Working doc for defining experimental spec

Input parameters:

: The parameters for generating SEQ, AND, OR, and LOOP constructs (‘SEQ’ is the base case of generating two consecutive activities). These are fixed, and are defined in Appendix-A, Algorithm 4 of [Bezerra’s prior work](http://www.sciencedirect.com/science/article/pii/S0306437912000567) on process mining and anomaly detection. Their respective probabilities are 0.4, 0.1, 0.3, and 0.2, at each step of model-generation. Some additional model complexity requirements are enforced, such that a generated model is not accepted unless it contains at least 9 unique activities, and at least 10 unique paths from start to end.

: These parameters are embedded at each node to determine the likelihood of selecting outgoing paths; hence they’re only defined for OR and LOOP free-choice behavior. Whereas determine the complexity of different models, defines the complexity of logs. For a given free-choice node, one may taken an OR/LOOP path with probability or not with probability . More uniform values of , nearer to 0.5, should create more regular behavior, such that the traces will more equally traverse different regions of the graph. More skewed values of should result in less diverse traces, and hence greater difficulty in distinguishing outliers from anomalies.

: Bezerra’s work manually injected anomalies of different types into the traces. Instead I generate anomalies probabilistically during trace generation, just as are embedded in the model. This just seems more faithfully-realistic, since the resulting traces may contain no anomalies, and I hope to show my method is robust when there are no anomalies.

: The anomaly threshold per the Bayesian metric we discussed previously. This parameter is strongly connected to the statistics of log generation, whereby we expect certain regular structure, and define anomalies as occurring in the context of normative substructure. I may also use a pre-filtering criterion, such as only evaluating substructures below some frequency threshold like one-tenth the number of traces, 0.1 \* |{traces}|.

Model generation:

parameters are replicated from Bezerra and fixed, by which I’ve generated 60 models. From each of these models, logs of 1000 traces can be generated for different values of .

Experiments:

Experiment 1: Evaluate the variance of with respect to . Vary , from uniform (0.5) to very skewed (0.9) in increments of 0.1, generating a complete log of 1000 traces for each value. For each log, run SUBDUE and compute accuracy, precision and recall. Plot these values for each value of under a static . (Depending on the results, re-run in different increment scales if needed, such as logarithmic increments: 0.2, 0.1, 0.05, etc.)

🡪Compute the value of sufficient to capture all anomalies, and the precision/recall/accuracy for that value?