Probabilistic Evaluation of Process Model Matching Techniques

Elena Kuss[1], Henrik Leopold[2], Han van der Aa[2], Heiner Stuckenschmidt[1], and Hajo A. Reijers[2]

[1] Research Group Data and Web Science
University of Mannheim, 68163 Mannheim, Germany
elena|heiner@informatik.uni-mannheim.de
[2] Department of Computer Science
Vrije Universiteit Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
h.leopold|j.h.vander.aa|h.a.reijers@vu.nl

Abstract. Process model matching refers to the automatic identification of corresponding activities between two process models. It represents the basis for many advanced process model analysis techniques such as the identification of similar process parts or process model search. A central problem is how to evaluate the performance of process model matching techniques. Often, not even humans can agree on a set of correct correspondences. Current evaluation methods, however, require a binary gold standard, which clearly defines which correspondences are correct. The disadvantage of this evaluation method is that it does not take the true complexity of the matching problem into account and does not fairly assess the capabilities of a matching technique. In this paper, we propose a novel evaluation method for process model matching techniques. In particular, we build on the assessment of multiple annotators to define probabilistic notions of precision and recall. We use the dataset and the results of the Process Model Matching Contest 2015 to assess and compare our evaluation method. We find that our probabilistic evaluation method assigns different ranks to the matching techniques from the contest and allows to gain more detailed insights into their performance.

Keywords: Process Model Matching, Non-binary Evaluation, Matching Performance Assessment

1 Introduction

Process models are conceptual models used for purposes ranging from the documentation of organizational operations [6] to the definition of requirements for information systems [19]. Process model *matching* refers to the automatic identification of corresponding activities between such models. The application scenarios of matching techniques are manifold. They include the analysis of model differences [12], harmonization of process model variants [13], process model

search [9], and the detection of process model clones [22]. The challenges associated with the matching task are considerable. Among others, process model matching techniques must be able to deal with heterogeneous vocabulary, different levels of granularity, and the fact that typically only a few activities from one model have a corresponding counterpart in the other. In recent years, a significant number of process model matching techniques have been defined to address these problems (cf. [4, 10, 11, 14, 23, 24]). One central question that concerns all these techniques is how to demonstrate that they actually perform well.

To demonstrate the performance of a matching technique, authors typically conduct evaluation experiments that consist of solving a concrete matching problem. So far, the basis of such evaluation experiments is a binary *gold standard* created by humans, which clearly defines which correspondences are correct. By comparing the correspondences generated by the matching technique against those from the gold standard, it is possible to compute the well-established metrics precision, recall, and F-measure [15]. In this way, the performance of an approach can be quantified and compared against others.

The disadvantage of this evaluation method is that it does not take the true complexity of the matching problem into account. This is, for instance, illustrated by the gold standards of the Process Model Matching Contests (PMMCs) 2013 and 2015. The organizers of the contests found that there was not a single model for which two independent annotators fully agreed on all correspondences [1, 3]. A binary gold standard, however, implies that any correspondence that is not part of the gold standard is incorrect and, thus, negatively affects the above mentioned metrics. This raises the question of why the performance of process model matching techniques is determined by referring to a single correct solution when human annotators may not even agree on what this correct solution is.

Recognizing the need for a more suitable evaluation strategy for process model matching techniques, we use this paper to propose a novel *process model matching evaluation method*. Instead of building on a binary gold standard, we define a non-binary gold standard that combines a number of binary assessments created by individual annotators. This allows us to express the *support* that exists for correspondences in the non-binary gold standard as the fraction of annotators that agree that a given correspondence is correct. The *probabilistic* precision and recall metrics we define take these support values into consideration when assessing the performance of matching techniques. As such, correspondences with high support values have a greater impact on precision and recall scores than correspondences with low support.

The rest of the paper is organized as follows. Section 2 illustrates the problems associated with the usage of binary gold standards for process model matching evaluation. In Section 3, we define the non-binary gold standard and probabilistic precision and recall metrics. In Section 4, we assess and compare the proposed probabilistic evaluation metrics by applying our method on the dataset of the PMMC 2015. Section 5 discusses related work on the evaluation of matching techniques in different application domains. Finally, we conclude the paper and discuss future research directions in Section 6.

