A Robust F-measure for Evaluating Discovered Process Models

Jochen De Weerdt*, Manu De Backer*†, Jan Vanthienen*, and Bart Baesens*‡

*Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven
Naamsestraat 69, B-3000 Leuven
Email: Jochen.DeWeerdt@econ.kuleuven.be

‡School of Management, University of Southampton
Highfield Southampton, SO17 1BJ, United Kingdom

†Department of HABE, Hogeschool Gent, Universiteit Gent
Voskenslaan 2, B-9000 Ghent, Belgium

***Abstract*—Within process mining research, one of the most important fields of study is process discovery, which can be defined as the extraction of control-flow models from audit trails or information system event logs. The evaluation of discovered process models is an essential but difficult task for any process discovery analysis. With this paper, we propose a novel approach for evaluating discovered process models based on artificially generated negative events. This approach allows for the definition of a behavioral F-measure for discovered process models, which is the main contribution of this paper.**

I. Introduction

Within the research domain of process mining, a lot of attention has been bestowed on process discovery. Process discovery can be best defined as extracting control-flow process models from information system event logs. Over the years, several different process discovery algorithms [1]–[8] have been proposed in literature. Many of these algorithms are able to deal with specific problems related to control-flow discovery: parallelism, loops, duplicate tasks, noise, and non-local dependencies. However, the effort spent in developing process discovery algorithms is not counterbalanced by the effort put into the evaluation of the discovered process models. As such, Rozinat et al. [9] and De Weerdt et al. [10] identified that a well-defined evaluation framework for process discovery is still missing. The lack of an evaluation framework is primarily due to the difficulty of combining metrics that capture different dimensions along which process models should be evaluated. More specifically, process models cannot be evaluated on recall or sensitivity only. Although this dimension is of utmost importance, other requirements for discovered process models such as precision and generalization beyond observed behavior should be included in any process discovery evaluation analysis.

With this paper, we propose an evaluation approach based upon artificially generated negative events that allows for the application of the well-known F-measure to discovered process models. Accordingly, this paper is structured as follows. Section II outlines the technique of inducing artificial negative events and the different approaches to measuring recall and precision. In Section III, we define our novel evaluation approach, which will be empirically validated in Section IV. Finally, in Section V, a number of key discussion points are commented on before the conclusions are formulated in Section VI.

II. Measuring recall and precision

A. Artificially generating negative events

Event logs rarely contain information about transitions that are not allowed to take place. This makes process discovery an inherently unsupervised learning problem. To make a tradeoff between overly general or overly precise process models, learners make additional assumptions about the given event sequences. Such assumptions are part of the inductive bias of a learner. Process discovery algorithms generally include the assumption that event logs portray the complete behavior of the underlying process and implicitly use this completeness assumption to make a tradeoff between overly general and overly precise process models.

In [8], Goedertier et al. describe a technique to artificially generate negative events based on an event log containing only positive events. The induction procedure is the foundation for their process discovery technique called AGNEs Miner. In this paper, we will make use of the same principle for generating negative events in order to build our evaluation approach.

1) Principle: The technique of inducing negative events is relatively straightforward. Negative events record that at a given position in an event sequence, a particular event cannot occur. At each position in each event trace in the log, it is examined which negative events can be recorded for this position. In a first step, the technique stipulates that the event log is made more compact, by grouping process traces that have identical sequences into grouped process instances. By grouping similar process instances, searching for similar behavior in the event log can be performed more efficiently.

In the second step, all negative events are induced for each grouped process instance. Negative examples can be introduced in grouped process instances by checking at any given positive event whether any other activity type in the event log could occur as an event at this position. For each of these events, it is tested whether there exists a similar sequence in the event log in which at that point the event under consideration occurs. If such an event does not occur in any other sequence, such behavior is not present in the event log. Consequently, a negative event can be added at this position in the event sequence. On the other hand, if a similar sequence is found with this behavior, no negative event is generated.

2) Example: In order to elucidate the principle of generating artificial negative events more clearly, the injection procedure is illustrated with a small example. Take the event log in Figure 1, which is perfectly represented by the process model in Figure 2. When we look for instance at the third positive event (activity c) in trace σ_1, it can be seen that five artificial negative events are induced at this position. Because activity d appears at the same position of event c_p in another similar trace in the event log, namely trace σ_2, activity d cannot be added as a negative event in trace σ_1 at the position of event c_p. In contrary, all the other activity types in this event log can be added as artificial negative events because there are no similar traces in the event log where these activities appear as a positive event at the position of event c_p in trace σ_1. In this way, artificial negative events are added to the event log for each positive event in a log trace (see Table I).

