A Probabilistic Evaluation Procedure for Process Model Matching Techniques

Elena Kuss[a], Henrik Leopold[b], Han van der Aa[b], Heiner Stuckenschmidt[a], Hajo A. Reijers[b]

[a] *Research Group Data and Web Science, University of Mannheim, Mannheim, Germany*
[b] *Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*

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. Current evaluation methods require a binary gold standard that clearly defines which correspondences are correct. The problem is that often not even humans can agree on a set of correct correspondences. Hence, evaluating the performance of matching techniques based on a binary gold standard 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 procedure for process model matching techniques. In particular, we build on the assessments of multiple annotators to define the notion of a non-binary gold standard. In this way, we avoid the problem of agreeing on a single set of correct correspondences. Based on this non-binary gold standard, we introduce probabilistic versions of precision, recall, and F-measure as well as a distance-based performance measure. We use a dataset from the Process Model Matching Contest 2015 and a total of 16 matching systems to assess and compare the insights that can be obtained by using our evaluation procedure. We find that our probabilistic evaluation procedure allows us to gain more detailed insights into the performance of matching systems than a traditional evaluation based on a binary gold standard.

Keywords: Probabilistic evaluation, Process model matching, Evaluation techniques

Email addresses: elena@informatik.uni-mannheim.de (Elena Kuss), h.leopold@vu.nl (Henrik Leopold), j.h.vander.aa@vu.nl (Han van der Aa), heiner@informatik.uni-mannheim.de (Heiner Stuckenschmidt), h.a.reijers@vu.nl (Hajo A. Reijers)

1. Introduction

Process models are conceptual models used for purposes ranging from the documentation of organizational operations [1] to the definition of requirements for information systems [2, 3]. 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 [4], harmonization of process model variants [5, 6], process model search [7, 8, 9], and the detection of process model clones [10, 11]. 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. [9, 12, 13, 14, 15, 16]). One central question that concerns all of 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 a matching technique against those from the binary gold standard, it is possible to compute the well-established performance measures precision, recall, and F-measure [17]. In this way, the performance of an approach can be quantified and compared against others. The disadvantage of this evaluation procedure 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 pair of process models for which two independent annotators fully agreed on the *correct* correspondences [18, 19]. 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 performance measures. 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 procedure*. Instead of requiring a binary gold standard, we define a *non-binary* gold standard that combines a number of binary assessments created by individual annotators. This enables the consideration of correspon-

dences on which some, but not all annotators agree. In particular, the non-binary gold standard can express the *support* that exists for correspondences as the fraction of annotators that agree that a given correspondence is correct. Based on these support values, we define probabilistic notions of precision, recall, and F-measure. Furthermore, we introduce an alternative performance measure that is based on the distance between the support value from the non-binary gold standard and the matcher output. The overall rationale of the new evaluation measures is that correspondences with high support values have a bigger impact on the matcher performance than correspondences with low support values.

Note that this paper is an extended version of an earlier conference paper [20]. We extend the work from [20] in three ways: (1) we introduce an additional distance-based performance measure, (2) we provide an analysis of the robustness of our evaluation procedure with respect to the number of required annotators for the non-binary gold standard, and (3) we include four additional matching systems in the evaluation.

The rest of the paper is organized as follows. Section 2 elaborates on the process model matching task and illustrates the problem of using a binary gold standard for process model matching evaluation. In Section 3, we present our new evaluation procedure. We define the notion of a non-binary gold standard and introduce probabilistic evaluation measures. In Section 4, we assess and compare the proposed probabilistic evaluation measures by applying our procedure on the dataset of the PMMC 2015. Section 5 discusses works related to matching in a broader context. Finally, we conclude the paper and discuss future research directions in Section 6.

2. Background

This section discusses the background of our work. Section 2.1 introduces the task of process model matching and gives an overview of existing matching techniques. Afterwards, Section 2.2 elaborates on the challenges associated with evaluating the performance of process model matching techniques and identifies the research gap.

2.1. Process Model Matching

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[1]. Formally, the correspondences between the sets of activities of A_1 and A_2 can be captured by a relation $match : \mathcal{P}(A_1) \times \mathcal{P}(A_2)$. An element $(A'_1, A'_2) \in match$ defines that the set of activities $A'_1 \subseteq A_1$ corresponds to the set of activities $A'_2 \subseteq A_2$. If $|A'_1| = 1$ and $|A'_2| = 1$, we refer to the correspondence as *elementary*, otherwise we call it *complex*.

