# Activity Matching with Human Intelligence

Carlos Rodríguez[1], Christopher Klinkmüller[2,3], Ingo Weber[3,4], Florian Daniel[5], and Fabio Casati[1]

[1] University of Trento, Via Sommarive 9, 38123, Povo (TN), Italy
{crodriguez,casati}@disi.unitn.it
[2] Department of Computing, Macquarie University, Sydney, Australia
[3] Data61, CSIRO, Sydney, Australia
{christopher.klinkmuller,ingo.weber}@data61.csiro.au
[4] University of New South Wales, Sydney, Australia
[5] Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy
florian.daniel@polimi.it

**Abstract.** Effective matching of activities is the first step toward successful process model matching and search. The problem is nontrivial and has led to a variety of computational similarity metrics and matching approaches, however all still with low performance in terms of precision and recall. In this paper, instead, we study how to leverage on human intelligence to identify matches among activities and show that the problem is not as straightforward as most computational approaches assume. We access human intelligence (i) by crowdsourcing the activity matching problem to generic workers and (ii) by eliciting ground truth matches from experts. The precision and recall we achieve and the qualitative analysis of the results testify huge potential for a human-based activity matching that contemplates disagreement and interpretation.

**Keywords:** Activity matching, label matching, crowdsourcing

## 1 Introduction

Organizations with sizable process model collections encounter several use cases where matching activities of process models (deciding which of the activities of the process models are similar or even the same) is important, including search over the collection [9, 19, 15, 21] or identifying cloned models or fragments in models [10]. This problem has been addressed with a multitude of automated approaches over the last decade [3, 7, 17, 20, 23, 25, 24]. However, the success of fully automated, one-size-fits-all approaches is very limited when applied to heterogeneous process model collections, as observed in the Process Model Matching Contest of 2013 and 2015 [2, 5]. In some earlier work of ours we thus pursued a semi-automated approach, where user feedback was collected and the matching was improved based on corrections provided by the users [16]. Based on this input, the f-measure could be increased by around 40-50% in comparison to earlier works. The limiting factor of the approach is however the low availability of users with the necessary skills and time to invest.

In this paper, we start from the observation that deciding if two activities are similar or even the same is nontrivial, that purely computational approaches are not always able to correctly interpret the activities' textual labels, and that *human intelligence* (like in the case of user feedback [16]) can indeed make a difference. One of the reasons for the low performance of automated techniques is that often process models are not correctly formalized and, at best, come in the form of semiformal process models that, for instance, lack proper definitions of actors (e.g., no pools or swim lanes in the model), don't explicitly model data objects, use different activity labeling conventions (e.g., with or without mentioning the actor, the data object or the actual action to be performed), and so on. As a consequence, matching activities requires interpretation, an interpretation we claim needs to comprise also the context of the activities to be matched (e.g., the surrounding activities and the respective control flow structure). In line with the approach pushed forward in [16], we further believe this interpretation requires human intelligence, while the specific challenge we approach in this paper is to match activities by relaxing the assumption that this human intelligence necessarily comes from experts. We thus show how to match activities with the help of the crowd by crowdsourcing and studying different task designs oriented to generic, non-expert workers (the members of the crowd).

*Crowdsourcing* in fact provides convenient access to human intelligence via the Web, thanks to dedicated crowdsourcing platforms connecting workers with requesters who offer work. While there is a multitude of platforms supporting different crowdsourcing models, such as marketplaces [14], contests [4] and auctions [22], we specifically concentrate on marketplace platforms for *micro-tasks* with fixed rewards, as assessing the similarity of two activities is fine-grained enough to be formulated as a micro-task. Other examples of typical micro-tasks are annotating images, translating text or performing search activities on the Web. Prominent platforms supporting micro-tasks are Amazon Mechanical Turk (`https://www.mturk.com`) or CrowdFlower (`http://crowdflower.com`).

Designing effective crowdsourcing micro-tasks is however known to be challenging [1]. For instance, if too little information is given on a task, workers may not be able to complete the task; if too much information is given, they may abort the task or give arbitrary answers. Understanding if and how crowdsourcing can be leveraged to match activities in a way that indeed allows workers to bring in their human intelligence, as well as understanding if and how matching decisions by the crowd differ from those computed by algorithms or, instead, from those provided by process modeling experts, has therefore no immediate answer. We answer these questions by making the following contributions:

- A *conceptual model* of how the activity matching problem can be mapped to micro-tasks with basic, built-in quality controls;
- The design and implementation of a *ground truth elicitation experiment* with process modeling experts to study expert agreement inside a given domain;
- The design and implementation of *three crowdsourcing experiments* to study the performance (precision and recall) of the crowd compared to automated algorithms and the experts;

– A discussion of the *effect of human intelligence* and of the *effect of context visibility* on the quality of matches.

Next, we discuss the difference between machine- and human-based matching and review related works. In Section 3, we introduce crowdsourcing and a conceptual framework for task design, which we use in Section 4 to implement three tasks, along with an exercise to elicit ground truth mappings. In Section 5, we report on the outputs by experts, two automated matchers, and the crowd and discuss the results and findings in Section 6.

