Bayesian evaluation and comparison of ontology alignment systems

Mohammadi, Majid

DOI
10.1109/ACCESS.2019.2903861

Publication date
2019

Document Version
Final published version

Published in
IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 12, 2019, accepted March 5, 2019, date of publication March 13, 2019, date of current version May 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903861

Bayesian Evaluation and Comparison of Ontology Alignment Systems

MAJID MOHAMMADI
Faculty of Technology, Policy, and Management, Delft University of Technology, 2628BX Delft, The Netherlands
e-mail: m.mohammadi@tudelft.nl

ABSTRACT Ontology alignment systems are evaluated by various performance scores, which are usually computed by a ratio related directly to the frequency of the true positives. However, such ratios provide little information regarding the uncertainty of the overall performance of the corresponding systems. The comparison is also drawn merely by the juxtaposition of computed scores, and specify that one system is superior to one another provided that its score is higher. Nonetheless, the comparison based solely on two figures would not quantify the significance of difference and would not determine the extent to which one system is better. The problem compounds for comparison over multiple benchmarks since averages and micro-averages of performance scores are considered. In this paper, the evaluation of alignment systems is translated into a statistical inference problem by introducing the notion of *risk* for alignment systems. The risk with respect to a performance score is shown to follow a binomial distribution and is equivalent to the complement of the score, e.g., precision risk $= 1 -$ precision. It is also demonstrated that the maximum likelihood estimation (MLE) is precisely equivalent to the conventional evaluation by using ratios. Instead of using the MLE, the Bayesian model is developed to estimate the risk with respect to a score (or equivalently, the score itself) as a probability distribution from the performance of the systems over single or multiple benchmarks. As a result, the evaluation outcome is a distribution instead of a figure, which provides a broader view of the overall system performance. A Bayesian test is also devised to compare various systems based on their estimated risks, which can compute the confidence that one system is superior to one another. We report the result of applying the proposed methodology to multiple tracks from the ontology alignment evaluation initiative (OAEI).

INDEX TERMS Bayesian, evaluation, ontology alignment, precision, recall.

I. INTRODUCTION

The semantic web technologies have been impressively advanced since their genesis. The principal goal of the semantic web technologies is to present information in a way that a machine can understand and construe them. To this end, ontologies are introduced as a standard tool to formally model a domain's objects and their relations [1].

One essential hurdle for machines to understand is when the information comes from various sources. Although it is likely that distinct sources use the same standard to present their information, there is no guarantee that they use similar terminologies or the identical way of designing the concepts. Thus, information sources are heterogeneous by nature even though they state the same piece of data.

The associate editor coordinating the review of this manuscript and approving it for publication was Wajahat Ali Khan.

The ontology alignment is the process that reconciles the difference between various sources of information, which state the same concepts in distinct ways. Given two ontologies, the outcome of the ontology alignment is a set of correspondences (i.e., mappings) each of which maps entities in the first ontology to those in the second [2]. Having the set of correspondences is the prerequisite to various tasks such as ontology integration [3]–[5], semantic web service discovery [6]–[8], peer-to-peer information sharing [9], and linked data [10].

Due to its diverse applicability, the ontology alignment has been the topic of many research so that numerous alignment systems are put forward to discover the mappings of two given ontologies [11]–[22]. Also, the OAEI has taken place for more than a decade whose primary objectives are to monitor the progress of the field, bring together various alignment systems, and compare them systematically. Thus, it is

of the essence to have a reliable means of evaluation and comparison.

A. MOTIVATION

The alignment systems are typically evaluated by various performance scores such as precision, recall, and F-measure. The scores count the frequency of the true discovered correspondences in alignment based on which a ratio is computed as a performance measure. However, summarizing the overall performance of an alignment system in one figure would not reflect many facets of its achievements. For instance, assume that System A discovers four correspondences, two of which are correct. System B, on the other hand, identifies 100 correspondences half of which are correct. It is evident that precision of both systems is 0.5, but System B is more likely to have the same precision on other benchmarks since its precision is computed across a more substantial number of correspondences. In other words, the uncertainty in the precision of the second system is much less than that of the first which is not reflected solely by computing precision.

The comparison of two systems on one benchmark is also drawn by the juxtaposition of the obtained performance scores. The decision is made quite simple: if the score (e.g., precision, recall, etc.) of the first system is higher than the second, then it has better performance. The comparison based merely on the scores would provide little information regarding the overall discrepancy among systems. For instance, it does not quantify the extent to which one system is better than the other. Let A, B, and C be three alignments with precision 0.8, 0.79, and 0.5, respectively. Such a comparison would only indicate that A is better than B and C, but does not provide any more information on how significantly A is better than B or C. However, precision of A is approximately the same as B, and it is significantly superior to C. On top of that, one cannot state if two systems are practically identical unless they have exactly the same score, a rare circumstance to happen.

In the case of having more than one benchmark, the comparison is made based on the average of a score: the higher the average value, the better the system. It is identical to the case of having one benchmark since this decision is also made on the basis of two figures. From the statistical point of view, however, averaging is not safe for many reasons. First and foremost, it is sensitive to outliers, thus the fair performance of a system can be deteriorated if its performance is not good enough even on one single benchmark [23]. Another drawback of averaging is the commensurability which makes averaging meaningless if the results on various benchmarks are not comparable [24].

The null hypothesis testing has been already considered to compare various alignment systems on single or multiple benchmarks [23], [25], but they also suffer from various drawbacks. The inference is based on p-values, which is the probability of observing two alignments given the null hypothesis (e.g., the equivalence of two alignment systems) is correct. The decision based on p-values is fallacious since the p-value is not the probability of interest, i.e., the probability of the null hypothesis given the alignments [26], [27].

In addition, the statement of significance using the null hypothesis testing would not necessarily mean that the alignment systems are significantly different in practice [28]. This can be particularly seen when the difference between two given alignment system is quite subtle and imperceptible but the resulting p-value is quite small due to the usage of non-parametric statistics (e.g., Wilcoxon or Friedman test). The null hypothesis will also be rejected if large sample size is provided as well, no matter how different the samples are.

Yet another breakpoint of the null hypothesis testing is that it does not provide any information if the null hypothesis is not rejected [27]. In this case, one cannot claim any statement about the equivalence of two given alignment systems. There is also no principled way to decide the value of significance level α based on which a p-value would be claimed as significant [26].

B. CONTRIBUTIONS

In this study, we first demonstrate that the ontology alignment evaluation is a statistical inference problem. In this regard, the notion of *risk* for an ontology alignment system is introduced, which is broad enough to accommodate any performance score. The alignment risk cannot be computed based merely on its definition; thus, we study two strategies, i.e., maximum likelihood estimation (MLE) and Bayesian estimation, to approximate it. We show that the MLE of risk with respect to a performance score, e.g., precision risk, is equivalent to the complement of the same performance measure, i.e., precision risk $=$ 1 - precision. The byproduct of estimation of the precision risk, as a result, is that the precision is obtained as well. We also prove that the MLE of risk regarding a performance measure in the case that there are multiple benchmarks is tantamount to the complement of its micro-average. These results corroborate that the evaluation of alignment systems is indeed a statistical problem.

We further provide a Bayesian model to estimate the risk. The underlying idea behind the MLE is that there is an unknown parameter which has a *precise* probability value, and the goal is to estimate that value in a way that it maximizes the odds of observing the data (here, the performance of alignments). The notion of having a precise probability is why it provides little information regarding the performance of alignment systems and is thus the source of pitfalls in the current practice of alignment evaluation and comparison.

The Bayesian paradigm, on the other hand, would estimate the unknown parameter using a distribution, which is its crucial difference with the MLE. Approximation of the alignment risk using a distribution not only contains the MLE's precise value as its central tendency (e.g., mean, median, or mode), but it also takes into account the uncertainty that the observed performance might entail. Thus, a Bayesian model is developed to approximate the risk distributions in

the presence of single or multiple datasets. Similar to the MLE, the estimation of the performance distribution could be easily obtained based on the estimation of the risk distribution, i.e., if the precision risk distribution is estimated, then the precision distribution is its complement.

As a result of the Bayesian model, we have a distribution with respect to each performance score instead of having a score for representing the performance of alignment. Such distributions take into account the uncertainty of the alignment performance; hence, the precision of two alignment systems with the same ratio of true positives to true negative would have different distributions if the number of true positive alters. Considering the example in the previous subsection, two alignment systems have the precision of 0.5 while they discover four and a hundred correspondences, respectively. The reason that these two systems are deemed equivalent by the conventional evaluation, which we refer to as MLE in this paper, gets back to the nature of the used statistical strategy, i.e., MLE. The Bayesian estimation, on the other hand, gives a probability distribution so that two systems with the same precision ratio (or any other score than precision) would have totally distinct distributions if its number of correspondences are different.

In addition, a new Bayesian test is devised based on the estimated risk to compare different alignment systems. The test computes the probability that one alignment system is better than one another hinged on their estimated risk. In particular, the probability that System A is superior to System B is the probability that the risk of System A is less than the risk of System B. The probability can be computed as the mathematical expectation of their risk differences. We can further use the region of practical equivalence (rope) [29], and consider that two alignment systems are identical if their risk difference is less than the rope length. The Bayesian test does not suffer from the pitfalls of the decisions based on p-values since it computes the probability of interests for inference. e.g., the probability that two systems have distinct performance given their alignments over single or multiple benchmarks. The Bayesian tests also avoid other pitfalls of the p-values. Another advantage of the proposed Bayesian model is that it can also be used for the evaluation, in contrast to the null hypothesis testing which is used for comparison only.