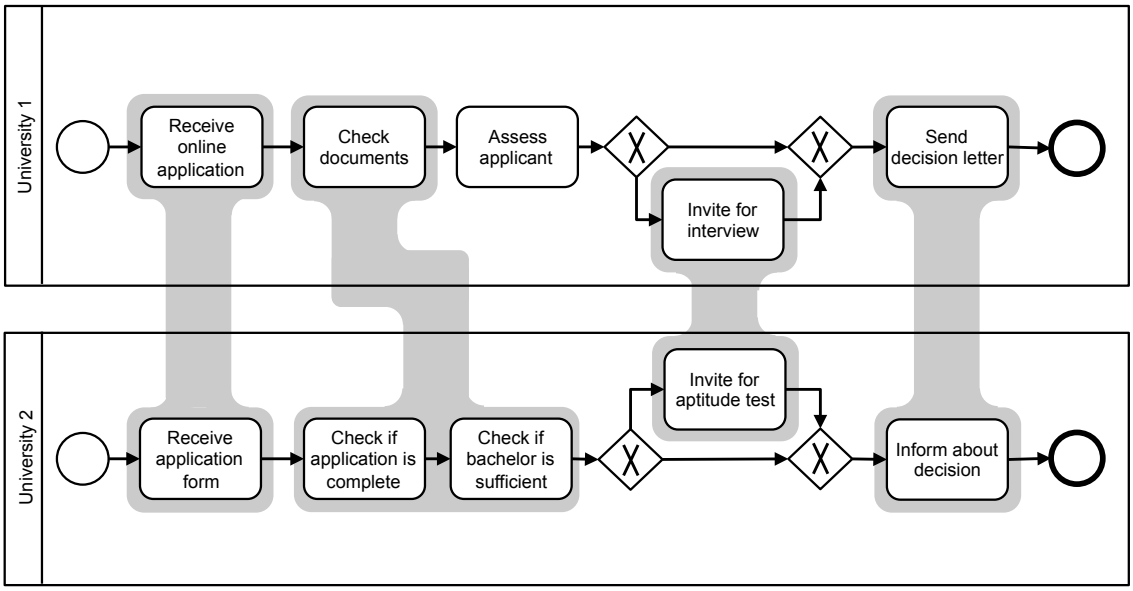


Figure 1: Two process models and possible correspondences

To illustrate the goal of the process model matching task, consider the example depicted in Figure 1. It shows two process models describing the steps students have to take to be admitted for the graduate programs of two different universities. Although both processes are quite similar, the identification of the illustrated correspondences is far from trivial. Consider, for instance, the complex correspondence between "*Check documents*" and "*Check if application is complete*" as well as "*Check if bachelor degree is sufficient*". To automatically recognize that the latter two activities relate to a stream of action that can be referred to as "*Check documents*", the recognition of complex semantic relationships is required. This also applies to the correspondence between the "*Invite for interview*" and "*Invite for aptitude test*" activities. Here, a matching technique must be capable to automatically recognize that both an "*interview*" as well as an "*aptitude test*" is a means of evaluating the suitability of a student.

[1]Note that the notion of *similar behavior* is not formally defined in the domain of process model matching. In some cases, a correspondence relates to a part-of relationship between two activities, in some cases it relates to an alternative way of achieving the same objective. In this paper, we accept that the notion of similarity is subjective and address it with the concept of a non-binary gold standard as discussed below.

To address such challenges associated with process model matching, many different matching techniques have been proposed in recent years. Typically, these techniques combine different measures to quantify the structural as well as the textual similarity between the considered process models. The first matching techniques that have been defined combined structural measures such as the graph edit distance with syntactic text similarity measures such as the Levenshtein distance [8, 15]. More recent techniques also consider semantic relationships between words, most commonly by building on the lexical database WordNet [13, 14, 12]. A few techniques also employ alternative strategies. Examples include matching techniques incorporating human feedback [21], techniques selecting the most promising similarity measures based on prediction [22], techniques selecting the best correspondences based on voting [23], and techniques that employ machine learning [24].

Considering the variety of matching techniques that have been defined in prior work, a key question is how to *evaluate* the performance of these techniques. While the specific technologies or model-related aspects exploited by the matching technique do not change how a matching technique needs to be assessed, the question is how to fairly quantify to what extent the generated correspondences are correct. In the next section, we discuss the challenges that are associated with this and how it relates to the notion of correctness.

2.2. The Challenge of Evaluating Process Model Matching Performance

Currently, the evaluation of matching techniques almost exclusively relies on precision, recall, and F-measure [25]. These are standard metrics from the information retrieval field that can be used to quantify the performance of matchers alongside different dimensions. The reliance on these metrics applies to process model matching techniques (cf. [18, 19, 14, 15, 16]) as well as to the related fields of schema matching and ontology matching techniques, (cf. [26, 27]). Available alternatives mainly focus on relaxing the strict notion of precision and recall in order to better reflect the performance of matching techniques. For instance, Ehrig and Euzenat [28] propose alternative notions for these measures that take the closeness of results in ontology matching into account. 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 to the expected result. Sagi and Gal [29] adapt precision and recall to evaluate non-binary confidence values produced by schema matching techniques. Despite the existence of these different measures, what they all have in common is that they rely on the existence of a binary gold standard, i.e. on a single set of correct correspondences.

To illustrate the challenge associated with defining such a single set of correspondences, again consider the correspondences from Figure 1. Upon close inspection, it becomes clear that many of the identified 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. For all these disputable cases, it is well-imaginable that some annotators indeed agree that these cases represent correct correspondences, whereas other annotators may disagree with this. This makes the selection of a single set of correct correspondences a, partially, subjective task. In this paper, we therefore argue that a binary evaluation of process model matching techniques does not account for the full complexity of the process model matching task. In particular, such a binary evaluation does not consider disagreements that may exist regarding the correctness of correspondences. Hence, a 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 procedure. Our procedure builds on a non-binary 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 procedure for the probabilistic evaluation of process model matching. Section 3.1 introduces the notion of a non-binary gold standard. Then, Section 3.2 defines probabilistic versions of the metrics precision, recall, and F-measure. Finally, Section 3.3 defines

an alternative measure for matching evaluation based on the distance between the matcher output and the non-binary gold standard.

3.1. Defining the Notion of a Non-binary Gold Standard

The starting point of our evaluation procedure 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 three specific details related to this definition. First, Definition 1 also allows for one-to-many and many-to-many relationships, i.e., complex correspondences. 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. The advantage of capturing a complex correspondence based on several elementary correspondences is that the matching technique is not required to identify the entire complex correspondence. If it, for instance, identifies (a_1, a_2) but not (a_1, a_3), it would at least get credit for having identified (a_1, a_2). Second, the information that is available for deciding about a possible correspondence may vary from model to model. In general, we assume that the decision will be mainly based on the labels. If available, however, also data objects can provide valuable input. Third, 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, $\mathcal{H}$, σ) 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*
- *$\sigma : A_1 \times A_2 \to \mathbb{R}$ is a function assigning to each $(a_1, a_2) \in A_1 \times A_2$ 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.

3.2. Probabilistic Precision, Recall, and F-Measure

Based on the support values provided by a non-binary gold standard, we define probabilistic versions of precision, recall, and F-measure, which take the uncertainty of correspondences into account. For convenience, we introduce $\mathcal{C}$ to refer to the set of all correspondences that have a support value above 0.0.