## 2 Activity Matching: Background

An *activity* is commonly interpreted as an action performed by an actor on some data object and represented by a textual label that describes the activity, $a = \langle act, role, obj, lab \rangle$. For example, an activity "Submit online form" may express a student submitting an online application form through some admission system. Typical *actions* are "create", "read", "update" and "delete" for documents, "send" and "receive" for messages, and "decide" for decisions. The *roles* depend very much on the domain of the process; for instance, a university admission process may involve a student, an admin and an examiner. The *data object* varies too, depending on the documents/artifacts worked on during the process; typical data objects are virtual/physical documents or entries in a database.

### 2.1 Machine- vs. human-based activity matching

Given two business processes models $BP_1$ and $BP_2$ and two activities $a^1 \in BP_1$ and $a^2 \in BP_2$, the purpose of *activity matching* is to decide whether the two activities match, that is, if they have the same or similar actor, role and data object, respectively (note that, for conciseness, *prop* is used to iterate over properties, and *pmatch* matches properties):

$$match(a^1, a^2) \iff \bigwedge_{prop \in \{act, role, obj\}} pmatch(prop^1, prop^2)$$

The basic problem is that of identifying 1:1 matches of activities of type $match(a^1, a^2)$. In general, however, matching activities is a 1:n or even an m:n problem: $match(a^1, \{a_j^2\})$ or $match(\{a_i^1\}, \{a_j^2\})$. For example, while one process may use an activity "Send documents," another one may split the group of documents into the individual documents to be sent and use the activities "Send form" and "Send ID" to represent the same activity. This would correspond to a $match($*"Send documents"*$, \{$*"Send form"*, *"Send ID"*$\})$. In practical settings that ask for the matching of process models that stem from different organizations and/or different modelers, 1:n and m:n correspondences are unavoidable.

The presence of 1:n and m:n correspondences, in turn, implies for activities that actions may have sub-actions, roles may have sub-roles, and documents may have sub-documents. Thus there may also exist $partof(a^1, a_j^2)$ relationships

between two activities that, for instance, qualify $a_j^2 \in \{a_j^2\}$ as part of $a^1$, starting from a *ppartof* relationship among the individual properties of the activities (we assume $ppartof(a, b) = true \iff a = b$ or $b = subelement(a)$):

$$partof(a^1, a_j^2) \iff \bigwedge_{prop \in \{act, role, obj\}} ppartof(prop^1, prop_j^2)$$

A 1:n activity match can thus be defined as a match of an activity $a^1$ with a set of activities $\{a_j^2\}$ that perform parts of $a^1$:

$$match(a^1, \{a_j^2\}) \iff \bigwedge_{j \in |\{a_j^2\}|} partof(a^1, a_j^2)$$

Merging the activity matches and part-of relationships from $BP_1$ to $BP_2$ with those from $BP_2$ to $BP_1$ identifies the m:n matches between the processes.

Now, asserting an exact match both among activities and their individual properties is generally hard, and the use of *similarity metrics* that assess a degree of matching is common practice [8]. In the case of automated matching algorithms, similarity is typically based on objective, syntactic or semantic features of the *labels* describing the activities ($t^a$ is a threshold value) [20, 17]:

$$match^a(a^1, a^2) \iff sim^a(lab^1, lab^2) > t^a$$

If instead of by machines, activity matches are to be identified by human actors, such as process modelers or domain experts, subjective similarity metrics are applied. The respective criteria are based on the personal experience and expertise of the human actor, and typically don't consider only the labels of activities in an isolated fashion, but also interpret parts of or the full process models containing the activities to be matched. That is, humans don't simply assert similarity based on labels, but naturally also take into account the context of the activities, i.e., other surrounding model constructs (activities, data objects, control flow constructs, etc.). Activity labels are the starting point of the analysis, while the objective is the identification of the real meaning of activities in the process models, that is, the actual action, role and data object an activity refers to. Two activities therefore match if the perceived similarity of these properties exceeds some subjective threshold:

$$match^h(a^1, a^2) \iff \sum_{prop \in \{act, role, obj\}} \alpha_{prop} * sim^h(prop^1, prop^2) > t^h$$

The exact values of $t^h$ and of the weights $\alpha_{prop} \in [0, 1]$ are subjective, and only the expert himself/herself can judge how and when he/she wants to assert a match or not. The expert might – depending on his view – consider also other properties, e.g., resources, process context, dependencies, or similar that help him/her in the decision process. Analogous considerations hold also for the *partof* relationship that allows the identification of 1:n and m:n matches if assessed by human actors. To the best of our knowledge, computational approaches

do not focus so far on *partof* relationships with the meaning defined above; existing matching techniques are not limited to 1:1 matches only, but identified 1:n or m:n matches are the result of label similarity not of a reasoning on the actual meaning of activities.

In this paper, we are particularly interested in eliciting the interpretation represented by the $match^h$ function (including possible *partof* relations) and less in that of the $match^a$ function. The intuition is that humans reason on the essence of the problem, while machines do so only on a proxy of it (the labels).

### 2.2 Related work

The identification of correspondences between models has been studied in the field of ontology and schema matching [11]. However, the applicability of such approaches to process model matching is limited as process models depict actions and their execution order instead of concepts and their relations. Accordingly, a poor performance was observed when applying schema and ontology matchers to process models [2, 7]. Furthermore, process similarity search techniques [8] which measure the overall similarity of process models provide basic concepts for comparing process models on a fine-grain level. Such techniques rely on textual [18], structural [6, 12], and behavioral information [19, 26].

