# Stochastic Process Mining (Extended Abstract)

Adam Burke
Queensland University of Technology
at.burke@qut.edu.au

There are many information systems in the world, used by many organizations to build things and help people. In process mining [1], models of organizations at work are automatically constructed and analyzed. By describing large amounts of data quickly, process mining accelerates understanding of what an organizations does, and how it may improve. For example, a medical doctor may note which common hospital intake tasks are bottlenecks, or an auditor may see evidence that certain regulatory checks are carried out as expected. These processes may be explicitly defined by the organization, as in an insurance claim process, or implicit, as in a hospital emergency room.

The successes of process mining to date have largely been with control-flow models. In a control-flow model, causality is represented, but probability is not. Stochastic process models are potentially more powerful tools for some types of organizational analysis and optimization where frequency and variation are key, because one way of understanding organizations is to look at the patterns of what they repeatedly do, and what they treat as exceptional. Stochastic models may also be used in prediction and performance. This project then investigates:

*How can processes in organizations be understood using stochastic models mined from organizational data?*

The relatively small body of existing work on this topic is reviewed with research sub-questions and methods in Sections I-III. This project extends and applies this existing work while contributing novel techniques and analysis for the construction and use of these models on real-world event data. The process models notations used are Generalized Stochastic Petri Nets (GSPNs) [2] and closely related extensions. Section I summarizes already conducted research into stochastic process discovery, while Sections II and III relate to ongoing and planned research into quality dimensions and concept drift.

## I. Discovering Stochastic Models (RQ1)

*RQ1 How may stochastic process models be discovered automatically?*

This project work investigated composition of existing control-flow discovery techniques with new weight estimation [3] and direct stochastic process discovery [4] techniques. Direct Process Discovery algorithms output stochastic models without an intermediate control-flow discovery step. The techniques for generating stochastic models are not themselves necessarily stochastic: they may also include analytic methods.

*RQ1.1 How may control-flow discovery techniques be leveraged for stochastic process model discovery?*

A stochastic process discovery framework was developed in the investigation of this question [3]. In it, a control flow discovery algorithm is first used to discover a Petri net [1, p60], and the result is combined with a weight estimation step to produce a GSPN [2]. Six estimators fitting the framework were implemented, and evaluated experimentally against stochastic conformance measures, using established discovery algorithms and real-life public event logs[1]. The framework is a generalization of existing stochastic discovery techniques that compose control-flow discovery with a pipelined stochastic estimation step [5], [6]. It is also a specialization, in that it produces GSPNs with immediate transitions, rather than GDT_SPNs [5].

Stochastic quality measures of Entropy Recall and Precision [7] and Earth-Movers' Distance [8] were used for the evaluation, together with real-life logs from BPI challenges and a variety of discovery algorithms. Estimation techniques found were of comparable quality, were applicable to a broader range of event logs, and were generally faster than GDT_SPN discovery [5].

*RQ1.2 How may stochastic models be discovered directly from event logs?*

Techniques based on a computing pipeline of control-flow model discovery followed by some inference of stochastic data have some inherent design limitations. The control-flow-only nature of the initially discovered model may introduce a representational bias toward the structures of that output formalism. The multiple passes through the log is also awkward, and perhaps inefficient. Accordingly, this project introduced the novel *Toothpaste miner* framework, a set of direct discovery algorithms where control-flow and stochastic aspects are discovered in concert, through a process of reduction and abstraction, in polynomial time [4]. To do this, it introduces an intermediate abstraction targeted at stochastic process discovery, the Probabilistic Process Tree (PPT). The algorithms start with a trace model and reduce it to a target model using formally defined rules, "squeezing" trace information into a usable form. A prototype was implemented[2] and evaluated empirically against existing techniques with promising results.

A PPT is an extension of Process Trees that includes relative probabilities in the form of node weights, and is related formally to GSPNs. Toothpaste miner is inspired by region-based miners for control-flow models [9], [10] and the Alergia [11] algorithm.

[1] Java code and data at https://github.com/adamburkegh/spd_we

[2] Haskell code and data at https://github.com/adamburkegh/toothpaste

## II. Quality Dimensions (RQ2)

*RQ2 What quality dimensions are empirically observable in stochastic process models and logs?*

The quality of process models is often quantified with conformance measures, and those measures related to four standard control-flow quality dimensions [1, p188]. Recent research into stochastic process measures suggests both connections and challenges to these quality dimensions. For example, entropy measures have been related to both precision and recall (fitness) [7], but Earth-Movers Distance does not obviously align with an existing dimension [12]. Challenges range from needing to translate the technical definition of e.g., fitness, to a stochastic context, to differences in how adjacent fields theorize the role of simplicity and generalization.

At least one empirical and quantitative study compares process measures and dimensions for control-flow models [13]. Factor analysis showed Fitness and Precision as observable, orthogonal dimensions. Simplicity was excluded, and there was insufficient evidence to support a Generalization dimension. For stochastic process models, this suggests experimental investigation of how models may be distinguished can inform and foster new formalized measures, and understand relations between them. Process mining data is ultimately derived from real-world social activity, and that will narrow the vast space of candidate formalisms.