σ_1	abcdeg
σ_2	abdceg
σ_3	abcdefg
σ_4	abdcefg

Fig. 1. Example event log

σ_1	a_p	b_p	c_p	d_p	e_p	g_p
	b_n	a_n	a_n	a_n	a_n	a_n
	c_n	c_n	b_n	b_n	b_n	b_n
	d_n	d_n	e_n	c_n	c_n	c_n
	e_n	e_n	f_n	e_n	d_n	d_n
	f_n	f_n	g_n	f_n	f_n	e_n
	g_n	g_n		g_n	g_n	f_n

TABLE I
ARTIFICIAL NEGATIVE EVENTS FOR TRACE σ_1

The availability of an event log supplemented with artificial negative events allows for the construction of a confusion matrix for a mined control-flow model. By replaying the event log in the model, each and every positive and artificial negative event can be evaluated. In this way, a matrix as in Table II can be constructed denoting whether the positive and negative events are predicted correctly or not by the model. The availability of such a confusion matrix is a substantial advance of the evaluation method based on the artificial negative event generation proposed by Goedertier et al. [8].

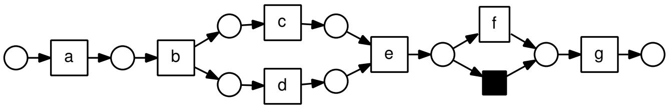


Fig. 2. Process Model for the event log in Figure 1

	Actual pos.	Actual neg.
Pred. pos.	True Pos. (TP)	False Pos. (FP)
Pred. neg.	False Neg. (FN)	True Neg. (TN)

TABLE II
CONFUSION MATRIX

B. Recall

Recall or sensitivity is undoubtedly reckoned as the most important evaluation dimension of discovered process models. A recall metric reflects how much behavior present in the event log is captured by the model. For every process discovery algorithm, it is of utmost importance to render models with good recall because representing the control-flow behavior in an event log is the major objective of any discovery technique.

In recent years, a number of authors proposed metrics for quantifying the recall of a discovered process model in respect to the event log. A well-known recall metric is fitness f [11]. This Petri net based metric punishes for missing and remaining tokens when replaying the event log in the discovered process model. Although often used, there exist different alternatives for the fitness metric. For example, Weijters et al. [2] proposed the Parsing Measure PM, a much more coarse grained metric, that quantifies the percentage of traces that can be replayed in the discovered process model. Other valuable recall metrics are Completeness ($PF_{complete}$) [4] and an alternative Completeness metric as defined by Greco et al. [12].

Originating from their technique allowing to induce artificial negative events into an event log, Goedertier et al. [8] defined two evaluation metrics. One of these metrics is Behavioral recall (r_B^p), which captures the percentage of correctly classified positive events in the event log by the discovered process model. In the next section, we will make use of this recall metric in order to define an F-measure for discovered process models.

C. Precision

The key challenge for any process discovery technique is to come up with accurate process models that at the same time find the right balance between underfitting (overly general process model) and overfitting (overly precise process model). The precision evaluation dimension gauges whether a mined process model does not underfit the behavior present in the event log. As illustrated in Figure 3, a flower model (Figure 3(c)) is very sensitive because it allows any sequence of activities, nevertheless this model does not deliver any knowledge with respect to the control-flow behavior in the event log. This is the main motivation why process models should also be evaluated along the precision dimension.

In the literature, few precision metrics have been proposed. Greco et al. [12] defined Soundness. Soundness is the percentage of traces compliant with the process model that have been registered in the log. Calculating Soundness is not straightforward because enumerating all possible paths in a process model is hard. Even for smaller process models, it might be impossible to determine all the traces that are compliant with a process model.

A by far more used precision metric is the advanced behavioral appropriateness ($a_B^{'}$) as defined by Rozinat et al. [11]. Although this metric is theoretically sound in order to evaluate the precision of a process model, we have illustrated previously that there are a number of drawbacks [10]. For instance, the calculation requires an exhaustive simulation which is computationally very demanding. Moreover, the implementation of this exhaustive simulation for calculating the metric within the ProM framework [13] is only approximate.

D. Precision based on artificially generated negative events

As explained before, with the availability of an event log supplemented with artificial negative events, a confusion matrix can be composed. Drawing upon this confusion matrix, a novel precision metric is defined. By evaluating the true positive events (TP) in respect to all predicted positive events ($TP+FP$), the precision dimension of a mined process model can be assessed. Therefore we define Behavioral Precision p_B as in Equation 1.

$$p_B = \left(\frac{\sum_{i=1}^{k} n_i TP_i}{\sum_{i=1}^{k} n_i TP_i + \sum_{i=1}^{k} n_i FP_i} \right) \tag{1}$$

Note that k is the total number of different grouped process instances in the event log. Index i runs over all different grouped process instances. n_i is the number of instances within one group of similar process instances. TP denotes the correctly predicted positive events, while FP denotes the incorrectly predicted artificial negative events.

This precision metric, which fully coincides with the standard definition of precision within the field of data mining, has the advantage of requiring much less computational resources in respect to the other precision metrics mentioned earlier. Furthermore, having now both a recall metric and a precision metric based on artificially generated negative events, we are able to define an F-measure for evaluating discovered process models, as discussed in the next section.