Based on these approaches, a variety of process model matching techniques has been proposed [3, 7, 17, 20, 23, 25]. Essentially, all these techniques determine correspondences based on the comparison of activity labels, i.e., they try to estimate the functional overlap of activities based on their textual description. Additionally, some approaches integrate structural and behavioral information to decide whether activities correspond or not [3, 7, 20, 23]. However, comparative evaluations based on different data sets revealed that the quality of these approaches is too low to be applicable in practice [2, 5].

Human intervention has been recognized as a source for improving the performance of matchers [16, 24]. In [24] experts are required to provide correspondences for a subset of the model pairs in a model collection. With regard to these correspondences the quality of different matchers is determined. Then, a prediction model that correlates process characteristics to the quality of the matchers is trained and used to select matchers for the remaining model pairs. Similarly, an approach that exploits expert feedback to learn the domain specific vocabulary used in a model collection is introduced in [16]. Correspondences that were automatically determined and manually corrected by experts are analyzed and the textual similarity assessment is adopted. This way improvements with regard to the f-measure of up to 53% compared to the state-of-the-art were achieved. We pursue the idea of relying on human intelligence, in particular utilizing the crowd, to reduce the workload for experts and speed up the matching process.

## 3 Crowdsourcing the Activity Matching Problem

*Crowdsourcing* (CS) is the outsourcing of a unit of work to a crowd of people via an open call for contributions [13]. A *worker* is a member of the crowd (a human)

that performs work, and a *requester* is the organization, company or individual that crowdsources work. For the purpose of this paper, we specifically leverage on work expressed as *micro-tasks*, where crowdsourcing a micro-task (simply "task" in the following) involves the following steps: The requester publishes a description of the task to be performed in a crowdsourcing platform. The crowd inspects and possibly expresses interest for tasks. The requester also defines the reward workers will get for performing the task and how many answers (task instances) should be collected (instantiated) per task. Not everybody of the crowd may, however, be eligible to perform a given task, either because the task requires specific capabilities (e.g., language skills) or because the workers should satisfy given properties (e.g., only female workers). Deciding which workers are allowed to perform a task is called pre-selection, is optional, and may be done either by the requester manually or by the platform automatically (e.g., via gold data). Once workers are enabled to perform a task, the platform creates as many task instances as necessary to collect the expected number of answers. Upon completion of a task instance (or a set thereof), the requester may inspect the collected answers and validate the respective quality. Work that is not of sufficient quality is not useful, and the requester may not reward it.

The major challenge in designing a crowd task is to ensure that the requester can rely on the results. That means the results obtained from the crowd have to be of a high quality and should only contain a small portion of imprecise or incorrect answers. To achieve this goal, the task designer has to bring together both worlds, that of the requester and that of the crowd. On the one hand, it is therefore necessary to design tasks in such a way that (i) workers obtain sufficient insights into the context, (ii) they can conveniently express their decisions, and (iii) quality is adequate in order to leverage the potential of the crowd. On the other hand, requirements imposed by the requester, like time or cost constraints as well as the confidentiality of information, need to be taken into account.

In this paper, we are specifically interested in studying opportunities to crowdsource the task of activity matching as an instance of the more general problem of correspondence identification. We thus started this study by structuring the problem space, in order to be able to discuss task design alternatives and guide our research. As a result, we developed a *conceptual crowdsourcing design framework for activity matching*, which decomposes the overall task into several fine-grained aspects that need to be considered. The framework is the result of a discussion on how to relate, combine or slice the aspects.

As shown in Table 1, on an abstract level the framework is concerned with (i) how *questions* are posed to workers, (ii) which options workers have when *answering*, and (iii) how *answer quality* is controlled. In the following, we discuss the complete framework with all of its dimensions in more detail.

**Question group:** This group defines what specific tasks the contributors are asked to perform, in order to enable the matching of activities from different process models, and which information is provided.

*Task description* — It is important to describe well the task and its purpose to clarify what the requester wants to obtain from the workers. *Correspondence*

| Groups | Dimensions | Options | | |
|---|---|---|---|---|
| Question | Task description | Correspondence identification | Activity cluster identification | Activity annotation |
| | Representation | Whole process | Process fragment | Activity label |
| | Documentation | Additional | None | |
| Answer | Modality | Fixed | Free | Combination |
| | Range | Binary | Numeric | Semantic |
| | Direction | Unidirectional | Bidirectional | |
| Quality | Audience | External | Internal | Team |
| | Timing | Before | During | After |
| | Test nature | Gold questions | Ad-hoc questions | |

**Table 1.** The conceptual crowdsourcing design framework for activity matching

*identification* asks for feedback on the relations between activities or sets thereof to separate corresponding from non-corresponding activities. *Activity cluster identification* addresses relations of activities within the same model to identify activities that relate to a same higher-level activity. *Activity annotation* solicits feedback regarding a single activity to enable an indirect alignment of activities, e.g., by mapping them to a taxonomy that could be a set of harmonized labels, a set of semantic annotations or a reference process.

*Representation* — As process models show the internals of how an organization operates, there may exist privacy concerns in showing them to public workers. Instead of showing the *whole process model*, only a *process fragment* may be shown, or even only *activity labels* without any further information. This dimension is also characterized by a tradeoff between complexity and quality: showing large models at full may overwhelm workers, while it might be necessary for workers to have sufficient information to take decisions.

*Documentation* — *Additional* documentation, such as a short explanation or even process handbooks or glossaries, might be presented to workers to provide help and instructions on how to perform the task. Yet, it could be a choice to provide *no* documentation, if the task is self-explaining or the documentation might again overwhelm the worker.