2 Problem Illustration

Given two process models with their respective sets of activities A_1 and A_2, the goal of process model matching is to automatically identify the activities (or sets of activities) from A_1 and A_2 that represent similar behavior. The result of conducting process model matching, therefore, is a set of activity correspondences. One of the central questions in the context of process model matching is how to assess whether the correspondences identified by a matching technique are correct. To illustrate the problems associated with the *evaluation* of process model matching, consider the example depicted in Figure 1. It shows two simplified process models from the PMMC 2015 [1], as well as possible correspondences between them.

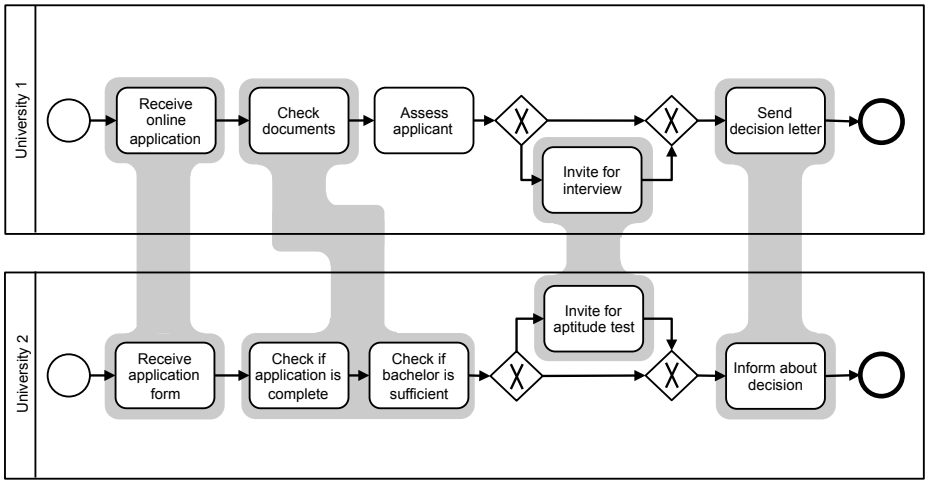


Fig. 1: Two process models and possible correspondences

Upon closer inspection of the correspondences shown in Figure 1 it becomes clear that many of the correspondences are actually disputable. Consider, for instance, the correspondence between "*Receive online application*" from University 1 and "*Receive application form*" in the process of University 2. On the one hand, we can argue in favor of this correspondence because they both describe the receipt of an application document. On the other hand, we can argue that these activities do not correspond to each other because the former relates to an online procedure, whereas the second refers to a paper-based step. We can bring forward similar arguments for the correspondence between "*Invite for interview*" and "*Invite for aptitude test*". Both activities aim to assess whether an applicant is suitable for a university. However, an interview is clearly a different assessment instrument than an aptitude test, which makes the correspondence disputable. Lastly, also the correspondence between "*Check documents*" from University 1

and the two activities "*Check if application is complete*" and "*Check if bachelor is sufficient*" from University 2 is controversial. If we consider the activity "*Check documents*" to solely relate to the completeness of the documents, then the activity "*Check if bachelor is sufficient*" should not be part of the correspondence. These examples illustrate that it may be hard and, in some cases, even impossible to agree on a single *correct* set of correspondences. Despite this, the evaluation of process model matching techniques currently depends on the definition of such a single set of correct correspondences, i.e. a binary gold standard. This binary gold standard is needed to compute precision, recall, and F-measure, which are traditionally used to evaluate process model matching techniques (cf. [1, 3, 14, 23, 24]).

In this paper, we argue that a binary evaluation of process model matching techniques does not account for the full complexity of the process model matching task. Binary evaluation does not consider disagreements that may exist regarding the correctness of correspondences. Therefore, binary evaluation does not provide a fair assessment of the output generated by a matching technique. We address this problem by defining the first non-binary process model matching evaluation method. We build on a gold standard that has been defined by several annotators and, in this way, allows to account for the subjectivity associated with identifying correspondences.

3 Probabilistic Evaluation of Process Model Matching

In this section, we define our method for non-binary matching evaluation. The starting point of our method is formed by binary assessments created by individual human annotators. Each of these *binary human assessments* captures the correspondences that a single annotator identifies between two given process models.

