IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 May 2023, accepted 29 May 2023, date of publication 8 June 2023, date of current version 21 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284053

RESEARCH ARTICLE

Step-by-Step Case ID Identification Based on Activity Connection for Cross-Organizational Process Mining

KAZUKI TAJIMA[1], BOJIAN DU[1], YOSHIAKI NARUSUE[1], (Member, IEEE), SHINOBU SAITO[2], YUKAKO IIMURA[2], AND HIROYUKI MORIKAWA[1], (Member, IEEE)

[1]Graduate School of Engineering, The University of Tokyo, Tokyo 113-8654, Japan
[2]Computer and Data Science Laboratories, NTT Corporation, Tokyo 108-8585, Japan

Corresponding author: Kazuki Tajima (den.maharuda.030405@gmail.com)

ABSTRACT Cross-organizational process mining aims to discover an entire process model across multiple organizations where their identifier (ID) systems are not managed uniformly, and each organization has an independent ID system. Cross-organizational process mining has been gaining popularity as information systems increase in complexity. However, previous methods have limitations in that they do not work well for event logs that contain only common items, or cyclic orchestrations, which indicates that the model contains loops. In this paper, we propose an accurate cross-organizational process mining technique based on a step-by-step case ID identification mechanism that uses only common items in event logs and can handle cyclic orchestrations. Step-by-step case ID identification repeats the following steps: 1) identification of case IDs based on activity connection of adjacent event pairs, and 2) extraction of additional activity connections by leveraging the newly identified case IDs. We alternately identify the most probable case ID pairs and remove events belonging to these identified case IDs from the event log, which contributes to extracting additional activity connections and narrowing down the candidates of case ID pairs. Evaluation using real-world event logs showed that the proposed method generates the process model with more than 98.4% precision and more than 94.2% recall for two datasets, outperforming previous methods.

INDEX TERMS Process mining, cross-organizational process mining, integrating event logs, identifying case IDs.

I. INTRODUCTION

Process mining is a technology that visualizes a business process from an event log generated by information systems like e-commerce systems. It is essential in information system digital transformation (DX). The insights provided by process mining make the operation process more transparent and efficient, and help in the governance within ESG. Process mining outperforms manual process visualization in terms of cost, speed, comprehensiveness, and objectivity [1]. Recently, process mining has been used not only for improving operations by understanding a process model [2], [3] but also for the development of new information systems and business process outsourcing [4]. It has steadily been used in practical applications [5], [6], [7].

Cross-organizational process mining, a technology for discovering process models that span multiple systems, is becoming increasingly appealing as information systems evolve. In cross-organizational process mining, case identifiers (IDs) assigned to events in the same trace vary from organization to organization, that is, each organization has its own ID system, making inappropriate to use general process mining techniques as is. It is recognized as a pressing issue and is even positioned as an ''important challenge that needs to be addressed'' in the Process Mining Manifesto [3].

Much of the previous research on cross-organization process mining [2], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] uses items that are not necessarily recorded in

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

the common event log, even though the only items that are always included in the event log are the timestamp, case ID, and activity name [20], [21], [22]. Bayomie et al. proposed event-case correlation methods that use activity names and timestamps in addition to an underlying process model as an input [23]. The methods following the above method also require an underlying process model as an input [24], [25]. Pourmirza et al. proposed more general methods [20], [21] based only on timestamps and activity names. However, these two methods have some limitations. First, they do not work well for a cyclic orchestration, which means that the model contains loops. Second, threshold values must be set by human operators. Further, there is room for performance improvement.

To address these issues, we propose step-by-step case ID identification based on activity connection, which is 1) highly accurate, 2) based on only three common items, that is, timestamps, case IDs, and activity names in an event log, and 3) applicable to cyclic orchestrations. Instead of proposing a new type of process model or process mining technique, we converted a cross-organizational process mining problem into a single-organization process mining problem by identifying case IDs and integrating event logs. This study assumes event logs are complete and error-free. However, even in cases where this assumption does not hold, event logs generated by our proposed method can be applied to process mining techniques [26], [27] for a single organization to generate a sound process model.

In order to achieve cross-organizational process mining from common items, the following steps are repeated step-by-step: 1) identification of case IDs based on the activity connection of adjacent event pairs, and 2) extraction of additional activity connections by leveraging the newly identified case IDs. In other words, we alternately identify the most probable case ID pairs and remove events belonging to these identified case IDs from the event log. This makes the additional activity connections extracted and the candidates of case ID pairs narrowed down. In addition, by further extracting new activity connections and time differences based on identified case IDs, we can identify case IDs under more reliable conditions.

Our study primarily focuses on horizontal cross-organizational process mining and assumes that event logs have no missing or incorrect data. Our method is built on the assumption that the activities sharing the same case ID tend to have a smaller time difference than those with different case IDs. The proposed method can be applied to processes where this assumption holds. Additionally, our method uses only three common attributes to determine the presence or absence of a connection between activities: case ID, activity name, and timestamp. Hence, our approach is applicable to heterogeneous event logs, provided the attributes corresponding to the case ID, activity name, and timestamp are known for each log.

We evaluated our proposal on the dataset used in the Business Process Intelligence Challenge (BPIC) 2012 [28] and 2017 [29], which are the event logs of loan applications of Dutch financial institutions. Correct process models are not provided in these datasets, but unified case IDs are. In this study, we generate a process model based on the unified case IDs by using Disco [30], which is a software for general process mining provided by Fluxicon. Additionally, we regard the generated models as a baseline for evaluation, which is the same as previous research [21]. The evaluation results show that our method outperforms previous research by generating the process model with more than 98.4% precision and more than 94.2% recall for both datasets. Our code is publicly available at https://github.com/maharu-39/step-by-step_case_id_identifier.

The rest of this paper is structured as follows: In Section II, we summarize previous research. In Section III, we explain the basic concepts of process mining. In Section IV, we propose a new case ID identification mechanism to achieve cross-organizational process mining. In Section V, we apply our approach to real-world event logs and evaluate their performance. We also compare the obtained results with those of previous research [20], [21] and discuss the effectiveness of our method. In Section VI, we conclude this study and present a roadmap for future work.

II. RELATED WORKS

A. CROSS-ORGANIZATIONAL PROCESS MINING WITH ADDITIONAL ITEMS

Q. Zeng et al. proposed a method based on the data structure called RM_WF_Net, which uses resource information and message information [8]. This method has been extended to consider privacy [9] and heterogeneous relationships between organizations [10]. Recently, methods for jointing process models for an entire process model using information about the messages exchanged via communication activities are proposed [2], [11]. However, these methods can only be effective in information systems that specifically record resource and message information in the event log. In addition, some proposed methods apply artificial immune systems [12], [13], genetic algorithms [14], rule-based algorithms [15], [16], linguistic processing of event log entries [17], [18], and the similarity function between events [19]. However, the above methods are restricted in that they cannot function well when the event logs lack additional items. The event-case correlation method is proposed by Bayomie et al [23]. However, it requires an underlying process model as an input, and the following methods [24], [25] have the same drawback.

B. CROSS-ORGANIZATIONAL PROCESS MINING WITH ONLY COMMON ITEMS

S. Pourmirza et al. proposed correlation mining [20] which aims to visualize process models across organizations based on only timestamp and activity names. Several types of consistent process models are first generated, with the condition that the number of cases that flow into and out of an activity be equal to the number of events that occur for that activity, except for the process model's starting and ending points. Next, the most appropriate process model is selected using

two indices called the P/S matrix and Duration matrix. These matrices are square matrices of order n, where n represents the number of the kind of activities. P/S_{uv} denotes the ratio of the cases where events from Activity u occur before events from Activity v. The larger P/S_{uv} is, the more likely it is that u and v are connected, that is, in the same trace. D_{uv} represents the average time difference between u and v. If D_{uv} is small, u and v are likely to be connected.

