the participant *Shipper*, who is responsible for continuing the transportation of the merchandise. The *Shipper* then generates an invoice corresponding to the service and requests the corresponding payment from the participant *Consigner*, who processes the request. Finally, the *Consigner* manages the fee for the service made with his client *Sender and Buyer* and notifies the data to track the merchandise transfer.
6. **Manufacturing process**. This artificial collaborative business process describes a supply chain management scenario, considering the manufacturing and delivery process of product orders, and involves six business partners [49,50]. First, the *Bulk Buyer* orders a set of products from the *Manufacturer*. The manufacturing of these products requires that different suppliers supply the raw materials. In this scenario, assume that *Supplier A* and *Supplier B* are raw materials suppliers named A and B. Based on the order, the *Manufacturer* calculates the demand for materials A and B

(for example, the fuselage and engines). *Supplier B* can supply raw material B, while material A is sent to *Middleman*. The *Middleman* forwards the order to *Supplier A*, who obtains permission from the authority and coordinates with the participant's *Special Carrier* to deliver material A to the *Manufacturer*. When the delivery process starts, the *Special Carrier* informs the *Manufacturer*, so that he can prepare the pre-processing procedure for material A. When the raw material is received, the *Manufacturer* performs a quality test, and if it is favorable, pre-processes matter A. In the case of matter B, the quality test is carried out by *Supplier B*. When the pre-processing of material A is completed and the test results of material B have been validated, the *Manufacturer* begins manufacturing the product. Additionally, the *Manufacturer* sends status reports to the *Bulk Buyer* before and after production, with a final testing process and product delivery completing the process.

Table 2. Characteristics of the event log of each inter-organizational scenario.

Scenario	Participants	Collaborations	Cases	Events	Unique Events
Air quality system	IoT Air Quality Monitor (IAQM)	2	9180	511,049	24
	Repository Management Service (RMS)		9180	215,651	8
	System Access Service (SAS)		9180	64,260	7
Healthcare	Patient	4	100	250	4
	Laboratory		100	300	3
	Hospital		50	200	5
	Gynecologist		100	700	9
Travel Agency	Travel Agency	1	100	869	5
	Customer		100	607	5
Purchase order	M-Repair	1	100	714	19
	M-Parts		100	686	18
Transfer of goods	Shipper	4	2	24	12
	Sender and Buyer		2	20	10
	Consigner		2	40	20
	Carrier		2	20	10
Manufacturing process	Bulkbuyer	8	5	60	4
	Middleman		5	56	4
	Manufacturer		5	202	20
	Supplier A		5	78	8
	Supplier B		5	19	4
	Special Carrier		5	39	8

5.2. Experiment Results

This section presents a detailed evaluation (for each scenario) of implementing the proposed methodology for discovering process choreography by identifying message-type events. The performance achieved by the implementation of the approach was measured using the metrics *relationship Precision (PR)* (Equation (3)), *relationship Recall (RR)* (Equation (4)), and *F-score of the relationship (FsR)* (Equation (5)). Moreover, a quantitative evaluation of the quality of the intra-organizational and inter-organizational process models discovered by the proposed approach is presented using the *precision*, *Recall*, and *Generalization* metrics.

Table 3 presents the values obtained for each evaluation metric of the process choreography discovered for each inter-organizational scenario. The second column shows the pairs of participants for which a message exchange was identified. Columns 3, 4, and 5 display the metrics *PR*, *RR*, and *FsR* achieved in the evaluation of the discovered choreography. In addition, Table 3 shows the U_t and U_a thresholds defined for each scenario. The sensitivity of each threshold can be interpreted as follows: If the value of the threshold

U_t is close to 0, a greater number of cases from the event log are analyzed to find the message-type events. On the contrary, if the value of U_t is close to 1, the method analyzes fewer cases, reducing the search space. The U_a threshold is highly sensitive, because a change in this parameter's value will impact the values achieved in the RP, RR, and FsR metrics. So, if the threshold U_a is decreased close to 0, this will allow more relevant relationships to be recovered, making the value of the RP metric high. However, this will cause more non-relevant relationships to be recovered, which will harm (reduce) the RR metric. Otherwise, if the U_a threshold is increased close to 1, this can cause the relationships found to be true and relevant relationships. Therefore, the RR metric would increase, while the RP metric would decrease, because some relevant relationships would not be recovered due to a low value in the similarity measure. In our experiment, the value of U_a was assigned by maximizing the FsR metric, in order to recover the greatest number of relevant relationships and the least number of non-relevant relationships, since FsR represents the weighting of the RP and RR metrics.

Table 3. Results of the evaluation of the proposed methods to identify the message-type tasks and their subtype.