Definition 1 (Binary Human Assessment). *Let A_1 and A_2 be the sets of activities of two process models. Then, a binary human assessment can be captured by the relation $H : A_1 \times A_2$. Each element $(a_1, a_2) \in H$ specifies that the human assessor considers the activity a_1 to correspond to the activity a_2.*

Note that Definition 1 also allows for one-to-many and many-to-many relationships. If, for instance, the elements (a_1, a_2) and (a_1, a_3) are both part of H, then there exists a one-to-many relationship between the activity a_1 and the two activities a_2 and a_3. Further note that a binary human assessment according to Definition 1 should be created independently and solely reflect the opinion of a single assessor. Based on a number of such independently created binary human assessments, we can then define a non-binary gold standard.

Definition 2 (Non-binary Gold Standard). *A non-binary gold standard is a tuple $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ where*

- *A_1 and A_2 are the sets of activities of two process models,*

- $\mathcal{H} = \{H_1, \ldots, H_n\}$ *is a set of independently created binary human assessments, and*
- $\sigma : \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$ *is a function assigning to each* $(a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ *a* support value*, which is the number of binary human assessments in* $\mathcal{H}$ *that contain the correspondence* (a_1, a_2) *divided by the total number of binary human assessments* $|\mathcal{H}|$*.*

The overall rationale of the non-binary gold standard from Definition 2 is to count the individual opinions from the binary human assessments as votes. In this way, we obtain a *support value* σ for each correspondence according to the number of votes in favor of this correspondence. In this way, any correspondence with a support value $0.0 < \sigma < 1.0$ can be regarded as an uncertain correspondence. For these correspondences, there is no unanimous vote about whether or not it is a correct correspondence. Based on these support values, we define non-binary notions of the well-established metrics precision, recall, and F-measure that take the uncertainty of correspondences into account. For convenience, we introduce $\mathcal{C}$ as the set of all unique correspondences based on the union of all binary human assessments from $\mathcal{H}$.

Definition 3 (Probabilistic Precision, Recall, and F-Measure). *Let* A_1 *and* A_2 *be the sets of activities of two process models,* $M : A_1 \times A_2$ *the correspondences identified by a matching technique, and* $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ *a non-binary gold standard. Then, we define probabilistic precision, recall, and F-measure as follows:*

$$\text{Probabilistic Precision (ProP)} = \frac{\sum_{m \in M} \sigma(m)}{\sum_{m \in M} \sigma(m) + |M \setminus \mathcal{C}|} \tag{1}$$

$$\text{Probabilistic Recall (ProR)} = \frac{\sum_{m \in M} \sigma(m)}{\sum_{c \in \mathcal{C}} \sigma(c)} \tag{2}$$

$$\text{Probabilistic F-Measure (ProFM)} = 2 \times \frac{\text{ProP} \times \text{ProR}}{\text{ProP} + \text{ProR}} \tag{3}$$

Probabilistic precision and recall are adaptations of the traditional notions of precision and recall that incorporate the support values from a non-binary gold standard $\mathcal{GS}$. We define *probabilistic precision* ProP as the sum of the support values of the correspondences identified by the matching technique (M) divided by the same value plus the number of correspondences that are not part of the gold standard ($|M \setminus \mathcal{C}|$). This definition gives those correspondences that have been identified by many annotators a higher weight than those that have only been identified by a few. Therefore, it accounts for the uncertainty associated with correspondences in the non-binary gold standard. As a result, the impact of false positives, i.e. correspondences that have been identified by the matching technique but are not part of the gold standard, result in a strong penalty of 1.0. We justify this high penalty by the high coverage of uncertain correspondences

included in non-binary gold standards. These gold standards can be expected to contain a broad range of potential correspondences, including those identified by only a single annotator. Any correspondence not included in this broad range can be considered as incorrect with certainty, which is reflected in the penalty of 1.0 for false positives.

Probabilistic recall ProR follows the same principle as the probabilistic precision. It resembles the traditional definition of recall, but incorporates the support values from the non-binary gold standard respectively. As a result, identifying correspondences with a higher support has a higher influence on the recall than identifying correspondences with a low support. The probabilistic F-measure ProFM presents the harmonic mean of probabilistic precision and recall. It is computed in the same way as the traditional F-measure, though it is here based on ProP and ProR.