Moreover, S. Pourmirza et al. extended correlation mining by proposing a correlation miner [21], which not only visualizes the process model but also identifies case IDs, that is, match a certain case ID in one organization to a certain case ID in another.

However, these two methods have some limitations. First, indicators do not work well for cyclic orchestrations, because P/S_{uv} focus on the overall relationship of u and v and don't function well. Furthermore, threshold values must be set by human operators. Further, there is room for improving performance.

III. PRELIMINARIES

A. EVENT LOGS

Data about the execution of a process is recorded as an event log $\mathcal{L}$ [31], and an event e_i is a record in an event log, where i represents an index. In this paper, i and j are used as indices in the event log. Each sequence of activities is referred to as a trace, and a unique ID is assigned as a case ID to identify the trace. It is assumed that e_i is described by at least the following three elements [20], [21], [22].

- *Activity name*: Name of the operation.
- *Case ID*: ID to identify the trace.
- *Timestamp*: Time when the activity was completed.

An example of an event log is shown in Table 1. Thus, an event e_i is expressed as follows:

$$e_i \in \mathcal{A} \times \mathcal{C} \times \mathcal{T}, \tag{1}$$

where $\mathcal{A}$ is the set of activities, $\mathcal{C}$ is the set of case IDs, and $\mathcal{T}$ is the set of timestamps.

The activity, case ID, and timestamp of an event e_i are represented as $e_i^{\mathcal{A}}$, $e_i^{\mathcal{C}}$, and $e_i^{\mathcal{T}}$, respectively. On the other hand, a certain activity like "send an e-mail" is represented as A_p, where p is an activity name. In this paper, p, q, and r are used as activity names. Similarly, a certain case ID like "40" is represented as C_x, where x is a case ID number. In this paper, x and y are used as case IDs. As can be seen, the activity name contains the contents of the operation, such as "send an e-mail". Table 1 contains two traces, named C_{40} and C_{50}. Here, C_{40} refers to the flow of $\langle Task1,\ Task2,\ Task4,\ Task5,\ Task3\rangle$, and C_{50} refers to the flow of $\langle Task1,\ Task2,\ Task3\rangle$.

In cross-organizational process mining, case IDs are assigned to events by each organization independently. In other words, it differs from general process mining in that case IDs are not unique. To distinguish such case IDs from common case IDs, we call case ID "local case ID"

TABLE 1. An example event log.

Timestamp	Activity Name	Case ID
2016-01-01 08:51:15	Task1: send an e-mail	40
2016-01-01 08:51:16	Task2: receive e-mail	40
2016-01-01 09:00:49	Task1: send an e-mail	50
2016-01-01 09:00:50	Task2: receive e-mail	50
2016-01-01 11:15:42	Task4: call incomplete files	40
2016-01-02 15:40:51	Task5: send incomplete files	40
2016-01-03 09:11:01	Task3: permission granted	50
2016-01-03 18:13:54	Task3: permission granted	40

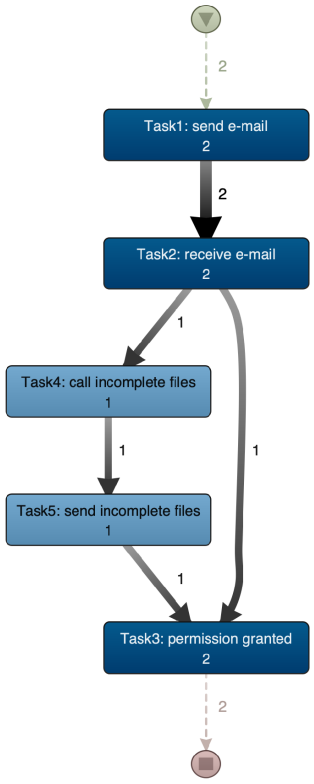

FIGURE 1. Process model generated from Table 1.

TABLE 2. An event log of organization α.

Timestamp	Activity Name	Local Case ID
2016-06-04 11:53:53	Task1: create application form	64
2016-06-04 11:53:58	Task2: send an e-mail	64
2016-06-04 12:00:49	Task1: create application form	31
2016-06-04 12:01:03	Task2: send an e-mail	31
2016-06-07 11:15:42	Task3: permission granted	64
2016-06-07 15:40:51	Task3: permission granted	64

in Table 2 and Table 3. In Table 2 and Table 3, organization α and β work together. Though the procedure is $\langle Task1,\ Task2,\ Task4,\ Task5,\ Task3\rangle$, case ID is not unique across organizations. Note that case IDs are not completely removed. Therefore, it is beyond the scope of our proposal that no case IDs are allocated, such as when they are extracted from non-process-aware information systems, or when case IDs are recorded in the log erroneously.

TABLE 3. An event log of organization β.

Timestamp	Activity Name	Local Case ID
2016-06-07 11:14:59	Task4: add materials	53
2016-06-07 11:15:40	Task5: send an e-mail	53
2016-06-07 15:40:01	Task4: add materials	92
2016-06-07 15:40:50	Task5: send an e-mail	92

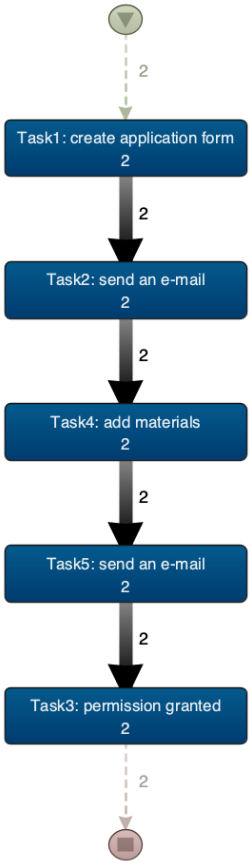

FIGURE 2. Process model generated from Table 2 and Table 3.

In this study, we show the log integrated by our proposal as follows:

$$\alpha + \beta, \tag{2}$$

$$(\alpha + \beta) + \gamma, \tag{3}$$

where α, β, γ are organization names. (2) means the log of α is integrated into the log of β. (3) means the log of α is first integrated into the log of β, and then the integrated logs of α and β are integrated into the log of γ.

B. PROCESS MODEL

A process model is the flow of all system operations obtained through process mining. It specifies which activities have a valid execution order [32]. There are three typical representations of process models: directly-follows graph [33], petri nets [34], and BPMN [35]. A directly-follows graph is the simplest representation of process models. In a directly-follows graph, each node represents an activity and each edge represents the flow. Petri net and BPMN are higher-level representations for the effective expression of process models. In Section V, we create a directly-follows graph from an integrated event log by Disco [30] to evaluate the effectiveness of our method compared with that of previously reported methods.

The example of a directly-follows graph created from the event logs in Table 1 is shown in Fig. 1 An appropriate process model should represent both traces. For example, in Fig. 1, the model represents both traces which are assigned C_{40} and C_{50}. Fig. 2 is a process model created from Table 2 and Table 3, which shows an entire process across organizations.

IV. APPROACH

This method aims to identify case IDs for logs recorded in different organizations. Process mining algorithms designed for a single organization can be applied to the integrated log created by case ID identification to visualize a process.

Our proposal focuses on two organizations at first. Then, to visualize a process that spans more than two organizations, an overall process model can be generated by repeating this method. The final model reflects processes that span all the organizations.

