Challenge Paper: Data Quality Issues in Queue Mining

Avigdor Gal, Technion–Israel Institute of Technology
Arik Senderovich, Technion–Israel Institute of Technology
Matthias Weidlich, Humboldt-Universität zu Berlin

1. QUEUEING NETWORKS AND QUEUE MINING

Queues represent a setting where agents compete over a scarce resource: people wait for public transportation, jobs wait to be processed, patients await treatment at a hospital, *etc.* While data logs record many aspects of our lives, information about queues is rarely recorded. *Queue mining* [Senderovich et al. 2015] is the process of revealing queue information from data logs for the purpose of discovering queueing models, conformance checking, and optimization. As such, queue mining enables bottleneck detection and delay prediction [Gal et al. 2017].

A *queueing network* is the most general form of a queueing model, represented as a directed graph with nodes being the queueing stations (corresponding to types of resources), edges corresponding to routing between stations, and node attributes corresponding to station dynamics (*e.g.*, arrival patterns, service time distributions, station capacity, and service policy– for example, first-come first-served).

Customers arrive into a queueing station, wait (enqueued) before being served by the station and then leave to the next station (or exit the system). Queueing networks are often assumed to have a single customer type and an immediate Markovian routing (after completion at a station a customer appears in the next station with some probability). Also, simple queueing networks typically forbid concurrent activities. To overcome these limitations, multi-class state-dependent fork/join networks may be a better alternative [Senderovich et al. 2016].

Mining the nodes and edges of a queueing network is based on well-established techniques from process mining [van der Aalst 2016], which elicit the corresponding control-flow structure from event logs of a business process. Queue mining, in turn, targets performance analysis of processes, rendering the dynamic part of the network, *e.g.*, the characterization of arrival rates, service times and policies, its main concern.

Virtually all analysis techniques grounded in data logs face issues related to data quality. While, on the abstract level, the types of issues may be similar, their implications on analysis results and the approaches taken to overcome them are highly specific to the applied analysis techniques. In the related field of process mining, for instance, the tasks to find, merge, and clean data logs and dealing with their heterogeneity has been identified as one of the core challenges in the Process Mining Manifesto [van der Aalst et al. 2011]. However, even if considering a generic data quality issue such as missing data in the log, the implications and remedy strategies are specific to an analysis technique. For instance, missing data may lead to *directly-follows incompleteness*, which refers to missing observations of possible transitions between two activities of a process and is motivated by a specific class of process mining algorithms [Leemans et al. 2013].

We illustrate the setting of queue mining with a case from healthcare, *i.e.*, the Dana-Farber Cancer Institute (DFCI), a large outpatient cancer hospital in the United States. Patients arrive at DFCI according to an appointment book. However, during the day, patient schedules may change due to, *e.g.*, no-shows, unscheduled arrivals, and long service times. DFCI is equipped with a Real-Time Locating System (RTLS) that records about 1.4 Million events per day (positions of medical staff, patients, and resources). Based on this data, the operations at DFCI can be analysed using queue

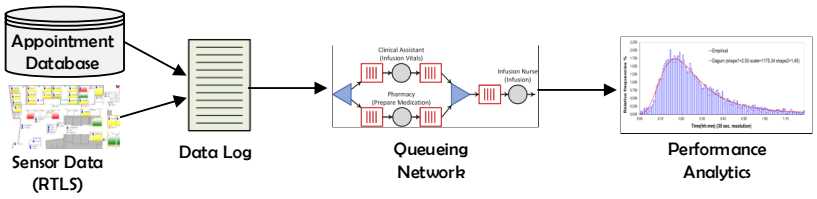

Fig. 1. Queue mining in healthcare, based on appointment databases and data sensing.

mining, as depicted in Figure 1: Data on appointments and locations is consolidated into a data log, from which a queueing network is mined. This model is then utilised to derive performance analytics, which enables operational improvements.

Against this background, we focus on two data quality issues, namely *missing events and attribute values* and *data granularity*, and the resulting challenges when mining queue information from data logs. We highlight the state-of-the-art and outline research challenges.

2. DATA QUALITY ISSUES AND RESULTING CHALLENGES IN QUEUE MINING

To elicit queueing networks from data, queue mining techniques make several assumptions on the originating data log. Specifically, the log is a set of quadruples with a unique case identifier (*e.g.*, a patient), resource type (*e.g.*, a doctor) corresponding to a station in the network, and start and completion time. Note that queueing time is implicit, as one is routed immediately after completion to the next station. Thus, queueing time is the time between subsequent completion times and start times.