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 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 non-binary 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 non-binary 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 to be certainly incorrect, 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 5 out of 6 correspondences are considered to be correct correspondences by one or more binary human assessments. Matcher $\mathcal{M}_1$ identifies exactly these 5 correspondences. Therefore, $\mathcal{M}_1$ achieves ProP and ProR scores of 1. By contrast, matcher $\mathcal{M}_2$ identifies only 3 of the 5 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 4 correspondences, instead of the 3 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 shows that correspondences with a high support value have a greater contribution to the metrics than those with low support.

Non-binary gold standards also 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 $\mathcal{C}_\tau$ refers to the set of correspondences with a support level $\sigma \geq \tau$. Then, we define bounded probabilistic precision, recall,*

Table 1: Exemplary matcher output and metrics

$\mathcal{C}$	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
ProP		**1**	**0.71**	**0.71**
ProR		**1**	**0.77**	**0.77**
ProFM		**1**	**0.74**	**0.74**

and F-measure as follows:

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

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

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

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. The bounded scores for $\mathcal{M}_3$ reveal that it identifies a higher number of correspondences with lower support values.

3.3. Probabilistic Distance

The previously introduced notions of ProP, ProR, and ProFM implicitly build on the premise that matchers should also identify correspondences with low support values. In fact, they reward matchers that identify correspondences with low support values and penalize matchers that fail to identify them. As an illustration, consider a correspondence for which 2 out of 5 human annotators agree that this is a correct correspondence. If identified by a matcher, the ProP, ProR,

and ProFM scores of the matcher will increase, because the correspondence has a non-zero support value. However, it is important to recognize that also 3 out of the 5 annotators agree that this is *not* an actual correspondence, i.e. the majority of the annotators disagree with the correspondence. The previously introduced metrics do not fully take such a majority of disagreements into account. Recognizing this characteristic, we also introduce an alternative performance measure that explicitly considers agreements and disagreements in a non-binary gold standard. This performance measure builds on the notion of *distance* between the matcher output and the support values from the non-binary gold standard. The overall rationale is to explicitly account for agreements and disagreements with the annotators of the non-binary gold standard. Intuitively, this means that correspondences with low support values are no longer favorable since most annotators disagree with these correspondences. We define the measure *Probabilistic Distance (ProD)* as follows.

Definition 5 (Probabilistic Distance). *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, $\mu : A_1 \times A_2 \to \{0, 1\}$ a function that returns 1 if a correspondence $m \in M$ and 0 if a correspondence $m \notin M$, and $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ a non-binary gold standard. Then, we define the Probabilistic Distance as follows:*

$$\text{Probabilistic Distance (ProD)} = \sum_{m \in (M \cup \mathcal{C})} (\mu(m) - \sigma(m))^2 \tag{7}$$

The core idea underlying the ProD measure is to compute the distance between the matcher output (which can be 1 or 0) and the support value σ from the non-binary gold standard. We square the values to obtain a lower penalty for correspondences that have a high support. To illustrate the mechanism of ProD, consider Table 2. It shows how the output of the three matchers from Table 1 is evaluated by ProD.

The example depicted in Table 2 illustrates three key characteristics of ProD. First, matchers identifying a correspondence that is not part of the non-binary gold standard or fail identifying a correspondence with a support of 1, receive a penalty of 1. Second, it does not matter whether a matcher identifies a correspondence with a support of 0.5 (see c_4). The distance in both cases is identical. This is a reasonable approach taking into account that the matcher agrees/disagrees with half of the annotators. Third, the penalty for identifying a correspondence with a low support is higher than for not identifying it (see c_5). This is again in line with the argument of taking agreements into account. Given a support of 0.25 of c_5, a matcher that does not identify c_5, disagrees with 25% of the annotators. A matcher that does identify c_5, disagrees with 75% of the annotators.

		$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$	
$\mathcal{C}$	$\sigma(c_n)$	$\mu(\mathbf{c_n})$	$\mathbf{ProD(c_n)}$	$\mu(\mathbf{c_n})$	$\mathbf{ProD(c_n)}$	$\mu(\mathbf{c_n})$	$\mathbf{ProD(c_n)}$
c_1	1.00	1	0	1	0	1	0
c_2	0.75	1	0.063	1	0.063	1	0.063
c_3	0.75	1	0.063	1	0.063	0	0.563
c_4	0.50	1	0.25	0	0.25	1	0.25
c_5	0.25	1	0.563	0	0.063	1	0.563
c_6	0.00	0	0	1	1	1	1
Total			**0.938**		**1.438**		**2.438**

Table 2: Illustration of Probabilistic Distance

In the next section, we apply the above introduced evaluation procedure to the dataset of the Process Model Matching Contest 2015.

4. Evaluation Experiments

In this section, we apply our probabilistic evaluation procedure to the University Admission dataset, which is a matching problem that was first introduced in the context of the Process Model Matching Contest 2015 [18]. To this end, we created a non-binary gold standard, based on correspondences identified by 8 individual annotators, and compute the probabilistic measures for 16 different matchers that solved this matching problem. 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 procedure delivers. Section 4.1 first describes the setup of our experiments. Section 4.2 then elaborates on the results. Section 4.3 discusses the robustness of our results from the perspective of the required number of annotators. Finally, Section 4.4 reflects on the findings in the context of a discussion.

4.1. Setup

To demonstrate the usefulness of our evaluation procedure, we apply the procedure to the University Admission dataset of the PMMC 2015 [18]. 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 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 8 individuals to identify the correspondences for the 36 model pairs from the dataset. We prepared respective templates for each model pair and asked the annotators to complete this task model pair by model pair. We instructed them to not spend more than two hours in a row on this task to avoid low quality results caused by depletion. The group of involved annotators was heterogeneous and included 4 researchers being familiar with process model matching and 4 student assistants from the University of Mannheim in Germany. The student assistants were introduced to the problem of process model matching, but they were not influenced in the way they identified correspondences. The result of this step, was a non-binary gold standard based on 8 binary assessments. On average, the annotators spent around one hour per model pair (i.e, approximately 36 hours per annotator). 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 ProP, ProR, ProFM, and ProD for a total of 16 matchers. Twelve matchers solved this matching problem in the context of the PMMC 2015 and 4 matchers solved it in the context of a subtrack of the Ontology Alignment Evaluation Initiative (OAEI) 2016 [30]. In line with the report from both the PMMC 2015 and OAEI 2016, 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. Section 4.2.1 elaborates on the characteristics of the non-binary gold standard we created. Section 4.2.2 presents the results from the evaluation with ProP, ProR, and ProFM and compares them to the results of the non-binary evaluation. Section 4.2.3 discusses the insights from the evaluation with the bounded versions of ProP, ProR, and ProFM. Finally, Section 4.2.4 presents the results from the evaluation with ProD.

