An uncertainty-aware event log of network traffic

Gal Engelberg[1,2,*], Moshe Hadad[1,2,*], Marco Pegoraro[3,*], Pnina Soffer[1], Ethan Hadar[2] and Wil M.P. van der Aalst[3]

[1]*Department of Information Systems, University of Haifa, Haifa, Israel*
[2]*Accenture Labs, Tel Aviv, Israel*
[3]*Chair of Process and Data Science, RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany*

Abstract

Business Process Management (BPM) heavily relies on event logs for process mining. However, traditional event logs may not always be available or may be harder to obtain for unlogged or unconventionally logged activities. To overcome these limitations, network traffic data can be used as an alternative source for constructing event logs. However, incorporating network traffic data poses its own set of challenges. These challenges include dealing with the large volume and diverse nature of network packets, as well as the uncertainty in mapping low-level events in a stream to specific activity types and border points, namely, the start and the end of an activity. In this paper, we introduce novel datasets that have been constructed from an enterprise network simulation environment. These datasets consist of two types of event logs: network traffic-level event logs and abstracted business-level event logs. Both types of logs exhibit various forms of uncertainty. These labeled datasets can serve as valuable benchmarks for a range of process mining tasks, such as event abstraction, process discovery, and conformance checking from uncertain event data.

Keywords

Event log, process mining, network traffic, uncertainty, XES, supervised training

1. Introduction

In recent decades, there has been an increasing adoption of Business Process Management (BPM) in organizations. BPM enables organizations to explore, analyze, monitor, and continuously enhance work processes [1]. Within the realm of BPM, process mining techniques are utilized to discover and monitor process models, identify bottlenecks in the processes, detect deviations from expected execution, and more [2]. Process mining heavily relies on event logs, which are generated by information systems or by recording the actions performed during process execution. As per the XES standard [3], a log consists of traces that depict the execution of a process instance. A trace represents an ordered collection of events, each event comprising attributes such as an activity label, a timestamp, and a case ID.

International Conference on Business Process Management, September 11–15, 2023, Utrecht, the Netherlands
*Corresponding author.
gal.engelberg@accenture.com (G. Engelberg); moshe.hadad@accenture.com (M. Hadad); pegoraro@pads.rwth-aachen.de (M. Pegoraro); spnina@is.haifa.ac.il (P. Soffer); ethan.hadar@accenture.com (E. Hadar); wvdaalst@pads.rwth-aachen.de (W. M.P. v. d. Aalst)
http://mpegoraro.net/ (M. Pegoraro); https://vdaalst.com/ (W. M.P. v. d. Aalst)
0000-0001-9021-9740 (G. Engelberg); 0000-0002-9315-6260 (M. Hadad); 0000-0002-8997-7517 (M. Pegoraro); 0000-0003-4659-883X (P. Soffer); 0000-0002-0955-6940 (W. M.P. v. d. Aalst)

Nevertheless, there are instances where conventional event logs are either unavailable or exist in diverse formats [4]. Consequently, alternative sources such as databases, machine logs, SOAP messages, and others are examined to generate standardized event logs [4]. For instance, in previous studies, event logs were extracted from relational databases [5], and low-level database transaction logs were utilized to uncover process models [6]. Furthermore, processes may encompass multiple systems or incorporate activities that are not adequately documented in the log. As such, we suggest incorporating network traffic data as an additional data source for the purpose of process event log creation.

The potential utilization of network traffic data for process mining remains relatively unexplored, offering opportunities to address limitations in existing event logs by capturing the entire process and incorporating unlogged actions. However, leveraging network traffic data for BPM presents significant challenges due to the gap between technical network operations and business activities. The considerable volume of network traffic data, consisting of packets originated by diverse network protocols, further complicates the reconstruction of messages, with a substantial number of packets generated for each business activity [7]. Moreover, uncertainty represents a fundamental challenge in utilizing network traffic-based event logs, as the continuous nature of network traffic obscures clear indications of individual activity start and end. In addition, the concurrent execution of multiple activities makes the mapping of network traffic packets to specific business activity traces highly challenging and subject to an inherent uncertainty. In our recent work, we addressed these challenges by discovering a network traffic behavioral model of an activity [7], and by proposing a log abstraction method that transforms network traffic data to a business-level event log [8], which holds uncertainty to some extent.