Fig. 3 shows an overview of the main case ID identification mechanism. Here, two organizations, named orgA and orgO, are used as examples. First, in the left section, two activities from different organizations that occur consecutively and within a short period are determined to be connected, that is, occurring within the same trace (Section IV-A). Next, in the middle, when activities that are considered to be connected occur consecutively, the case ID in one organization is matched to the case ID in another, that is, case IDs are identified (Section IV-B). Finally, by inspecting the process of the identified case IDs, the connection between the activities whose time difference is short is extracted (Section IV-C). The information obtained is displayed in the upper right corner. This information also helps identify the case IDs in the bottom right corner (Section IV-D). The case IDs are gradually identified by repeating the above steps for the unidentified event logs. The process is terminated when there are no more identifiable case IDs, that is, when all case IDs have been identified, or when the remaining case IDs cannot be identified with high reliability (Section IV-E). Furthermore, the above steps are performed with multiple thresholds; therefore, the appropriate thresholds are automatically set. Then, edges regarded as noise are removed (Section IV-F). When applied to more than two organizations, the bias of the integrated order is removed (Section IV-G).

A. IDENTIFICATION OF THE CONNECTION OF ACTIVITIES

In the first step, two activities from different organizations that occur consecutively and with a short time difference threshold named $th1$ are determined to be connected, that is,

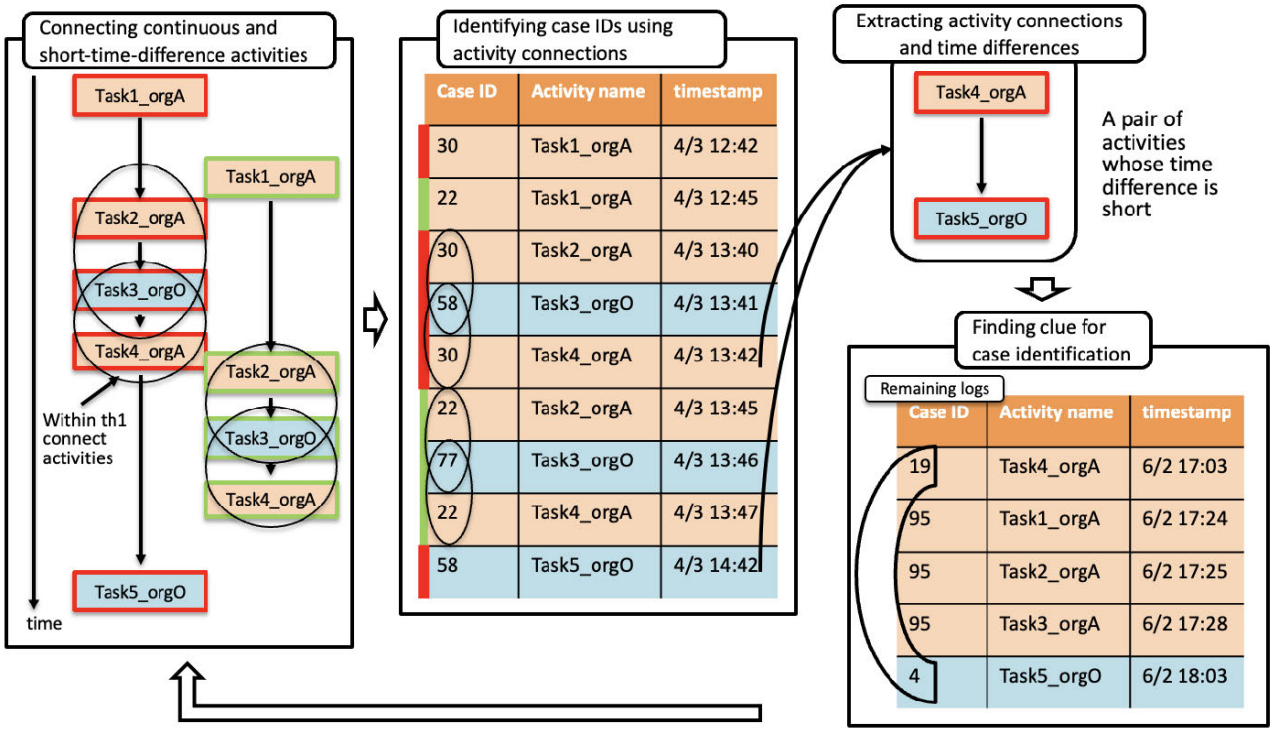

FIGURE 3. Overview of the main case ID identification mechanism (Section IV-A∼IV-E).

belonging to the same trace, to provide a highly reliable clue for case ID identification.

In detail, the first step is to combine the event logs of both organizations and sort them in time order. Next, if a pair of activities of adjacent events occur within $th1$ of each other, the number of occurrences is counted. The count of activity pairs under the above conditions is as follows:

$$N_1(A_pA_q) = \sum_{i=1}^{L-1} f_1(i; A_p, A_q), \tag{4}$$

$$f_1(i; A_p, A_q) = \begin{cases} 1 \text{ if } \begin{cases} e_{i+1}^{\mathcal{T}} - e_i^{\mathcal{T}} \leq th1 \\ e_i^{\mathcal{A}} = A_p \\ e_{i+1}^{\mathcal{A}} = A_q \end{cases} \\ 0 \text{ if } otherwise \end{cases}, \tag{5}$$

where $N_1(A_p, A_q)$ represents the count of $\langle A_p, A_q \rangle$, and L represents the number of events in a log. Pairs of activities counted that exceed a certain threshold named th2 are determined to be connected. This procedure corresponds to the leftmost part of Fig. 3, and two traces are displayed with a time axis. An orange square denotes organization A's activity and a light blue square denotes organization O's. Activities surrounded by lines of the same color belong to the same trace. Fig. 3 shows the connection, $\langle Task2\_orgA,\ Task3\_orgO \rangle$, $\langle Task3\_orgO,\ Task4\_orgA \rangle$ in the two traces.

When compared to previous methods [20], [21], this approach has the advantage of being applicable to cyclic orchestrations.

B. IDENTIFICATION OF CASE IDs

In the second step, if a pair of activities that are thought to be connected occur, the case IDs which are assigned to these activities' events are identified. If there is a connection between two activities, it means that the two events are in the same trace, so it is highly likely that the case IDs of the two events are the same.

As a specific procedure, in the sorted event logs created by integrating two event logs, if a pair of activities of adjacent events are connected, the number of pair of case IDs that are assigned to events of these activities is counted. The following formula expresses the number of case ID pairs under the aforementioned conditions:

$$N_2(C_xC_y) = \sum_{i=1}^{L-1} f_2(i; C_x, C_y), \tag{6}$$

$$f_2(i; C_x, C_y) = \begin{cases} 1 \text{ if } \begin{cases} (e_i^{\mathcal{A}}, e_{i+1}^{\mathcal{A}}) \in \mathcal{CN} \\ e_i^{\mathcal{C}} = C_x \\ e_{i+1}^{\mathcal{C}} = C_y \end{cases} \\ 0 \text{ if } otherwise \end{cases}, \tag{7}$$

where $N_2(C_x, C_y)$ represents the count of $\langle C_x, C_y \rangle$, and $\mathcal{CN}$ represents the set of activity pairs that are considered connected. Only the pairs of case IDs with a count greater than a certain threshold $th3$ are extracted as pairs to be identified. In the middle part of Fig. 3, the event log shows how "C_{30}, C_{58}" and "C_{22}, C_{77}" are identified to each other using the activity connections $\langle Task2\_orgA,\ Task3\_orgO \rangle$ and

$\langle Task3\_orgO,\ Task4\_orgA\rangle$ obtained in the leftmost part. In other words, $N_2(C_{30}, C_{58}) = 2$ and $N_2(C_{22}, C_{77}) = 2$.