4.2.1. Non-binary Gold Standard Creation

The non-binary gold standard resulting from the 8 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.

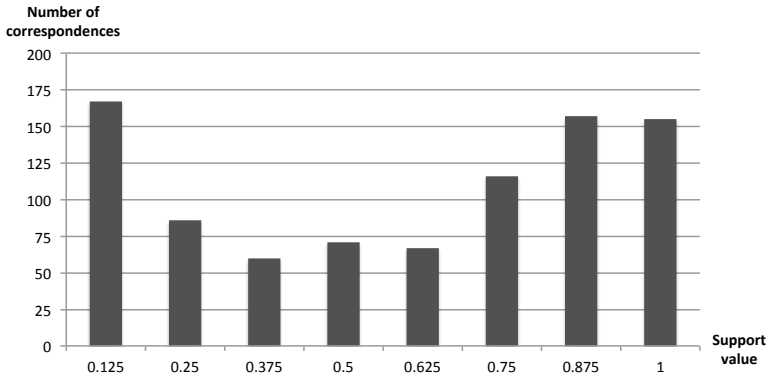


Figure 2: Distribution of support values in the non-binary gold standard

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 6 or more votes (support value ≥ 0.75). On the other hand, there is also a high number of correspondences with three votes or less (support value ≤ 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 evaluation procedure. Instead of excluding a high number of possible correspondences, we include them with a respective support value. This avoids a loss of information.

Figure 3 further illustrates the average number of correspondences that are added to the non-binary gold standard by an additional annotator. The numbers from Figure 3 emphasize that the number of correspondences added by an additional annotator decreases very quickly. While the second annotator, on average, adds about 145 new correspondences to the non-binary gold standard, the 8th annotator only adds 24 new correspondences. Note that the correspondences that are newly introduced by the 8th annotator only have a support of 0.125, since none of the previous annotators agreed with these correspondences. Overall, these numbers show that we quickly reach

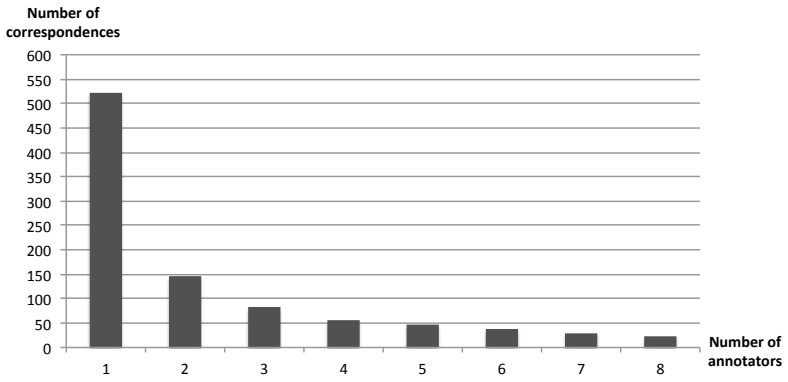


Figure 3: Average increase of number of correspondences with additional annotators

a point where hardly new reasonable correspondences are added. This is in line with the notion of *theoretical saturation* in qualitative research settings [31]. In this context, theoretical saturation describes the point where no new insights can be obtained from analyzing additional data.

4.2.2. Evaluation Using Probabilistic Precision, Recall, and F-Measure

Table 3 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), recall (ProR) for each of the 16 matchers that participated in the PMMC 2015 or the OAEI 2016. The column *Rank - New* indicates the rank the matcher has achieved according to the ProFM micro value. The column *Rank - Old* shows the rank the system has achieved according to the binary evaluation.

The results in the table illustrate that the probabilistic evaluation has notable effects on the ranking. Although 4 matchers remain on the same rank, the ranking changes dramatically for other matchers. For instance, the matcher *AML-PM* moves from rank 14 to 5 and the matcher *RMM-NLM* moves from rank 3 to rank 14. A brief analysis of the matchers' inner workings 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 4 illustrates this effect by showing an excerpt from the correspondences generated by the matcher *AML-PM* and the respective entries from the binary and the non-binary gold standard. We can see that from the 5 correspondences from Table 4 only two were included in the binary gold standard. In the context of an evaluation

Rank			Approach	ProFM		ProP		ProR	
New	Old	Δ		mic	mac	mic	mac	mic	mac
1	2	+1	RMM-NHCM	.432	.391	.83	.777	.292	.297
2	11	+9	LogMap	.42	.366	.683	.676	.304	.301
3	1	-2	AML	.419	.376	.795	.728	.284	.289
4	6	+2	Know-Match-SSS	.411	.358	.679	.788	.295	.297
5	14	+9	AML-PM	.408	.395	.411	.46	.406	.408
6	13	+7	KnoMa-Proc	.406	.345	.573	.594	.314	.302
7	5	-2	OPBOT	.369	.318	.669	.676	.254	.248
8	12	+4	BPLangMatch	.361	.327	.559	.505	.267	.265
9	7	-2	RMM-SMSL	.358	.325	.6	.712	.255	.256
10	9	-1	DKP-lite	.347	.284	.895	.911	.215	.219
11	8	-3	DKP	.341	.285	.759	.691	.22	.223
12	15	+3	RMM-VM2	.318	.307	.333	.337	.304	.306
13	4	-9	Match-SSS	.315	.249	.827	.814	.194	.203
14	3	-11	RMM-NLM	.312	.253	.73	.583	.198	.203
15	10	-5	TripleS	.301	.21	.518	.498	.212	.216
16	16	±0	pPalm-DS	.275	.261	.229	.289	.345	.344

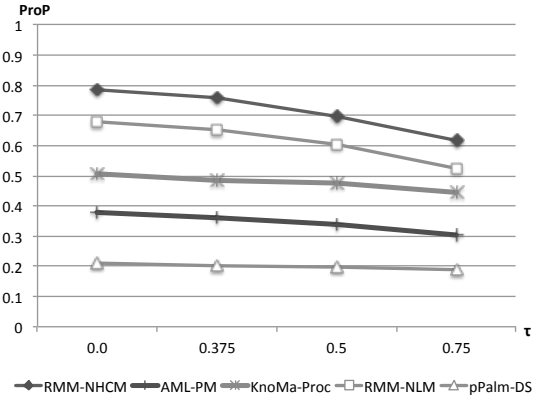
Table 3: Results of probabilistic evaluation with non-binary gold standard

based on this binary 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.