Careful readers might question the necessity of having a new test since there are multiple Bayesian tests in the literature, in particular in the machine learning literature [26], [28], [30]. In machine learning, the validation is often conducted by resampling techniques such as K-fold cross-validation, which results in having multiple performance metrics for each dataset coming from each fold. Thus, the comparison can be made by using the statistical comparison [30]. In ontology alignment, on the other hand, the comparison must be based on the generated alignment and the reference with no use of resampling techniques. Thus, the tests based on the resampling technique cannot be applied to the alignment comparison.

There is another family of tests which can be used for alignment comparison in the case that there are multiple benchmarks [26]. In this regard, we need to summarize the performance of each dataset by score, and then compute the extent to which one system is superior to one another based on their scores on multiple benchmarks. The proposed test, on the other hand, takes all the correspondence from multiple benchmarks as the input, and calculate the overall performance without summarizing. Therefore, the proposed test can better indicate the difference between given alignments rather than the tests in [26] since its error estimation is less. Another drawback of these tests is that they cannot be applied to make the comparison on one benchmark, while the proposed test can be readily used for comparing systems on one benchmark as well.

Finally, we visualize the outcomes of the Bayesian analysis. For evaluation, precision, recall, and F-measure distributions are displayed. For comparison, the results of the Bayesian test are visualized by a weighted directed graph. The proposed statistical analysis of alignment systems are applied to the anatomy and conference tracks of the OAEI, and the participating systems are evaluated and compared accordingly.

In a nutshell, the contributions of this research can be summarized as follows:

- The ontology alignment evaluation is formulated as a statistical inference problem by introducing the notion of *risk*, and two widely-used schools for inference are explained and compared. In particular, the current practice of ontology evaluation is proved to be precisely identical to the MLE.
- A Bayesian model is especially-tailored to estimate the risk, which yields a distribution of risk with respect to a particular performance score. Accordingly, a distribution for the same score is also obtained.
- A Bayesian test is developed based on the risk distribution to compare two different alignment systems. The test computes the extent to which one system is superior to one another, and avoids the pitfalls of the decisions based on p-values. Also, the probability of two systems being equivalent can be calculated, thanks to the Bayesian notion of *rope*.
- The overall analysis of the proposed methodology is visualized using the performance distributions and a weight directed graph, displaying the overall comparison of multiple alignment systems.

The main focus of this article is on the evaluation and comparison of ontology alignment systems, but the methodology proposed in this article can be directly applied to other domains of information retrieval where precision, recall, and F-measure are extensively used.

C. ORGANIZATION

The remainder of this article is structured as follows. The basic concepts regarding ontology alignment and its evaluation are discussed in Section 2. Section 3 contains the

introduction of the ontology alignment risk and discusses the MLE and Bayesian estimations in details. We present the Bayesian hierarchical model in Section 4 and devise a Bayesian test in Section 5 according to the estimated risks. The experiments regarding the Bayesian model is presented in Section 6, and the article is concluded in Section 7.

II. PRELIMINARIES

In this section, the basic definitions required for the remainder of the article are presented.

Ontologies are the tools to describe a domain formally by its objects and the relations therein. In this research, an ontology is regarded as a set of classes, data and object properties, and instances (or individuals) for a particular domain of interest. The set of classes, properties, and individuals are also referred as entities of the given ontology.

A typical ontology matching system takes two ontologies (usually referred as the source and the target) as the input and tries to find the similar entities of the source to those of the target [2]. To find the identical entities, one might use several similarity metrics, e.g., string, linguistic, structural similarity measures [2], [31], [32].

A correspondence is the mapping of one entity from the source ontology to one in the target. A correspondence might contain some extra information about the mapping such as the type and the confidence of mapping. A simple correspondence would comprise of $< e, e', r >$ where e and e' are, respectively, two entities from source and target ontologies, and r denotes the relationship of entities (e.g., equivalence, subsumption, etc.). Correspondences typically have a degree of confidence which indicates their reliability. In this research, we solely take into account the correspondences and not their confidence.

For the given source and target ontologies, the alignment is the set of correspondences between the pairs of their entities. Based on this definition, the alignment is the typical outcome of the ontology alignment systems.

After the discovery of alignments, the performance of systems is typically measured by computing several performance scores, which require a reference alignment containing the ground truth of the mappings between given ontologies.

Given an alignment A and a reference R, precision is defined as the ratio of the true positives to the total number of discovered correspondences, e.g.,

$$Pr(A, R) = \frac{|A \cap R|}{|A|} \tag{1}$$

where $Pr(.,.)$ denotes the precision, and $|.|$ is the cardinality operator. Since the false negative would not influence it, the precision is called the measure of correctness. As the complement to the precision, recall is defined as the ratio of the true positive to the total number of correspondences in the reference, e.g.,

$$Re(A, R) = \frac{|A \cap R|}{|R|} \tag{2}$$

where $Re(.,.)$ denotes the recall. In contrast to the precision, the false positive would not have any impact on the recall. That is why it is called the measure of completeness.

One might be interested in the trade-off of the precision and recall. In this case, the F-measure could be utilized and is defined as

$$\begin{aligned} F(A, R) &= \frac{2Pr(A, R) \times Re(A, R)}{Pr(A, R) + Re(A, R)} \\ &= \frac{2|A \cap R|}{|R| + |A|} \end{aligned} \tag{3}$$

There are some other performance scores such as relaxed precision and recall [33], semantic precision and recall [34], and weighted precision and recall [2]. The current way of the alignment evaluation is to compute the above scores for single or multiple mapping tasks. However, such an approach would not reflect the overall performance of the alignment systems. On top of that, the comparison based on such scores provides little information regarding the difference between the alignments. In further sections, we evaluate and compare systems based on a novel Bayesian model which provides much more information regarding the performance of alignment systems.

III. RISK OF ONTOLOGY ALIGNMENT SYSTEMS

The section begins by presenting the formal definition of the alignment risk. We then discuss the potential MLE and Bayesian estimation along with their advantages/pitfalls.

The risk is related to the error of a system, which can be seen as a complement to a performance measure. For instance, *silence* is the complement to recall; thus, the recall risk is indeed equivalent to silence. The following definition concisely presents the core definition of the alignment risk.

Definition 1 (Alignment Risk): The risk of an ontology alignment system is the probability that the system makes an error.

Definition 1 is broad enough to accommodate different performance measures since "error" can have distinct interpretations in different circumstances. We consider the error of a given alignment with respect to a performance measure. For instance, if precision is the sought score, then the precision risk is the probability of having a false positive. If the comparison is based on recall, then the recall risk is the probability of having a false negative. F-measure would be a little intricate, but the F-measure risk could be defined as the probability of having a false positive or a false negative, thanks to the equation (3).

The risk of an ontology alignment system is not an observed variable, but it is a parameter to be estimated based on the outcomes of a system over single or multiple benchmarks. The estimation of such a parameter would seem formidable at the beginning, but the following critical yet straightforward observation would pave the way for doing so.

For a moment, we focus on the estimation of the precision risk. Assume that we know the precision risk τ_{Pr} of an alignment system, hence the probability that one correspondence in a given alignment A is false would be τ_{Pr}. Besides,

the probability of a correspondence being true is $1-\tau_{Pr}$. As a result, it is a Bernoulli trial with the failure probability τ_{Pr} and success probability $1-\tau$. Thus, the probability of having K false positives out of N correspondences in the alignment would follow the binomial distribution.

Definition 2: Given the alignment A with the risk τ, the probability of observing K errors out of N trials would follow a binomial distribution, e.g.,

$$Pr(K, N; \tau) = \begin{pmatrix} \mathrm{N} \\ \mathrm{K} \end{pmatrix} \tau^K (1-\tau)^{N-K}.$$

The number of errors would vary from one performance score to one another. For precision, the number of trials is the number of correspondences in the alignment, i.e., $\mathrm{N} = |A|$, and the number of errors is the false positives. For recall, on the other hand, the number of trials is the number of correspondences in the reference, e.g., $N = |R|$, and the number of errors is the false negatives.

For F-measure, equation (3) follows

$$\begin{aligned} Risk_F(A, R) &= 1 - \text{F-measure(A,R)} \\ &= 1 - \frac{2|A \cap R|}{|A| + |R|} \\ &= \frac{|A - R| + |R - A|}{|R| + |A|}. \end{aligned} \tag{4}$$

According to this equation, the number of trials for F-measure is the sum of correspondences in A and R, e.g., $\mathrm{N} = |A| + |R|$, and the number of errors is the sum of false positives and false negatives.

Having known the number of trials and errors, one can estimate the risk of an alignment based on Definition 2. A straightforward way of estimating the risk is to use the maximum likelihood estimation (MLE). The risk estimation using the MLE would be the fraction K/N, e.g.,

$$\tau = \frac{K}{N} \qquad 1 - \tau = \frac{N - K}{N} \tag{5}$$

For the precision risk, for instance, $N - k$ is the number of true correspondences in the alignment and N is the total number of correspondences. Thus, $1 - \tau$ is exactly the precision score. A Similar argument follows for recall and F-measure. The MLE also reveals the fact that any estimation would bear both precision and the precision risk if the precision is the desired criteria. Thus, one can simply consider $1 - \tau$ to compute directly the desired performance score, and not its risk.

The MLE in equation (5) is merely for one single benchmark. The micro-average of precision and recall on q benchmarks are defined as

$$\hat{Pr} = \frac{\sum_{i=1}^{q} |TP_i|}{\sum_{i=1}^{q} |A_i|} \qquad \hat{Re} = \frac{\sum_{i=1}^{q} |TP_i|}{\sum_{i=1}^{q} |R_i|} \tag{6}$$

where TP_i, A_i, and R_i are the true positives, identified alignment, and the reference of the i^{th} benchmark, respectively, and $\hat{Pr}$ and $\hat{Re}$ are the micro-average precision and recall. The following theorem proves that the MLE of risk for a specific score on multiple benchmarks is equivalent to the complement of the micro-average of the same score.