In the proposed methodology, case IDs are identified by statistically analyzing the difference in timestamps between two activities. Hence, only the statistical trend in time differences impacts the identification of case IDs. This implies that a small number of errors in the order of event logs are insignificant in our methodology, making the order of event logs less crucial when integrating multiple logs into a single log.

If more than one case in an organization is identified to a certain case in the other organization, the pair with the greater number of occurrences is identified. For example, if $N_2(C_p, C_q) = 10$, $N_2(C_p, C_r) = 6$, then C_p and C_q are identified. C_r is not identified to anything. Events that are not identified remain in the event log.

C. EXTRACTION OF ACTIVITY CONNECTIONS AND TIME DIFFERENCES BASED ON IDENTIFIED CASE IDs

In the third step, we further extract information on activity connections by inspecting the process from the identified case IDs, assuming that the case IDs have been correctly identified.

The concrete procedure is to extract the event logs assigned to a set of identified case IDs and sort them in time order. For example, in the middle part of Fig. 3, C_{30} and C_{58} are identified. Therefore, they are considered the same ID and we can see the existence of a certain process $\langle Task1\_orgA,\ Task2\_orgA,\ Task3\_orgO,\ Task4\_orgA,\ Task5\_orgO\rangle$. Next, we refer to the consecutive activities in the inspected process, for example, $\langle Task4\_orgA,\ Task5\_orgO\rangle$ in Fig. 3. If they are not yet considered connected, the number of occurrences of a pair of activities is counted, and the time difference is also recorded as follows:

$$N_3(A_p, A_q) = \sum_{\mathcal{CI}} \sum_{\mathcal{AS}} f_3(i, j; A_p, A_q), \tag{8}$$

$$f_3(i, j; A_p, A_q) = \begin{cases} 1 \text{ if } \begin{cases} (e_i^{\mathcal{C}}, e_j^{\mathcal{C}}) \in \mathcal{CI} \\ \langle e_i^{\mathcal{A}}, e_j^{\mathcal{A}}\rangle \in \mathcal{AS} \\ e_i^{\mathcal{A}} = A_p \\ e_j^{\mathcal{A}} = A_q \end{cases} \\ 0 \text{ if } otherwise \end{cases}, \tag{9}$$

$$D(A_p, A_q) = \{e_j^{\mathcal{T}} - e_i^{\mathcal{T}} \mid f_3(i, j; A_p, A_q) = 1\}, \tag{10}$$

where $N_3(A_p, A_q)$ represents the count of $\langle A_p, A_q\rangle$, $D(A_p, A_q)$ represents the set of the time difference of $\langle A_p, A_q\rangle$, $\mathcal{CI}$ represents the set of the identified case IDs, $\mathcal{AS}$ represents the set of a pair of continuous activities in a certain trace. In Fig. 3, the connections between the activities $\langle Task2\_orgA, Task3\_orgO\rangle$ and $\langle Task3\_orgO, Task4\_orgA\rangle$ have already been found by the method described in Section IV-A (the leftmost of Fig. 3). However, this method discovers a connection $\langle Task4\_orgA, Task5\_orgO\rangle$ that has not yet been discovered (the upper right of Fig. 3).

After extracting information from all identified case IDs, activity pairs with a count greater than a certain threshold $th4$ are considered to be connected activities. The exception occurs when the time difference of more than 5% of newly found activity pairs is more than th1.

D. ASSISTANCE FOR IDENTIFICATION OF CASE IDs

In the fourth step, we use the connected activity pairs and their time difference obtained in IV-C to aid case ID identification in non-contiguous portions of the event log, as the method introduced in IV-B only identifies case IDs of adjacent events.

We examine the remaining event log in time order, beginning at the top. If the activity in the event log is the first activity of a pair of connected activities obtained in IV-C, the second activity is sought where the second activity probably occurred according to the common time difference between the first activity and the second activity. The number of occurrences of the case IDs in the first and second activities is counted as follows:

$$N_4(C_x, C_y) = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} f_4(i, j; C_x, C_y), \tag{11}$$

$$f_4(i, j; C_x, C_y) = \begin{cases} 1 \text{ if } \begin{cases} e_j^{\mathcal{T}} - e_i^{\mathcal{T}} < upper_{ij} \\ e_j^{\mathcal{T}} - e_i^{\mathcal{T}} > lower_{ij} \\ e_i^{\mathcal{C}} = C_x \\ e_j^{\mathcal{C}} = C_y \end{cases} \\ 0 \text{ if } otherwise \end{cases}, \tag{12}$$

where $N_4(C_x, C_y)$ represents the count of $\langle C_x, C_y\rangle$, $upper_{ij}$ represents the upper bounds of the 95% confidence interval of the time difference of connected activities, and $lower_{ij}$ represents the lower bounds. In Fig. 3, if $Task4\_org1$ and $Task5\_orgO$ exist in the remaining event log with a time difference within the 95% confidence interval, C_{19} and C_4 are identified. In other words, $N_4(C_{19}, C_4) = 1$. When identifying the case ID again, these counts are taken into account.

E. REPEATING OF THE PROCEDURES

The case IDs are gradually identified by repeating steps IV-A to IV-D for the remaining event logs that have not been identified. When there are no more identifiable case IDs, the process is terminated, implying that all case IDs have been identified or that the remaining case IDs cannot be identified with high reliability.

F. AUTOMATIC THRESHOLD SETTING AND NOISE REMOVAL

The issue with using thresholds is that the optimal threshold at which the method works well depends on the dataset used. Therefore, the threshold is determined by the experimenter's heuristic, which necessitates an experienced experimenter.

In this study, the threshold is set automatically based on the number of noise edges. First, several candidate thresholds are prepared. For each set of thresholds, case IDs are identified as described in IV-A∼IV-E. The process model is generated

based on the identified case IDs, and the edges between two activities are calculated. The edge between activities that occurs less than 1% of the number of the total case IDs is considered noise. The number of such edges (noise) is counted. The threshold set with the fewest number of such edges is adopted. This is based on the assumption that the more infrequent edges, the less likely is the accuracy of the process model.

The process model finally generated also removes edges that could be considered noise, and only the essential processes are extracted. Correct edges that occur infrequently may be removed as well. However, the purpose of process mining is to find mainstream flows, so removing infrequent processes as noise is considered quite natural.

G. ELIMINATION OF BIAS DUE TO INTEGRATION ORDER

When three or more organizations are targeted, the order of integration produces different results. The reason is that if less connected organizations are integrated first, only a small number of case IDs are identified and the many remaining logs are discarded. For example, in the condition that there are three organizations α, β, γ, and α has few connections to β, if we integrate α and β first, a large amount of case IDs will not be identified and discarded.

We eliminate the bias by integrating the multiple results. After identifying case IDs and setting appropriate thresholds, the number of edges obtained in each integration order is summed without removing noise. Then, after summing up, edges that are less than "1% of the total number of case IDs $\times$ the number of integration patterns", are considered noise and removed. This improves the robustness of the method.

V. EVALUATION

To verify the effectiveness of the proposed method, we evaluate this approach using real-world event logs.