To illustrate these metrics, consider the correspondences, their support values, and the output of three matchers depicted in Table 1. The support values reveal that five out of six correspondences are considered to be correct correspondences in one or more binary human assessments. Matcher $\mathcal{M}_1$ identifies exactly these five correspondences. Therefore, $\mathcal{M}_1$ achieves ProP and ProR scores of 1.0. By contrast, matcher $\mathcal{M}_2$ identifies only three of the five correct correspondences. The matcher also includes the incorrect correspondence c_6 in its output. This results in a ProP value of 0.71 and a ProR value of 0.77. Although matcher $\mathcal{M}_3$ correctly identifies four correspondences, instead of the three identified by $\mathcal{M}_2$, it achieves the exact same ProP and ProR values. This occurs because $\mathcal{M}_3$ identifies c_4 and c_5, which have a combined support value of 0.75, i.e. the same support value as correspondence c_3 that is identified by $\mathcal{M}_2$. This demonstrates that correspondences with a high support value have a greater contribution to the metrics than those with low support.

Table 1: Exemplary matcher output and metrics

Corr.($\mathcal{C}$)	**Supp.**(σ)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
c_1	1.00	1	1	1
c_2	0.75	1	1	1
c_3	0.75	1	1	0
c_4	0.50	1	0	1
c_5	0.25	1	0	1
c_6	0.00	0	1	1

Furthermore, non-binary gold standards allow us to obtain more fine-granular insights into the performance of matchers. We can achieve this by computing probabilistic precision and recall scores for correspondences with a minimal support level. By adapting the equations from Definition 3 in this way, we can differentiate between matchers that identify correspondences with a broad range

of support values and those that focus on the identification of correspondences with high support values. We capture this notion of *bounded* probabilistic precision, recall, and F-measure in Definition 4.

Definition 4 (Bounded Probabilistic Precision, Recall, and F-measure). *Let A_1 and A_2 be the sets of activities of two process models, $M : A_1 \times A_2$ the correspondences identified by a matching technique, $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ a non-binary gold standard, and $\mathcal{C}_\tau$ refer to the set of correspondences with a support level $\sigma \geq \tau$. Then, we define bounded probabilistic precision, recall, and F-measure as follows:*

$$\text{ProP}(\tau) = \frac{\sum\limits_{m \in M} \sigma(m)}{\sum\limits_{m \in M} \sigma(m) + |M \setminus \mathcal{C}_\tau|} \tag{4}$$

$$\text{ProR}(\tau) = \frac{\sum\limits_{m \in M} \sigma(m)}{\sum\limits_{c \in \mathcal{C}_\tau} \sigma(c)} \tag{5}$$

$$\text{ProFM}(\tau) = 2 \times \frac{\text{ProP}(\tau) \times \text{ProR}(\tau)}{\text{ProP}(\tau) + \text{ProR}(\tau)} \tag{6}$$

By computing bounded precision and recall values, we can directly gain insights into the differences between the results obtained by matchers $\mathcal{M}_2$ and $\mathcal{M}_3$. For instance, $\mathcal{M}_2$ and $\mathcal{M}_3$ respectively achieve ProP(0.75) scores which only consider correspondences with $\sigma \geq 0.75$, i.e. 0.71 and 0.50. Similarly, they achieve ProR(0.75) scores of 0.77 and 0.54. These metrics indicate that matcher $\mathcal{M}_2$ is more successful in identifying correspondences with high support values. By contrast, the bounded scores reveal that $\mathcal{M}_3$ identifies more correspondences, although it also includes those with lower support values.

4 Evaluation Experiments

In this section, we apply our probabilistic evaluation method to a dataset from the Process Model Matching Contest 2015. To this end, we create a non-binary gold standard and compute the probabilistic metrics for the matchers that participated in the contest. The overall goal of our experiments is to demonstrate the usefulness of the non-binary perspective and the value of the insights that our evaluation method delivers. Section 4.1 first describes the setup of our experiments. Then, Section 4.2 elaborates on the results.