Theorem 3: Let S be the alignment system operated on q benchmarks, and the alignments $A_{1:q}$ are identified. The MLE of $1 - \tau$ with respect to various scores is tantamount to micro-averaging of the same score over q benchmark.

Proof: Let $R_{1:q}$ be the reference alignments with respect to q benchmarks and assume that the system S has independently discovered the alignments $A_{1:q}$. The MLE entails

$$arg\max_{\tau} log[p(\tau; A_{1:q}, R_{1:q})]$$

where log is the logarithm function, and $p(\tau; A_{1:q}, R_{1:q})$ is the likelihood of τ based on q benchmarks, and is defined as

$$p(\tau; A_{1:q}, R_{1:q}) = \prod_{i=1}^{q} \begin{pmatrix} N_i \\ K_i \end{pmatrix} \tau^{K_i} (1-\tau)^{N_i - K_i}$$

where K_i and N_i are the numbers of errors and trails for the i^{th} alignment, respectively. It follows

$$\begin{aligned} &arg\max_{\tau} logp(\tau; A_{1:q}, R_{1:q}) \\ &= arg\max_{\tau} \quad \sum_{i=1}^{q} K_i log(\tau) + (N_i - K_i) log(1-\tau) \\ &= arg\max_{\tau} \quad log(\tau) \left(\sum_{i=1}^{q} K_i \right) + log(1-\tau) \left(\sum_{i=1}^{q} N_i - K_i \right). \end{aligned}$$

The point τ^* is the optimal value of the above minimization if and only if its derivation with respect to τ is zero. Thus,

$$\begin{aligned} \frac{\partial}{\partial \tau} log(p) = 0 \Rightarrow \frac{\sum_{i=1}^{q} K_i}{\tau} - \frac{\sum_{i=1}^{q} N_i - K_i}{1-\tau} &= 0 \\ \Rightarrow \left(\sum_{i=1}^{q} K_i \right) \left(\frac{1}{\tau} + \frac{1}{1-\tau} \right) &= \left(\sum_{i=1}^{q} N_i \right) \frac{1}{1-\tau} \\ \Rightarrow \left(\sum_{i=1}^{q} K_i \right) \left(\frac{1}{\tau(1-\tau)} \right) &= \left(\sum_{i=1}^{q} N_i \right) \frac{1}{1-\tau} \\ \Rightarrow \tau = \frac{\sum_{i=1}^{q} K_i}{\sum_{i=1}^{q} N_i} \quad &\text{and } 1 - \tau = \frac{\sum_{i=1}^{q} N_i - K_i}{\sum_{i=1}^{q} N_i}. \end{aligned}$$

For precision, $N_i - K_i = |TP_i|$ and $N_i = |A_i|$, and for recall $N_i - K_i = |TP_i|$ and $N_i = |R_i|$. Similarly, the MLE of F-measure will follow. Thus, the MLE of $1 - \tau$ with respect to a particular score over multiple benchmarks is precisely identical to the micro-average of the same score, and the proof is complete. ■

So far, it is shown that the evaluation of alignment systems is a statistical inference problem, and the current evaluation using various performance scores could indeed obtain by the MLE, thanks to the notion of risk. It is further discussed that the pitfalls regarding the current evaluation approach are coming from the nature of the MLE.

In the MLE, the parameters of interest are assumed to be fixed but unknown, and the optimization procedure would find the optimal values as the precise point estimate. Thus, the evaluation and comparison using the MLE boil down to one figure for the former and the juxtaposition of two

figures for the latter. The Bayesian estimation, on the other hand, the parameters are not assumed to be fixed but rather a random variable. Thus, the outcome of the Bayesian analysis would result in the distribution estimation of parameters instead of a sharp point estimate. Having such distribution would enable us to take into account the uncertainty regarding alignment performance and to compare various systems more meaningfully by inferring over the risk posterior distribution.

One Bayesian approach for the risk estimation is to use the beta-binomial conjugate. In this conjugate, the beta prior with parameters a and b $beta(a, b)$ is considered, and the posterior for a given alignment with K errors out of N trials is computed as

$$p(\tau|N, K) = beta(a + K, b + N - K). \tag{7}$$

The mode of the posterior distribution is

$$Mode = \frac{a + K - 1}{a + b + N - 2}.$$

If the uninformative prior $beta(1, 1)$ is selected, then the mode of beta-binomial would be equivalent to the MLE estimate, e.g., Mode = K/N. However, the variance of the beta distribution would be different for larger values of N and K. Such uncertainty is not reflected if the MLE is utilized. This simple example shows that the Bayesian estimation not only contains the MLE estimate as the central tendency, but also provides more information regarding the uncertainty of alignment performance.

The simple beta-binomial distribution would suffice if there were only one benchmark for evaluation. However, the generalization to multiple benchmarks cannot be performed merely using this model. In the next section, we develop a Bayesian hierarchical model to estimate the risk based on the outcome of an alignment system across multiple benchmarks.

IV. RISK APPROXIMATION: A BAYESIAN HIERARCHICAL MODEL

The risk of a system is a latent variable which must be approximated using a methodology. The MLE and a simple Bayesian model are discussed in the previous section, and their drawbacks are explicated. In this section, we develop a Bayesian hierarchical model to estimate the risk of an alignment system for q benchmarks. Further, the model would estimate the final risk of an alignment system based on its risk over multiple benchmarks.

Assume that the ontology alignment system has performed over q benchmarks, and we obtain N_i and K_i for each $i = 1, \ldots, q$. We show the set of all N_i and K_i obtained from q benchmarks as $N^{1:q}$ and $K^{1:q}$, respectively. The objective is to estimate the risk of a system over every benchmark, e.g., $\hat{\tau}_i$, and the overall risk τ^*. Therefore, the Bayes rule follows

$$\begin{aligned} P(\tau^{1:q}, \tau^*|N^{1:q}, K^{1:q}) &\propto P(N^{1:q}, K^{1:q}|\tau^{1:q}, \tau^*)P(\tau^{1:q}, \tau^*) \\ &= P(\tau^*)\prod_{i=1}^{q} P(N_i, K_i|\tau_i)P(\tau_i|\tau^*) \end{aligned} \tag{8}$$

where the last equality obtained since the results of different benchmarks are independent of each other. The graphical model associated with equation (8) is depicted in Figure 1. The rectangular shape denotes the observed variables, and the circles depict the random variables which need to be approximated. As a convention, the variables τ_i, N_i and K_i are contained in another rectangular which means that the same model is repeated for different benchmarks.

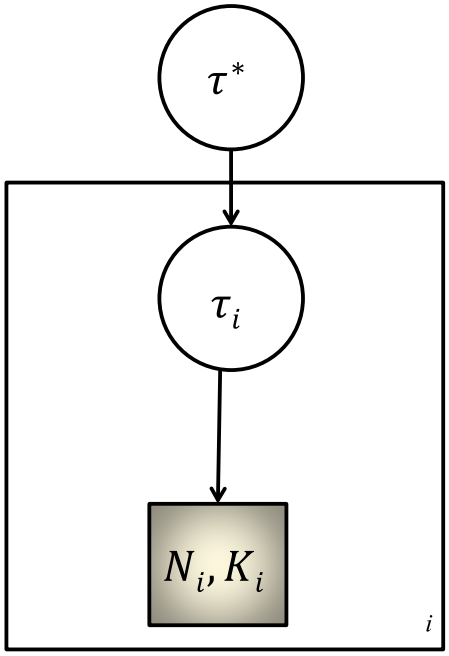

FIGURE 1. The graphical representation of the Bayesian model for estimating the risk.

We now need to specify the distribution of all elements in equation 8. So far, the number of errors has been modeled as the binomial distribution, e.g.,

$$K_i \sim binomial(\tau_i, N_i). \tag{9}$$

The parameter τ_i is unknown and must be estimated, hence we need to model it as another distribution. The τ_i distribution could be

$$\tau_i \sim beta(a, b) \tag{10}$$

where $beta(.,.)$ is the beta distribution, and a and b are its corresponding shape parameters. To make the model more meaningful, we reformulate the beta distribution with two other parameters. Let τ^* be the mean of the beta distribution and the concentration parameter be $\gamma = a + b$, we have

$$\begin{aligned} \tau^* &= \frac{a}{a+b} \quad \& \quad \gamma = a + b \\ &\Rightarrow a = \tau^*\gamma \quad \& \quad b = (\gamma - 1)\tau^* \\ &\Rightarrow \tau_i \sim beta\big(\tau^*\gamma, \tau^*(\gamma - 1)\big) \end{aligned} \tag{11}$$

The equation (11) means that the risk τ_i follows a beta distribution whose mean is τ^*. Thus, the values of τ_i are at

the neighborhood of τ^*, and their proximity is controlled by the parameter γ.

The parameters γ and τ^* are also unknown, hence we again model them as a distribution. According to equation (11), the values of γ must be greater than one since the concentration parameter cannot be negative. There are many distributions for non-negative variables, and we use here the gamma distribution for γ, e.g.,

$$\gamma - 1 \sim gamma(\alpha, \beta) \tag{12}$$

where *gamma* is the gamma distribution, and α_i and β_i are its shape and rate parameters, respectively.

The τ^* is yet another parameter to be estimated. Thus, we model it as a beta distribution as well

$$\tau^* \sim beta(a^*, b^*) \tag{13}$$

The final step is to identify the remaining parameters. For the gamma distributions, we need to specify $\alpha_i, \beta_i, \alpha^*$, and β^*. The parameters can be stated in a way to be completely uninformative. However, we let the data learn the parameters. Thus, we model them as the uniform distribution

$$\alpha, \beta \sim uniform(l, u)$$

where *uniform*(., .) is the uniform distribution with the parameters l and u. We particular set $l = 0$ and $u = 1000$ to cover a broad spectrum of values.

