

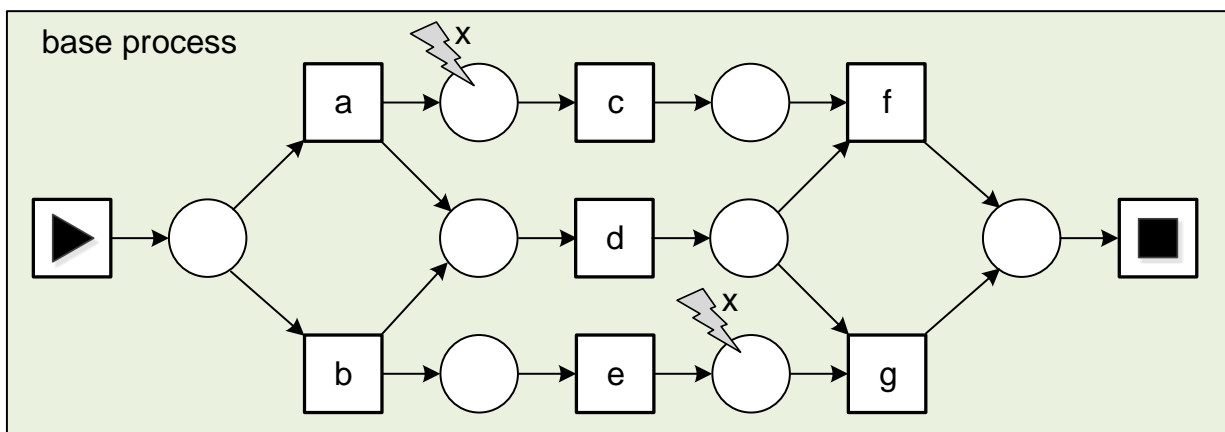
Benchmarking logs to test scalability of process discovery algorithms

Prof.dr.ir. Wil van der Aalst

www.vdaalst.com

There are dozens, if not hundreds, of process discovery algorithms. The time complexity of these algorithms is very different and may depend on various characteristics of the event log. The set of event logs included aims to support the evaluation of the performance of process discovery algorithms.

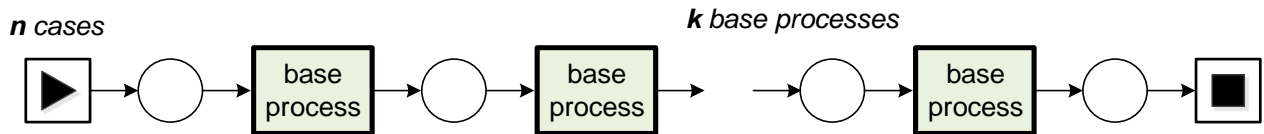
Starting point is the following *base process*:



The places marked with **x** identify states where an exceptional activity **x** can take place. Depending on the algorithm this infrequent and irregular behavior (compared to the rest) can be handled in different ways. The base process is used generate more complex processes using sequential, parallel, choice, and loop composition. In each process the base process is replicated **k** times. The activity name is extended with the number of the base process. Example names are **a_5**, **b_4**, **x_7**, etc. Each process also generates a predefined number of **n** cases (process instances). The files are named accordingly.

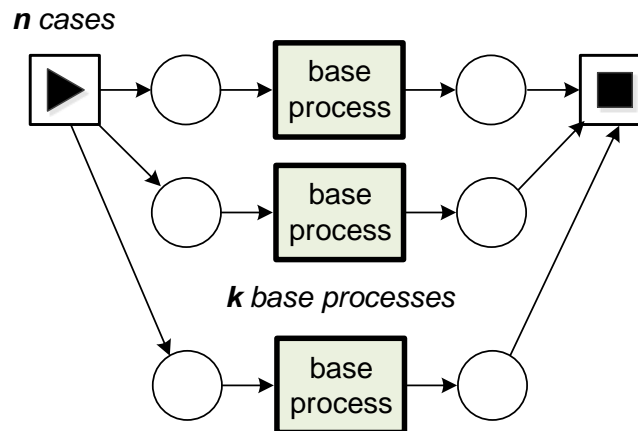
event_log_sequence_k_n

The files **event_log_sequence_k_n** refer to data sets involving the sequence composition. Each event log is available as a csv and xes file. Per event log **n** cases are generated and the base model is replicated **k** times. The model has **k*8** unique activities. The event log has **n*k*4** normal events and a smaller fraction of exceptional events.