**Answer group:** While the question group refers to the presentation of the task, this group defines how workers can answer questions.

*Modality* — This concerns the degree of freedom a worker has in answering. The workers might be asked to select from a *fixed* set of options or to enter a *free* text answer. Furthermore, *combined* versions where workers can select from a set of options or enter a new answer are conceivable.

*Range* — Requesters might be interested in different aspects of relations between two activities or one activity and a taxonomy element. In the most simple case, workers are expected to give a *binary* value indicating whether a relation holds or not. Alternatively, the degree to which a relation holds can be measured on a *numeric* scale, e.g., $0 - 100\%$. Relations might also be assigned to a *semantic* class, such as "unrelated", "A subsumes B", or "equal".

*Direction* — This dimension specifies if relations among activities expressed by a worker by relating one activity to another are *unidirectional* or *bidirectional*. The use of bidirectional relations may reduce the effort needed to match activities.

**Quality group.** This group characterizes the methods adopted to ensure that the answers by the crowd are reliable and useful to the requester.
*Audience* — In general, tasks may be crowdsourced to different audiences. If the requester is an organization with own employees, *internal* workforce might be considered, while *external* crowds can be involved by any kind of requester. The involvement of *teams* of workers, which have proven to promise better results (e.g., experts that work together on the alignment of processes), is harder.
*Timing* — Quality control methods can be applied *before* (e.g., by excluding workers based on skill tests), *during* (e.g., by incorporating test questions to validate the experts answers) or *after* (e.g., by removing inconsistent and unreliable answers) feedback collection. Several methods can be used in an experiment.
*Test nature* — Tests can come in the form of so-called *gold questions*, that is, questions that workers are asked to answer but for which the answers are already known, or in the form of *ad-hoc questions*, which are added to the task only for testing purposes (e.g., skills test or CAPTCHA-like tests to tell workers and robots apart) without any real use for the requester.

Jointly, these dimensions span a space of potential task designs. A particular crowdsourcing experiment can be understood as a point in this space. Without considering that selected options can be implemented in different ways or that certain combinations might be impractical, the space has $2^3 \times 3^6 = 5832$ points. Yet, the framework can still be extended with additional dimensions, e.g., we do not specifically study the effect of different rewards in this paper. Nevertheless, the framework serves as a useful tool for taking informed decisions about task designs and for comparing them. In the next section, we will use the framework to describe the three task designs we adopt in our study.

## 4 Study Design

### 4.1 Dataset

The *dataset* we use for the experiments in this paper is a subset of that introduced in [20], which consists of nine models (36 different model pairs) of the study admission processes at different German universities. The models were created by graduate students from Humboldt-Universität zu Berlin within a research seminar on process modeling in three semesters. We use the respective BPMN models with 10 to 44 activities and an average of 21 activities per model.

The *subset* of process models we selected for the study described in this paper consists of four models (Frankfurt (F), TU Munich (M), Cologne (C), FU Berlin (B)) and three model pairs. Models were paired to represent different levels of syntactical label similarity, so as to enable a representative comparison of the crowd with automated algorithms: F/M has 10 activities with exactly the same activity labels, C/F only 6, and C/B 0 (none). Limiting the study to three model pairs was necessary to contain the cost of the crowd and expert experiments.

All process models express semi-formal, high-level views on the processes and are not executable without further refinement. For instance, the models do

not make use of pools and swim lanes, follow different activity naming conventions (they stem from different modelers), are characterized by ambiguity (for instance, it is very hard to assess what action and/or role the activities "Keep in Applicant pool" or "document" refer to), and gateways partly lack conditions. Yet, this is the typical situation of process repositories that contain models whose purpose is documentation rather than execution. The dataset has already been used for a comparative evaluation of matching approaches in [5] as well as for the evaluation of a matching approach based on expert feedback in [16] and represents a convenient choice for the comparison with prior works.

Since in this study we are particularly interested in understanding the effect of the human interpretation of models by both the crowd and experts, we opted not to reuse the *ground truth* mappings proposed in [20]. On the one hand, these mappings turned out to be too restrictive in our initial trials and mostly focused on exact matches (no separation of the *match* and *partof* relationships); on the other hand, without insight into the individual researchers' mappings prior to the consolidation it is not possible to assess inter-expert agreement. The creation of a new ground truth is thus part of the experiment described next.

### 4.2 Expert-based activity matching exercise

In order to (i) be able to study the agreement among experts about activity matchings and (ii) have a ground truth for the comparison of the crowd with automated algorithms, we set up an activity matching exercise that involved five process modeling experts (one PhD candidate, three PhDs and one assistant professor, all with BPMN expertise). The exercise aimed to produce four *individual mappings* for each of the three model pairs, plus one *consolidated mapping* that integrates the other four according to the judgement by the most senior participant. All participants were provided with the BPMN models of the four chosen processes and asked to identify all possible *match* and *partof* relationships for each of the three process pairs (F/M, C/F, C/B – see Section 4.1). Data were collected using a Google Spreadsheet (`https://goo.gl/N3xNgb`), and activities and relationships could be selected from suitable dropdown lists; the spreadsheets also contained links to the graphical BPMN models and allowed the experts to express a similarity degree for identified matches using a 7-point Likert scale ranging from "somehow similar" to "the same" as well as to provide informal feedback. All experts concluded the exercise within 30-60 minutes and were rewarded with a free lunch for their effort.