Finally, the parameters a^* and b^* must be specified. We assign $a^* = b^* = 0.1$ since it is an uninformative prior distribution.

The specified model should be solved using Markov-chain Monte Carlo (MCMC) techniques [35]. The model was written in JAGS [36], and the required sampling was performed accordingly.

The Bayesian model has been intuitively developed based on the assumption that the risk of the alignment system on one benchmark is in the neighborhood of the overall alignment risk. We further validate the model and obtained distributions with the scores obtained by traditional way, e.g., the MLE. The experimental investigation supports the reasonable outcome of the approximated distributions since the distributions are centered around the MLE in all cases.

V. COMPARISON OF ALIGNMENT SYSTEMS: BAYESIAN TEST

Having the risk distributions of two alignment systems, it is also possible to compare their performance using a Bayesian test. The risk distributions allow us making the comparison more meaningfully since we can compute the probability (or confidence) that one system is better than one another. Thus, the comparison is not drawn based solely on the juxtaposition of two scores. Further, we can define the region of practical equivalence (rope) to identify the systems with identical performance.

There is no principled way to determine the length of the rope, shown by r, and it is an expert decision to asses. The idea of the rope is quite simple: if the difference between posterior risk distributions of two alignment systems is less than r, then the understudy systems are practically equivalent. Based on this notion, one can compute the probability that two systems are practically equal. If one is interested in determining the better systems even with a subtle difference, then $r = 0$ and the outcome of the test would indicate the superiority of one system over one another.

The probability of the alignment A_1 with the risk $\hat{\tau}_1^*$ is better than the alignment A_2 with the risk $\hat{\tau}_2^*$ can be computed as

$$\begin{aligned} P(A_1 > A_2) &= P(\hat{\tau}_1^* < \hat{\tau}_2^*) \\ &= \iint \mathcal{I}_{\hat{\tau}_1^* - \hat{\tau}_2^* > r} P(\hat{\tau}_1^* | data) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^* \end{aligned} \tag{14}$$

where $P(\hat{\tau}_i^* | data)$ is the posterior risk distribution of the i^{th} system, and $\mathcal{I}$ returns one if the condition specified in its subscript is satisfied, and zero otherwise.

Similarly, one can compute the probability that A_1 is better than A_2 and the probability that they are equivalent as

$$\begin{aligned} P(A_1 < A_2) &= \int\int \mathcal{I}_{\hat{\tau}_2^* - \hat{\tau}_1^* > r} P(\tau | \hat{data}_1^*) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^* \\ P(A_1 = A_2) &= \int\int \mathcal{I}_{|\hat{\tau}_1^* - \hat{\tau}_2^*| < r} P(\tau | \hat{data}_1^*) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^* \end{aligned} \tag{15}$$

The above-mentioned probabilities could also be obtained from the MCMC samples. As an instance, the equation (14) is estimated by t samples of the MCMC chains as follows

$$P(A_1 > A_2) = \frac{1}{t} \sum_{i=1}^{t} \mathcal{I}_{\hat{\tau}_2^{*i} - \hat{\tau}_1^{*i} > r} \tag{16}$$

where $\hat{\tau}_j^{*i}$ is the i^{th} sample of $\hat{\tau}_j^*$ drawn by the MCMC, and $j = 1, 2$. Other probabilities are also computed in a similar way.

VI. EXPERIMENTAL RESULTS

This section dedicates to the experiments regarding the proposed Bayesian hierarchical model. We consider the results of conference and anatomy tracks from the OAEI 2017 to display the applicability of the Bayesian model.

The experiments on each track are twofold. The first one is the evaluation in which we display the distribution of precision, recall, and F-measure, and show that the obtained distributions are meaningful since they are centered around the MLE. The second part is the comparison where we conduct the proposed Bayesian test and visualize the overall outcome by a weighted directed graph.

The results of the OAEI 2017 are publicly available and can be downloaded from the OAEI web site.[1] Then, we use the Alignment API [37] to find the numbers required for the hierarchical model. In particular, the numbers K and N for

[1] http://oaei.ontologymatching.org/2017/results/index.html

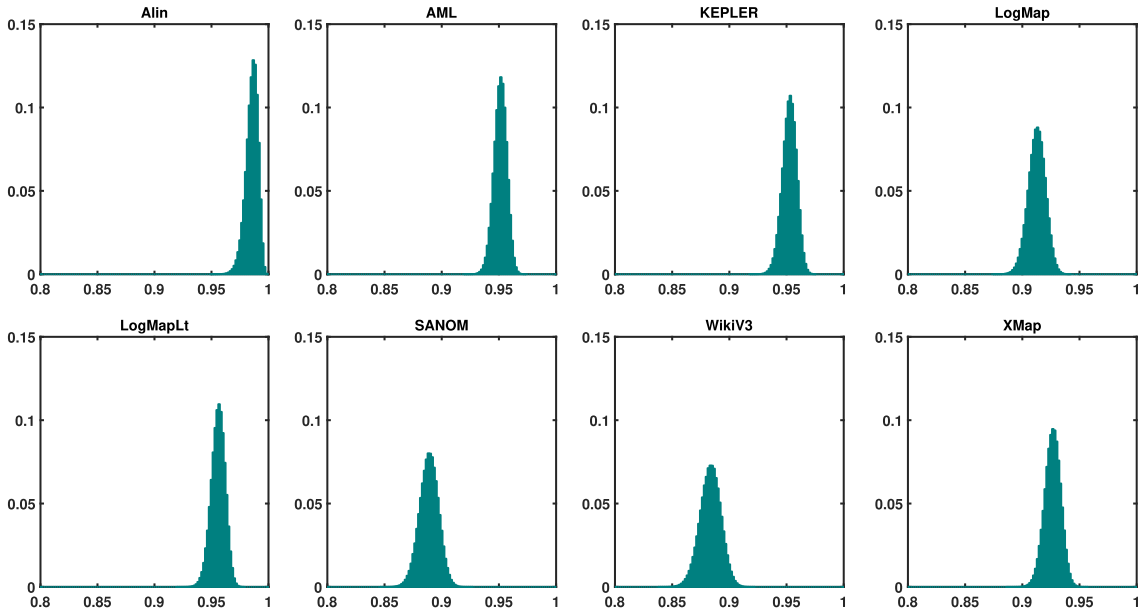

FIGURE 2. The estimation of the precision performance distribution $1-\tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

precision risk of the alignment A could be obtained by the Alignment API as

$$K_{PR} = nbFound - nbCorrect \quad N_{PR} = nbFound \tag{17}$$

where $nbFound = |A|$, $nbCorrect = |A \cap R|$, and the subscript PR represents the precision risk. The functions *nbCorrect* and *nbFound* in equation (17) are provided by functions with similar names in the Alignment API. Similarly, these numbers could be obtained for recall and F-measure as follows

$$\begin{aligned} K_{RR} &= nbExpected - nbCorrect \\ N_{RR} &= nbFound \\ K_{FR} &= nbExpected + nbFound - 2 \times nbCorrect \\ N_{FR} &= nbExpected + nbFound \end{aligned}$$

where $nbExpected = |R|$, and subscripts RR and FR represent the recall risk and F-measure risk.

We considered the results of two tracks of the OAEI and compared the participating systems together. The systems which were evaluated are Alin [16], AML [13], KEPLER [38], LogMap and LogMapLite [12], SANOM [14], [21], WikiV3 [39], and XMap [15].

A. ANATOMY TRACK

The anatomy track is about the matching of Adult Mouse anatomy to a part of the NCI Thesaurus. The first ontology has around 2,400 classes while the second contains approximately 3,400 classes. The principal task in this track is merely the alignment of classes.

The alignments discovered by systems were downloaded to which the Bayesian hierarchical model was applied. On account of the clarity of results, the distribution of $1-\tau$ was considered since it could be directly related to the performance scores themselves. We refer to this distribution as the *performance distribution* of alignments as opposed to the risk distribution.

TABLE 1. The precision, recall, and F-measure of various systems on the OAEI anatomy track. The maximum likelihood estimation (MLE) is equivalent to that of the traditional way of reporting scores, and the other one is the mean of the distribution obtained by the proposed Bayesian hierarchical model (BHM).

System	Precision		F-measure		Recall	
	MLE	BHM	MLE	BHM	MLE	BHM
AML	0.95	0.95	0.943	0.943	0.936	0.936
XMap	0.926	0.925	0.893	0.893	0.863	0.862
KEPLER	0.958	0.951	0.836	0.833	0.741	0.741
LogMap	0.918	0.911	0.88	0.877	0.846	0.846
LogMapLite	0.962	0.954	0.829	0.826	0.728	0.728
SANOM	0.888	0.888	0.870	0.870	0.853	0.852
WikiV2	0.883	0.882	0.802	0.801	0.734	0.734
Alin	0.996	0.984	0.506	0.504	0.339	0.339

We first compare the outcomes obtained from the Bayesian model to those of the traditional way of reporting results. To this end, the means of the performance distributions were compared with performance scores. The traditional performance scores were referred as the MLE since we showed that they are the maximum likelihood of the risk. Table 1 tabulates the MLE and the mean of Bayesian hierarchical model (BHM) estimation for each of the three performance scores. This table proves that the MLE and the mean of the BHM estimation are very close to each other. Thus, the proposed model would yield all information provided by the traditional way of the evaluation.