A. EVENT LOGS

The BPIC 2012 and 2017 datasets, which are both obtained from Dutch financial institutions, are used. BPIC 2012 is the event log for personal loans or overdraft application processes with 13087 case IDs and 262200 events from October 1, 2011, to March 14, 2012. With 31509 case IDs and 1202267 events, the BPIC 2017 dataset is the event log for loan applications from January 1, 2016, to February 2, 2017. These event logs consist of three sub-processes, organizations are named A, O, and W, and each activity name has an ID at the beginning to indicate the sub-process in which it was recorded. For example, an activity name of the work performed in sub-process A is "A_Create application". The case IDs identified for the whole organization are given as the correct answer. These case IDs are not used in our proposed process mining procedure.

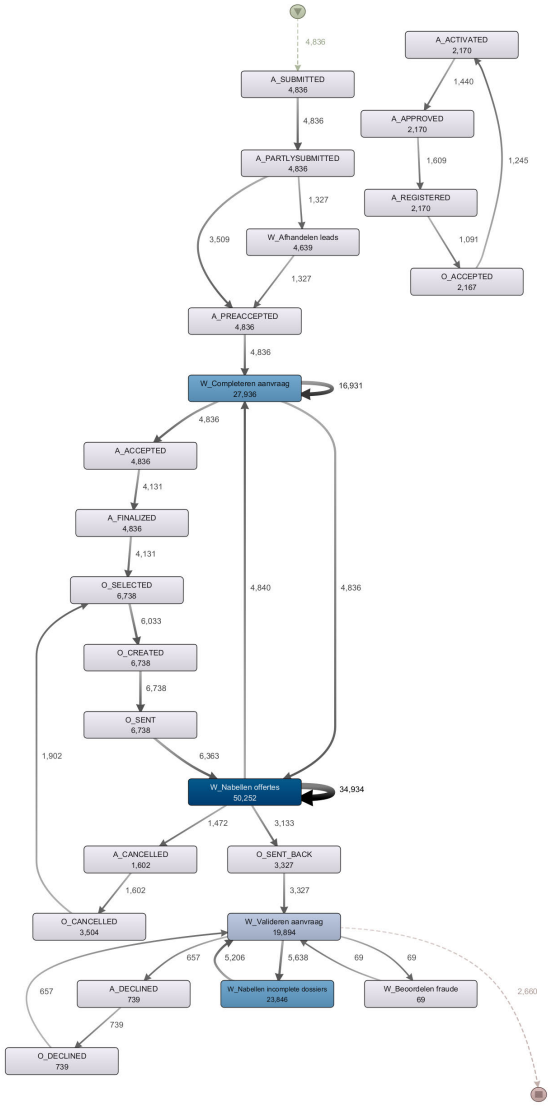

FIGURE 4. BPIC 2012 dataset process model.

B. EXPERIMENTAL SETUP

First, the event logs are divided into three sections so that each organization can be considered mutually independent. The proposed method then integrates them and compares them to the correct answers. The performance of the proposed method on the BPIC 2012 dataset was compared with that of previously reported methods [20], [21]. In addition, we evaluated our method on the BPIC 2017 dataset as well. We used the set of candidate threshold values as follows:

$$th1 \in \{1s, 10s, 100s, 1000s, 10000s\}, \quad (13)$$

$$th2 \in \{1, 4, 16, 64, 256, 1000, 4000, 16000\}, \quad (14)$$

$$th3 \in \{0, 1, 2, 3, 4, 6, 8\}, \quad (15)$$

$$th4 \in \{1, 4, 16, 64, 256, 1000, 4000, 16000\}. \quad (16)$$

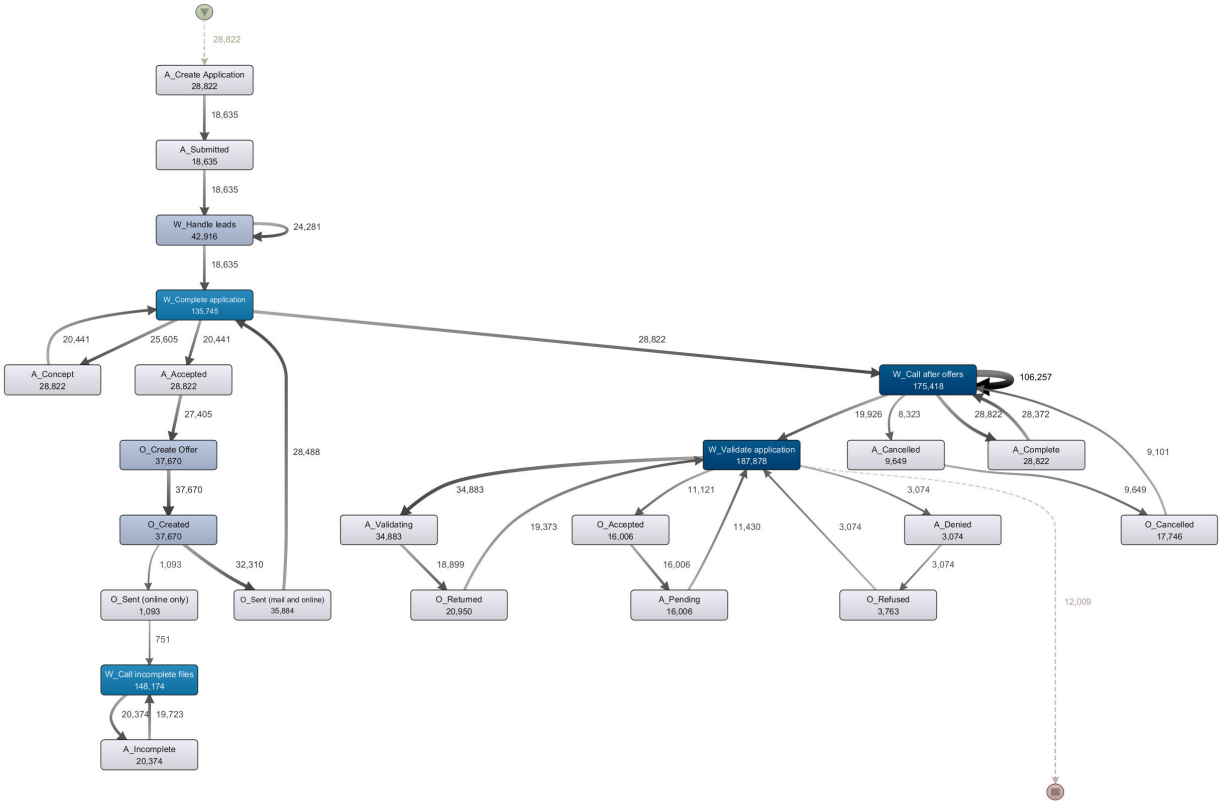

FIGURE 5. BPIC 2017 dataset process model.

In this experiment, precision and recall were calculated for the two aspects of how accurately the case IDs were identified and how accurately the process models were generated. In order to evaluate how accurately the process model was generated, we generated the process models in a directly-follows graph from integrated event logs by Disco, similar to previous research [20], [21].

The definitions of precision and recall are as follows:

$$Precision = \frac{TP}{(TP + FP)}, \tag{17}$$

$$Recall = \frac{TP}{(TP + FN)}, \tag{18}$$

where TP means True Positives, FP means False Positives, and FN means False Negatives. For the evaluation of the case ID identification, TP is the number of correctly identified case IDs, FP is the number of incorrectly identified case IDs, and FN is the number of not identified case IDs. For the evaluation of the generated process model, TP is the number of correct edges generated by the proposed method, FP is the number of incorrect edges generated by the proposed method, and FN is the number of correct edges not generated by the proposed method. $F1$ is an index that summarizes the two indices of precision and recall as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{19}$$

When comparing process models, the edges of a correct process model that occur for less than 1% of the total number of case IDs are removed as noise. For the BPIC 2012 dataset, the number of case IDs which span all organizations in A+W+O is 5015. Therefore, the reference value for noise removal of our proposal (50.15) is almost the same as that of previous methods (50) [20], [21].