Scenario	Collaboration	RP	RR	FsR	Parameters	Time (s)
Air quality system	IAQM-SAS	1	1	1	$U_t \geq 0.5$, $U_a \geq 0.98$	105,371
	IAQM-RMS	1	1	1	$U_t \geq 0.5$, $U_a \geq 0.98$	83,419
Healthcare	Patient-Gynecologist	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.96$	60
	Gynecologist-Laboratory	1	0.5	0.66	$U_t \geq 0.2$, $U_a \geq 0.50$	46
	Hospital-Gynecologist	0.5	0.5	0.5	$U_t \geq 0.2$, $U_a \geq 0.96$	57
	Hospital-Patient	–	–	–	$U_t \geq –$, $U_a \geq –$	21
Travel agency	Customer-Travel agency	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.89$	158
Purchase order	M-Repair-M-Parts	0.83	1	0.90	$U_t \geq 0.5$, $U_a \geq 0.9$	203
Transfer of goods	Sender and Buyer-Consigner	0.8	1	0.88	$U_t \geq 0.5$, $U_a \geq 0.6$	1
	Carrier-Consigner	0.8	0.8	0.8	$U_t \geq 0.5$, $U_a \geq 0.6$	1
	Consigner-Shipper	0.75	0.75	0.75	$U_t \geq 0.5$, $U_a \geq 0.6$	1
	Carrier-Shipper	1	1	1	$U_t \geq 0.5$, $U_a \geq 0.4$	1
Manufacturing process	Bulkbuyer-Manufacturer	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.9$	1
	Supplier B-Manufacturer	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.95$	1
	Middleman-Manufacturer	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.98$	1
	Middleman-Special Carrier	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.9$	1
	Middleman-Supplier A	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.9$	1
	Special Carrier-Supplier A	0.28	1	0.43	$U_t \geq 0.2$, $U_a \geq 0.9$	1
	Supplier A-Manufacturer	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.98$	2
	Special Carrier-Manufacturer	1	1	1	$U_t \geq 0.2$, $U_a \geq 0.98$	2

Moreover, the time costs required for the identification of the message tasks were very acceptable. In event logs with a low level of complexity and a low number of cases and events, the approach identified the message-type tasks in approximately 200 s, using the threshold U_t and U_a parameters presented in Table 3. For the collaborations of the Air Quality System scenario, which are characterized by real-life event logs with a high number of cases and events, as well as a medium-high complexity, the algorithm required 105,371 and 83,419 s for collaborations IAQM-SAS and IAQM-RMS, respectively. The time consumed in this scenario encouraged us to continue experimenting with large event logs with greater complexity. The experimentation was carried out on a personal computer with an AMD Ryzen 5 3400G processor, 3700 MHz, 4 cores, with Radeon Vega Graphics, 16 Gb RAM, and 1 Tb SDD with Windows 10 operating system and the Python 3.7 programming language. The time consumed by the approach to the processing of message task identification in the event logs of the Air Quality System scenario could be decreased, and the first action would be focused on improving the processing characteristics of the computing equipment.

Table 4 shows the average value of the metrics *RP*, *RR*, and *FsR* per scenario in the identification of message-type events. These averages are derived from the values achieved for the metrics presented in Table 3. The overall evaluation of this task was 0.86 for the *RP* metric, 0.89 for the *RR* metric, and a *FsR* of 0.86. The results presented are acceptable, indicating that the approach had a performance greater than 89% in the task of finding message-type events in an inter-organizational scenario; that is, it correctly identified (89%) the interaction of messages between the participants, which is the basis for the specification of the process choreography. Furthermore, 86% of the relationships found were genuinely relevant.

Table 4. Average evaluation of the identification of events at the collaboration level.

Scenario	RP	RR	FsR
Air quality system	1	1	1
Healthcare	0.62	0.50	0.54
Travel agency	1	1	1
Purchase order	0.83	1	0.90
Transfer of goods	0.83	0.88	0.83
Manufacturing process	0.91	1	0.92
Average	0.86	0.89	0.86

Table 5 shows the validation of the quality of the intra-organizational process models discovered, which were recovered from the behavior identified in the event logs through implementing the Split-miner algorithm. The level of quality of the recovered intra-organizational process models is of great importance in the proposed approach, since these models are the basis for discovering inter-organizational processes. Table 5 shows the average value of the evaluation metrics per scenario based on the intra-organizational process models discovered in each scenario. For the *precision* metric, a value between 0.85 and 1.00 was obtained, considering the six scenarios. For the *recall* metric, a value of 1.00 was obtained in all the scenarios evaluated. Values of 0.30 and 0.54 were obtained for the *generalization* metric in the scenarios *transfer of goods* and *manufacturing process*; the evaluation for the rest of the scenarios was between 0.79 and 0.96 for this metric. Overall, these results indicate that the intra-organizational models replicated all behavior from event logs and that the models accounted for 5% of behaviors not included in the event logs. However, it is observed that there was a large number of infrequent events, according to the values obtained for the *generalization* metric.

Table 5. Quality assessment of discovered intra-organizational business process models.