The difference of two approaches, however, is that the BHM estimation would suggest more insights about the performance of the system under study. In particular, we plot the *performance* distributions for each of the scores. Figures 2-4 display the performance distributions of precision, recall, and F-measure, respectively. It is readily seen that

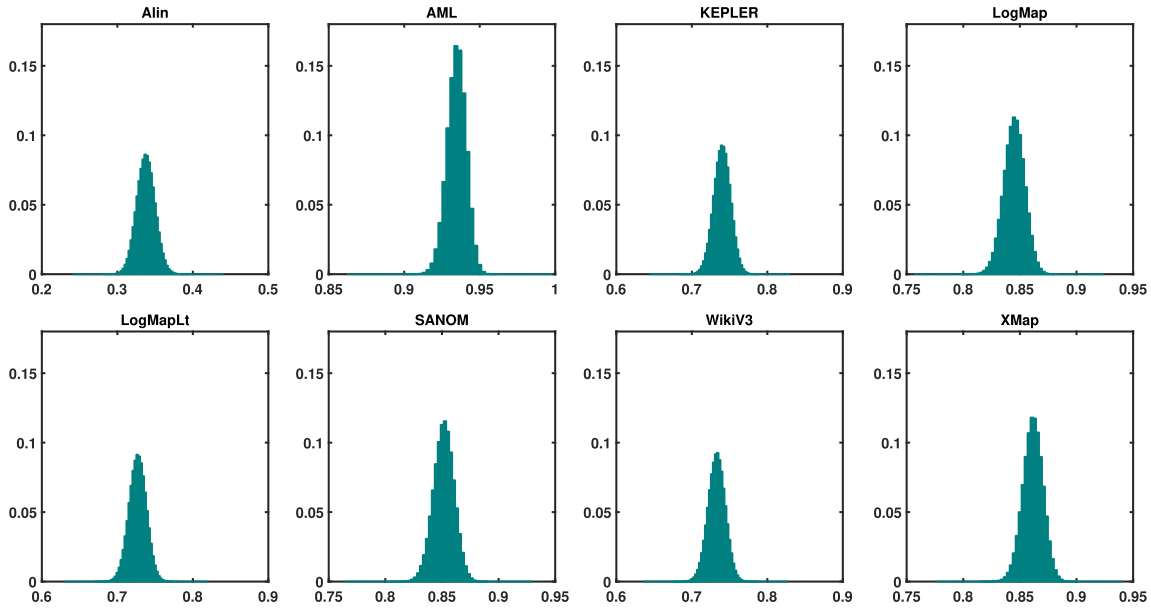

FIGURE 3. The estimation of the recall performance distribution $1 - \tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

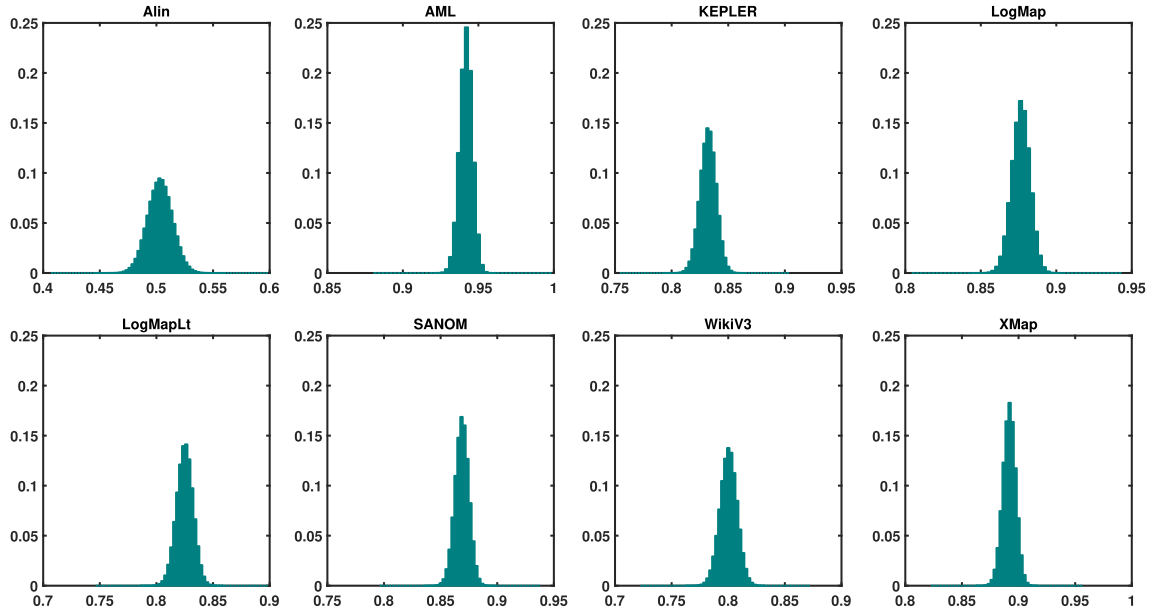

FIGURE 4. The estimation of the F-measure performance distribution $1 - \tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

the peaks of the distributions are over the corresponding MLE with some variations.

We further compare the systems over the anatomy track using the Bayesian test introduced in Section V. For each pair of systems, the probability of one of them being superior to one another is computed with the size of rope equals to zero. Thus, the equivalence of two systems is not considered in this experiment.

The comparison is drawn from three points of view, each of which related to precision, recall, and F-measure. Figures 5-7 are the weighted directed graphs demonstrating the outcomes of comparison. The nodes in these graphs are the systems under comparison, and each edge $A \xrightarrow{w} B$ means that A is superior to B with the probability w.

It is understandable from Figure 5 that Alin is the best performing system in terms of precision, followed by LogMapLite and KEPLER. At the other extreme, SANOM and WikiV3 are the ones with poor performance concerning precision. Regarding recall, however, AML, XMap, and SANOM are the systems with superior performance, thanks

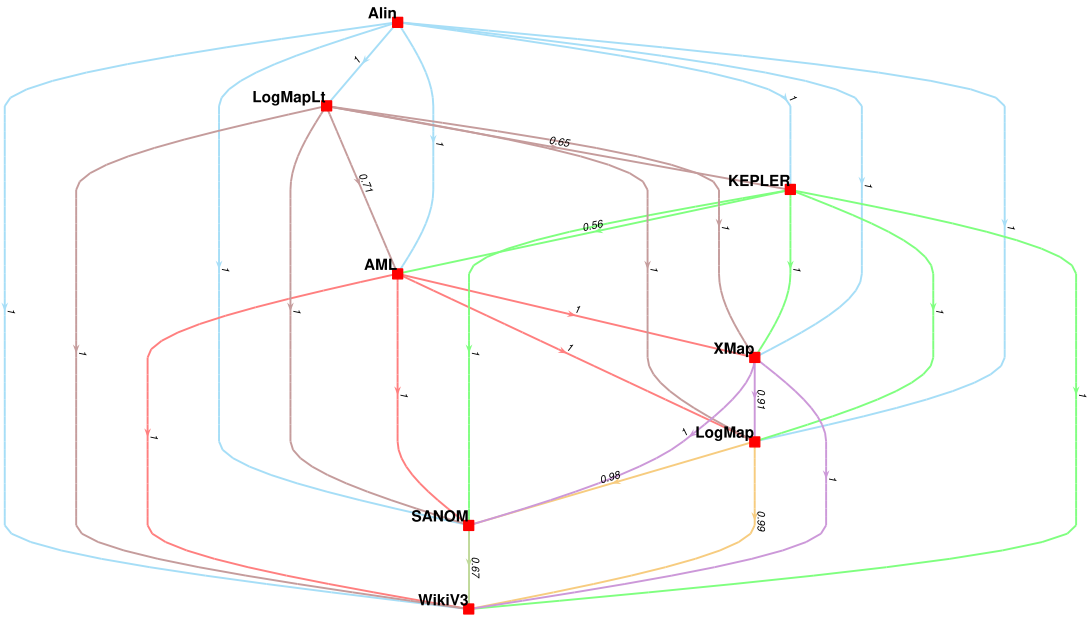

FIGURE 5. Comparison of eight alignment systems with respect to their precision on the OAEI anatomy track using the proposed Bayesian test.

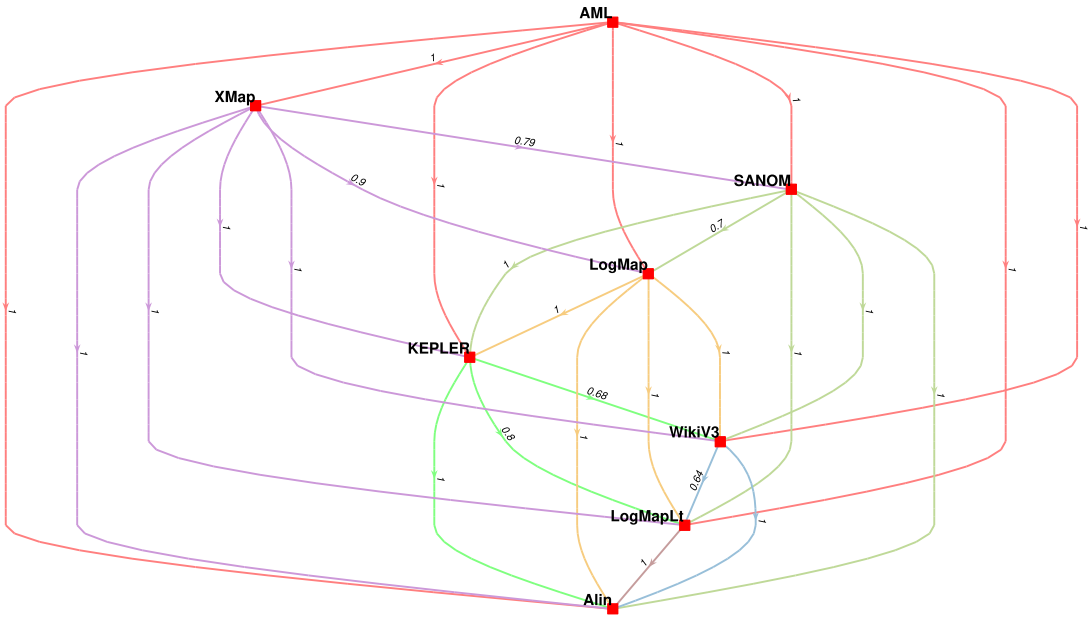

FIGURE 6. Comparison of eight alignment systems with respect to their recall on the OAEI anatomy track using the proposed Bayesian test.

to Figure 6. In contrast to precision, Alin has poor performance with respect to recall.