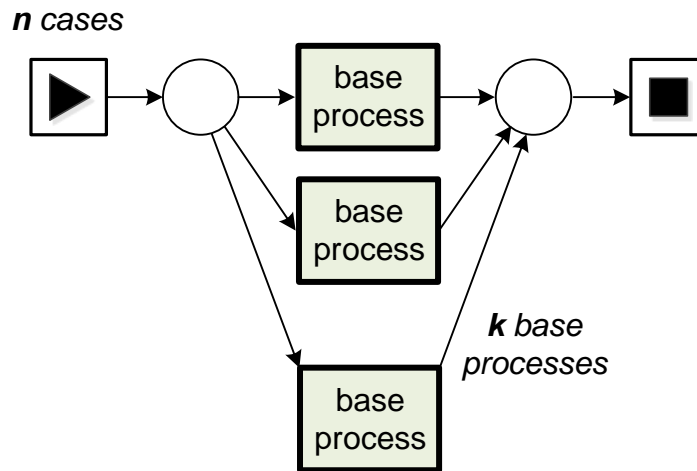
event_log_parallel_k_n

The files **event_log_parallel_k_n** refer to data sets involving the parallel composition. Each event log is available as a csv and xes file. Per event log **n** cases are generated and the base model is replicated **k** times. The model has **k*8** unique activities. The event log has **n*k*4** normal events and a smaller fraction of exceptional events.



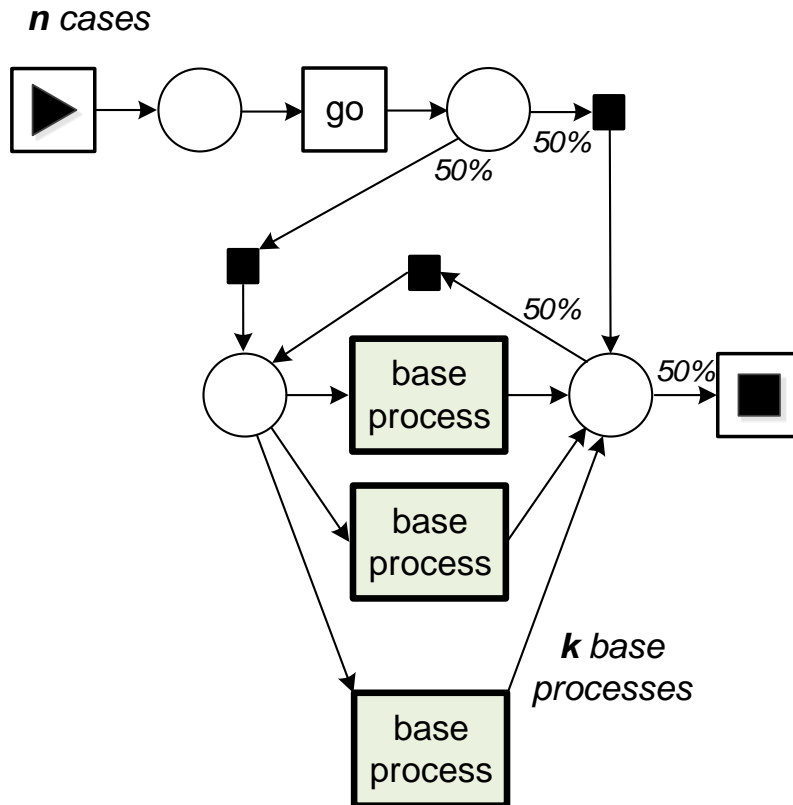
event_log_choice_k_n

The files **event_log_choice_k_n** refer to data sets involving the choice composition. Each event log is available as a csv and xes file. Per event log **n** cases are generated and the base model is replicated **k** times. The model has **k*8** unique activities. Each instance of the base process has the same probability. The event log has **n*4** normal events and a smaller fraction of exceptional events.



event_log_choice_loop_k_n

The files **event_log_parallel_k_n** refer to data sets involving a combination of a loop and choice composition. Each event log is available as a csv and xes file. Per event log **n** cases are generated and the base model is replicated **k** times. The model has **$k*8+1$** unique activities. Activity **go** has been added to avoid empty traces. The event log has **$n*(1.5*4+1)=7*n$** normal events and a smaller fraction of exceptional events.



CPN Files

The largest event logs in this data set have millions of events. If you need even bigger datasets, you can generate these yourself using the CPN Tools sources files included (*.cpn). Each file has two parameters **nofcases** = **n** and **nofdupl** = **k**.

I hope that these event logs will contribute to even more scalable algorithms and tools.

Best regards,

Trento October 2017

Prof.dr.ir. Wil van der Aalst
www.vdaalst.com