Uncertainty in event logs refers to recorded executions of specific activities in a process, accompanied by an indication of uncertainty in the event attributes. Uncertainty affecting the attributes that define the control-flow of a process (case ID, timestamp, and activity label) is particularly critical, since it directly affects most process mining techniques, including the discovery of a process model and conformance checking. [9] introduced a taxonomy of uncertain event data, and a method for conformance checking under uncertainty conditions. The logs presented in this paper are affected by uncertainty, which we classify and describe.

In this paper, building on the results obtained in [7, 8], we provide the low-level network traffic data, and the resulted business-level event logs [1] to the BPM community. These datasets could serve as a benchmark for various of process mining tasks, such as event abstraction, process discovery and conformance checking from uncertain event data. In addition to the provided datasets, we perform an analysis of their uncertain behavior. The remainder of the paper is structured as follows. Section 2 specifies the dataset extraction steps, Section 3 describes the uncertain behavior of the datasets; finally, Section 4 presents the conclusions of this paper.

2. Data Extraction

This section describes the steps taken to extract the event logs. The event logs were generated with an enterprise network simulation environment, that represents a typical enterprise system executing business process cases. The simulated environment consisted of endpoint machines,

[1]The resource, a usage guide, and a license: https://github.com/HaifaUniversityBPM/traffic-data-to-event-log

each associated with a human participant engaged in specific activities. These endpoint machines communicated with an Odoo ERP web application [10] through an HTTP application layer protocol. The Odoo ERP application, in turn, communicated with a PostgreSQL database server[2] and a mail server via the PGSQL and SMTP application layer protocols, respectively. We applied the following steps to extract the datasets, a detailed description can be found in [7, 8].

Simulation of business process cases. The simulation environment executed cases of two business processes: an HR (human resources) recruitment process, and a purchase-to-pay process. For each execution, the network communication between the different devices was recorded. We created two types of data sets: one for training, in which each activity was running in isolation, and another for evaluations, in which process cases were running in parallel.

Pre-processing and filtering. At this stage, we performed an iterative process of filtering the data to remove noise and irrelevant information to obtain more compact, cleaner, and slightly abstracted logs for further analysis. In the resulting datasets, each packet is represented as an event in a stream.

Event classification and case correlation. At this stage, we trained two sequence models on the training data set using Conditional Random Fields (CRF) [11]: The first was used for identifying the events that act as border points of an activity, representing an activity start and end. The second was used to identify the activity type that correlates to a sequence of events. We then applied the trained models to the evaluation dataset, where packets generated by activities that were performed in parallel, as part of process cases that run in parallel to one another, were interleaved in one stream. Then, we assigned a case ID for each activity type by correlating attribute values between events. Activity timestamps were defined as the recording time of their correlated events.

3. Features and Uncertain Behavior

The extraction procedure for the network traffic logs we describe relies on probabilistic methods for the recognition of some of the event attributes. As mentioned in Section 2, the CRF learning model is able to read network packages marked as activity start or end, and to provide an estimate for the activity label of the corresponding business-level event. Such estimate is in the form of a discrete probability distribution over an alphabet of labels. It is possible to extrapolate a single final label from the distribution, such as by following the maximum likelihood principle and selecting the label with the highest probability. This, however, leads to a loss of information.

An improved approach is the *uncertain event log* [9], which contains probabilistic descriptions of event attributes, and is therefore able to represent attribute values through probability distributions. Such distributions are contained in the log as meta-attributes; for instance, events are connected to maps that link every activity label to a probability value. This enables the application of a family of analysis techniques especially designed for uncertain event logs.

Some of the logs we share contain a discrete probability distribution representing activity labels; according to the taxonomy proposed in [9], they are $[A]_{\mathrm{W}}$-type logs. Specific analysis techniques are available for this class of data, such as recent conformance checking approaches by Van der Aa et al. [12] and Bogdanov et al. [13], as well as an XES data standard definition [14].

[2] https://www.postgresql.org/

Table 1
Some statistics of the HR and purchase-to-pay network traffic logs with interleaving; based on a 10 cases simulation. The number of realizations refer to the possible real-life scenarios represented by the stochastic attributes in the uncertain traces [9].