As a combination of both precision and recall, one can compare the systems in terms of F-measure using Figure 7. From this figure, One can realize that the overall performance of AML and XMap are superior, followed by LogMap and SANOM.

B. CONFERENCE TRACK

The conference track consists of sixteen ontologies form the conference organizations. Since the domain of all ontologies is identical, it seems to be an excellent benchmark to verify the performance of alignment systems. For the OAEI, there are usually 21 mapping tasks from matching seven ontologies altogether.

The evaluation and comparison of the OAEI conference track are different from the anatomy track since there are multiple benchmarks to conduct the comparison. This would help show the performance of the proposed hierarchical model with respect to the traditional way of the evaluation and comparison.

Table 2 displays the evaluation of eight systems on the OAEI conference track. The scores with the subscript *MLE* are the averages of performance scores over all benchmarks,

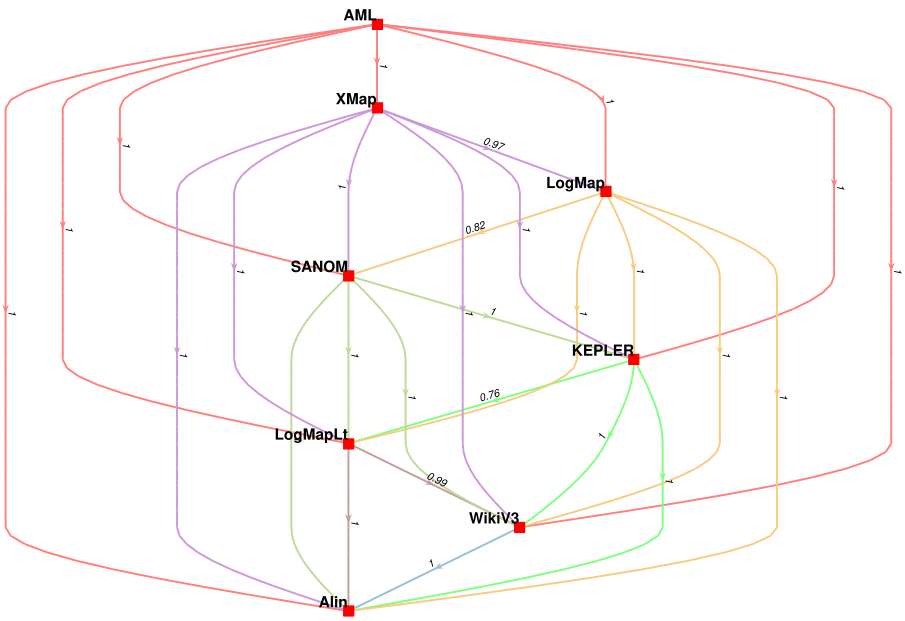

FIGURE 7. Comparison of eight alignment systems with respect to their F-measure on the OAEI anatomy track using the proposed Bayesian test.

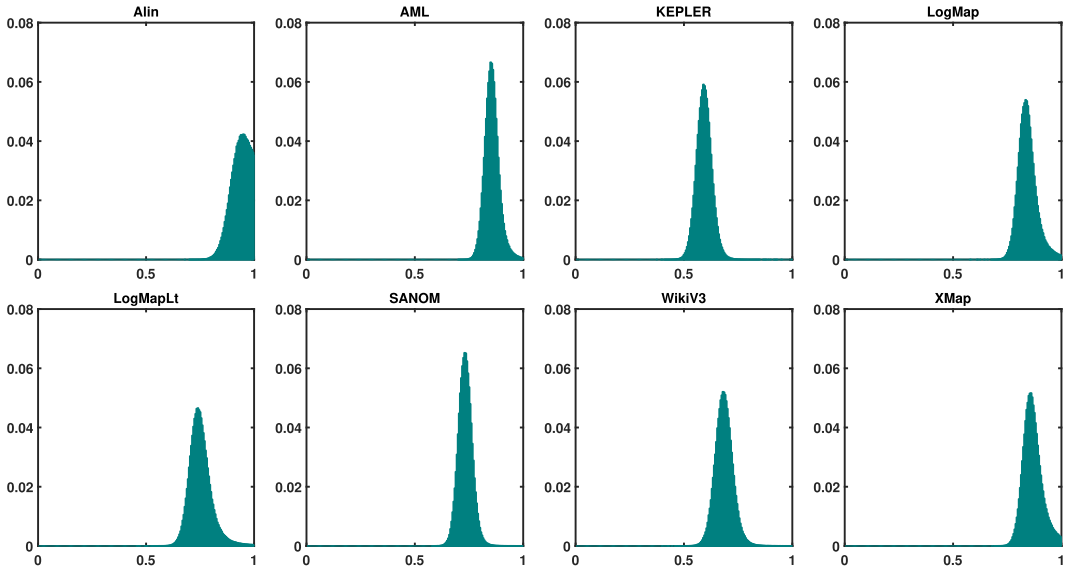

FIGURE 8. The estimation of the precision distribution $1 - \tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

which is the traditional way of evaluating the alignment systems. We also place the standard deviations (SD) of each score over multiple benchmarks which will yield benefits for the interpretation of the estimated distributions by the proposed model. Besides, the averages of the estimated distributions are also shown for the interest of comparison. The acronyms $\hat{Pr}$, $\hat{F}$, and $\hat{Re}$ represent precision, F-measure, and recall, respectively, and their subscripts indicate if they either the MLE or the Bayesian estimation (BHM).

It is readily seen from Table 2 that the means of the estimated distributions are mostly close to the average performance. However, there are some discrepancies as well. For instance, the average F-measure of AML is 0.74 while the mean of its F-measure distribution is around 0.764. We further compute the median, another measure of central tendency, which is known to be more robust in dealing with outliers. Interestingly, the median of F-measures for AML is around 0.762, which is entirely close to what is estimated by the proposed model. The same argument holds for the AML precision estimation, and for other systems with other performance scores, i.e., Alin recall, KEPLER precision, SANOM recall. This experiment supports the validity of the proposed Bayesian model since the average of the estimated distributions is at the proximity of the MLE. The experiment

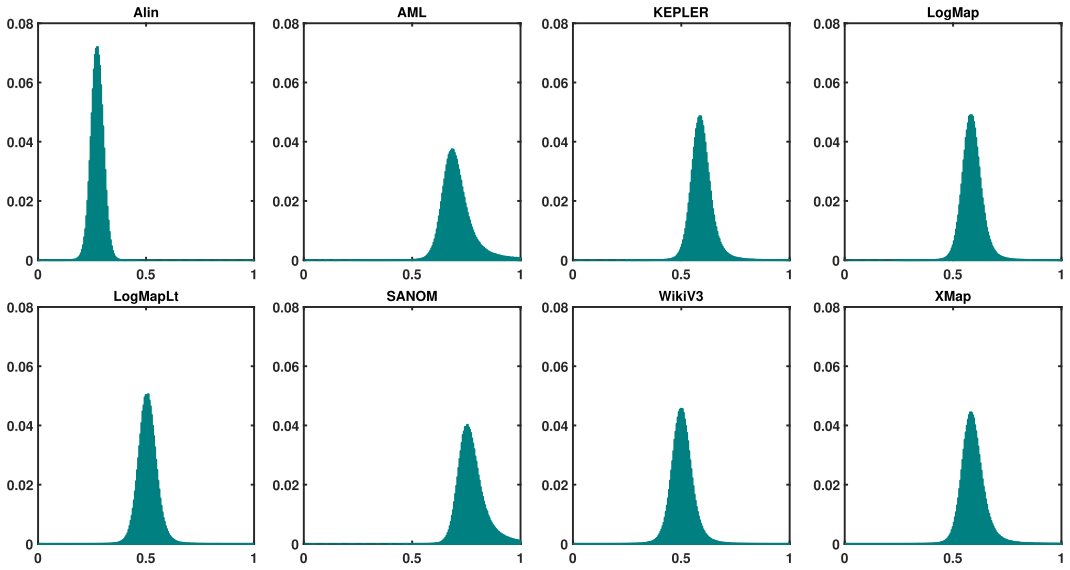

FIGURE 9. The estimation of the recall distribution $1-\tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

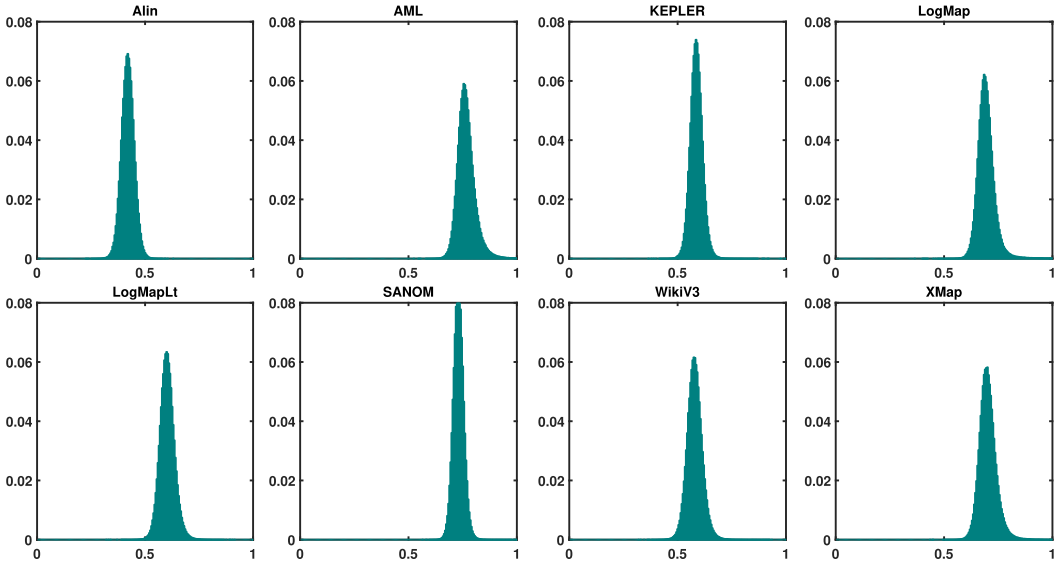

FIGURE 10. The estimation of the F-measure distribution $1-\tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

also confirms the sensitivity of averaging to outliers, which is one of its most important drawbacks, and corroborates the appropriateness of the proposed Bayesian model.