Intra-Organizational Business Process	Precision	Recall	Generalization
Air quality system	1	1	0.96
Healthcare	0.98	1	0.87
Travel agency	0.85	1	0.89
Purchase order	1	1	0.79
Transfer of goods	1	1	0.30
Manufacturing process	0.88	1	0.54
Average	0.95	1	0.72

On the other hand, Table 6 presents the evaluation of the quality level of the IOBPs discovered through the metrics of precision, recall, and generalization. In the experiment, 21 organizations participating in 20 peer collaborations were identified, derived from the event logs analyzed. Considering a general evaluation, a value of 0.94 was obtained for the *precision* metric, 0.99 for the *recall* measure, and a value of 0.63 for the *generalization* indicator.

At that level, the results obtained are acceptable and indicate that the inter-organizational process models discovered were capable of reflecting 99% of the behavior found in the merged collaborative event logs. Furthermore, the discovered models only reflected 6% of behaviors not seen in the event logs. However, low values remained for the *generalization* metric, which denotes infrequent activities in the event log. It is essential to mention that this metric is only informative for knowing the conformation of the event log and does not express a performance value for the discovered model. Due to the nature of business processes, it is common for there to be infrequent activities or behaviors in an event log.

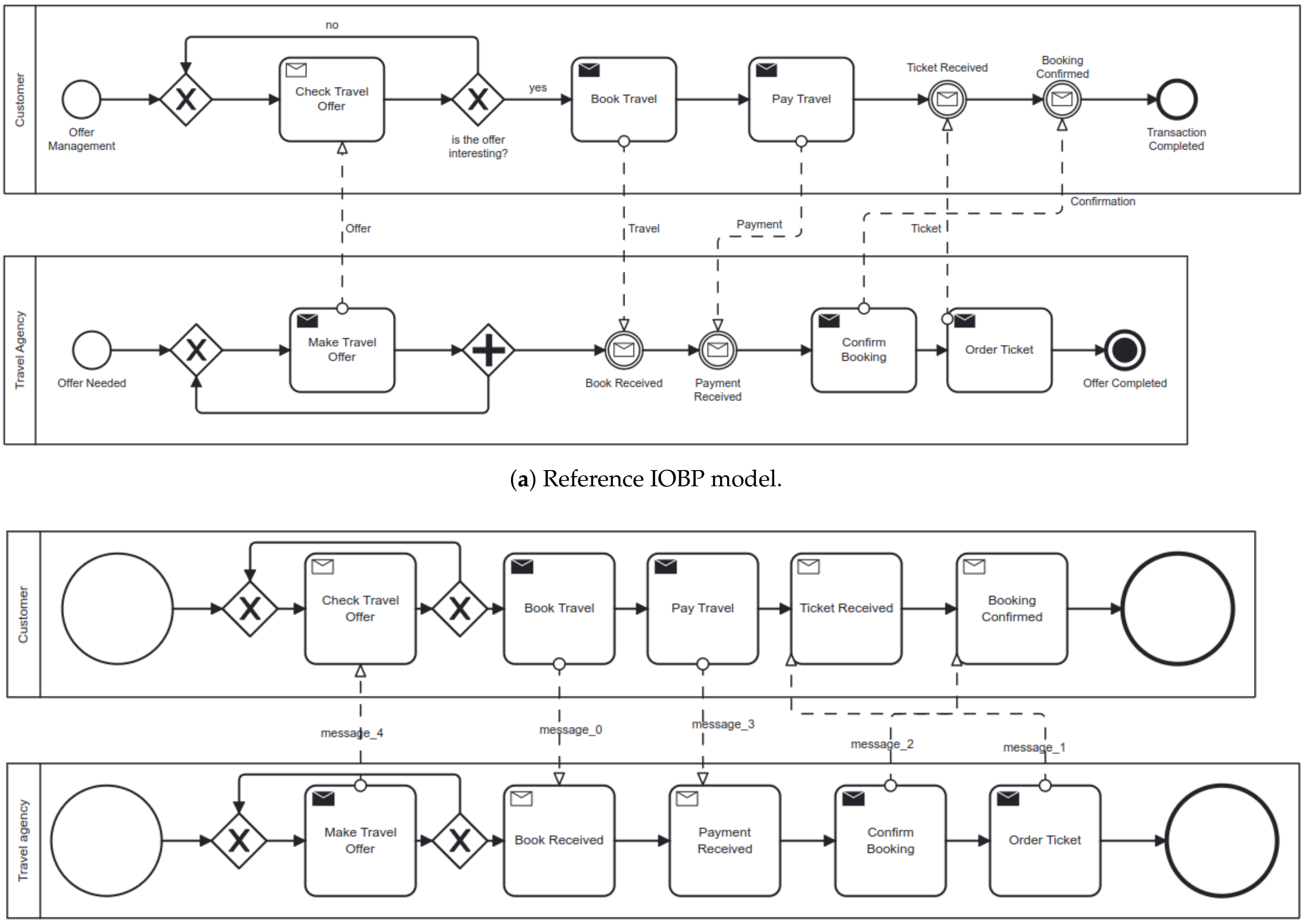
Table 6. Quality assessment of discovered IOBP models.