III. Applying the F-measure to Process Discovery

A. Definition

Originally, the F-measure (Equation 2) was proposed by van Rijsbergen [14] in the context of information retrieval. In the fields of machine learning and data mining [15], the F-measure is often used as a standard balance between precision and recall for evaluating point classifiers.

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

In fact, the F-measure can be seen as a point classifier alternative of the AUC (Area Under the ROC-curve) [16], a popular evaluation metric for rank classifiers. AUC cannot be used as an evaluation approach for process discovery because a discovered process model can only be seen as a point classifier for each individual event in the event log. In particular, a discovered process model determines whether a positive or artificial negative event is correctly classified or not, it does not assign a probability to this classification.

In order to take into account the typicalities of the process discovery evaluation setting, we define the Behavioral F-measure F_B for a discovered process model in Equation 3, entirely founded upon the formula in Equation 2.

$$F_B = 2 \times \frac{p_B \times r_B^p}{p_B + r_B^p}, \quad with \tag{3}$$

$$p_B = \left(\frac{\sum_{i=1}^{k} n_i TP_i}{\sum_{i=1}^{k} n_i TP_i + \sum_{i=1}^{k} n_i FP_i} \right)$$
$$r_B^p = \left(\frac{\sum_{i=1}^{k} n_i TP_i}{\sum_{i=1}^{k} n_i TP_i + \sum_{i=1}^{k} n_i FN_i} \right)$$

Note that the meaning of the symbols remain exactly the same as in Equation 1, with FN denoting the falsely predicted positive events.

B. Advantages

The key advantage of this novel evaluation approach consists of a transparent and robust method to combine two important evaluation dimensions for discovered process models: recall and precision. More precisely, our approach allows for careful comparison of different process models obtained from the same event log. As such, benchmarking state-of-the-art process discovery techniques can be carried out in an understandable and effective way. What is more, we think that this novel evaluation approach is an important step towards a well-defined evaluation framework for process discovery.

Also, the availability of a new precision metric, quantifying whether a process model does not underfit the data, is of major importance. The Behavioral Precision (p_B) is theoretically sound and can be calculated swiftly. In this way, this metric is a useful alternative for the currently available precision metrics, that suffer from computational inefficiencies.

In order to empirically validate our proposed evaluation approach, we will demonstrate the application of the Behavioral F-measure (F_B) within a benchmarking experiment based on 20 artificial event logs. This analysis is presented in the next section.

IV. Empirical validation

This section reports on the empirical validation of the proposed evaluation approach. Hence, we will first illustrate the approach with a very simple process discovery example before the evaluation approach is compared to traditional process discovery evaluation metrics in a benchmarking experiment of process discovery techniques making use of 20 artificially constructed event logs.

A. An illustrative example

Figure 3 illustrates the novel evaluation approach with a simple example. For an event log 3(a) and three corresponding control-flow models 3(c) 3(b) 3(d), two evaluation approaches are compared. Approach A is our novel evaluation approach based on artificially generated negative events enabling the computation of the F-measure. The other evaluation approach consists of a procedure described by Rozinat et al. [11], to equally weigh fitness and advanced behavioral appropriateness in order to evaluate mined process models along two dimensions. As can be seen from Table 3(e), both approaches correctly evaluate the best process model and punish the imprecise flower model and the incomplete model 3(d). However, there exist some small differences between the two approaches. First of all, the F-measure punishes the imprecise and incorrect process models more severely, which can be judged advantageous. Furthermore, there is also a discrepancy between the two precision metrics p_B and $a_B^{'}$. This discrepancy is due to the differences in how precision is quantified. p_B depends on replaying the event log in the discovered Petri net model, while $a_B^{'}$ only takes into account the ratios of sometimes follows and sometimes precedes relations in the model and in the event log.

B. Benchmarking state-of-the-art process discovery techniques with artificial event logs

1) Techniques: This benchmarking experiment compares six state-of-the-art process discovery techniques (see Table III) in order to validate the novel evaluation approach. Next to the six techniques, a flower model is included in the benchmark as a reference model.

Name	Author	Year
α^+	van der Aalst et al. [1], [17]	2004
HeuristicsMiner	Weijters et al. [2]	2006
α^{++}	Wen et al. [3]	2007
GeneticMiner	Alves de Medeiros et al. [4]	2007
DTGeneticMiner	Alves de Medeiros [18]	2007
AGNEsMiner	Goedertier et al. [8]	2009

TABLE III
PROCESS DISCOVERY TECHNIQUES

2) Artificial event logs: In order to benchmark the six selected process discovery techniques, an experiment with 20 event logs has been set up. These event logs have previously been used by Alves de Medeiros et al. [18] to evaluate the GeneticMiner algorithm. The characteristics of the artificial event logs are presented in Table IV. In order to validate the robustness of the algorithms, we conducted two different experiments: once we applied the selected discovery techniques on the event logs without any addition of noise and once we randomly injected 20% noise using the available noise injection filter in the ProM framework.