We further plot the estimated distributions by the proposed Bayesian model on the OAEI conference track. Figures 8-10 display the estimated performance distributions of precision, recall, and F-measure, respectively. Table 2 confirms that the central tendencies of distributions are in the proximity of the mean or median of the performance scores. The standard deviations of these distributions are proportionate to the standard deviations of the scores, and to the number of false positives and false negatives. As an instance, the standard deviation of AML precision is less than that of Alin, thanks to Table 2. Similarly, the standard deviation of the AML precision distribution is evidently less than that of Alin, according to Figure 8. As a result, if the performance scores had little variation over various benchmarks, then the resulting estimated distribution would have a lower standard deviation.

As another example, consider the precision performance distribution of Alin and WikiV3 whose scores' standard deviations are approximately identical (see Table 2). However, the performance distribution of WikiV3 is more focused than that of Alin. The reason is that WikiV3 has discovered 222 correspondences overall, of which 149 is correct. Alin, on the other hand, has identified 93 correspondences over

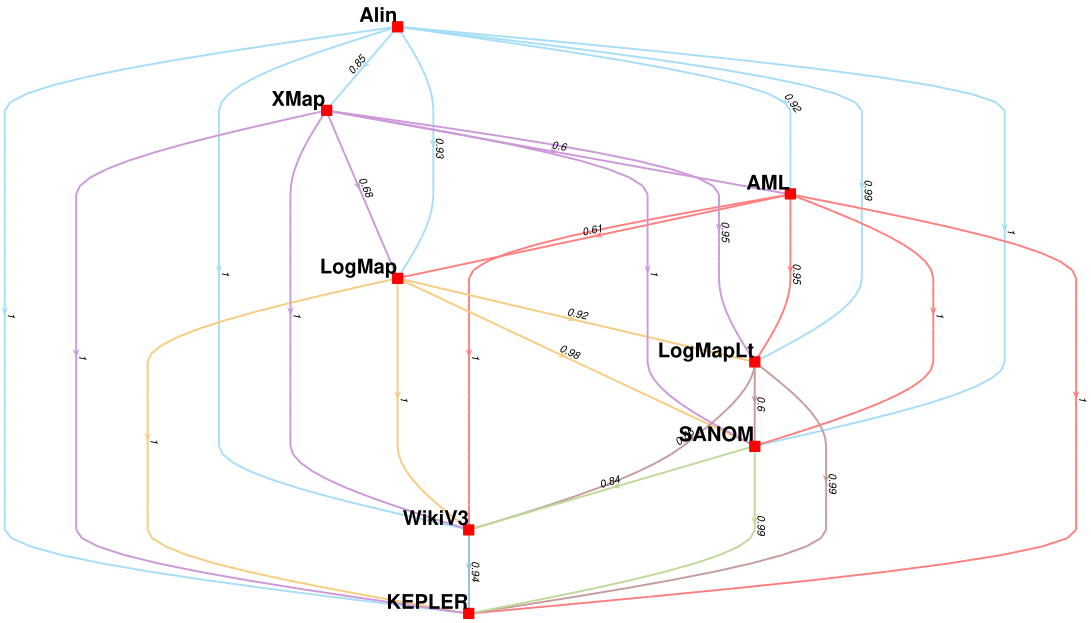

FIGURE 11. Comparison of alignment systems with respect to their precision on the OAEI conference track using the proposed Bayesian test.

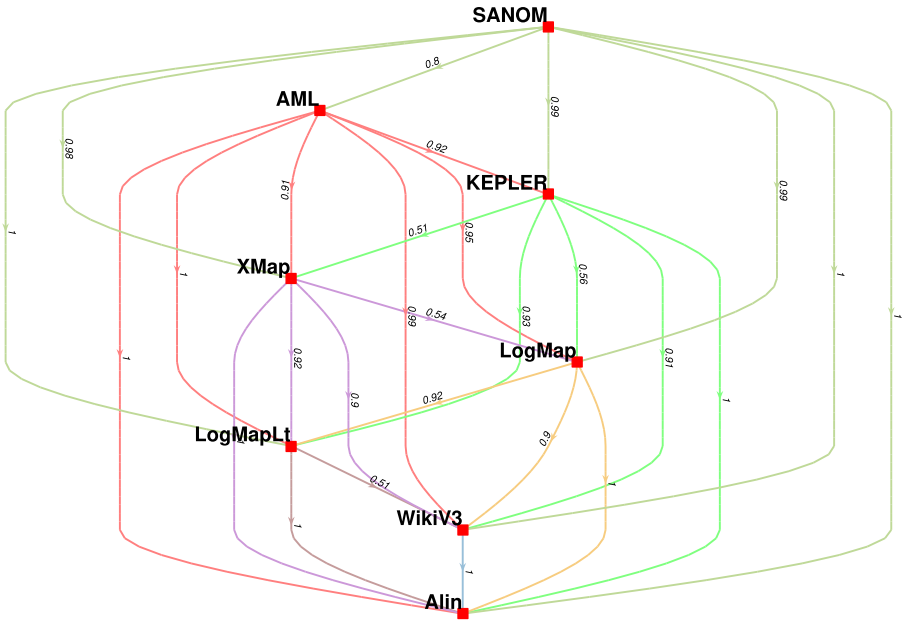

FIGURE 12. Comparison of alignment systems with respect to their recall on the OAEI conference track using the proposed Bayesian test.

all tasks, 83 of which are correct. It is thus expected that the performance distribution of WikiV3 precision is more concentrated than that of Alin. Similar arguments hold for those of other performance scores and other systems.

Having conducted the evaluation of systems, we can now compare them with respect to various performance scores using the proposed Bayesian test. Figures 11-13 show the graphs summarizing the comparison of various systems on the OAEI conference track regarding precision, recall, and F-measure, respectively.

Figure 11 indicates that Alin is the best performing system in terms of precision while KEPLER and WikiV3 are those with poor precision. Figure 12 supports that SANOM is the top system concerning recall and it is followed by AML and KEPLER, while Alin and WikiV3 are at the other extreme.

The comparison concerning F-measure is summarized in Figure 13. According to this figure, AML and SANOM are the top performing systems while Alin and WikiV3 are at the other end of the graph.

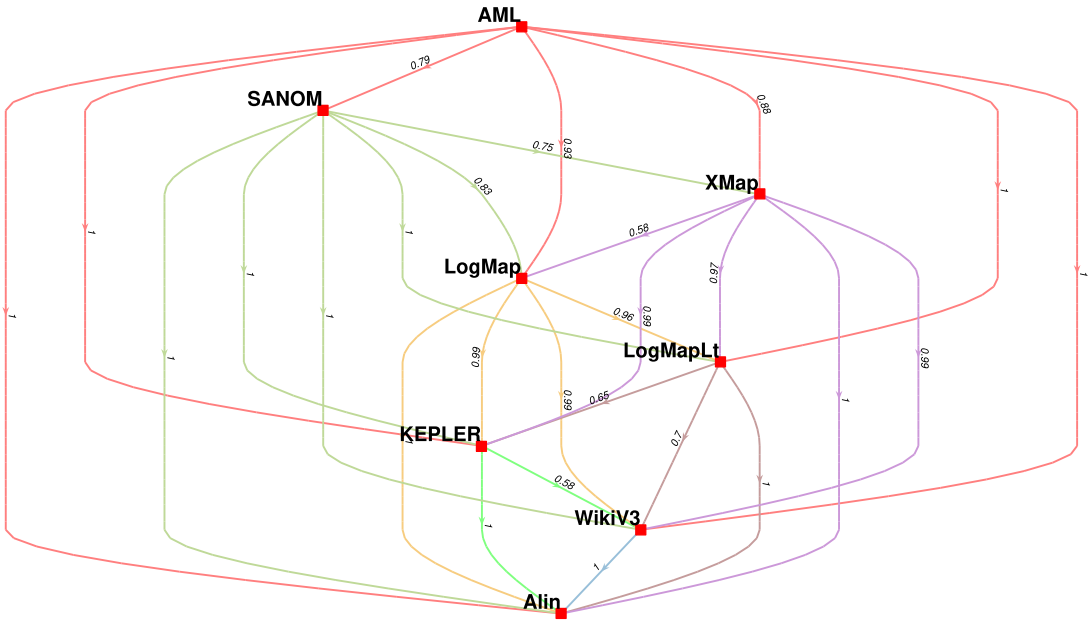

FIGURE 13. Comparison of alignment systems with respect to their F-measure on the OAEI Conference track using the proposed Bayesian test.

TABLE 2. Comparison of alignment systems on the OAEI conference track. The performance scores with the subscript *MLE* are the scores obtained by the average over all benchmarks while those with the subscript *BHM* are the means of the estimated distributions by the proposed Bayesian hierarchical model (BHM). We further tabulate the standard deviations (SD) of the performance measures which help us analyze the estimated distributions. The acronyms $\hat{Pr}$, $\hat{F}$, and $\hat{Re}$ stand for precision, F-measure, and recall, respectively.