IOBP	Precision	Recall	Generalization
Air quality system	1	0.99	0.86
Healthcare	1	1	0.67
Travel agency	0.80	1	0.93
Purchase order	1	1	0.50
Transfer of goods	1	1	0.35
Manufacturing process	0.85	0.99	0.48
Average	0.94	0.99	0.63

Figure 2 shows a comparison between the reference IOBP model (see Figure 2a) and the IOBP model discovered (see Figure 2b) for the scenario *Travel Agency*. The figure shows that the process logic, tasks, and gateways discovered coincided with the behavior contained in the reference IOBP. The process choreography deployed through the interaction of messages between pools, represented by the message flow connectors, was similar among the reference and discovered models, with the logic of the matching process and the logic and sequence of the process choreography. In the validation of the discovered process choreography, a value of 1.00 was obtained for all metrics (see Table 4), which indicates that the proposed approach could identify all interactions between participants. On the other hand, in the evaluation of the quality of the collaborative model, values of 0.80 and 1.00 were obtained for the precision and recall metrics, respectively, demonstrating that the collaborative model could reproduce most of the behavior found in the merged event log, without adding behaviors not included in the event log (see Table 6).

The process choreography between the participants *Customer* and *Travel agency* in both models (see Figure 2a,b) contained interactions through five message flow connectors, as described below.

1. The participant *Travel agency* sends a message using the task *Make Travel Offer* of the subtype *send* and the participant *Customer* receives the message using the task *Check Travel Offer* of the subtype *receive*. The interaction is represented by a message flow connector named *offer*, deployed as *message_4* in the discovered process model (see Figure 2b).
2. The *Customer* confirms his interest in the travel proposal through the task *Book Travel* of the subtype *send*, which establishes a communication with the participant *Travel agency* using the message connector *Travel* (*message_0*). The message is received by the company *Travel agency* in the task *Book Received* of the subtype *receive*.
3. The next interaction of *Customer* with the *Travel agency* is presented through the message flow connector *Payment* (*message_3*), which links the task *Pay Travel* of the subtype *send* with the *Payment Received* of subtype *receive* contained in the *pool* of the participant *Travel agency*.
4. The participant *Travel agency* confirms the travel reservation through a message sent by the task *Confirm Booking*, which is received with the *Booking Confirmed* of the subtype *receive* from the participant *Customer*.
5. Finally, the participant *Travel agency* sends a business document about the paid order, sending it in the message flow connector *Ticket* (*message_1*) using the task *Order Ticket* of subtype *send* and the task (*Ticket Received* of the subtype *receive*.



(**a**) Reference IOBP model.

(**b**) Discovered IOBP model.

Figure 2. Comparison of the reference IOBP/process choreography model versus the discovered IOBP/process choreography model for the *travel agency* scenario.

6. Discussion

The approaches for discovering IOBP models and process choreography presented in [39–41] exhibited a similar objective to our proposal. In [39], the authors described the discovery of an IOBP model and the interaction of messages between the participants using a healthcare scenario, as used in our experimentation. Their experiment obtained values of 0.4 and 1.00 for the fitness and precision metrics, respectively, utilizing an extended version of the event log to identify messages between participants. For their part [40], the authors reported the discovery of an IOBP model using the same Healthcare scenario and an extended event log to manage the message data, reporting independent diagrams for the IOBP model and the choreography process discovered. In our experimentation, the quality assessment of the IOBP model of the healthcare scenario obtained a value of 1.00 for the precision and recall metrics. Furthermore, in the quality assessment of the discovery of the intra-organizational models that made up the IOBP model, values of 0.98 and 1.00 were achieved for the precision and recall, respectively.

On the other hand, ref. [41] obtained results between 0.94 and 1.00 for the precision metric and 0.905 and 1.00 for the recall metric in their discovery of a collaborative model using a classic event log (BPIC 2012) in the process mining domain. In their approach, no additional information is required to determine which tasks can be correlated, applying a technique of adjacent activities, and identifying the minimum execution time between the tasks to assess their link. We presented an experiment with the event log of the Air Quality System scenario, which had characteristics and complexity similar to the BPIC 2012 event log. The results of the identification evaluation of the message task achieved a value of 1 for the precision and recall metrics. In discovering the IOBP model, a precision of 1 and 0.99 for recall was obtained. Our approach demonstrated a high performance on most event logs considered in the experimentation, without including additional information in the event log to identify message-type tasks.