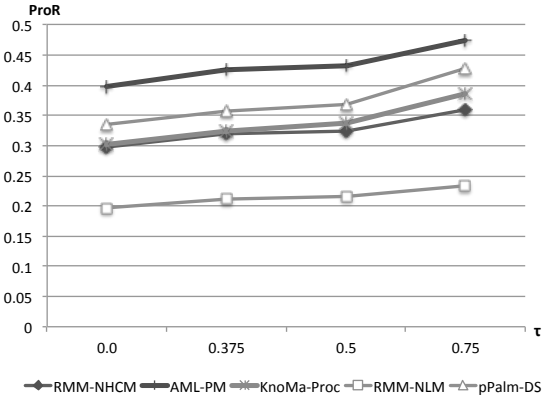
Table 4: Effect of gold standard on assessment of output of matcher *AML-PM*

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

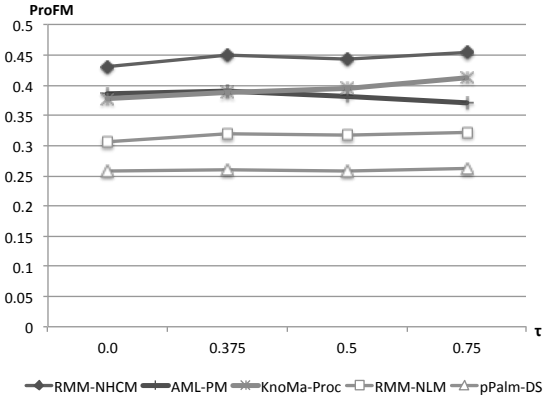
For the matcher *RMM-NLM* we observe the opposite effect. In the context of the evaluation with the non-binary gold standard it misses a huge range of correspondences. Consequently, the



(a) Bounded probabilistic Precision



(b) Bounded probabilistic Recall



(c) Bounded probabilistic F-Measure

Figure 4: ProP, ProR, and ProFM for different values of τ

ProR of this matcher decreases considerably.

4.2.3. Evaluation Using Bounded Probabilistic Precision, Recall, and F-Measure

The bounded variants of ProP, ProR, and ProFM provide the possibility to obtain more detailed insights into the performance of the matchers. Figure 4 illustrates this by showing the values of ProP, ProR, and ProFM for $\tau = 0.0$, $\tau = 0.375$, $\tau = 0.5$, and $\tau = 0.75$ for 5 selected matchers from the PMMC 2015.

The results from Figure 4 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 correspondences 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).

4.2.4. Evaluation Using Probabilistic Distance

The probabilistic distance ProD explicitly takes the number of agreements and disagreements with the annotators from the gold standard into account. As a result, matching systems that identify correspondences with low support values are penalized. Table 5 gives an overview of the results obtained using this distance measure. It shows for each matcher the ProD value, the ProFM value, the ranks based on the respective measures, and the delta between the ranks.

The results depicted in Table 5 illustrate that the use of ProD has notable effects on the ranking. We can identify several matchers whose rank changed considerably. For instance, the matcher AML-PM went from rank 5 to rank 14 and the matcher DKP-lite went from rank 10 to rank 2. However, it is also interesting to note that the first and the last rank did not change. The matcher RMM-NHCM has both the lowest ProD value as well as the highest ProFM value. The matcher pPalm-DS has both the highest ProD value as well as the lowest ProFM value. As a result, they remain on the first and the last rank respectively.

To better understand these results, it is necessary to look into the specific correspondences that the matchers identify. An analysis of the correspondences identified by the matcher AML-PM reveals, for instance, that this matcher establishes a high number of correspondences with low support values. This means that the fairly good ProFM value of AML-PM results from a high number of small rewards for low-support correspondences. Since ProD does not reward but penalizes the identification of such correspondences, ProD is rather high in comparison to other matching systems. For the matcher DKP-lite, which moved 8 ranks up, we observe the opposite

Rank			Approach	ProD	ProFM	
ProD	ProFM	Δ			mic	mac
1	1	±0	RMM-NHCM	261.1	.432	.391
2	10	+8	DKP-lite	265.6	.347	.284
3	3	±0	AML	269.8	.419	.376
4	13	+9	Match-SSS	276.6	.315	.249
5	11	+6	DKP	288.6	.341	.285
6	2	-4	LogMap	295.2	.42	.366
7	14	+7	RMM-NLM	297.6	.312	.253
8	4	-4	Know-Match-SSS	298.8	.411	.358
9	7	-2	OPBOT	313.9	.369	.318
10	9	-1	RMM-SMSL	340.6	.358	.325
11	8	-3	BPLangMatch	343.4	.361	.327
12	6	-6	KnoMa-Proc	344.9	.406	.345
13	15	+2	TripleS	347.4	.301	.21
14	5	-9	AML-PM	510	.408	.395
15	12	-3	RMM-VM2	533.8	.318	.307
16	16	±0	pPalm-DS	815.7	.275	.261