	$\hat{Pr}_{MLE}$	SD	$\hat{Pr}_{BHM}$	$\hat{F}_{MLE}$	SD	$\hat{F}_{BHM}$	$\hat{Re}_{MLE}$	SD	$\hat{Re}_{BHM}$
ALIN	0.93	0.228	0.933	0.43	0.105	0.418	0.29	0.173	0.271
AML	0.84	0.170	0.853	0.74	0.123	0.764	0.67	0.343	0.703
KEPLER	0.61	0.260	0.591	0.59	0.105	0.584	0.60	0.329	0.591
LogMap	0.84	0.197	0.841	0.68	0.118	0.692	0.59	0.317	0.585
LogMapLite	0.76	0.278	0.750	0.61	0.127	0.600	0.53	0.326	0.504
SANOM	0.74	0.188	0.730	0.72	0.085	0.728	0.73	0.307	0.771
Wiki3	0.69	0.221	0.682	0.58	0.124	0.576	0.52	0.349	0.501
XMap	0.86	0.200	0.865	0.69	0.122	0.699	0.59	0.338	0.590

VII. CONCLUSION AND FUTURE WORKS

This paper presented a new way for both evaluation and comparison of alignment systems. The traditional way of the evaluation was to summarize the system performance in a figure which was a score for one benchmark, or its average in the case of multiple. The paper introduced the notion of risk and showed that the MLE of risk with respect to a performance score is exactly the same as the complement of the same score. Instead, we presented a new Bayesian model to estimate a distribution for each of performance scores. Such a model would give more information about the alignment performance and would help compare the alignment systems more meaningfully. We applied the proposed model to the OAEI anatomy and conference tracks and contrasted the results with those of the traditional way. We further compared the systems in those tracks and summarized the overall outcome using a weighted directed graph.

One of the drawbacks of the proposed methodology is that it does not consider the uncertainty regarding each correspondence. Right now, there is an uncertain version for the conference track to which the proposed model cannot be applied since the correspondences are considered to be only true or false, e.g., the confidence value is one for each discovered correspondence. It is an interesting avenue for improving the proposed model to enable it to estimate the risk of alignment systems in the presence of uncertain correspondences.

In addition, the proposed model can only consider one performance score at a time. For evaluation and comparison, however, it is necessary that multiple performance measures are taken into account. In this regard, a practical way is to use multi-criteria decision making (MCDM) methods to elicit a weight vector for each track of the OAEI, and then combine the distributions of various metric together. As a result, a final aggregated distribution is obtained for each system, which can also be utilized for comparison. This has left for future research.

REFERENCES

[1] W. Li, L. Yan, F. Zhang, and X. Chen, "A formal approach of construction fuzzy XML data model based on OWL 2 ontologies," *IEEE Access*, vol. 6, pp. 22025–22033, 2018.
[2] J. Euzenat *et al.*, *Ontology Matching*. New York, NY, USA: Springer, 2007, vol. 18.

[3] E. Jiménez-Ruiz, B. C. Grau, I. Horrocks, and R. Berlanga, "Ontology integration using mappings: Towards getting the right logical consequences," in *Proc. Eur. Semantic Web Conf.* New York, NY, USA: Springer, 2009, pp. 173–187.
[4] N. F. Noy, "Semantic integration: A survey of ontology-based approaches," *ACM Sigmod Rec.*, vol. 33, no. 4, pp. 65–70, Dec. 2004.
[5] H. Wache *et al.*, "ubner, "Ontology-based integration of information-a survey of existing approaches," in *Proc. IJCAI Workshop Ontologies Inf. Sharing*, 2001, pp. 108–117.
[6] J. Wang, P. Gao, Y. Ma, K. He, and P. C. Hung, "A web service discovery approach based on common topic groups Extraction," *IEEE Access*, vol. 5, pp. 10193–10208, 2017.
[7] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, "Ontology alignment for linked open data," in *Proc. Int. Semantic Web Conf.* New York, NY, USA: Springer, 2010, pp. 402–417.
[8] Z. Duo, L. Juan-Zi, and X. Bin, "Web service annotation using ontology mapping," in *Proc. IEEE Int. Workshop Service-Oriented Syst. Eng.*, 2005, pp. 235–242.
[9] J. J. Jung, "Reusing ontology mappings for query routing in semantic peer-to-peer environment," *Inf. Sci.*, vol. 180, no. 17, pp. 3248–3257, Sep. 2010.
[10] C. Pedrinaci *et al.*, "Toward the next wave of services: Linked services for the web of data," *J. Ucs*, vol. 16, no. 13, pp. 1694–1719, Jul. 2010.
[11] X. Hu, Z. Feng, S. Chen, K. Huang, J. Li, and M. Zhou, "Accurate identification of ontology alignments at different granularity levels," *IEEE Access*, vol. 5, pp. 105–120, 2017.
[12] E. Jiménez-Ruiz and B. C. Grau, "LogMap: Logic-based and scalable ontology matching," in *Proc. Int. Semantic Web Conf.* New York, NY, USA: Springer, 2011, pp. 273–288.
[13] D. Faria *et al.*, "Results of AML in OAEI 2017," in *Proc. 12th Int. Workshop Ontology Matching*, Aug. 2017, p. 122.
[14] M. Mohammadi, A. Atashin, W. Hofman, and Y.-H. Tan, "Sanom results for OAEI 2017," in *Proc. Twelfth Int. Workshop Ontology Matching*, Sep. 2017, p. 185.
[15] M. Achichi *et al.*, "Results of the ontology alignment evaluation initiative 2016," in *Proc. OM: Ontology Matching*. Oct. 2016, pp. 73–129.
[16] J. da Silva, F. A. Baiao, and K. Revoredo, "Alin results for OAEI 2017," in *Proc. Twelfth Int. Workshop Ontology Matching*, 2017, p. 114.
[17] J. Wang, Z. Ding, and C. Jiang, "GAOM: Genetic algorithm based ontology matching," in *Proc. IEEE Asia-Pacific Conf. Services Comput.*, Dec. 2006, pp. 617–620.
[18] X. Xue and Y. Wang, "Optimizing ontology alignments through a memetic algorithm using both matchfmeasure and unanimous improvement ratio," *Artif. Intell.*, vol. 223, pp. 65–81, Jun. 2015.
[19] J. Bock and J. Hettenhausen, "Discrete particle swarm optimisation for ontology alignment," *Inf. Sci.*, vol. 192, pp. 152–173, Jun. 2012.
[20] J. Martinez-Gil, E. Alba, and J. F. Aldana-Montes, "Optimizing ontology alignments by using genetic algorithms," in *Proc. Workshop Nature Based Reasoning Semantic Web*, Karlsruhe, Germany, Oct. 2008, pp. 521–535.
[21] Y.-H. T. Majid Mohammadi and W. Hofman, "Simulated annealing-based ontology matching," *ACM Trans. Manage. Inf. Syst.*, to be published.
[22] X. Xue, Y. Wang, and W. Hao, "Optimizing ontology alignments by using NSGA-II," *Int. Arab J. Inf. Technol.*, vol. 12, no. 2, 2015.
[23] M. Mohammadi, W. Hofman, and Y.-H. Tan, "A comparative study of ontology matching systems via inferential statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 615–628, Apr. 2018.
[24] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[25] M. Mohammadi, A. A. Atashin, W. Hofman, and Y. Tan, "Comparison of ontology alignment systems across single matching task via the McNemar's test," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 4, p. 51, Jun. 2018.
[26] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, "Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 2653–2688, Jan. 2017.
[27] E.-J. Wagenmakers, "A practical solution to the pervasive problems ofp values," *Psychonomic Bull. Rev.*, vol. 14, no. 5, pp. 779–804, Oct. 2007.
[28] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, "Statistical comparison of classifiers through Bayesian hierarchical modelling," *Mach. Learn.*, vol. 106, no. 11, pp. 1817–1837, Nov. 2017.
[29] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: CRC Press, 2013.
[30] G. Corani and A. Benavoli, "A Bayesian approach for comparing cross-validated algorithms on multiple data sets," *Mach. Learn.*, vol. 100, nos. 2–3, pp. 285–304, Sep. 2015.
[31] M. Cheatham and P. Hitzler, "The properties of property alignment," in *Proc. OM*, Oct. 2014, pp. 13–24.
[32] M. Cheatham and P. Hitzler, "String similarity metrics for ontology alignment," in *Proc. Int. Semantic Web Conf.* New York, NY, USA: Springer, 2013, pp. 294–309.
[33] M. Ehrig and J. Euzenat, "Relaxed precision and recall for ontology matching," in *Proc. K-Cap Workshop Integrating Ontology*, Oct. 2005, pp. 25–32.
[34] Q. Ji, Z. Gao, Z. Huang, and M. Zhu, "Semantic precision and recall for evaluating incoherent ontology mappings," in *Proc. Int. Conf. Active Media Technol.* New York, NY, USA: Springer, 2012, pp. 338–347.
[35] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton, FL, USA: CRC Press, 1995.
[36] M. Plummer *et al.*, "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," *Proc. 3rd Int. Workshop Distrib. Stat. Comput.*, Vienna, Austria, vol. 124, no. 125.10, 2003.
[37] J. David, J. Euzenat, F. Scharffe, and C. Trojahn dos Santos, "The alignment API 4.0," *Semantic web*, vol. 2, no. 1, pp. 3–10, Jan. 2011.
[38] M. Kachroudi, G. Diallo, and S. B. Yahia, "OAEI 2017 results of kepler," in *Proc. Twelfth Int. Workshop Ontology Matching*, 2017, p. 138.
[39] S. Hertling, "Wikiv3 results for OAEI 2017," in *Proc. Twelfth Int. Workshop Ontology Matching*, 2017, p. 190.

Authors' photographs and biographies not available at the time of publication.