In this way, the proposal to discover the choreography of a process in an inter-organizational collaboration environment is governed by a set of configurable methods. For example, the values of the variables U_t and U_a allow filtering cases with similar information and selecting events that are potentially considered message-type tasks, respectively. The *word embedding* representation used to calculate cosine similarity at the case and event level is highly effective. However, it is limited to the quantity and quality of information within the event logs, to generate a robust model that allows the discovery of the choreography between the participants of an IOBP. Furthermore, the patterns established in Table 1, as well as the models of intra-organizational processes discovered by the Split-miner algorithm, were fundamental elements for identifying the subtype in the message-type tasks, enabling the discovery of the choreography of the process.

According to the results obtained in the evaluation of the discovery of the choreography of the process (see Tables 3 and 4), the following classification can be defined based on the characteristics identified in the experimentation:

- **Complete choreography**. This refers to the fact that the proposed approach can find the complete process choreography in an inter-organizational environment, with the same number and relationships as found in the reference choreography. In conclusion, case-level and event-level representations of the event log only allow the discovery of message-type events.
- **Under-complete choreography**. This refers to an approach that has the ability to find a percentage of the relationships in the choreography of the process. This situation may be because there was insufficient information for the representation obtained from the word embedding model to obtain a high similarity, making it difficult to relate all message-type events.
- **Over-complete choreography**. This refers to the fact that the method finds part of the process's choreography but also recovers irrelevant relationships not found in the reference choreography. This behavior is because the information used to obtain the representation from the event logs is very general. The above issue causes more relationships to be recovered than the existing ones and to meet the condition that the calculated similarity value exceeds the threshold U_a.
- **Partially correct choreography**. This refers to the model identifying a percentage of the process choreography correctly. In addition, with the ability to find partially correct relationships; that is, in a relationship of two identified events (a, b'), an event a or b' is incorrect in the relationship, due to the relationship that is expected to be recovered, according to the reference model, whether (a, c') or (e, b'). The above may be because the identified relationship (a, b') has a higher degree of similarity than the expected relationship (a, c') or (e, b'). This behavior is caused by the fact that the information used to obtain the representation is not sufficiently discriminating to separate the relationships correctly and that the word-embedding model did not correctly learn from the information in the event logs, causing the generation of relationships with high similarity between message-type tasks and other event-types.

In the experimentation carried out, the scenario with the greatest complexity in identifying message-type tasks was *Healthcare*, according to the weighted value of 0.54 in the *FsR* metric. In the *Helthcare* scenario, the process choreographies discovered had the characteristics of a *partially correct choreography* and a *under-complete choreography*. In the *Air Quality System* and *Travel Agency* scenarios, the process choreographies were classified as a *complete choreography*, indicating that the information in the event logs, as well as the patterns defined in the proposed methodology, supported the construction of a process choreography similar to the expected one. Moreover, in the scenarios *Purchase order* and *Manufacturing process*, process choreographies were generated with characteristics of

Complete choreography and *Over-complete choreography*, which indicates that the complete choreography was recovered but relationships that were not part of the choreography were also recovered, as seen in the *RP* metric of 0.83 and 0.91, respectively. Finally, in the *Transfer of goods* scenario, choreographies with characteristics of *over-complete choreography* and *under-complete choreography* were obtained, which were reflected in the *PR* and *RR* metrics, indicating that true relations and relations that were not part of the choreography of the process were recovered.

7. Conclusions

This paper describes a data-driven methodology for discovering inter-organizational business processes (IOBP) and process choreography. The methodology comprises a set of methods and rules that allow information analysis from event logs and intra-organizational models generated by each participant involved in a collaboration. The above enabled the generation of the knowledge necessary for constructing an IOBP model and the interaction between participants through message flow connectors, facilitating the discovery of the process choreography. The results demonstrated the effectiveness of the methods and rules for discovering the choreography of the process, and generating the IOBP models obtained high values in the quality metrics, verifying the ability of the approach to faithfully reproduce the behaviors found in the merged event logs. In summary, we contribute to the process mining domain with formal methods that identify message-type tasks without requiring information to be added to the event log. We provide a set of rules that support defining message task subtypes. Additionally, we contribute a method for merging intra-organizational event logs, which enables the creation of an inter-organizational event log. Finally, three measurement indicators derived from the precision, recall, and f-score metrics are provided to evaluate the quality of the process choreography discovered.

Finally, in future work, the thresholds U_a and U_t will be optimized, since the proposed solution is parametric and the assigned values are individually functional through the analyzed collaboration. Furthermore, we plan to implement a tool that supports the proposed approach as a complement to the ProM process mining framework, as well as incorporating into our tool the Inductive Miner [51] and the Evolutionary Tree Miner [52] procedural algorithms for business process model discovery, which are based on the extraction of process trees from the event log.