Statistic	Log	
	HR	PTP
Number of traces	10	10
Number of events	74	126
Number of unique activity labels	7	7
Average number of events per trace	7.4	12.6
Median number of events per trace	6.0	14.0
Minimum number of events per trace	4	8
Maximum number of events per trace	14	14
Taxonomic classification	$[A]_{\mathrm{W}}$	$[A]_{\mathrm{W}}$
Average number of realizations per trace	10,084,532,298.4	7,344,521,182.6
Median number of realizations per trace	412,972.0	7,909,306,972.0
Minimum number of realizations per trace	2,401	823,543
Maximum number of realizations per trace	96,889,010,407	13,841,287,201
Total number of trace realizations	100,845,322,984	73,445,211,826

In addition to the datasets in CSV format, we also share data in the uncertainty-extended XES standard. Table 1 shows the complexity of the data by summarizing some of the features (traditional and uncertainty-related) of the interleaving event logs. A complete resource's schema is specified in 1. An example of uncertain attributes is shown in the following listing:

```
<string key="uncertainty:classification_prob_windows_start-15_end-15_action_1" value="discrete_weak">
   <float key="ResumeReviewActivity" value="0.384322532584433" />
   <float key="GenerateJobApplicationActivity" value="0.1843477633451853" />
   <float key="PerformAnInterviewMeeting" value="4.7399447586012736e-14" />
   <float key="PerformAnInterviewCall" value="2.407567933598791e-34" />
   <float key="ScheduleAnInterviewActivityCall" value="1.280111885616518e-21" />
   <float key="ScheduleAnInterviewMeeting" value="1.6039484157515345e-17" />
   <float key="ContractProposal" value="8.29100569383624e-35" />
</string>
```

4. Conclusion

In developing beyond its native applications in business and logistics, process analysis expanded to encompass data paradigms structurally distinct from traditional event logs. Examples are logs of network traffic, and techniques customized for network traffic data that are able to uncover novel insights regarding a business process.

In this paper, we provide insight into our publicly accessible synthetic datasets derived from the logging of network data in an organization. We described the extraction procedure, the features of the dataset, and we show the uncertain behavior that appears in the obtained log as a result of the probabilistic output of the extraction method. We hope that the availability of these logs will be useful to the BPM community, and will stimulate research within the field of

BPM on network traffic data.

Acknowledgments

This work is a collaboration with Accenture Labs, Israel. The authors gratefully acknowledge the support by the Alexander von Humboldt (AvH) Stiftung.

References

[1] M. Dumas, M. La Rosa, J. Mendling, H. A. Reijers, et al., Fundamentals of business process management, Springer, 2013.
[2] W. M. P. van der Aalst, Process mining: data science in action, Springer, 2016.
[3] W. M. P. van der Aalst, C. Günther, J. Bose, et al., IEEE standard for extensible event stream (XES) for achieving interoperability in event logs and event streams, IEEE Std 1849 (2016) 1–50.
[4] V. Huser, Process mining: Discovery, conformance and enhancement of business processes, 2012.
[5] E. González López de Murillas, H. A. Reijers, W. M. P. van der Aalst, Connecting databases with process mining: a meta model and toolset, Software & Systems Modeling 18 (2019) 1209–1247.
[6] W. M. P. van der Aalst, Extracting event data from databases to unleash process mining, in: BPM-Driving innovation in a digital world, Springer, 2015, pp. 105–128.
[7] G. Engelberg, M. Hadad, P. Soffer, From network traffic data to business activities: a process mining driven conceptualization, in: International Conference on Business Process Modeling, Development and Support, Springer, 2021, pp. 3–18.
[8] M. Hadad, G. Engelberg, P. Soffer, From network traffic data to a business-level event log, in: International Conference on Business Process Modeling, Development and Support, Springer, 2023, pp. 60–75.
[9] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, Conformance checking over uncertain event data, Information Systems 102 (2021) 101810.
[10] A. Ganesh, K. N. Shanil, C. Sunitha, A. M. Midhundas, OpenERP/Odoo-An open source concept to ERP solution, in: 2016 IEEE 6th International Conference on Advanced Computing (IACC), IEEE, 2016, pp. 112–116.
[11] C. Sutton, A. McCallum, et al., An introduction to conditional random fields, Foundations and Trends® in Machine Learning 4 (2012) 267–373.
[12] H. van der Aa, H. Leopold, H. A. Reijers, Efficient process conformance checking on the basis of uncertain event-to-activity mappings, IEEE Transactions on Knowledge and Data Engineering 32 (2019) 927–940.
[13] E. Bogdanov, I. Cohen, A. Gal, Conformance checking over stochastically known logs, in: International Conference on Business Process Management, Springer, 2022, pp. 105–119.
[14] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, An XES extension for uncertain event data, in: International Conference on Business Process Management (BPM 2021), volume 2973 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 116–120.