The organizations listed in the leftmost column of the table indicate integrated organizations. The $(A+O+W)_{sum}$ row displays the results of removing bias due to integration order. Three significant digits are assumed.

Finally, we compare the process models produced by the proposed approach to the correct process models visualized by Disco. We remove whole traces that contain edges considered as noise to load the event log to Disco and Disco's setting is 100% for Activity and 0% for Path. This setting displays all activities, but flows with a small number of appearances are automatically removed as noise. The process models visualizing the respective data are shown in Fig. 4 and Fig. 5.

TABLE 4. Results of identifying case IDs using the BPIC 2012 dataset.

Method	Organization	Precision	Recall	F1
Proposed method	A+O	1.000	0.930	0.964
	A+W	1.000	0.518	0.682
	O+W	1.000	0.271	0.426
	(A+O)+W	1.000	0.930	0.964
	(A+W)+O	1.000	0.907	0.951
	(O+W)+A	0.999	0.271	0.426
Correlation miner [21]	A+O+W	0.70	0.65	0.674
Correlation miner (noise removed) [21]	A+O+W	0.74	0.71	0.725

TABLE 5. Results of generated process models using the BPIC 2012 dataset.

Method	Organization	Precision	Recall	F1
Proposed method	A+O	1.000	1.000	1.000
	A+W	1.000	0.867	0.929
	O+W	1.000	0.892	0.943
	(A+O)+W	0.984	1.000	0.992
	(A+W)+O	0.984	1.000	0.992
	(O+W)+A	1.000	0.905	0.950
	$(A+O+W)_{sum}$	0.984	0.984	0.984
Correlation miner [20]	A+O+W	0.85	0.63	0.724
Correlation miner (noise removed) [20]	A+O+W	0.79	0.68	0.731

C. EVALUATION USING THE BPIC 2012 DATASET

Table 4 shows the results of the case ID identification and Table 5 shows the results of the process model when the experiment was conducted on the BPIC 2012 dataset.

In Table 4, in all cases, the precision is almost 1.00, which indicates that our method outperforms the previously reported method [21]. For the combinations (A+O)+W and (A+W)+O, the recalls are over 0.9, which also implies that our method outperforms the previously reported method [21]. In the case of O+W or (O+W)+A, the recalls of the case ID identification are low. This is because the number of identified case IDs is small, and a large number of edges that occur only a few times in the integrated event logs O+W and A+W are considered as noise. However, the goal of process mining is to obtain the main process, so edges that appear only a few times should be avoided. Further, this does not impair the performance of generating the process model this time. Note that in correlation miner, the evaluation results of case ID identification are obtained by randomly selecting 90 case IDs, running them four times, and taking the average, due to their performance limitation [21].

As shown in Table 5, the proposed method displays high precision and recall that surpass those of the previously reported method [20] regardless of the order in which the organizations are combined. The approach to eliminate the bias in the order of integration, that is, $(A+O+W)_{sum}$, also shows higher precision and recall.

Disco visualizes the process model for the BPIC 2012 dataset. Fig. 4 shows the process model from the original correct data, and Fig. 6 shows the data integrated by the proposed approach. Edges are summed up when bias is eliminated, so their value is greater than the correct data. Comparison of Fig. 4 with Fig. 6 shows that the proposed method accurately generates the correct process.

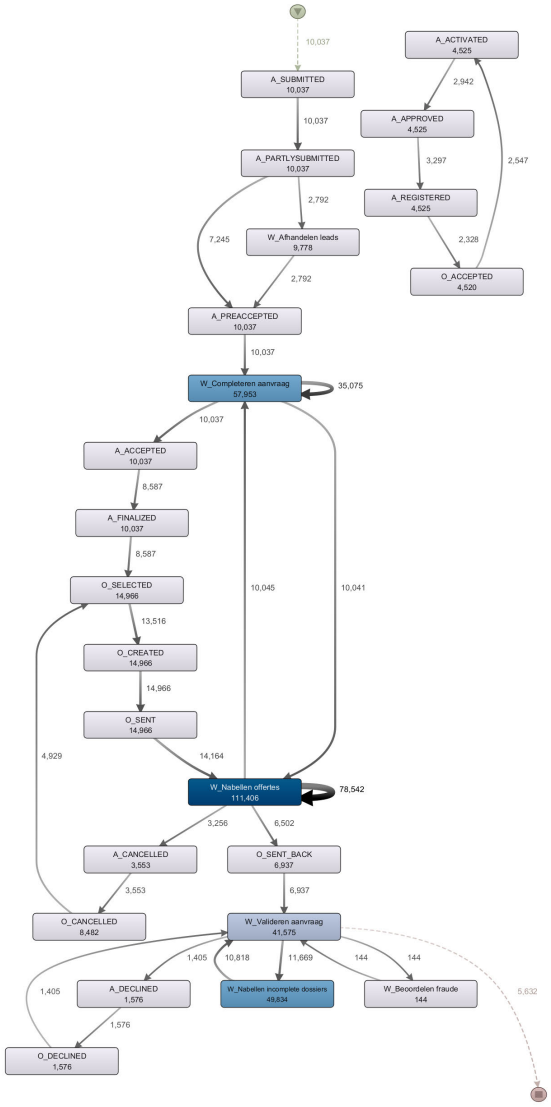

FIGURE 6. BPIC 2012 dataset process model generated by proposed approach.

D. EVALUATION USING BPIC 2017 DATASET

Table 6 shows the results of the case ID identification and Table 7 shows the results of the process model when the experiment was conducted on the BPIC 2017 dataset.

High precision and recall were obtained almost regardless of the order in which the organizations were combined. Although in the case of (A+W)+O, (O+W)+A, the recall of the case ID identification is relatively low, similar to V-C, this does not impair the performance of generating a process model. The method showed high precision and recall regardless of the order of integration, and further performance