3) Statistical tests: A procedure described in Demšar [19] is followed to statistically test the results of the benchmarking experiment. In a first step of this procedure, the Friedman test [20] is performed which is a non-parametric equivalent of the well known ANOVA test (ANalysis Of VAriance). The null hypothesis of the Friedman test states that all techniques perform equivalent. The test statistic is defined as:

$$\chi_F^2 = \frac{12P}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

	activity types	**$\neq$ process inst.**	**process inst.**	**$\|$ activity types**	**loop**	**skip**	**non-free choice**	**duplicate tasks**
a10skip	12	6	300	1		✓		
a12	14	5	300	2				
a5	7	13	300	1	✓			
a6nfc	8	3	300	1			✓	
a7	9	14	300	4				
a8	10	4	300	1				
betaSimplified	13	4	300	0		✓	✓	✓
choice	12	16	300	0				
DriversLicense	9	2	300	0				
DriversLincensel	11	87	350	1	✓	✓	✓	✓
herbstFig3p4	12	32	300	3	✓			
herbstFig5p19	8	6	300	1				✓
herbstFig6p18	7	153	300	0	✓			
herbstFig6p31	9	4	300	0				✓
herbstFig6p36	12	2	300	0			✓	
herbstFig6p38	7	5	300	3				✓
herbstFig6p41	16	12	300	4				
l2l	6	10	300	0	✓			
l2lOptional	6	9	300	0	✓			
l2lSkip	6	8	300	0	✓			

TABLE IV
EVENT LOG PROPERTIES

with R_j the average rank of algorithm $j = 1, 2 \ldots k$ over P data sets. Under the null hypothesis, the Friedman test statistic is distributed according to χ_F^2 with $k-1$ degrees of freedom, at least when P and k are big enough ($P > 10$ and $k > 5$). Otherwise, exact critical values are used based on an adjusted Fisher z-distribution.

If the null hypothesis of equivalent performing techniques is rejected by the Friedman test, a post-hoc Bonferroni-Dunn test [21] is applied to compare the process discovery techniques. The post-hoc Bonferroni-Dunn test is a non-parametric alternative of the Tukey test and is defined as:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6P}}$$

with critical value q_α based on the Studentized range statistic divided by $\sqrt{2}$, and an additional Bonferroni correction by dividing the confidence level α by the number of comparisons made, $\frac{\alpha}{(k-1)}$, to control for family wise testing. This results in a lower confidence level and thus in higher power. The difference in performance of the best performing technique and other techniques is significant if the corresponding average ranks differ by at least the Critical Distance (CD).

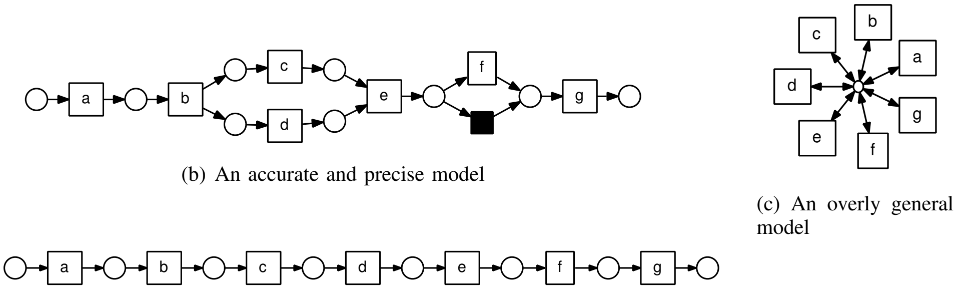
4) Results: Results of the benchmarking experiment are presented in Figure 4 and Table V. In what follows, the most important conclusions are discussed.

The aim of this benchmarking study is twofold. First of all, we use this experiment in order to evaluate the novel evaluation approach proposed in Section III. Secondly, the results of this benchmarking experiment allow for assessing the performance of different process discovery techniques on artificial data sets. Because in the second part of the experiment, we subjected the techniques to the same event logs but with the injection of noise, the noise robustness of the different techniques is also analyzed.