4.1 Setup

To demonstrate the usefulness of our evaluation method, we apply it to the University Admission dataset from the PMMC 2015 [1]. This dataset consists of nine BPMN process models describing the admission processes for graduate study programs of different German universities. The size of the models varies

between 10 and 44 activities. The task in the context of the Process Model Matching Contest 2015 was to match these models pairwise, resulting in a total number of 36 matching pairs. Our experiments with this dataset consist of two steps:

1. *Non-binary gold standard creation*: To define a non-binary gold standard, we asked eight different individuals to identify the correspondences for the 36 model pairs from the dataset. The group of annotators involved was heterogeneous and included four researchers being familiar with process model matching and four student assistants from the University of Mannheim in Germany. The student assistants were introduced to the problem of process model matching but not influenced in the way they identified correspondences. The result of this step, was a non-binary gold standard based on eight binary assessments. Note that we did not apply any changes to the individual assessments. We included them in their original form into the non-binary gold standard.
2. *Probabilistic evaluation*: Based on the non-binary gold standard, we calculated probabilistic precision, probabilistic recall, and F-measure for each of the 12 matchers that participated in the PMMC 2015. In line with the report from the PMMC 2015 we distinguish between micro and macro average. Macro average is defined as the average precision, recall, and F-measure of all 36 matching pairs. Micro average, by contrast, is computed by considering all 36 pairs as one matching problem. The micro average scores take different sizes of matching pairs (in terms of the correspondences they consist of) into account. As a result, a poor recall on a small matching pair has only limited impact on the overall micro average recall score.

4.2 Results

This section discusses the results of our experiments. We first elaborate on the characteristics of the non-binary gold standard we created. Then, we present the results from the probabilistic evaluation and compare them to the results of the non-binary evaluation from the PMMC 2015. Finally, we present the insights from the bounded probabilistic evaluation.

Non-binary Gold Standard Creation

The non-binary gold standard resulting from the eight binary assessments consists of a total of 879 correspondences. The binary gold standard from the PMMC 2015 only consisted of 234 correspondences, which is less than a third. The average support value per model pair ranges from 0.33 to 0.91. This illustrates that the models considerably differ with respect to how obvious the contained correspondences are.

Figure 2 illustrates the distribution of the support values. It shows that there are two extremes. On the one hand, there is a high number of correspondences with six or more votes (support value ≥ 0.75). On the other hand, there is also

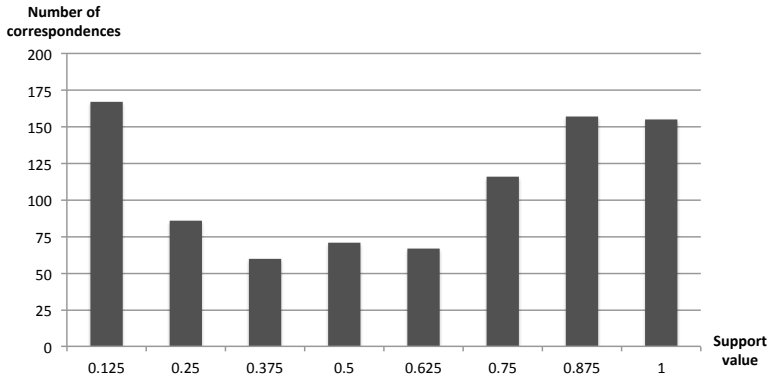


Fig. 2: Distribution of support values in the non-binary gold standard

a high number of correspondences with three votes or less (support value $\leq$ 0.375). Overall, the number of correspondences that would be included based on a majority vote (support value ≥ 0.5) amounts to 495, which is only a little more than half of the correspondences from the non-binary gold standard. These numbers illustrate the complexity associated with defining a binary gold standard and highlight the risks of a purely binary method. Instead of excluding a high number of possible correspondences, we include them with a respective support value.

Probabilistic Evaluation

Table 2 presents the probabilistic evaluation results based on the non-binary gold standard. It shows the micro and macro values of probabilistic F-measure (ProFM), precision (ProP), and recall (ProR) for each matcher that participated in the PMMC 2015. The column *Rank - New* indicates the rank the matcher has achieved according to the probabilistic F-measure micro value. The column *Rank - Old* shows the rank the systems has achieved according to the binary evaluation from the PMMC 2015 [1].