Author Contributions: Conceptualization, J.D.H.-R. and E.T.-L.; methodology, J.D.H.-R. and E.T.-L.; software, J.D.H.-R.; validation, J.D.H.-R., E.T.-L. and M.S.; formal analysis, J.D.H.-R., E.T.-L. and M.S.; investigation, J.D.H.-R., E.T.-L. and M.S.; resources, E.T.-L.; data curation, J.D.H.-R.; writing—original draft preparation, J.D.H.-R. and E.T.-L.; writing—review and editing, J.D.H.-R., E.T.-L. and M.S.; visualization, J.D.H.-R.; supervision, E.T.-L. and M.S.; project administration, E.T.-L.; funding acquisition, E.T.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Universidad Autónoma de Tamaulipas (México). In addition, this work was supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT) of México under grant number 719519.

Data Availability Statement: The event logs presented in this study are available in the following research papers [31,46–50].

Acknowledgments: The authors are grateful to the Autonomous University of Tamaulipas, Mexico, for supporting this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BPMN	Business process model and notation
VRM	Vector representation matrix
SCP	Set of case pair
ESM	Event score matrix
PME	Pair of message-type events
DFG	Directly-follows graph
IOBP	Inter-organizational business process
RIOBP	Reference inter-organizational business process
RP	Relationship precision
RR	Relationship recall
FsR	F-score of the relationship
IAQM	IoT air quality monitor
RMS	Repository management service
SAS	System access service

References

1. Gutiérrez, B.R.; Quintero, A.M.R.; Parody, L.; López, M.T.G. When business processes meet complex events in logistics: A systematic mapping study. *Comput. Ind.* **2023**, *144*, 103788. [CrossRef]
2. Khan, I.S.; Kauppila, O.; Fatima, N.; Majava, J. Stakeholder interdependencies in a collaborative innovation project. *J. Innov. Entrep.* **2022**, *11*, 38. [CrossRef]
3. Bazan, P.; Estevez, E. Industry 4.0 and business process management: State of the art and new challenges. *Bus. Process. Manag. J.* **2022**, *28*, 62–80. [CrossRef]
4. Rafiei, M.; Van Der Aalst, W.M. An Abstraction-Based Approach for Privacy-Aware Federated Process Mining. *IEEE Access* **2023**, *11*, 33697–33714. [CrossRef]
5. Shi, J.; Jiang, Z.; Liu, Z. Digital Technology Adoption and Collaborative Innovation in Chinese High-Speed Rail Industry: Does Organizational Agility Matter? *IEEE Trans. Eng. Manag.* **2024**, *71*, 4322–4335. [CrossRef]
6. Wang, C.; Hu, Q. Knowledge sharing in supply chain networks: Effects of collaborative innovation activities and capability on innovation performance. *Technovation* **2020**, *94–95*, 102010. [CrossRef]
7. Fernandes, J.; Reis, J.; Melão, N.; Teixeira, L.; Amorim, M. The Role of Industry 4.0 and BPMN in the Arise of Condition-Based and Predictive Maintenance: A Case Study in the Automotive Industry. *Appl. Sci.* **2021**, *11*, 3438. [CrossRef]
8. Ribeiro, V.; Barata, J.; da Cunha, P.R. Modeling Boundary-Spanning Business Processes in Industry 4.0: Incorporating Risk-Based Design. In *Advances in Information Systems Development: Crossing Boundaries between Development and Operations in Information Systems*; Insfran, E., González, F., Abrahão, S., Fernández, M., Barry, C., Lang, M., Linger, H., Schneider, C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 143–162. [CrossRef]
9. Erasmus, J.; Vanderfeesten, I.; Traganos, K.; Grefen, P. Using business process models for the specification of manufacturing operations. *Comput. Ind.* **2020**, *123*, 103297. [CrossRef]
10. Czvetkó, T.; Kummer, A.; Ruppert, T.; Abonyi, J. Data-driven business process management-based development of Industry 4.0 solutions. *CIRP J. Manuf. Sci. Technol.* **2022**, *36*, 117–132. [CrossRef]
11. Bernabei, M.; Eugeni, M.; Gaudenzi, P.; Costantino, F. Assessment of Smart Transformation in the Manufacturing Process of Aerospace Components Through a Data-Driven Approach. *Glob. J. Flex. Syst. Manag.* **2023**, *24*, 67–86. [CrossRef]
12. Chauhan, A.; Kaur, H.; Mangla, S.K.; Kayikci, Y. Data driven flexible supplier network of selfcare essentials during disruptions in supply chain. In *Annals of Operations Research*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–31. [CrossRef]
13. Jans, M.; Laghmouch, M. Process Mining for Detailed Process Analysis. In *Advanced Digital Auditing: Theory and Practice of Auditing Complex Information Systems and Technologies*; Berghout, E., Fijneman, R., Hendriks, L., de Boer, M., Butijn, B.J., Eds.; Springer International Publishing: Cham, Switzerland, 2023; pp. 237–256. [CrossRef]
14. Chapela-Campa, D.; Dumas, M. From process mining to augmented process execution. *Softw. Syst. Model.* **2023**, *22*, 1977–1986. [CrossRef]
15. Camargo, M.; Dumas, M.; González-Rojas, O. Automated discovery of business process simulation models from event logs. *Decis. Support Syst.* **2020**, *134*, 113284. [CrossRef]
16. Van Der Aalst, W. *Process Mining: Data Science in Action*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 2. [CrossRef]
17. Zerbino, P.; Stefanini, A.; Aloini, D. Process science in action: A literature review on process mining in business management. *Technol. Forecast. Soc. Chang.* **2021**, *172*, 121021. [CrossRef]
18. Berti, A.; Schuster, D.; van der Aalst, W.M.P. Abstractions, Scenarios, and Prompt Definitions for Process Mining with LLMs: A Case Study. In *International Conference on Business Process Management*; De Weerdt, J., Pufahl, L., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 427–439. [CrossRef]