As explained in Section IV-B3, the results of the different process discovery techniques are compared by first applying

σ_1	abcdeg
σ_2	abdceg
σ_3	abcdefg
σ_4	abdcefg

(a) Event log



(b) An accurate and precise model

(c) An overly general model

(d) An incomplete model

	Evaluation approach A			Evaluation approach B		
	r_B^p	p_B	F_B	f	$a_B^{'}$	$\frac{1}{2}f + \frac{1}{2}a_B^{'}$
Best Model	1	1	1	1	1	1
Flower Model	1	$0,15$	$0,26$	1	$0,10$	$0,55$
Incomplete Model	$0,85$	$0,76$	$0,80$	$0,87$	1	$0,93$

(e) Evaluation results

Fig. 3. Illustration of the novel evaluation approach

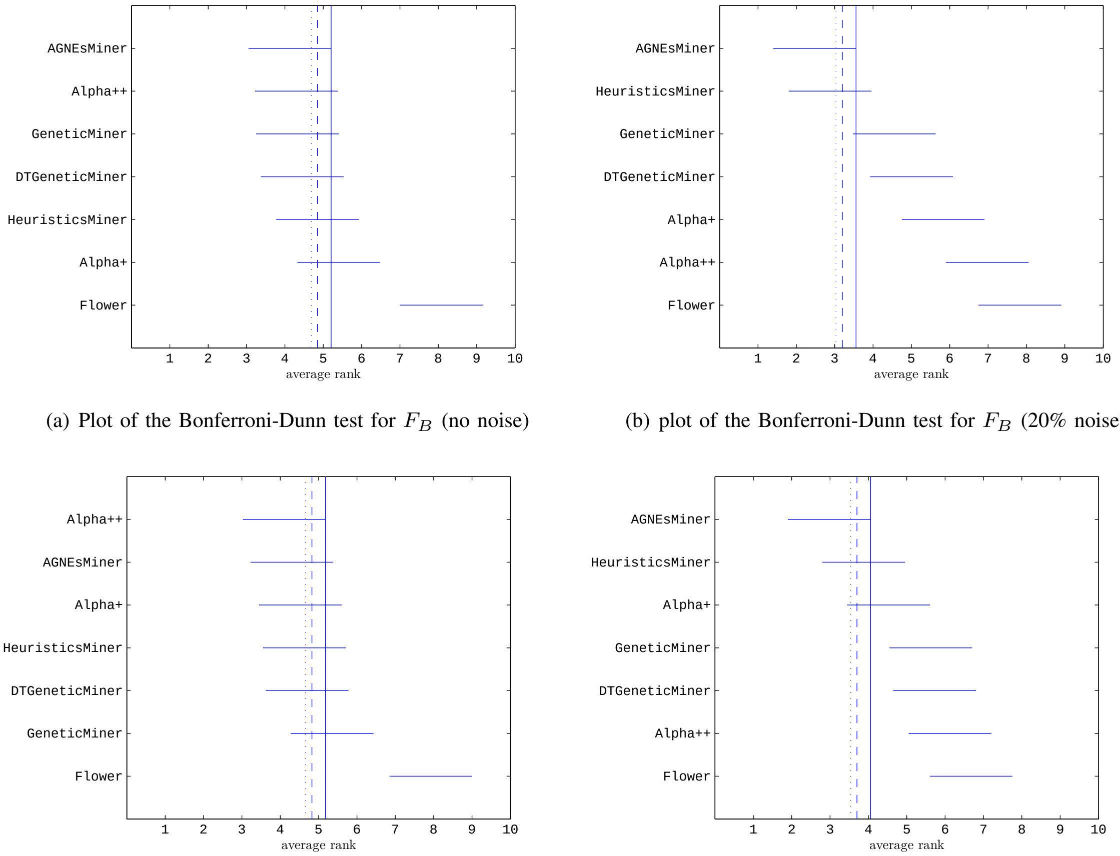
a Friedman test, followed by a Bonferroni-Dunn test. We performed these statistical tests for both evaluation approaches as presented in Section IV-A. The Friedman test resulted in a p-value close to zero (p values between 0.0000 and 0.0005) indicating the existence of significant differences across the applied techniques, both in case no noise was added and in case 20% noise was added. In a next step, the Bonferroni-Dunn test to compare the performance of all the models with the single best performing model is applied. The results are plotted in Figure 4. The horizontal axis in these figures corresponds to the average rank of a technique across the different artificial event logs. The techniques are represented by a horizontal line; the more this line is situated to the left, the better performing a technique is. The left end of this line depicts the average ranking while the length of the line corresponds to the critical distance for a difference between any technique and the best performing technique to be significant at the 99% confidence level. The dotted, dashed and full vertical lines in the figures indicate the critical difference at respectively the 90%, 95% and 99% confidence level. A technique is significantly outperformed by the best technique if it is located at the right side of the vertical line.

Taking this into account, it can be concluded from Figures 4(a) and 4(c) that without noise, the techniques do not significantly differ both in terms of F-measure and in terms of $\frac{1}{2}f + \frac{1}{2}a_B^{'}$. However, when introducing noise, the picture shifts profoundly. When taking into account $\frac{1}{2}f + \frac{1}{2}a_B^{'}$ as well as F_B, AGNEsMiner and HeuristicsMiner clearly outperform the other algorithms. For α^+ and α^{++}, this is not surprising because it is known that these algorithms are not robust to noise. However, regarding the genetic algorithms, the underperformance is quite surprising. Generally speaking, these algorithms are described as noise robust. However, the 20% noise addition in our experiment causes problems for the genetic algorithms in terms of discovering the correct process model.