The results from the table illustrate that the probabilistic evaluation has notable effects on the ranking. Although four matchers remain on the same rank, the ranking changes dramatically for other matchers. For instance, the matcher *AML-PM* moves from rank 10 to 2 and the matcher *RMM-NLM* moves from rank 2 to rank 9. A brief analysis of how the matchers work provides an explanation for this development. The matcher *AML-PM* does not impose strict thresholds on the similarity values it uses for identifying correspondences. As a result, it also identifies correspondences with low support values. In the binary gold standard, however, these correspondences were simply not included and resulted in a decrease of precision. Table 3 illustrates this effect by showing an excerpt from the correspondences generated by the matcher *AML-PM* and the

Rank			Approach	ProFM		ProP		ProR	
New	Old	Δ		mic	mac	mic	mac	mic	mac
1	1	±0	RMM-NHCM	**.431**	.387	**.783**	.751	.297	.302
2	10	+8	AML-PM	.387	.365	.377	.390	**.398**	.399
3	9	+6	KnoMa-Proc	.378	.312	.506	.493	.302	.286
4	4	±0	OPBOT	.369	.322	.648	.666	.258	.256
5	5	±0	KMSSS	.368	.313	.563	.623	.274	.276
6	8	+2	BPLangMatch	.360	.325	.532	.475	.272	.272
7	11	+4	RMM-VM2	.329	.293	.516	.643	.242	.240
8	3	-5	MSSS	.307	.238	.761	.772	.192	.201
9	2	-7	RMM-NLM	.306	.244	.681	.565	.197	.203
10	6	-4	RMM-SMSL	.301	.289	.309	.306	.294	.297
11	7	-4	TripleS	.293	.200	.486	.473	.210	.214
12	12	±0	pPalm-DS	.258	.235	.210	.249	.335	.332

Table 2: Results of probabilistic evaluation with new gold standard

respective entries from the binary and the non-binary gold standard. We can see that from the five correspondences from Table 3 only two were included in the binary gold standard. In the context of an evaluation based on this gold standard these three correspondence would therefore reduce the precision of this matcher. An evaluation based on the non-binary gold standard, however, would come to a different assessment. The non-binary gold standard does not only include the two correspondence from the binary gold standard, but also includes the three other correspondences. It is obvious that this positively affects the ProP of the matcher and improves its overall ProFM respectively.

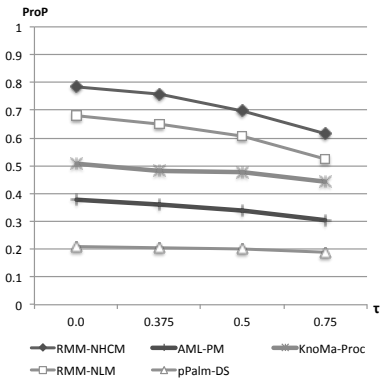
Table 3: Effect of gold standard on assessment of output of matcher *AML-PM*

Correspondence ($\mathcal{C}$)		Gold Standard	
Activity 1	Activity 2	Binary	Non-binary
Send documents by post	*Send appl. form and documents*	0	0.750
Evaluate	*Check and evaluate application*	0	0.500
Apply online	*Complete online interview*	0	0.375
Wait for results	*Waiting for response*	1	0.875
Rejected	*Receive rejection*	1	0.625

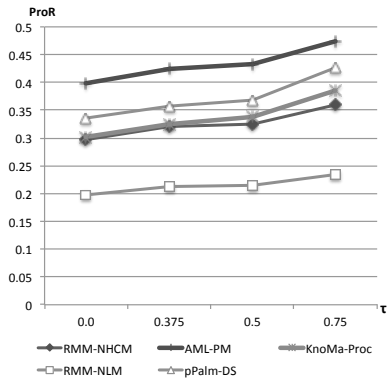
For the matcher *RMM-NLM* we observe the opposite effect. In the context of the evaluation with the non-binary gold standard it misses a huge range of correspondences. Consequently, the ProR of this matcher decreases considerably.

Bounded Probabilistic Evaluation

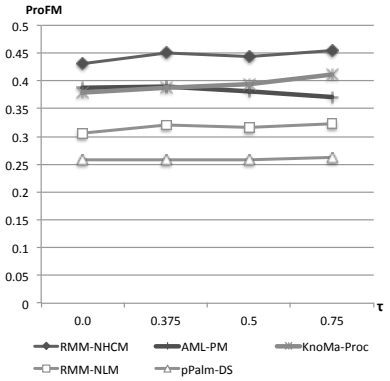
The bounded variants of probabilistic precision, recall, and F-measure provide the possibility to obtain more detailed insights into the performance of the matchers. Figure 3 illustrates this by showing the values of ProP, ProR, and



(a) Bounded Probabilistic Precision



(b) Bounded Probabilistic Recall



(c) Bounded Probabilistic F-Measure

Fig. 3: ProP, ProR, and ProFM for different values of τ

ProFM for $\tau = 0.0$, $\tau = 0.375$, $\tau = 0.5$, and $\tau = 0.75$ for five selected matchers from the PMMC 2015.