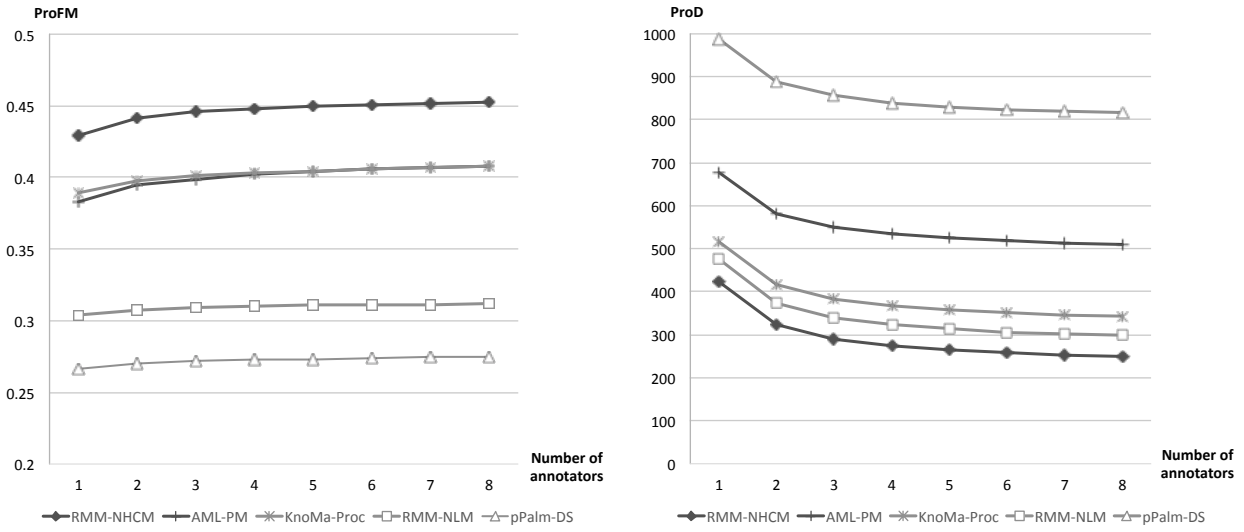
Table 5: Results of probabilistic evaluation with non-binary gold standard

effect. This matcher mainly produces correspondences with high support values. While this resulted in a rather moderate ProFM value because of all the unidentified low-support correspondences, the ProD value of this matcher is very low, resulting in a good rank.

The two extreme cases of AML-PM and DKP-lite illustrate that ProD penalizes matchers that identify a high number of correspondences with low support values and rewards matchers that do not. This also reveals the specific characteristics of the matching systems on the first and the last rank. The matcher RMM-NHCM identifies a considerable number of correspondences with high support values. As a result, both ProFM as well as ProD yield good results. The matcher pPalm-DS, by contrast, simply produces a considerable amount of noise. The high number of false positives results in a bad performance from the perspective of both measures.

4.3. Robustness of Results

The advantage of the probabilistic evaluation procedure presented in this paper is that it builds on the individual assessments of a number of annotators. In this way, we circumvent the almost unfeasible task of defining a single set of correct correspondences. However, building on the assessments of annotators also raises the question when the evaluation results actually become robust, i.e. how many annotators are required before the presented performance measures stabilize. Fig-



(a) Probabilistic F-Measure

(b) Probabilistic Distance

Figure 5: Development of probabilistic evaluation measures with increasing number of annotators

ure 5 illustrates how the ProFM and ProD develop for 5 representative matching systems with an increasing number of annotators. To avoid a bias resulting from the order of the annotators (including someone as the 8th annotator who identified a lot of correspondences, would lead to a non-representative movement in the graph), we computed the average values for both evaluation measures based on all possible annotator combinations. For example, the values for 4 annotators are obtained by computing and averaging ProFM and ProD for all possible combinations of 4 annotators.

The values in Figure 5a show that ProFM converges after only including 4 annotators, i.e. the inclusion of additional annotators has a negligible effect on the results. For instance, the additional correspondences included by the 7th annotator do not even change the third decimal place for most matching systems. For ProD, we observe that more annotators are required. We see that ProD changes quite drastically when including additional annotators. This can be explained by the strong effect of low-support correspondences on this measure. Additional annotators are likely to include more correspondences, which reduces the number of correspondences that are considered as false positives. Despite this rather strong decrease, we still observe that ProD converges. After including 7 annotators, the change is below 2% for all matching systems.

To get insights into the differences between the two annotator groups (student assistants and researchers), we also analyzed the binary assessments from both groups and compared the corre-

spondence they created. We found that the student group came up with more correspondences than the researcher group (825 versus 615). The total number of correspondences where the entire subgroup agreed on a correspondence was, however, slightly higher for the researcher group (242 versus 211). These numbers indicate that the student group had a more diverse view on the correspondences and, as a result, had a higher degree of disagreement. These insights emphasize once again that the idea of consulting several annotators is a promising strategy. The higher the number of annotators, the less individual opinions affect the evaluation.

Altogether, we can state that the presented performance measures stabilize after including 4 to 7 annotators. While we cannot give a general recommendation about the number of annotators that is required, our analysis showed that this number is likely to be below 10. Taking into account that annotators only need to be familiar with the domain and not with process model matching, this is a feasible number.

4.4. Discussion

The evaluation experiments in this section illustrate that the presented performance measures have a different focus.

ProFM (together with ProP and ProR) is based on the well-know measure from information retrieval and, therefore, might be considered as intuitive by many people. A specific characteristic of this measure is that it rewards matching systems that also recognize correspondences with low support values. Whether this is a desired outcome, largely depends on the application scenario of the matching system. If the output of a matching system is used as input for humans, i.e. the matching system's task is to suggest *possible* correspondences, identifying a larger number of correspondences is helpful. If the matching system is applied without any further human intervention, it is not. A notable advantage of this measure is the low number of annotators that is required for the non-binary gold standard. We found that ProFM already converges after including 4 annotators.

ProD takes the number of disagreements with the annotators of the non-binary gold standard explicitly into account. As a result, it rather favors matchers that focus on identifying high-support values. If this is a desired feature of a matcher, ProD will give a better impression of the performance than ProFM. A slight disadvantage of ProD is that it requires more annotators than ProFM to produce stable results. Our analysis showed that ProD only converged after including 7 annotators, as opposed to 4 for the ProFM metric.