2.1. Missing Events and Attribute Values

Missing attribute values, an extreme case of which is a missing tuple (or event), introduce biases to queue mining techniques. Below, we discuss situations where one of the attributes is missing, detailing the impact of the lack of each element on queue mining.

Case identity. In data logs where the identity attribute is missing, one cannot distinguish cases. In the above use case of DFCI, for example, one encounters situations in which physicians enter an examination room, stay there for a certain duration, and leave without a corresponding patient entry. These situations are referred to as *ghost patients*, since there is a high certainty that a patient actually visited the examination room. However, there is uncertainty regarding the identity of the ghost patient. While data quality issues related to case identity have also been described for process mining [Ferreira and Gillblad 2009], the resulting challenges are different. In queue mining, the ability to fit service time distributions per station is not affected. However, the ability to learn service policies [Senderovich et al. 2016], namely the ordering of cases in a queue, is crippled.

In process mining, a lack of case identity has been addressed by taking up ideas from the related realms of probabilistic databases and uncertain data management. Techniques such as Expectation-Maximization that proved valuable in these domains, e.g., for duplicate detection [Bilenko and Mooney 2003], may also help to obtain a probabilistic case assignment [Ferreira and Gillblad 2009]. However, the former technique also illustrates the additional challenges imposed by the process context as it is not robust in the presence of repetitive and concurrent behaviour. To overcome this limitation, a technique for labelling events by employing behavioural profiles was proposed in [Bayomie et al. 2016], assuming that data is clean, and does not contain infrequent realizations of the process—mostly non-realistic assumptions. Even if these limitations are ignored, there is a notable research gap: In queue mining, the reconstruction of the case identity is not necessarily needed, *e.g.*, for mining of service protocols.

Start/Completion times. The lack of start and/or completion times impacts severely the ability to learn the dynamic components of a queueing network. In the DFCI setting, inferring the start and completion times of an activity using location information is uncertain. For example, a patient who sits in an infusion room waiting for the chemotherapy medication, has not actually started their infusion. Trivially, when one of the temporal attributes is missing, the problem of estimating service times is based on a censored sample. To overcome censoring and make estimation unbiased, one may employ techniques from survival analysis [Kalbfleisch and Prentice 2011].

Further, missing start times make the inference of service policies difficult, since the order between served customers is lost. For single-server queueing nodes, it is plausible to assume that the order of departure is the same as the order of service start. However, for multi-server nodes, one needs additional information on a specific resource that served the customer to infer the correct ordering of customers.

A few works have been devoted to imputation of missing or erroneous start times. First, common techniques for anomaly detection [Chandola et al. 2009] may be adapted for the setting of data logs. As detailed in [Rogge-Solti and Kasneci 2014], erroneous start times are then identified using a Bayesian model to detect abnormal durations of subsequent log entries.

Other work learns a Bayesian phase-type model to impute missing start times, when completion times are known [Senderovich et al. 2015]. The above approaches, however, are not generally applicable, since they assume an accurate model of the routing of customers or are limited to scenarios where each resource type performs a single activity. Going beyond these works, we foresee that temporal regularities as observed in many processes, see [Lanz et al. 2016], provide a promising angle for more powerful imputation of missing start and completion times.

2.2. Data Granularity

To illustrate data granularity issues, we consider situations where resource types are missing in tuples of a data log, yet can still be inferred using other attributes. As detailed above, at DFCI, the deployed RTLS tracks medical staff, patients, and resources in real-time. Consequently, locations can be traced back to specific rooms, and resource types can be inferred from the matching between sensor identifier and location name. In [Senderovich et al. 2016], external process knowledge (that establishes a relation between locations and activities) is used to infer such resource types.

On a more general level, the question of how to aggregate low-level data to obtain data logs that are suitable for analysis of the operations of a system has been addressed using semi-automated matching [Baier et al. 2014] and pattern-based approaches [Mannhardt et al. 2016].

The extent to which heterogeneity in data granularity can be bridged with these approaches is limited, though, by the need for manual input or the limited expressiveness of considered patterns, respectively. Another direction could therefore be the application of classification and regression techniques, as suggested for the discovery of predictive process models from low-level multi-dimensional data [Folino et al. 2014]. Finally, sensor-based activity [Chen et al. 2012] and event recognition [Artikis et al. 2012] may help to overcome granularity issues by identifying situations of interest in low-level data. However, state-of-the-art recognition techniques neglect the perspective of customers traversing through a network of stations, each having specific and potentially time-varying dynamics.