This analysis is confirmed by the aggregated, nonparametrical results displayed in Table V. Note that, in this table, the **best** average performance over the 20 event logs is underlined and denoted in bold face for each metric. A paired t-test was used to test the significance of the performance differences. Performances that are **not significantly different at the 95% level** from the top-ranking performance with respect to a one-tailed paired t-test are tabulated in bold face. Statistically *significant underperformances at the 99% level* are emphasized in italics. Performances significantly different at the 95% level but not at the 99% level are reported in normal font.

The absolute figures confirm the good performance of the AGNEsMiner algorithm, both in terms of recall and precision. However, we should recognize the evaluation of this technique might be slightly biased because it is founded upon the same principle as the novel evaluation approach itself. Nevertheless, it should be concluded that without noise, many different algorithms perform more or less equally. Once noise is introduced, only HeuristicsMiner and AGNEsMiner remain able to render good process models. This last conclusion is founded upon results from both evaluation approaches considered.

Although the analysis of the process discovery algorithms is very useful, the ultimate reason for setting up this experiment is evaluating our newly proposed evaluation approach. Focussing on recall first, it can be concluded that r_B^p and f are very similar. From Tables V(a) and V(b), it can be seen that there are slight differences, but in general, these two metrics yield the same result. In contrast, comparing the two precision metrics p_B and $a_B^{'}$, it is concluded that these metrics do not portray similar behavior in our experiment when evaluating the precision of the discovered process models. In our opinion, there exist important drawbacks regarding $a_B^{'}$, which are due to the exhaustive simulation that is required to calculate this metric. In our previous study [10], we already pinpointed a number of concerns regarding this precision metric, which are confirmed by the experiment presented. Even more, we found out that in case event logs portray loop behavior,



(a) Plot of the Bonferroni-Dunn test for F_B (no noise)

(b) plot of the Bonferroni-Dunn test for F_B (20% noise)

(c) plot of the Bonferroni-Dunn test for $\frac{1}{2}f + \frac{1}{2}a_B^{'}$ (no noise)

(d) plot of the Bonferroni-Dunn test for $\frac{1}{2}f + \frac{1}{2}a_B^{'}$ (20% noise)

Fig. 4. Ranking of process discovery techniques for F_B in case no noise was added 4(a), F_B with 20% noise added 4(b), $\frac{1}{2}f + \frac{1}{2}a_B^{'}$ without noise 4(c), and $\frac{1}{2}f + \frac{1}{2}a_B^{'}$ with 20% noise 4(d). The dotted vertical line represents the 90% significance level, the dashed line the 95% significance level and the full line the 99% significance level.

$a_B^{'}$ often resulted in a considerable underestimation of the precision of the discovered process model. This is again due to inconsistencies in the exhaustive simulation procedure that is behind this metric.

The aforementioned analysis of the precision metrics, highly influences the analysis of the evaluation approaches. As for the experiment without noise, represented in Table V(a), conclusions based upon the different approaches completely contradict. Our novel evaluation approach indicates that GeneticMiner is the best performing technique, whereas according to evaluation approach B, GeneticMiner is the least appropriate technique. As said, this is primarily due to the differences in the precision metrics. Regarding the experiment with 20% noise addition, both evaluation approaches present equivalent results, but this is caused by the fact that two of the considered techniques clearly outperform the other techniques.

C. Conclusion

The benchmarking experiment presented allows for the formulation of a number of important conclusions. First of all, there exist important differences between the considered evaluation approaches. These differences are primarily due to the discrepancy of the precision metrics that underly these approaches. Because of the reliability issues with $a_B^{'}$, we think that our novel evaluation approach in terms of the application of the F-measure for process discovery models is highly valuable.

Besides, our evaluation approach has the advantage to solidly combine two crucial evaluation dimensions.Furthermore, the use of artificially generated negative events allows for swift calculation of both recall and precision, which is an important feature in respect to the exhaustive simulation that is required to calculate the advanced behavioral appropriateness. Accordingly, we describe our approach to be a robust evaluation method for process discovery models.

Finally, regarding the process discovery techniques under study, we found that AGNEsMiner and HeuristicsMiner were confirmed to be robust to noise, whereas the genetic approaches appeared less prone than previously considered. This conclusion should be investigated further because only one noise percentage was considered.