19. Tavares, G.M.; Oyamada, R.S.; Barbon, S.; Ceravolo, P. Trace encoding in process mining: A survey and benchmarking. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107028. [CrossRef]
20. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22–24 June 2014; Xing, E.P., Jebara, T., Eds.; Volume 32, pp. 1188–1196.
21. De Koninck, P.; vanden Broucke, S.; De Weerdt, J. act2vec, trace2vec, log2vec, and model2vec: Representation Learning for Business Processes. In *Business Process Management: 16th International Conference, BPM 2018, Sydney, NSW, Australia, 9–14 September 2018*; Weske, M., Montali, M., Weber, I., vom Brocke, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 305–321. [CrossRef]
22. Luettgen, S.; Seeliger, A.; Nolle, T.; Mühlhäuser, M. Case2vec: Advances in Representation Learning for Business Processes. In *International Conference on Process Mining*; Leemans, S., Leopold, H., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 162–174. [CrossRef]
23. Peeperkorn, J.; vanden Broucke, S.; De Weerdt, J. Conformance Checking Using Activity and Trace Embeddings. In *Business Process Management Forum: BPM Forum 2020, Seville, Spain, 13–18 September 2020, Proceedings 18*; Fahland, D., Ghidini, C., Becker, J., Dumas, M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 105–121. [CrossRef]
24. Augusto, A.; Conforti, R.; Dumas, M.; La Rosa, M.; Polyvyanyy, A. Split miner: Automated discovery of accurate and simple business process models from event logs. *Knowl. Inf. Syst.* **2019**, *59*, 251–284. [CrossRef]
25. Augusto, A.; Conforti, R.; Dumas, M.; La Rosa, M.; Maggi, F.M.; Marrella, A.; Mecella, M.; Soo, A. Automated discovery of process models from event logs: Review and benchmark. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 686–705. [CrossRef]
26. Dunzer, S.; Stierle, M.; Matzner, M.; Baier, S. Conformance checking: A state-of-the-art literature review. In Proceedings of the 11th International Conference on Subject-Oriented Business Process Management, Seville, Spain, 26–28 June 2019; pp. 1–10. [CrossRef]
27. Cherni, J.; Martinho, R.; Ghannouchi, S.A. Towards Improving Business Processes based on preconfigured KPI target values, Process Mining and Redesign Patterns. *Procedia Comput. Sci.* **2019**, *164*, 279–284. [CrossRef]
28. Dumas, M.; La Rosa, M.; Mendling, J.; Reijers, H.A. *Fundamentals of Business Process Management*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
29. Ladleif, J.; Weske, M. A legal interpretation of choreography models. In *Business Process Management Workshops: BPM 2019 International Workshops, Vienna, Austria, 1–6 September 2019, Revised Selected Papers 17*; Springer International Publishing: Cham, Switzerland, 2019; pp. 651–663. [CrossRef]
30. Bala, S.; Mendling, J.; Schimak, M.; Queteschiner, P. Case and activity identification for mining process models from middleware. In Proceedings of the IFIP Working Conference on The Practice of Enterprise Modeling, Vienna, Austria, 31 October–2 November 2018; pp. 86–102. [CrossRef]
31. Corradini, F.; Re, B.; Rossi, L.; Tiezzi, F. A Technique for Collaboration Discovery. In *Proceedings of the International Conference on Business Process Modeling, Development and Support, International Conference on Evaluation and Modeling Methods for Systems Analysis and Development*; Springer International Publishing: Cham, Switzerland, 2022; pp. 63–78. [CrossRef]
32. Zeng, Q.; Duan, H.; Liu, C. Top-down process mining from multi-source running logs based on refinement of Petri nets. *IEEE Access* **2020**, *8*, 61355–61369. [CrossRef]
33. Elkoumy, G.; Fahrenkrog-Petersen, S.A.; Dumas, M.; Laud, P.; Pankova, A.; Weidlich, M. Secure Multi-party Computation for Inter-organizational Process Mining. In *Enterprise, Business-Process and Information Systems Modeling*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 166–181. [CrossRef]
34. Corradini, F.; Muzi, C.; Re, B.; Rossi, L.; Tiezzi, F. Animating multiple instances in BPMN collaborations: From formal semantics to tool support. In Proceedings of the International Conference on Business Process Management, Sydney, NSW, Australia, 9–14 September 2018; pp. 83–101. [CrossRef]
35. López-Pintado, O.; Dumas, M.; García-Bañuelos, L.; Weber, I. Interpreted execution of business process models on blockchain. In Proceedings of the 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC), Paris, France, 28–31 October 2019; pp. 206–215. [CrossRef]
36. Köpke, J.; Franceschetti, M.; Eder, J. Optimizing data-flow implementations for inter-organizational processes. *Distrib. Parallel Databases* **2019**, *37*, 651–695. [CrossRef]
37. Mo, Q.; Song, W.; Dai, F.; Lin, L.; Li, T. Development of collaborative business processes: A correctness enforcement approach. *IEEE Trans. Serv. Comput.* **2019**, *15*, 752–765. [CrossRef]
38. Klinkmüller, C.; Ponomarev, A.; Tran, A.B.; Weber, I.; Aalst, W.v.d. Mining blockchain processes: Extracting process mining data from blockchain applications. In Proceedings of the International Conference on Business Process Management, Vienna, Austria, 1–6 September 2019; pp. 71–86. [CrossRef]
39. Corradini, F.; Pettinari, S.; Re, B.; Rossi, L.; Tiezzi, F. A technique for discovering BPMN collaboration diagrams. In *Software and Systems Modeling*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21. [CrossRef]
40. Peña, L.; Andrade, D.; Delgado, A.; Calegari, D. An Approach for Discovering Inter-organizational Collaborative Business Processes in BPMN 2.0. In *Process Mining Workshops*; De Smedt, J., Soffer, P., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 487–498. [CrossRef]
41. Tajima, K.; Du, B.; Narusue, Y.; Saito, S.; Iimura, Y.; Morikawa, H. Step-by-Step Case ID Identification Based on Activity Connection for Cross-Organizational Process Mining. *IEEE Access* **2023**, *11*, 60578–60589. [CrossRef]