In summary, we can state that the choice of the performance measure mainly depends on the application scenario of the evaluated matching system. Nevertheless, as illustrated by the matchers

RMM-NHCM and pPalm-DS, extremely good or bad systems yield good values for both performance measures.

5. Related Work

The evaluation procedure presented in this paper focuses on the domain of process model matching. However, matching problems and the question of how to evaluate them occur in a variety of contexts. Most notably, they also occur when matching different types of process-oriented artifacts, database schemas, and ontologies.

Techniques that match *different types of process-oriented artifacts* exist for process models and taxonomies [32], process models and textual process descriptions [33, 34] as well as process models and event logs [35, 36, 37]. The alignments that results from such techniques have various use cases. For example, the alignments between process models and textual descriptions can be used to automatically detect conflicts between these two types of process descriptions [33], whereas techniques that match event logs and process models provide a basis for *conformance checking* [38].

Schema matching techniques take two database schemas as input and identify corresponding elements between the two schemas [26]. The resulting correspondences play a central role in schema integration [39, 40], data warehousing [41], and semantic query processing [42]. *Ontology matching* concerns the identification of correspondences between the elements of two ontologies [43]. Ontologies are abstract models that explicitly define concepts, their properties, and their inter-relations for a specific domain (cf. [44, 45]). Application scenarios for ontology matching techniques include instance translation [46], ontology extension [47], and ontology merging [48].

What all these techniques have in common is that they need to evaluate a matching problem that humans typically have deviating views on. Therefore, the ideas presented in this paper can also provide relevant input for these domains.

6. Conclusion

In this paper, we proposed a probabilistic procedure for assessing the performance of process model matching techniques. Our evaluation procedure 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 procedure builds on a number of independent assessments of the correspondences, which are combined into a single non-binary gold standard. By interpreting the number of votes for each correspondence as support, we defined

two types of evaluation measures. First, we introduced probabilistic notions of the well-established metrics precision, recall, and F-measure. Second, we introduced a distance-based performance measure that explicitly takes the number of disagreements and agreements with the annotators of the non-binary gold standard into account.

To gain insights into the usefulness of our probabilistic evaluation procedure, we applied it to the University admission dataset and a total of 16 matching techniques. We recruited eight annotators for the creation of a non-binary gold standard and then computed the introduced probabilistic performance measures 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 procedure 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. While the probabilistic F-Measure favors matchers that produce many reasonable correspondences, the probabilistic distance rewards matchers that focus on identifying high-support correspondences.

In future work, we plan to apply our method on additional datasets 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 Process Model Matching Contest.

References

[1] M. Dumas, M. Rosa, J. Mendling, H. Reijers, Fundamentals of Business Process Management, Springer, 2013.

[2] C. Rolland, N. Prakash, A. Benjamen, A multi-model view of process modelling, Requirements Engineering 4 (4) (1999) 169–187.

[3] H. Leopold, J. Mendling, A. Polyvyanyy, Supporting process model validation through natural language generation, IEEE Transactions on Software Engineering 40 (8) (2014) 818–840.

[4] J. M. Küster, J. Koehler, K. Ryndina, Improving business process models with reference models in business-driven development, in: Business Process Management Workshops, Springer, 2006, pp. 35–44.

[5] M. La Rosa, M. Dumas, R. Uba, R. Dijkman, Business process model merging: An approach to business process consolidation, ACM Transactions on Software Engineering and Methodology (TOSEM) 22 (2) (2013) 11.

[6] M. Weidlich, J. Mendling, M. Weske, A foundational approach for managing process variability, in: International Conference on Advanced Information Systems Engineering, Springer, 2011, pp. 267–282.

[7] T. Jin, J. Wang, M. La Rosa, A. Ter Hofstede, L. Wen, Efficient querying of large process model repositories, Computers in Industry 64 (1) (2013) 41–49.

[8] R. M. Dijkman, M. Dumas, L. García-Bañuelos, Graph matching algorithms for business process model similarity search, in: Business Process Management, Springer, 2009, pp. 48–63.

[9] M. Kunze, M. Weidlich, M. Weske, Behavioral similarity–a proper metric, in: Business Process Management, Springer, 2011, pp. 166–181.

[10] R. Uba, M. Dumas, L. García-Bañuelos, M. La Rosa, Clone detection in repositories of business process models, in: Business Process Management, Springer, 2011, pp. 248–264.

[11] C. C. Ekanayake, M. Dumas, L. García-Bañuelos, M. La Rosa, A. H. ter Hofstede, Approximate clone detection in repositories of business process models, in: Business Process Management, Springer, 2012, pp. 302–318.

[12] U. Cayoglu, A. Oberweis, A. Schoknecht, M. Ullrich, Triple-s: A matching approach for Petri nets on syntactic, semantic and structural level, Tech. rep., Karlsruhe Institute of Technology (KIT) (2013).

[13] C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, A. Ludwig, Increasing recall of process model matching by improved activity label matching, in: Business Process Management, Springer, 2013, pp. 211–218.

[14] H. Leopold, M. Niepert, M. Weidlich, J. Mendling, R. Dijkman, H. Stuckenschmidt, Probabilistic optimization of semantic process model matching, in: Business Process Management, Springer, 2012, pp. 319–334.

[15] M. Weidlich, R. Dijkman, J. Mendling, The ICoP framework: Identification of correspondences between process models, in: Advanced Information Systems Engineering, Springer, 2010, pp. 483–498.

[16] M. Weidlich, E. Sheetrit, M. C. Branco, A. Gal, Matching business process models using positional passage-based language models, in: Conceptual Modeling, Springer, 2013, pp. 130–137.

[17] C. D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Vol. 1, Cambridge university press Cambridge, 2008.