### 4.3 Machine-based matching algorithms

As a baseline for the assessment of the crowd's performance we use two automated matching techniques. First, we consider the *bag-of-words technique* (BOT) [17]. For a given pair of process models, BOT iterates over the set of all activity pairs where each pair contains one activity from each model being matched. It computes a similarity score based on the activity labels, and retains all activity pairs with a score higher than a predefined threshold. To compute the similarity

score, the labels are split into sets of words, and each word in one set is compared to each word in the other set using a word similarity function. The final similarity score is the average of the maximum similarity scores for each word. If the two sets of words are of a different size, the larger set is reduced to the size of the smaller set by removing the words with the lowest maximum word similarity. In this study, we specifically use the configuration we submitted to the first Process Model Matching Contest and that yielded the best results on the university admission dataset in this contest [5].

The second technique is the *order preserving bag-of-words technique* (OPBOT) which contains different BOT configurations that it applies to a model collection separately. For each configuration it predicts the quality by investigating structural relations between the proposed correspondences. OPBOT then selects the most promising configuration and proposes its results. Similar to BOT, we utilized the configuration that participated in the second contest and was named as one of two outstanding matching techniques [2].

### 4.4 Crowd-based micro-tasks

Crowdsourcing platforms have different built-in options that support the aspects of the design framework in Table 1 to different degrees. We use Crowdflower (`www.crowdflower.com`) and propose three different task designs that vary in terms of the contextual information provided (none vs. process fragments) and the freedom given to workers in choosing matches (none vs. free definition of matches). The intuition behind these design options is that (i) contextual information (surrounding activities) helps making better judgements about the similarity of tasks and (ii) freedom of choice allows us to match activities more cost-effectively. All task designs ask workers to (i) decide if one or more pairs of activities are similar (*yes/no* answer) and to provide, for each identified match, (ii) the type of relationship (*match/partof*), (iii) a similarity score using a 7-point Likert scale (1-Not similar at all – 7-Very similar or identical); and (iv) a free-text explanation of the judgment. The specific designs are (see design sketches in Figure 1):

- *LabelOnly* is the most simple task design. It applies the computational approach to the crowd: workers are only presented with two activity labels.
- *ContextOne* shows two fragments with 3-5 activities from two process models and highlights the activities to be matched (1 per fragment).
- *ContextSet* shows the same process fragments as *ContextOne*, without however highlighting any pair of activities. If the workers spot a similarity, they can freely choose the respective activities from dropdown lists; the design allows the identification of up to 10 matches. No explanations are required.

For *LabelOnly* and *ContextOne* we have a total of 989 activity pairs to be compared; for *ContexSet* we have 63 process fragment pairs, given how we split the models. For all three task designs, we collect 3 judgements per pair (to improve quality), which leads us to a total number of 6123 units of work to crowdsource. We also use Crowdflower's built-in quality control based on gold questions

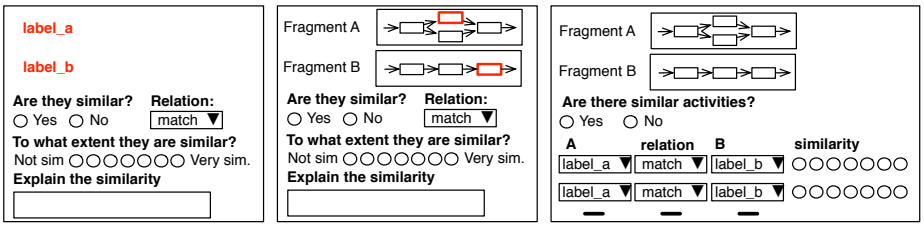


**Fig. 1.** Micro-task designs for activity matching. Actual screenshots of the tasks deployed in Crowdflower can be found in https://goo.gl/xjCHmv

(with known answers). For *LabelOnly* and *ContextOne*, the gold question asks whether or not the activities are similar; for *ContextSet* the gold question asks if there are similar activities in process fragments. As reward for *LabelOnly* and *ContextOne* we pay US$ 0.01 for each unit of work, while for *ContextSet* we pay US$ 0.05 as it requires more effort.

In terms of the design framework introduced in Section 3, all designs ask workers to identify *correspondences*, *without* providing additional documentation beyond an example. *LabelOnly* shows only *activity labels*, the other two designs use *process fragments*. All designs, except *ContextSet* (only *selections*) allow workers to input a *combination* of selections and free text in the form of *binary*, *numerical* and *semantic* inputs. Matchings are *bidirectional*, part-of relations *unidirectional*. The crowd is *external* (indep. of the authors), and quality control is done *during* task execution and *afterwards* with the help of *gold questions*.

### 4.5 Evaluation metrics

For the evaluation of the agreement between the experts in the creation of the ground truth and with the consolidated set of matchings, we use the *Jaccard similarity coefficient*, which expresses the similarity/diversity of sample sets:

$$J(M_i, M_j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|} \tag{1}$$

where $M_i$ and $M_j$ are the sets of correspondences identified by experts $i$ and $j$. If the experts agree on each match, $J(M_i, M_j) = 1$, otherwise $J(M_i, M_j) < 1$.