V. Discussion

A. Completeness Assumption

As explained, the proposed evaluation approach is entirely based on the principle of inducing artificial negative events into

(a) Without noise

	Evaluation approach A			Evaluation approach B		
	r_B^p	p_B	F_B	f	$a_B^{'}$	$\frac{1}{2}f + \frac{1}{2}a_B^{'}$
AGNEsMiner	**0,9979**	**0,9215**	**0,9507**	**0,9953**	**0,7965**	**0,8959**
α^+	0,9524	0,8616	0,8922	0,9685	**0,8082**	**0,8884**
α^{++}	0,9721	**0,9213**	**0,9407**	0,9838	**0,8642**	**0,9240**
DTGeneticMiner	**0,9991**	**0,9144**	**0,9493**	**0,9965**	0,7432	**0,8698**
GeneticMiner	0,9845	**0,9362**	**0,9538**	**0,9981**	0,7015	0,8498
HeuristicsMiner	*0,9586*	**0,9069**	0,9273	0,9733	**0,7742**	**0,8737**
Flower	1,0000	0,1174	0,2083	1,0000	0,1850	0,5925

(b) 20% noise

	Evaluation approach A			Evaluation approach B		
	r_B^p	p_B	F_B	f	$a_B^{'}$	$\frac{1}{2}f + \frac{1}{2}a_B^{'}$
AGNEsMiner	**0,9848**	**0,9229**	**0,9494**	**0,9808**	**0,8397**	**0,9103**
α^+	*0,6128*	*0,3574*	*0,4186*	*0,7626*	**0,8505**	*0,8066*
α^{++}	*0,3764*	*0,3711*	*0,3454*	*0,6996*	0,6294	*0,6645*
DTGeneticMiner	*0,8606*	*0,4148*	*0,5496*	*0,9071*	*0,4658*	*0,6865*
GeneticMiner	*0,8878*	*0,4550*	*0,5909*	*0,9330*	*0,4552*	*0,6941*
HeuristicsMiner	0,9302	0,7698	0,8607	**0,9619**	**0,7048**	0,8334
Flower	1,0000	0,1174	0,2083	1,0000	0,1996	0,5998

TABLE V
BENCHMARKING EXPERIMENT - AGGREGATED RESULTS OF BOTH EVALUATION APPROACHES

an event log. In order to induce these negative events, we make use of the assumption that an event log portrays the complete behavior of the underlying process. We acknowledge that in real-life situations, this completeness assumption might become problematic because an event log might not completely capture all possible behavior. When an event log does not capture all behavior, artificial negative events might be falsely introduced into the event log. However, this completeness assumption is an inherent problem for many unsupervised data mining tasks. When building a model based on a certain data set, it is always complicated to induce models that generalize towards unseen behavior that is not represented in the data. This is also the case for process discovery.

Furthermore, we argue that it is not always the case that one should conclude that an event log is incomplete with respect to all possible behavior. For example, incompleteness of an event log is strongly determined by the time frame of the event log under consideration and also by the number of cases. Often, only domain experts will be able to assess to what extent a certain event log is incomplete.

Nevertheless, we recognize that in highly flexible environments, it might be the case that an event log does not capture all possible behavior of a deployed process. However, we think that our novel evaluation approach is definitely of added value for the evaluation of discovered process models. Of course, when applying our approach, one should always bare in mind the consequences of the completeness assumption. When you are investigating highly unstructured processes, it might be necessary to think about other evaluation methods. However, currently available process discovery techniques face the same problem with respect to completeness. Günther et al. [5] identified that traditional process discovery approaches perform very poor when dealing with highly unstructured event logs. Accordingly, we think that our evaluation approach is definitely applicable to more structured cases of process discovery. When traditional process discovery techniques come up with reasonably interpretable results, we argue that this novel evaluation approach is able to assess discovered process models in a robust way.

What is more, we think that our approach might be dynamically adapted to counter the problem of the closed world assumption. In this paper we did not consider weighing recall and precision differently. However, by varying a parameter that controls the balance between recall and precision (typically called β), different F-measures can be defined weighing recall and precision unequally. By attaching less weight to the precision dimension, one reduces the impact of the completeness assumption because only this dimension takes into account falsely predicted negative events. In future research, it should be investigated whether the use of weights can be used advantageously in order to cope with more unstructured event logs. Preferably, this should be investigated using real-life event logs with different levels of unstructuredness.

Finally, we think that in order to deal with the strict completeness assumption, it should also be investigated whether we can transform a process discovery model from a point classifier into a rank classifier from an event viewpoint. When introducing probabilities in a smart way, it might be possible to alleviate the completeness issue within the limits of the evaluation approach discussed in this paper. Though, we consider this a thought-provoking challenge for future research.

B. Real-life event logs

Closely related to the aforementioned remark on the completeness assumption, we are convinced that both process discovery techniques and evaluation approaches should be assessed using real-life event logs. This is because such a setup allows for an analysis of the scalability of the algorithms and evaluation approaches. What is more, this type of study would yield important information concerning the practical application of certain process discovery techniques in general. Finally, the use of real-life event logs should allow an investigation on how to better cope with the completeness assumption in practice.

C. Root cause analysis

Identification of the root causes of control-flow inaccuracies of a certain discovered process model is an issue that is not addressed in this paper. Because the proposed evaluation method is not yet implemented as a ProM plug-in, we cannot provide a means for root cause analysis. However, it should be possible to devise a plug-in based on the novel evaluation approach that visually represents flaws in the process model under investigation. For the fitness and advanced behavioral appropriateness, there exists a plug-in in the ProM framework, called Conformance Checker, that allows to some extent the identification of root causes of control-flow inaccuracies.