42. Liu, C.; Li, H.; Zhang, S.; Cheng, L.; Zeng, Q. Cross-Department Collaborative Healthcare Process Model Discovery From Event Logs. *IEEE Trans. Autom. Sci. Eng.* **2023**, *20*, 2115–2125. [CrossRef]
43. Bernardi, M.L.; Cimitile, M.; Mercaldo, F. Cross-Organisational Process Mining in Cloud Environments. *J. Inf. Knowl. Manag.* **2018**, *17*, 1850014. [CrossRef]
44. Buijs, J. Flexible Evolutionary Algorithms for Mining Structured Process Models. Ph.D. Thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, April 2014. [CrossRef]
45. Janssenswillen, G.; Donders, N.; Jouck, T.; Depaire, B. A comparative study of existing quality measures for process discovery. *Inf. Syst.* **2017**, *71*, 1–15. [CrossRef]
46. Hernandez-Resendiz, J.D.; Tello-Leal, E.; Ramirez-Alcocer, U.M.; Macías-Hernández, B.A. Semi-Automated Approach for Building Event Logs for Process Mining from Relational Database. *Appl. Sci.* **2022**, *12*, 10832. [CrossRef]
47. Hernandez-Resendiz, J.D.; Tello-Leal, E.; Marin-Castro, H.M.; Ramirez-Alcocer, U.M.; Mata-Torres, J.A. Merging Event Logs for Inter-organizational Process Mining. In *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 3–26. [CrossRef]
48. Liu, C.; Duan, H.; Qingtian, Z.; Zhou, M.; Lu, F.; Cheng, J. Towards comprehensive support for privacy preservation cross-organization business process mining. *IEEE Trans. Serv. Comput.* **2016**, *12*, 639–653. [CrossRef]
49. Fdhila, W.; Rinderle-Ma, S.; Knuplesch, D.; Reichert, M. Change and Compliance in Collaborative Processes. In Proceedings of the 2015 IEEE International Conference on Services Computing, New York, NY, USA, 27 June–2 July 2015; pp. 162–169. [CrossRef]
50. Borkowski, M.; Fdhila, W.; Nardelli, M.; Rinderle-Ma, S.; Schulte, S. Event-based failure prediction in distributed business processes. *Inf. Syst.* **2019**, *81*, 220–235. [CrossRef]
51. Leemans, S.J.J.; Fahland, D.; van der Aalst, W.M.P. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In *Business Process Management Workshops: BPM 2013 International Workshops, Beijing, China, 26 August 2013, Revised Papers 11*; Lohmann, N., Song, M., Wohed, P., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 66–78. [CrossRef]
52. Buijs, J.C.A.M.; van Dongen, B.F.; van der Aalst, W.M.P. Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity. *Int. J. Coop. Inf. Syst.* **2014**, *23*, 1440001. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.