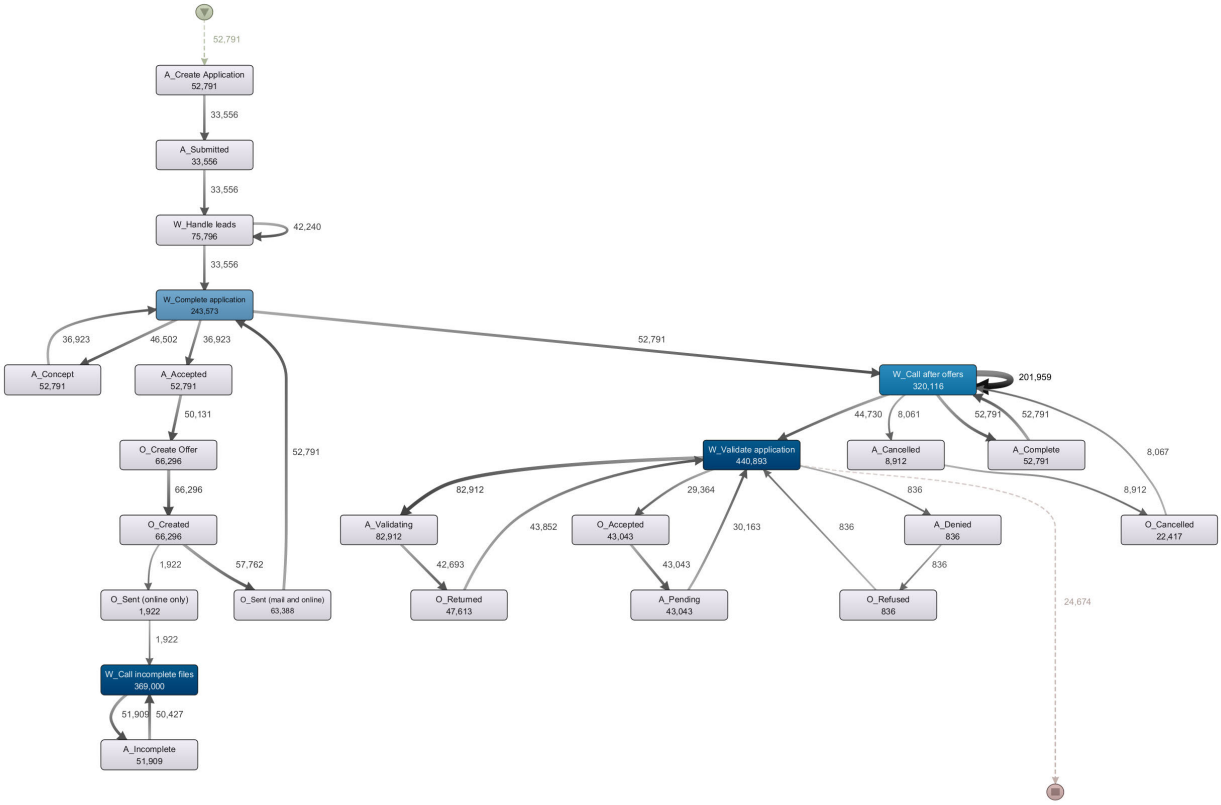

FIGURE 7. BPIC 2017 dataset process model generated by proposed approach.

TABLE 6. Results of identifying case IDs using the BPIC 2017 dataset.

Organization	Precision	Recall	F1
A+O	1.000	0.858	0.924
A+W	1.000	0.994	0.997
O+W	1.000	0.933	0.965
(A+O)+W	1.000	0.858	0.924
(A+W)+O	1.000	0.539	0.700
(O+W)+A	1.000	0.503	0.669

TABLE 7. Results of generated process models using the BPIC 2017 dataset.

Organization	Precision	Recall	F1
A+O	1.000	0.818	0.900
A+W	1.000	1.000	1.000
O+W	1.000	0.939	0.969
(A+O)+W	0.970	0.928	0.949
(A+W)+O	0.983	0.841	0.906
(O+W)+A	0.984	0.884	0.931
$(A+O+W)_{sum}$	0.985	0.942	0.963

improvement was observed when the bias in the order of integration was eliminated.

Similar to the BPIC 2012 dataset, the process model for the BPIC 2017 dataset is visualized by Disco. Fig. 5 shows the process model from the original correct data, and Fig. 7 shows the data integrated by the proposed approach. Edges are summed up when bias is removed, so the value of the edges is greater than the value of the correct data. Comparison of Fig. 5 with Fig. 7 shows that the proposed method accurately generates the correct process though there is a slight difference in the frequency of occurrences of activities, as observed in the evaluation of the BPIC 2012 dataset.

VI. CONCLUSION

In this paper, we propose an accurate cross-organizational process mining technique based on a step-by-step case ID identification mechanism that uses only common items in event logs and can deal with cyclic orchestrations. To the best of our knowledge, this method is the first to systematically identify pairs of reliable case IDs and extract new activity connections based on them. Our proposed method outperforms the existing techniques on the real-world event log, the BPIC2012 dataset, which has been used as a benchmark in related work. Furthermore, the proposed method also performs efficiently on the more data-rich real-world event log, the BPIC2017 dataset.

In the future, to further verify the effectiveness of our proposed method, it is necessary to combine it with other major process mining algorithms designed for single organizations. Evaluation of other process model representations is needed because of the limitations of the directly-follows graph. In addition, we will further explore the relationship between the recall of case ID identification and the performance of process mining, because the recall of the case ID identification is not necessarily high every time.

REFERENCES

[1] A. Rozinat, R. S. Mans, M. Song, and W. M. P. van der Aalst, "Discovering simulation models," *Inf. Syst.*, vol. 34, no. 3, pp. 305–327, May 2009.

[2] J. D. Hernandez-Resendiz, E. Tello-Leal, H. M. Marin-Castro, U. M. Ramirez-Alcocer, and J. A. Mata-Torres, "Merging event logs for inter-organizational process mining," in *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*, J. A. Zapata-Cortes, G. Alor-Hernández, C. Sánchez-Ramírez, and J. L. García-Alcaraz, Eds. Cham, Switzerland: Springer, 2021, pp. 3–26.

[3] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, and J. Buijs, "Process mining manifesto," in *Proc. Bus. Process Manage. Workshops*, Clermont-Ferrand, France, 2011, pp. 169–194.

[4] S. Saito, "Understanding key business processes for business process outsourcing transition," in *Proc. ACM/IEEE 14th Int. Conf. Global Softw. Eng. (ICGSE)*, Montreal, QC, Canada, May 2019, pp. 35–39.

[5] W. Van der Aalst, "Loosely coupled interorganizational workflows: Modeling and analyzing workflows crossing organizational boundaries," *Inf. Manage.*, vol. 37, no. 1, pp. 67–75, 2000.

[6] H. R'Bigui and C. Cho, "The state-of-the-art of business process mining challenges," *Int. J. Bus. Process Integr. Manage.*, vol. 8, no. 4, pp. 285–303, 2017.

[7] C. Liu, H. Li, S. Zhang, L. Cheng, and Q. Zeng, "Cross-department collaborative healthcare process model discovery from event logs," *IEEE Trans. Autom. Sci. Eng.*, early access, Aug. 3, 2022, doi: 10.1109/TASE.2022.3194312.

[8] Q. Zeng, S. X. Sun, H. Duan, C. Liu, and H. Wang, "Cross-organizational collaborative workflow mining from a multi-source log," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1280–1301, Feb. 2013.

[9] C. Liu, H. Duan, Q. Zeng, M. Zhou, F. Lu, and J. Cheng, "Towards comprehensive support for privacy preservation cross-organization business process mining," *IEEE Trans. Services Comput.*, vol. 12, no. 4, pp. 639–653, Jul. 2019.

[10] Q. Zeng, H. Duan, and C. Liu, "Top-down process mining from multi-source running logs based on refinement of Petri nets," *IEEE Access*, vol. 8, pp. 61355–61369, 2020.

[11] F. Corradini, B. Re, L. Rossi, and F. Tiezzi, "A technique for collaboration discovery," in *Proc. Int. Conf. Bus. Process Model., Develop. Support*, Leuven, Belgium, 2022, pp. 63–78.

[12] J. Claes and G. Poels, "Merging computer log files for process mining: An artificial immune system technique," in *Proc. Int. Conf. Bus. Process Manage.*, Clermont-Ferrand, France, 2011, pp. 99–110.

[13] Y. Xu, Q. Lin, and M. Q. Zhao, "Merging event logs for process mining with hybrid artificial immune algorithm," in *Proc. Int. Conf. Data Sci.*, Montreal, QC, Canada, 2016, p. 10.

[14] J. Claes and G. Poels, "Integrating computer log files for process mining: A genetic algorithm inspired technique," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, London, U.K., 2011, pp. 282–293.

[15] J. Claes and G. Poels, "Merging event logs for process mining: A rule based merging method and rule suggestion algorithm," *Exp. Syst. Appl.*, vol. 41, no. 16, pp. 7291–7306, Nov. 2014.