Now, given a ground truth, each correspondence identified by an activity matching approach can be classified as true positive ($TP$), false positive ($FP$) or false-negative ($FN$). This allows the computation of the common *precision*, *recall* and *f-measure*, as defined by the following formulas:

$$P = \frac{TP}{TP + FP} \quad (2) \qquad R = \frac{TP}{TP + FN} \quad (3) \qquad F = \frac{2 \times P \times R}{P + R} \quad (4)$$

Since in this study we explicitly distinguish between *match* and *partof* relationships among activities, we compute $P$, $R$ and $F$ for exact matches and

**Table 2.** Number of *match* (m) and *partof* (p) relations identified by the experts.

| | Expert 1 | | Expert 2 | | Expert 3 | | Expert 4 | | Consolidated | |
|---|---|---|---|---|---|---|---|---|---|---|
| | m | p | m | p | m | p | m | p | m | p |
| F/M | 6 | 5 | 9 | 12 | 8 | 5 | 18 | 5 | 7 | 12 |
| C/F | 5 | 9 | 6 | 8 | 6 | 3 | 7 | 8 | 4 | 12 |
| C/B | 3 | 14 | 4 | 19 | 4 | 6 | 5 | 18 | 3 | 21 |
| total | 14 | 28 | 19 | 39 | 18 | 14 | 30 | 31 | 14 | 45 |

**Table 3.** Jaccard similarity among experts of *match* (m), *partof* (p) and both together (b), averaged over the three process pairs; in bold the biggest similarities.

| | Expert 1 | | | Expert 2 | | | Expert 3 | | | Expert 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | m | p | b | m | p | b | m | p | b | m | p | b |
| Expert 1 | – | | | .435 | .340 | .515 | .391 | .077 | .345 | .333 | .311 | .431 |
| Expert 2 | **.435** | **.340** | **.515** | – | | | .609 | .205 | .452 | .441 | .400 | .566 |
| Expert 3 | .391 | .077 | .345 | .609 | .205 | .452 | – | | | **.500** | .286 | .431 |
| Expert 4 | .333 | .311 | .431 | .441 | .400 | .566 | .500 | **.286** | .431 | – | | |
| Consolidated | .400 | .304 | .464 | **.650** | **.615** | **.746** | **.684** | .229 | **.468** | .467 | **.551** | **.667** |

part-of relationships individually, as well as for the union of both relationships. This allows us to study the strengths and weaknesses of the approaches.

## 5 Experiment

### 5.1 Expert-based activity matching

The activity matching exercise with the experts produced a rich set of activity matchings, as reported in Table 2. Overall, the five experts identified 252 correspondences, 95 exact matches and 157 part-of relationships, with an average of 6.33 matches and 10.47 part-of relations per process model pair. The consolidation of the four individual results yielded 14 matches and 45 part-of relationships. Part-of relationships among activities are therefore so frequent that they cannot be neglected in practical activity matching exercises.

Table 3 analyzes the correspondences by the five experts in more detail with a cross-analysis of the respective Jaccard similarities, in order to understand the level of agreement or disagreement among the experts. We immediately note that there is no clear agreement among any of the experts. We also note that the consolidated mapping generally represents well the output by the four individual experts, especially if we compute similarity by merging both (b) matches and part-of relations; only expert 1 seems to have more affinity with expert 2 than with the consolidated mapping. This qualifies the consolidated mapping as the best choice for the evaluation of the performance of the crowd and the algorithms.

### 5.2 Machine-based activity matching

Table 4 presents the performance of the automated matchers with regard to the consolidated ground truth. The f-measures vary from 0.276 to 0.538 for BOT and from 0.276 to 0.621 for OPBOT across all three model pairs. OPBOT performs

**Table 4.** Average precision ($P$), recall ($R$, $R_m$, $R_p$) and f-measure ($F$) of BOT and OPBOT for the three process pairs separated by matching relation.

| BOT | | | | | OPBOT | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $P$ | $R$ | $R_m$ | $R_p$ | $F$ | $P$ | $R$ | $R_m$ | $R_p$ | $F$ |
| .700 | .359 | .536 | .258 | .448 | .900 | .338 | .536 | .230 | .481 |

**Table 5.** Average $P$, $R$, $R_m$, $R_p$, and $F$ values for LabelOnly, ContextOne and ContextSet as a function of worker agreement ($x$ out of 3 votes); best averages in bold.

| | LabelOnly | | | | | ContextOne | | | | | ContextSet | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $R_m$ | $R_p$ | $F$ | $P$ | $R$ | $R_m$ | $R_p$ | $F$ | $P$ | $R$ | $R_m$ | $R_p$ | $F$ |
| 1/3 votes | .194 | **.791** | .758 | **.393** | .207 | .207 | .781 | **.917** | .349 | .320 | .548 | .512 | .758 | .123 | **.530** |
| 2/3 votes | .410 | .558 | .758 | .274 | .453 | .467 | .600 | .869 | .222 | .509 | .635 | .321 | .647 | .059 | .417 |
| 3/3 votes | .582 | .491 | .758 | .190 | .515 | .631 | .460 | .758 | .147 | .515 | **.861** | .192 | .516 | .016 | .310 |

slightly better than BOT ($0.481 > 0.448$) with a better precision and a similar recall. Overall, it is interesting to note that the precision of the identified matches is generally high, while the recall is instead rather low. That is, if activity labels are similar, both algorithms are able to spot the similarity; if instead labels of similar activities are not similar enough, the algorithms fail. Also, computing recall over matches ($R_m$) and part-of relations ($R_p$) independently unveils that the algorithms are better in identifying exact matches than part-of relations.