D. Overfitting: generality dimension

As explained in Section II-C, precision gauges whether a model underfits the data. However, the proposed evaluation method does not allow for the verification to what extent a process model overfits the data. Our evaluation approach is not able to detect whether a process model overfits the data by for example just enumerating all the traces in the event log. Rozinat et al. [11] defined the advanced structural appropriateness as a metric to determine whether a process model overfits the data. This metric takes into account structural properties of the mined process models in terms of alternative duplicate tasks and redundant invisible tasks. Currently, quantifying the generality dimension based upon artificially generated negative events seems unfeasible. This causes our approach to be not fully applicable yet as a general evaluation framework for process discovery. In future work, we will investigate how we can quantify overfitting process models and thus further improve upon the important problem of finding a balance between generality and precision.

VI. Conclusion

With this paper, we have proposed a novel evaluation approach that allows for the application of the F-measure to the field of process discovery. Our approach is founded on a technique allowing generation of artificially generated negative events. The approach was discussed and evaluated in a benchmarking experiment with artificial event logs. This analysis yielded that our approach is definitely of added value and an interesting step towards a more holistic approach to process discovery evaluation.

Acknowledgment

The authors would like to thank the Flemish Research Council for financial support under Odysseus grant B.0915.09. Furthermore, we would like to thank Stijn Goedertier for his groundbreaking work on process discovery and Karel Dejaeger for his support on the statistical testing procedures for the benchmarking experiment.

References

[1] W. M. P. van der Aalst, A. J. M. M. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004.

[2] A. J. M. M. Weijters, W. M. P. van der Aalst, and A. K. Alves de Medeiros, "Process mining with the heuristicsminer algorithm," Eindhoven University of Technology, BETA Working Paper Series 166, 2006.

[3] L. Wen, W. M. P. van der Aalst, J. Wang, and J. Sun, "Mining process models with non-free-choice constructs," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 145–180, 2007.

[4] A. K. Alves de Medeiros, A. J. M. M. Weijters, and W. M. P. van der Aalst, "Genetic process mining: an experimental evaluation," *Data Mining and Knowledge Discovery*, vol. 14, no. 2, pp. 245–304, 2007.

[5] C. W. Günther and W. M. P. van der Aalst, "Fuzzy mining - adaptive process simplification based on multi-perspective metrics," in *BPM*, ser. Lecture Notes in Computer Science, G. Alonso, P. Dadam, and M. Rosemann, Eds., vol. 4714. Springer, 2007, pp. 328–343.

[6] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun, "A novel approach for process mining based on event types," *J. Intell. Inf. Syst.*, vol. 32, no. 2, pp. 163–190, 2009.

[7] D. R. Ferreira and D. Gillblad, "Discovering process models from unlabelled event logs," in *BPM*, ser. Lecture Notes in Computer Science, U. Dayal, J. Eder, J. Koehler, and H. A. Reijers, Eds., vol. 5701. Springer, 2009, pp. 143–158.

[8] S. Goedertier, D. Martens, J. Vanthienen, and B. Baesens, "Robust process discovery with artificial negative events," *Journal of Machine Learning Research*, vol. 10, pp. 1305–1340, 2009.

[9] A. Rozinat, A. K. A. de Medeiros, C. W. Günther, A. J. M. M. Weijters, and W. M. P. van der Aalst, "The need for a process mining evaluation framework in research and practice," in *Business Process Management Workshops*, 2007, pp. 84–89.

[10] J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens, "A critical evaluation study of model-log metrics in Process Discovery," in *Proceedings of the 6th International Workshop on Business Processes Intelligence (BPI2010)*, 2010.

[11] A. Rozinat and W. M. P. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Information Systems*, vol. 33, no. 1, pp. 64–95, 2008.

[12] G. Greco, A. Guzzo, L. Pontieri, and D. Saccà, "Discovering expressive process models by clustering log traces," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1010–1027, 2006.

[13] W. M. P. van der Aalst, B. F. van Dongen, C. W. Günther, A. Rozinat, E. Verbeek, and T. Weijters, "Prom: The process mining toolkit," in *BPM (Demos)*, ser. CEUR Workshop Proceedings, A. K. A. de Medeiros and B. Weber, Eds., vol. 489. CEUR-WS.org, 2009.

[14] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.

[15] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.

[16] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[17] A. K. Alves de Medeiros, B. F. van Dongen, W. M. P. van der Aalst, and A. J. M. M. Weijters, "Process mining: Extending the alpha-algorithm to mine short loops," Eindhoven University of Technology, BETA Working Paper Series 113, 2004.

[18] A. K. Alves de Medeiros, "Genetic process mining," Ph.D. dissertation, TU Eindhoven, 2006.

[19] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[20] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[21] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.