The results from Figure 3 show that the effect of a change in the minimum support level τ varies for the different matchers. In general, we observe a decreasing ProP and an increasing ProR for higher values of τ. This is intuitive because a higher value of τ results in the consideration of fewer correspondences. However, for some matchers this effect is stronger than for others. For instance, we observe hardly any change in ProP and a strong increase in ProR for the matcher *pPalm-DS*. This means that this matcher mainly identifies correspondences with high support. It therefore benefits from a stricter gold standard. The matcher *RMM-NLM* represents a contrasting case. The ProP of this matcher decreases dramatically with an increase of τ, while its ProR slightly increases. This reveals that this matcher also identifies a considerable number of correspon-

dences with low support. Since these correspondences turn into false positives when we increase τ, the ProP drops respectively.

The consideration of the bounded variants of ProP, ProR, and ProFM illustrate that an evaluation based on a non-binary gold standard facilitates a more detailed assessment of specific matchers. It is possible to identify whether a matcher focuses on rather obvious correspondences (with high support) or whether a matcher also identifies less apparent correspondences (with low support).

5 Related Work

Existing work on process model matching evaluate their approaches using precision, recall, and F-measures, see for example the reports of the Process Model Matching Contests [1, 3]. Thus, the used evaluation metrics compare an absolute correspondence list with a binary gold standard. Schema matching and ontology matching techniques are similar to process model matching techniques in the sense that these techniques all set out to identify relations between concepts in different conceptual models [8]. Research in the fields of schema and ontology matching (cf. [18, 21]) shows a similar tendency to evaluate the performance of matching techniques based on binary values. However, these fields use a broader range of evaluation metrics to suit needs related to specific applications. For example, aside from the F-measure [2], *error* [17], *information loss* [16], and *overall* [5] are all used to aggregate precision and recall values.

More recently, some metrics have been proposed that relax the binary evaluations of precision and recall metrics. Ehrig and Euzenat [7] propose alternative precision and recall metrics that take into account the closeness of results in ontology matching. Closeness can, for example, exploit the tree structure of ontologies, where the distance between elements in the tree can be computed to determine if a result is close or remote from the expected result. Sagi and Gal [20] adapt precision and recall metrics in order to support non-binary matching results. These metrics can, for instance, be directly applied on first-line-matching results that contain non-binary confidence values. Although this work also specifies that precision and recall could be adapted to support non-binary gold standards, to the best of our knowledge, no works have done this so far.

6 Conclusion

In this paper, we proposed a probabilistic method for assessing the performance of process model matching techniques. Our method is motivated by the insight that it is often hard and in many cases even impossible to define a sensible binary gold standard that clearly specifies which correspondences are correct. Therefore, our evaluation method builds on a number of independent assessments of the correspondences, which are combined into a single probabilistic gold standard. By interpreting the number of votes for each correspondence as support, we

defined probabilistic notions of the well-established metrics precision, recall, and F-measure.

To gain insights into the usefulness of our probabilistic evaluation method, we applied it to the University admission data set and the participating twelve matching techniques from the PMMC 2015. To this end, we recruited eight annotators for the creation of a non-binary gold standard and then computed the probabilistic metrics for each of the matching techniques. We found that the non-binary gold standard contained almost three times as many correspondences as the existing binary gold standard and that only for a fraction of these correspondences there was a unanimous agreement. This emphasizes the risk of using a purely binary evaluation method, which is also reflected in the considerable effect of our probabilistic evaluation method on the ranking of the matching techniques. Furthermore, we found that the probabilistic evaluation allows to obtain more detailed insights into the specific strengths and weaknesses of individual matchers.

In future work, we plan to apply our method on additional data sets and to investigate how human experts perceive the probabilistic results. Our overall goal is to establish the proposed method as a new standard for the evaluation of process model matching techniques and to apply it in the context of the next PMMC.