[16] A. Djedović, A. Karabegović, E. Žunić, and D. Alić, "A rule based events correlation algorithm for process mining," in *Proc. Int. Symp. Innov. Interdiscipl. Appl. Adv. Technol.*, Sarajevo, Bosnia, Herzegovina, 2020, pp. 587–605.

[17] L. Raichelson and P. Soffer, "Merging event logs with many to many relationships," in *Proc. Int. Conf. Bus. Process Manage.*, Haifa, Israel, 2014, pp. 330–341.

[18] L. Raichelson, P. Soffer, and E. Verbeek, "Merging event logs: Combining granularity levels for process flow analysis," *Inf. Syst.*, vol. 71, pp. 211–227, Nov. 2017.

[19] L. Cheng, B. F. Van Dongen, and W. M. P. Van Der Aalst, "Efficient event correlation over distributed systems," in *Proc. 17th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGRID)*, May 2017, pp. 1–10.

[20] S. Pourmirza, R. Dijkman, and P. Grefen, "Correlation mining: mining process orchestrations without case identifiers," in *Proc. Int. Conf. Service-Oriented Comput.*, Rome, Italy, 2015, pp. 237–252.

[21] S. Pourmirza, R. Dijkman, and P. Grefen, "Correlation miner: Mining business process models and event correlations without case identifiers," *Int. J. Cooperat. Inf. Syst.*, vol. 26, no. 2, pp. 1–32, May 2017.

[22] G. Park and M. Song, "Prediction-based resource allocation using LSTM and minimum cost and maximum flow algorithm," in *Proc. Int. Conf. Process Mining (ICPM)*, Jun. 2019, pp. 121–128.

[23] D. Bayomie, C. D. Ciccio, M. La Rosa, and J. Mendling, "A probabilistic approach to event-case correlation for process mining," in *Proc. Int. Conf. Conceptual Model.*, Salvador, Bahia, Brazil, 2019, pp. 136–152.

[24] D. Bayomie, C. Di Ciccio, and J. Mendling, "Event-case correlation for process mining using probabilistic optimization," *Inf. Syst.*, vol. 114, Mar. 2023, Art. no. 102167.

[25] D. Bayomie, K. Revoredo, C. Di Ciccio, and J. Mendling, "Improving accuracy and explainability in event-case correlation via rule mining," in *Proc. 4th Int. Conf. Process Mining*, Bolzano, Italy, 2022, pp. 24–31.

[26] C. Liu, L. Cheng, Q. Zeng, and L. Wen, "Formal modeling and discovery of hierarchical business processes: A Petri net-based approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 2, pp. 1003–1014, Feb. 2023.

[27] C. Liu, "Formal modeling and discovery of multi-instance business processes: A cloud resource management case study," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2151–2160, Dec. 2022.

[28] B. van Dongen. *4TU.Researchdata BPI Challenge 2012*. Accessed: Jun. 1, 2023. [Online]. Available: https://data.4tu.nl/articles/dataset/BPI_Challenge_2012/12689204

[29] B. van Dongen. *4TU.Researchdata BPI Challenge 2017*. Accessed: Jun. 1, 2023. [Online]. Available: https://data.4tu.nl/articles/BPI_Challenge_2017/12696884

[30] C. W. Günther and A. Rozinat, "Disco: Discover your processes," in *Proc. Demonstration Track 10th Int. Conf. Bus. Process Manage.*, vol. 940, Sep. 2012, pp. 40–44.

[31] H. Tong, *Non-Linear Time Series: A Dynamical System Approach*. Oxford, U.K.: Oxford Univ. Press, 1990.

[32] A. Awad, K. Raun, and M. Weidlich, "Efficient approximate conformance checking using trie data structures," in *Proc. 3rd Int. Conf. Process Mining (ICPM)*, Eindhoven, The Netherlands, Oct. 2021, pp. 1–8.

[33] W. M. P. van der Aalst, "A practitioner's guide to process mining: Limitations of the directly-follows graph," *Proc. Comput. Sci.*, vol. 164, pp. 321–328, Jan. 2019.

[34] W. Reisig and G. Rozenberg, *Lectures on Petri Nets I: Basic Models: Advances in Petri Nets*. Berlin, Germany: Springer, 1998.

[35] *Introduction to BPMN*, IBM Cooperation, S. A. White, Marietta, GA, USA, 2004.

KAZUKI TAJIMA received the B.E. degree in electrical and electronic engineering from the Graduate School of Engineering, The University of Tokyo, Tokyo, Japan, in 2022, where he is currently pursuing the master's degree in electronic information engineering with the Graduate School of Information Science and Technology.

BOJIAN DU received the B.E. degree in electronic information engineering from the Beijing University of Technology, Beijing, China, in 2017, and the M.S. and Ph.D. degrees in electrical engineering from The University of Tokyo, Tokyo, Japan, in 2020 and 2023, respectively. His research interest includes time-series data analysis.

YOSHIAKI NARUSUE (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, in 2012, 2014, and 2017, respectively.

Currently, he is an Associate Professor with the Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo. He is a member of the IEICE and IPSJ. He received the Second-Best Student Paper Award at the IEEE Radio and Wireless Symposium, in 2013, the Hiroshi Harashima Academic Encouragement Award, in 2013, the Best Paper Award at the IEEE Consumer Communications and Networking Conference, in 2018, and the ACM IMWUT Distinguished Paper Award, in 2020. His research interests include wireless power transfer, next-generation wireless communication systems, and the Internet of Things.

SHINOBU SAITO received the B.S. and M.S. degrees in administration engineering and the Ph.D. degree in systems engineering from Keio University, in 1999, 2001, and 2007, respectively.

From 2001 to 2015, he was a System Engineer with NTT Data Corporation, Japan. From 2015 to 2018, he was a Research Manager with NTT Corporation, Japan. Since 2018, he has been a Distinguished Researcher with NTT Computer and Data Science Laboratories. He was a Visiting Researcher with the Institute for Software Research (ISR), University of California at Irvine, Irvine, from 2016 to 2018. His research interests include software requirements engineering, design recovery, business modeling, and business process management.

YUKAKO IIMURA received the bachelor's degree in administrative studies from the Prefectural University of Kumamoto, Kumamoto, Japan, in 1998, and the M.E. degree in mathematical science and information systems from Kumamoto University, Kumamoto, in 2001.

Since 2001, she has been with Nippon Telegraph and Telephone Corporation, Japan, where she is currently a Research Engineer with NTT Computer and Data Science Laboratories. Her current research interests include requirements engineering and software engineering.

Ms. Iimura is a member of IPSJ.

HIROYUKI MORIKAWA (Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from The University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively.

He is currently a Full Professor with the School of Engineering, The University of Tokyo. From 2002 to 2006, he was a Group Leader with the NICT Mobile Networking Group. His research interests include ubiquitous networks, sensor networks, big data/IoT/M2M, wireless communications, and network services.

Prof. Morikawa is a fellow of IEICE. He has received more than 50 awards, including the IEICE Best Paper Award, in 2002, 2004, and 2010, the IPSJ Best Paper Award, in 2006, the JSCICR Best Paper Award, in 2015, the Info-Communications Promotion Month Council President Prize, in 2008, the NTT DoCoMo Mobile Science Award, in 2009, the Rinzaburo Shida Award, in 2010, the Radio Day Ministerial Commendation, in 2014, and the IEEE CCNC Best Paper Award, in 2018. He served as a technical program committee chair for many IEEE/ACM conferences and workshops, the Vice President of IEICE, the OECD Committee on Digital Economy Policy Vice Chair, and the Director of the New Generation M2M Consortium. He serves on numerous telecommunications advisory committees and frequently serves as a consultant to governments and companies.