3. CONCLUDING REMARKS

Queue mining is a novel research area that targets the extraction of queueing models from data logs. These models are then used for bottleneck detection and performance

prediction in systems, in which customers compete for scarce resources. This paper reviewed common data quality challenges and their implications for queue mining, as well as some existing solutions to these challenges. However, much work is required to cope with the aforementioned challenges, as current techniques make numerous unrealistic assumptions that need to be relaxed to improve the accuracy of queue mining methods.

REFERENCES

Alexander Artikis, Anastasios Skarlatidis, François Portet, and Georgios Paliouras. 2012. Logic-based event recognition. *Knowledge Eng. Review* 27, 4 (2012), 469–506.

Thomas Baier, Jan Mendling, and Mathias Weske. 2014. Bridging abstraction layers in process mining. *Inf. Syst.* 46 (2014), 123–139.

Dina Bayomie, Ahmed Awad, and Ehab Ezat. 2016. Correlating Unlabeled Events from Cyclic Business Processes Execution. In *CAiSE 2016. Proceedings*. 274–289.

Mikhail Bilenko and Raymond J. Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *SIGKDD 2003*, ACM, 39–48.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41, 3 (2009), 15:1–15:58.

Liming Chen, Jesse Hoey, Chris D. Nugent, Diane J. Cook, and Zhiwen Yu. 2012. Sensor-Based Activity Recognition. *IEEE Trans. Systems, Man, and Cybernetics, Part C* 42, 6 (2012), 790–808.

Diogo R. Ferreira and Daniel Gillblad. 2009. Discovering Process Models from Unlabelled Event Logs. In *BPM 2009. Proceedings*. 143–158.

Francesco Folino, Massimo Guarascio, and Luigi Pontieri. 2014. Mining Predictive Process Models out of Low-level Multidimensional Logs. In *CAiSE 2014. Proceedings (LNCS)*, Vol. 8484. Springer, 533–547.

Avigdor Gal, Avishai Mandelbaum, François Schnitzler, Arik Senderovich, and Matthias Weidlich. 2017. Traveling time prediction in scheduled transportation with journey segments. *Information Systems* 64 (2017), 266–280.

John D Kalbfleisch and Ross L Prentice. 2011. *The statistical analysis of failure time data*. Vol. 360. John Wiley & Sons.

Andreas Lanz, Manfred Reichert, and Barbara Weber. 2016. Process time patterns: A formal foundation. *Information Systems* 57 (2016), 38–68.

Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. 2013. Discovering Block-Structured Process Models from Event Logs - A Constructive Approach. In *PETRI NETS 2013. Proceedings (LNCS)*, Vol. 7927. Springer, 311–329.

Felix Mannhardt, Massimiliano de Leoni, Hajo A. Reijers, Wil M. P. van der Aalst, and Pieter J. Toussaint. 2016. From Low-Level Events to Activities - A Pattern-Based Approach. In BPM 2016. Proceedings *(LNCS)*, Vol. 9850. Springer, 125–141.

Andreas Rogge-Solti and Gjergji Kasneci. 2014. Temporal Anomaly Detection in Business Processes. In *BPM 2014. Proceedings (LNCS)*, Vol. 8659. Springer, 234–249.

Arik Senderovich, Sander J. J. Leemans, Shahar Harel, Avigdor Gal, Avishai Mandelbaum, and Wil M. P. van der Aalst. 2015. Discovering Queues from Event Logs with Varying Levels of Information. In BPM Workshops 2015, Revised Papers. 154–166.

Arik Senderovich, Andreas Rogge-Solti, Avigdor Gal, Jan Mendling, and Avishai Mandelbaum. 2016. The ROAD from Sensor Data to Process Instances via Interaction Mining. In CAiSE 2016. Proceedings. 257–273.

Arik Senderovich, Matthias Weidlich, Avigdor Gal, and Avishai Mandelbaum. 2015. Queue mining for delay prediction in multi-class service processes. *Information Systems* 53 (2015), 278–295.

Arik Senderovich, Matthias Weidlich, Liron Yedidsion, Avigdor Gal, Avishai Mandelbaum, Sarah Kadish, and Craig A. Bunnell. 2016. Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. *Information Systems* 62 (2016), 185–206.

Wil M. P. van der Aalst. 2016. *Process Mining - Data Science in Action, Second Edition*. Springer.

Wil M. P. at al. 2011. Process Mining Manifesto. In BPM Workshops 2011, Revised Papers *(LNBIP)*, Vol. 99. Springer, 169–194.