References

1. Antunes, G., Bakhshandeh, M., Borbinha, J., Cardoso, J., Dadashnia, S., Francescomarino, C.D., Dragoni, M., Fettke, P., Gal, A., Ghidini, C., Hake, P., Khiat, A., Klinkmüller, C., Kuss, E., Leopold, H., Loos, P., Meilicke, C., Niesen, T., Pesquita, C., Péus, T., Schoknecht, A., Sheetrit, E., Sonntag, A., Stuckenschmidt, H., Thaler, T., Weber, I., Weidlich, M.: The process model matching contest 2015. In: 6th International Workshop on Enterprise Modelling and Information Systems Architectures (2015)
2. Berlin, J., Motro, A.: Autoplex: Automated discovery of content for virtual databases. In: Cooperative Information Systems. pp. 108–122. Springer (2001)
3. Cayoglu, U., Dijkman, R., Dumas, M., Fettke, P., Garcıa-Banuelos, L., Hake, P., Klinkmüller, C., Leopold, H., Ludwig, A., Loos, P., et al.: The process model matching contest 2013. In: 4th International Workshop on Process Model Collections: Management and Reuse (PMC-MR'13) (2013)
4. Cayoglu, U., Oberweis, A., Schoknecht, A., Ullrich, M.: Triple-s: A matching approach for Petri nets on syntactic, semantic and structural level. Tech. rep., Karlsruhe Institute of Technology (KIT) (2013)
5. Do, H.H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In: Web, Web-Services, and Database Systems, pp. 221–237. Springer (2002)
6. Dumas, M., Rosa, M., Mendling, J., Reijers, H.: Fundamentals of Business Process Management. Springer (2013)
7. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: Proc. K-Cap 2005 workshop on Integrating ontology. pp. 25–32. No commercial editor. (2005)
8. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic matching. In: Encyclopedia of Database Systems, pp. 2561–2566. Springer (2009)

9. Jin, T., Wang, J., La Rosa, M., Ter Hofstede, A., Wen, L.: Efficient querying of large process model repositories. Computers in Industry 64(1), 41–49 (2013)
10. Klinkmüller, C., Weber, I., Mendling, J., Leopold, H., Ludwig, A.: Increasing recall of process model matching by improved activity label matching. In: Business Process Management, pp. 211–218. Springer (2013)
11. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity–a proper metric. In: Business Process Management, pp. 166–181. Springer (2011)
12. Küster, J.M., Koehler, J., Ryndina, K.: Improving business process models with reference models in business-driven development. In: Business Process Management Workshops. pp. 35–44. Springer (2006)
13. La Rosa, M., Dumas, M., Uba, R., Dijkman, R.: Business process model merging: An approach to business process consolidation. ACM Transactions on Software Engineering and Methodology (TOSEM) 22(2), 11 (2013)
14. Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R., Stuckenschmidt, H.: Probabilistic optimization of semantic process model matching. In: Business Process Management, pp. 319–334. Springer (2012)
15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
16. Mena, E., Kashyap, V., Illarramendi, A., Sheth, A.: Imprecise answers in distributed environments: Estimation of information loss for multi-ontology based query processing. International Journal of Cooperative Information Systems 9(04), 403–425 (2000)
17. Modica, G., Gal, A., Jamil, H.M.: The use of machine-generated ontologies in dynamic information seeking. In: Cooperative Information Systems. pp. 433–447. Springer (2001)
18. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. the VLDB Journal 10(4), 334–350 (2001)
19. Rolland, C., Prakash, N., Benjamen, A.: A multi-model view of process modelling. Requirements Engineering 4(4), 169–187 (1999)
20. Sagi, T., Gal, A.: Non-binary evaluation for schema matching. In: Conceptual Modeling, pp. 477–486. Springer (2012)
21. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on 25(1), 158–176 (2013)
22. Uba, R., Dumas, M., García-Bañuelos, L., La Rosa, M.: Clone detection in repositories of business process models. In: Business Process Management, pp. 248–264. Springer (2011)
23. Weidlich, M., Dijkman, R., Mendling, J.: The ICoP framework: Identification of correspondences between process models. In: Advanced Information Systems Engineering. pp. 483–498. Springer (2010)
24. Weidlich, M., Sheetrit, E., Branco, M.C., Gal, A.: Matching business process models using positional passage-based language models. In: Conceptual Modeling, pp. 130–137. Springer (2013)