### 5.3 Crowd-based activity matching

Table 5 reports on $P$, $R$ and $F$ for the crowdsourcing experiments, distinguishing between different levels of worker agreement on correspondences (recall that each activity pair was assessed 3 times). We consider two activities to be similar if either a *partof* or *match* relation was indicated by the crowd. For *LabelOnly* and *ContextOne*, the precision is lower than that of the algorithms, while the recall is higher. Interestingly, *ContextSet* shows a very good precision, up to 0.861 for 3/3 votes, however with a lower recall; the freedom given to workers seems to intrinsically favor precision, e.g, because workers only propose matches they are highly confident with. If we split $R$ into $R_m$ (matches only) and $R_p$ (part-of relations only), we see that the crowd is particularly good at recalling exact matches ($R_m \in [.758, .917]$ for *ContextOne*). Of course, the higher the agreement among workers, the higher the precision and the lower the recall.

Table 6 analyzes in more detail the correctness of the $TPs$ by model pair using the agreement level with the highest f-measure in Table 5. For instance, for F/M all matches proposed by the workers are correct, while only 30% of their part-of relations are correct. Overall, the proposed matches are very precise; the part-of relations less so.

**Table 6.** Correctness of workers' *match* and *partof* relations (true positives only).

| | LabelOnly (3/3 votes) | | ContextOne (3/3 votes) | | ContextSet (1/3 votes) | |
|---|---|---|---|---|---|---|
| | $m$ | $p$ | $m$ | $p$ | $m$ | $p$ |
| F/M | 1.00 | .300 | .947 | .400 | 1.00 | .333 |
| C/F | 1.00 | .480 | .900 | .364 | 1.00 | .100 |
| C/B | .571 | .552 | .875 | .615 | .500 | .471 |
| Avg | .912 | .459 | .919 | .471 | .926 | .333 |

A qualitative analysis of the $FNs$ (31) confirms the difficulties with the part-of relations, e.g., with the similarity between "Apply Online" and "add certificate

of bachelor degree," as well as with modeling ambiguity, e.g., with "Evaluate" (activity) vs. "less than 16 cp in mathematics" (condition). An analysis of the $FPs$ (15) reveals that the crowd may actually be right in some cases, e.g., "certificate received" vs. "documents received" (synonyms) or "Acceptance" vs. "accepted provisionally" (part-of), if the domain of the study was different. That is, most $FPs$ actually are plausible ground truth candidates.

The cost of the experiments was US$40.56, US$40.80 and US$28.32 for *LabelOnly*, *ContextOne* and *ContextSet*, respectively, including platform fees.

## 6 Discussion

We summarize the findings of this study as follows: (i) Process models can be *intrinsically ambiguous*, underspecified and even contradicting. Matching activities under these conditions requires an interpretation that goes beyond the scope of individual activity labels. (ii) Given this ambiguity, even *experts may not agree* on how to match activities. In fact, the disagreements we encountered in our experiments are both consistent and high among all experts. (iii) On the newly created ground truth data, the performance of the tested computational matchers was characterized by *high precision and low recall*, with a particular weakness in discovering part-of relationships among activities. (iv) Crowd-based activity matching outperformed the automated matchers by a margin of about 10%. Depending on the logic used for combining crowd worker answers, however, *high recall can be achieved when sacrificing precision*. The crowd was also able to elicit non-obvious part-of relationships by reasoning on activities like experts do, that is, trying to figure out the essence of activities (action, role, object). (v) The design of the micro-tasks for activity matching has a *strong effect on the quality* of the produced matchings. The three task designs we tested showed significant performance differences, depending on the level of insight into the context of activities as well as on the level of freedom (responsibility) given to workers. Asking the crowd to reactively judge a given activity pair tends to favor recall (*ContextOne*); asking it to proactively identify similar pairs tends to favor precision (*ContextSet*). (vi) Given the low agreement among the experts, the *P/R values reported here must however be handled and interpreted with care*. The less formal and complete models are, the more ambiguous they are, and the harder it is also to define a reliable ground truth and, hence, to reliably test approaches. The variance and disagreement in human feedback leads to the larger question: *is the assumption that an objective ground truth or "gold" standard exists valid?*

These findings advance the state of the art of activity and process matching with an original perspective on the problem compared to prior works on the topic, i.e., that of the human. To the best of our knowledge, this is the first study that proposes a crowd-based activity matching approach and compares it with state-of-the-art computational approaches. It is also one of the first studies that critically analyzes the (lack of) agreement among experts and that shows that performance tests based on ground truth data elicited from experts must be interpreted with care, perhaps more care than devoted to this aspect so far.

A consideration regarding the "noise" (spectrum and variety of matchings) produced by the crowd: while false positives (compared to the ground truth) by algorithms may not present useful information, the "false positives" by the crowd may even represent an *added value* in the context of process model matching. In fact, these matches may represent similarities the experts did not consider when creating the ground truth, e.g., because they simply were focused on a specific domain while the crowd was not. Especially in the context of exploratory search over process repositories (to search for similar practices, to understand how a given organization approaches typical problems, to identify processes that could be merged and consolidated, etc.) the different viewpoints and interpretations provided by the crowd may allow the discovery of unexpected models that indeed present semantic similarities not considered before. This kind of knowledge is hard if not impossible to elicit without the contribution of human intelligence.