[18] G. Antunes, M. Bakhshandeh, J. Borbinha, J. Cardoso, S. Dadashnia, C. D. Francescomarino, M. Dragoni, P. Fettke, A. Gal, C. Ghidini, P. Hake, A. Khiat, C. Klinkmüller, E. Kuss, H. Leopold, P. Loos, C. Meilicke, T. Niesen, C. Pesquita, T. Péus, A. Schoknecht, E. Sheetrit, A. Sonntag, H. Stuckenschmidt, T. Thaler, I. Weber, M. Weidlich, The process model matching contest 2015, in: 6th International Workshop on Enterprise Modelling and Information Systems Architectures, 2015.

[19] U. Cayoglu, R. Dijkman, M. Dumas, P. Fettke, L. Garcıa-Banuelos, P. Hake, C. Klinkmüller, H. Leopold, A. Ludwig, P. Loos, et al., The process model matching contest 2013, in: 4th International Workshop on Process Model Collections: Management and Reuse (PMC-MR'13), 2013.

[20] E. Kuss, H. Leopold, H. Van der Aa, H. Stuckenschmidt, H. A. Reijers, Probabilistic evaluation of process model matching techniques, in: Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35, Springer, 2016, pp. 279–292.

[21] C. Klinkmüller, H. Leopold, I. Weber, J. Mendling, A. Ludwig, Listen to me: Improving process model matching through user feedback, in: Business Process Management, Springer, 2014, pp. 84–100.

[22] M. Weidlich, T. Sagi, H. Leopold, A. Gal, J. Mendling, Predicting the quality of process model matching, in: Business Process Management, Springer, 2013, pp. 203–210.

[23] C. Meilicke, H. Leopold, E. Kuss, H. Stuckenschmidt, H. A. Reijers, Overcoming individual process model matcher weaknesses using ensemble matching, Decision Support Systems.

[24] A. Sonntag, P. Hake, P. Fettke, P. Loos, An Approach For Semantic Business Process Model Matching Using Supervised Machine Learning, in: Resarch in Progress Papers. 47., 2016.

[25] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern information retrieval, Vol. 463, ACM press New York, 1999.

[26] E. Rahm, P. A. Bernstein, A survey of approaches to automatic schema matching, the VLDB Journal 10 (4) (2001) 334–350.

[27] P. Shvaiko, J. Euzenat, Ontology matching: state of the art and future challenges, Knowledge and Data Engineering, IEEE Transactions on 25 (1) (2013) 158–176.

[28] M. Ehrig, J. Euzenat, Relaxed precision and recall for ontology matching, in: Proc. K-Cap 2005 workshop on Integrating ontology, No commercial editor., 2005, pp. 25–32.

[29] T. Sagi, A. Gal, Non-binary evaluation for schema matching, in: Conceptual Modeling, Springer, 2012, pp. 477–486.

[30] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, et al., Results of the ontology alignment evaluation initiative 2016, in: 11th ISWC workshop on ontology matching (OM), 2016, pp. 73–129.

[31] G. A. Bowen, Naturalistic inquiry and the saturation concept: a research note, Qualitative research 8 (1) (2008) 137–152.

[32] H. Leopold, C. Meilicke, M. Fellmann, F. Pittke, H. Stuckenschmidt, J. Mendling, Towards the automated annotation of process models, in: International Conference on Advanced Information Systems Engineering, Springer, 2015, pp. 401–416.

[33] H. van der Aa, H. Leopold, H. A. Reijers, Comparing textual descriptions to process models–the automatic detection of inconsistencies, Information Systems 64 (2017) 447–460.

[34] J. Sànchez-Ferreres, J. Carmona, L. Padró, Aligning textual and graphical descriptions of processes through ilp techniques, in: International Conference on Advanced Information Systems Engineering, Springer, 2017, pp. 413–427.

[35] T. Baier, C. Di Ciccio, J. Mendling, M. Weske, Matching of events and activities-an approach using declarative modeling constraints, in: International Conference on Enterprise, Business-Process and Information Systems Modeling, Springer, 2015, pp. 119–134.

[36] A. Senderovich, A. Rogge-Solti, A. Gal, J. Mendling, A. Mandelbaum, The road from sensor data to process instances via interaction mining, in: International Conference on Advanced Information Systems Engineering, Springer, 2016, pp. 257–273.

[37] H. van der Aa, A. Gal, H. Leopold, H. A. Reijers, T. Sagi, R. Shraga, Instance-based process matching using event-log information, in: International Conference on Advanced Information Systems Engineering, Springer, 2017, pp. 283–297.

[38] H. Van der Aa, H. Leopold, H. A. Reijers, Checking process compliance on the basis of uncertain event-to-activity mappings, in: International Conference on Advanced Information Systems Engineering, Springer, 2017, pp. 79–93.

[39] C. Batini, M. Lenzerini, S. B. Navathe, A comparative analysis of methodologies for database schema integration, ACM computing surveys (CSUR) 18 (4) (1986) 323–364.

[40] C. Parent, S. Spaccapietra, Issues and approaches of database integration, Communications of the ACM 41 (5es) (1998) 166–178.

[41] P. A. Bernstein, E. Rahm, Data warehouse scenarios for model management, in: International Conference on Conceptual Modeling, Springer, 2000, pp. 1–15.

[42] J. A. Wald, P. G. Sorenson, Explaining ambiguity in a formal query language, ACM Transactions on Database Systems (TODS) 15 (2) (1990) 125–161.

[43] N. F. Noy, Semantic integration: a survey of ontology-based approaches, ACM Sigmod Record 33 (4) (2004) 65–70.

[44] T. R. Gruber, et al., A translation approach to portable ontology specifications, Knowledge acquisition 5 (2) (1993) 199–220.

[45] M. Uschold, M. Gruninger, Ontologies and semantics for seamless connectivity, ACM SIGMod Record 33 (4) (2004) 58–64.

[46] M. Crubézy, M. A. Musen, Ontologies in support of problem solving, in: Handbook on ontologies, Springer, 2004, pp. 321–341.

[47] D. Dou, D. McDermott, P. Qi, Ontology translation on the semantic web, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, 2003, pp. 952–969.

[48] N. F. Noy, M. A. Musen, The prompt suite: interactive tools for ontology merging and mapping, International Journal of Human-Computer Studies 59 (6) (2003) 983–1024.