This research, which is ongoing, investigates empirically identifiable orthogonal quality dimensions for stochastic process models. The experiment design uses a dataset of stochastic process models for real-life processes, collected and evaluated against a set of computationally cheap measures, termed *exploration measures*. Exploration measures are based on existing control-flow and stochastic model measures and around fifteen measures are being considered. Existing stochastic conformance measures [7], [8] are considered *evaluation measures*. The dataset for the experiment comprises thousands of stochastic models, including randomly generated models and those from existing discovery algorithms. Evaluation measures will be used for a subset of discovered models where conformance measure tools were expected to terminate. The subset excludes random models and discovered models which perform poorly using exploration measures. Undermeasured phenomena and possible quality dimensions will be proposed based on a statistical analysis of the results.

## III. Long-Run Process Drift (RQ3)

*RQ3 How can we precisely describe the history of change in an organizations processes?*

A model describes a system. When the system changes, and an automatically discovered model needs to detect that, this problem is called *concept drift* [1, p320]. In process mining, the computational discovery and analysis of organizational process models, this becomes *process drift* [14]. Recent work on stochastic modelling for concept drift [15] suggests stochastic models can be a productively describe these phenomena.

The significant existing literature on concept drift in processes is focused on moments of drift, or anomaly detection. Building on this foundation, the evolution of organizational processes can itself be analysed computationally, yielding models of process change. This *long-run process drift* can be described with *second-order process models* [14].

This prospective research concerns novel algorithms, techniques and newly developed software for describing and understanding long-run process drift. The underlying formalisms are expected to be based on GSPNs and Reconfigurable Petri nets [16]. Models of long-run process drift will be constructed using a combination of stochastic process discovery and adapted concept drift detection techniques. The techniques will be evaluated experimentally against real-world event logs and stochastic quality measures. Initial exploration for suitable event log data is underway. Automatic techniques can make it easier to understand the history of change in a process, what is routine, and what is exceptional, and thereby make better organizations over time.

## References

[1] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin Heidelberg: Springer-Verlag, 2016.

[2] F. Bause and P. Kritzinger, *Stochastic Petri Nets: An Introduction to the Theory*. Vieweg+Teubner Verlag, 2002.

[3] A. Burke, S. J. J. Leemans, and M. T. Wynn, "Stochastic Process Discovery by Weight Estimation," in *Process Mining Workshops*. Cham: Springer, 2021, pp. 260–272.

[4] ——, "Discovering Stochastic Process Models By Reduction and Abstraction," in *Application and Theory of Petri Nets and Concurrency*, ser. Lecture Notes in Computer Science. Springer, 2021, pp. 312–336.

[5] A. Rogge-Solti, W. M. P. van der Aalst, and M. Weske, "Discovering Stochastic Petri Nets with Arbitrary Delay Distributions from Event Logs," in *BPM Workshops*, ser. LNBP. Springer, 2014, pp. 15–27.

[6] M. Camargo, M. Dumas, and O. Gonzlez-Rojas, "Automated discovery of business process simulation models from event logs," *Decision Support Systems*, vol. 134, p. 113284, Jul. 2020.

[7] S. J. J. Leemans and A. Polyvyanyy, "Stochastic-Aware Conformance Checking: An Entropy-Based Approach," in *Advanced Information Systems Engineering*, ser. LNCS. Springer, 2020, pp. 217–233.

[8] S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, and A. Polyvyanyy, "Stochastic process mining: Earth movers stochastic conformance," *Information Systems*, p. 101724, Feb. 2021.

[9] J. Carmona, J. Cortadella, and M. Kishinevsky, "New Region-Based Algorithms for Deriving Bounded Petri Nets," *IEEE Transactions on Computers*, vol. 59, no. 3, pp. 371–384, Mar. 2010, conference Name: IEEE Transactions on Computers.

[10] V. Liesaputra, S. Yongchareon, and S. Chaisiri, "Efficient Process Model Discovery Using Maximal Pattern Mining," in *BPM*, ser. LNCS. Springer, 2015, pp. 441–456.

[11] R. C. Carrasco and J. Oncina, "Learning stochastic regular grammars by means of a state merging method," in *Grammatical Inference and Applications*, ser. LNCS. Berlin: Springer, 1994, pp. 139–152.

[12] S. J. Leemans, A. F. Syring, and W. M. van der Aalst, "Earth movers stochastic conformance checking," in *International Conference on Business Process Management*. Springer, 2019, pp. 127–143.

[13] G. Janssenswillen, N. Donders, T. Jouck, and B. Depaire, "A comparative study of existing quality measures for process discovery," *Information Systems*, vol. 71, pp. 1–15, Nov. 2017.

[14] R. P. J. C. Bose, W. M. P. van der Aalst, I. liobait, and M. Pechenizkiy, "Dealing With Concept Drifts in Process Mining," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 154–171, Jan. 2014.

[15] T. Brockhoff, M. S. Uysal, and W. M. P. v. d. Aalst, "Time-aware Concept Drift Detection Using the Earth Movers Distance," in *ICPM*, Oct. 2020, pp. 33–40.

[16] M. Llorens and J. Oliver, "Structural and dynamic changes in concurrent systems: reconfigurable Petri nets," *IEEE Transactions on Computers*, vol. 53, no. 9, pp. 1147–1158, Sep. 2004.