Of course, the study described in this paper also comes with its very own limitations: The dataset we used contains processes that are very similar; results might change for more heterogeneous datasets. The micro-task designs we used represent a reasoned best effort, and we did not yet try to optimize results, for example by varying the reward of workers. Our experiments exemplarily analyzed three process model pairs, and obtaining statistical relevance of the results would require more data; due to resource restrictions, we opted for a more qualitative analysis. Finally, even though the results are promising, given the crowd costs reported in our study and the efforts in setting up an experiment like this, there is a trade-off that needs to be considered before opting for a crowd-based approach.

In our future work, we intend to extend the presented work in several directions. Different approaches from crowd workers and algorithms have different strengths: while some approaches have a high recall, others achieve high precision. We thus plan to investigate how we can combine approaches into novel matching workflows that combine the benefits of several approaches. For instance, we could use a crowd task design that yields high recall values at the expense of precision, and use an automated matcher to filter the crowd results. We also see as highly interesting understanding the human perception of similarity better. Such research would likely benefit from an interdisciplinary approach, in collaboration with psychologists, linguists, or sociologists.

**Acknowledgement**. We would like to thank M. Vitali, G. Meroni, P. Plebani (Politecnico di Milano) and S. Tranquillini and J. Stevovic (Chino, Trento) for their help with the creation of the ground truth matchings for the experiments.

## References

1. M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. Motahari-Nezhad, E. Bertino, and S. Dustdar. Quality control in crowdsourcing systems: Issues and directions. *Internet Computing, IEEE*, 17(2):76–81, March 2013.
2. G. Antunes et al. The process model matching contest 2015. In *EMISA 2015*.
3. M. Castelo Branco, J. Troya, K. Czarnecki, J. Küster, and H. Völzer. Matching business process workflows across abstraction levels. In *Model Driven Engineering Languages and Systems*, pages 626–641, 2012.

4. R. Cavallo and S. Jain. Efficient Crowdsourcing Contests. In *Proceedings of AAMAS 2012 - Volume 2*, pages 677–686, 2012.
5. U. Cayoglu et al. The process model matching contest 2013. In *PMC-MR*, 2013.
6. R. Dijkman, M. Dumas, and L. García-Bañuelos. Graph matching algorithms for business process model similarity search. In *BPM*, pages 48–63, 2009.
7. R. Dijkman, M. Dumas, L. Garcia-Banuelos, and R. Kaarik. Aligning business process models. In *EDOC 2009*, pages 45–53, 2009.
8. R. Dijkman, M. Dumas, B. van Dongen, R. Käärik, and J. Mendling. Similarity of business process models: Metrics and evaluation. *Inf. Syst.*, 36(2):498–516, 2011.
9. M. Dumas, L. García-Bañuelos, and R. M. Dijkman. Similarity search of business process models. *IEEE Data Eng. Bull.*, 32(3):23–28, 2009.
10. C. C. Ekanayake, M. Dumas, L. García-Bañuelos, M. L. Rosa, and A. H. M. ter Hofstede. Approximate clone detection in repositories of business process models. In *BPM*, pages 302–318, 2012.
11. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag New York, Inc, Secaucus, NJ, USA, 2007.
12. D. Grigori, J. C. Corrales, and M. Bouzeghoub. Behavioral Matchmaking for Service Retrieval. In *IEEE ICWS*, pages 145–152, 2006.
13. J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1st edition, 2008.
14. P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17(2):16–21, Dec. 2010.
15. T. Jin, J. Wang, M. L. Rosa, A. H. ter Hofstede, and L. Wen. Efficient querying of large process model repositories. *Computers in Industry*, 64(1):41–49, 2013.
16. C. Klinkmüller, H. Leopold, I. Weber, J. Mendling, and A. Ludwig. Listen to me: Improving process model matching through user feedback. In *BPM 2014*, pages 84–100, 2014.
17. C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig. Increasing Recall of Process Model Matching by Improved Activity Label Matching. In *BPM 2013*, pages 211–218, 2013.
18. A. Koschmider and E. Blanchard. User assistance for business process model decomposition. In *IEEE RCIS*, pages 445–454, 2007.
19. M. Kunze, M. Weidlich, and M. Weske. Behavioral similarity - a proper metric. In *BPM*, pages 166–181, 2011.
20. H. Leopold, M. Niepert, M. Weidlich, J. Mendling, R. M. Dijkman, and H. Stuckenschmidt. Probabilistic Optimization of Semantic Process Model Matching. In *BPM 2012*, pages 319–334, 2012.
21. S. Sakr, A. Awad, and M. Kunze. Querying process models repositories by aggregated graph search. In *BPM 2012*, pages 573–585, 2012.
22. B. Satzger, H. Psaier, D. Schall, and S. Dustdar. Auction-based crowdsourcing supporting skill management. *Inf. Syst.*, 38(4):547–560, June 2013.
23. M. Weidlich, R. Dijkman, and J. Mendling. The icop framework: Identification of correspondences between process models. In *CAiSE 2010*, pages 483–498, 2010.
24. M. Weidlich, T. Sagi, H. Leopold, A. Gal, and J. Mendling. Predicting the quality of process model matching. In *BPM 2013*, pages 203–210, 2013.
25. M. Weidlich, E. Sheetrit, M. C. Branco, and A. Gal. Matching business process models using positional passage-based language models. In *ER 2013*, pages 130–137, 2013.
26. H. Zha, J. Wang, L. Wen, C. Wang, and J. Sun. A workflow net similarity measure based on transition adjacency relations. *Computers in Industry*, 61(5):463–471, 2010.