A Comparative Study of Ontology Matching Systems via Inferential Statistics

Majid Mohammadi, Wout Hofman, and Yao-Hua Tan

Abstract—Ontology matching systems are typically compared by comparing their average performances over multiple datasets. However, this paper examines the alignment systems using statistical inference since averaging is statistically unsafe and inappropriate. The statistical tests for comparison of two or multiple alignment systems are theoretically and empirically reviewed. For comparison of two systems, the Wilcoxon signed-rank and McNemar's mid-p and asymptotic tests are recommended due to their robustness and statistical safety in different circumstances. The Friedman and Quade tests with their corresponding post-hoc procedures are studied for comparison of multiple systems, and their [dis]advantages are discussed. The statistical methods are then applied to *benchmark* and *multifarm* tracks from the ontology matching evaluation initiative (OAEI) 2015 and their results are reported and visualized by critical difference diagrams.

Index Terms—Ontology alignment evaluation, paired t-test, Wilcoxon signed-rank, McNemar, Friedman, Quade, post-hoc, Nemenyi, Holm, Shaffer, Bergmann

1 Introduction

There has been an increasing interest in ontology matching (or alignment) over the last years. As data come from various sources these days, the heterogeneity among them is inevitable. One solution to such an issue is to align the ontologies, which has a broad range of applications from data integration and agent interoperability in computer science [1], [2] to matching ontologies in biomedical and geoscience [3], [4]. Therefore, plenty of research has been dedicated to finding the correspondences between two different ontologies stating the same concepts. As a result, numerous alignment systems have been proposed claiming that they are better than, or competitive with, other state-of-the-art systems.

To recognize the alignment systems with superior performance, the ontology alignment evaluation initiative (OAEI) has taken place which makes it possible to compare ontology alignment systems in various conditions precisely. In a typical ontology matching paper, a new alignment method or a pre- or post-processing has been proposed, and an implicit hypothesis has been made that such an approach might have an enhanced performance over the existing ones. The comparison is typically based on the straightforward measures - *precision*, *recall* or *F-measure* - and a common way of reporting such measures is to put the performance scores of various systems over different datasets (typically OAEI datasets) in a table. The problem with such an approach, however, is that it is impossible to claim if one system is better than one another (of course not by 100 percent guarantee, but with reliable confidence.) Therefore, the remaining step, which is to statistically verify if there is a significant difference among systems, is the primary motivation of this paper.

Currently, the average performance of the ontology alignments systems over multiple datasets is the only yardstick toward which various ontology matching systems are compared. However, averages are sensitive to outliers. The existence of outliers is seemingly inevitable in the ontology matching since some systems have poor performance on particular datasets due to either the difficulty of datasets or their incapability to produce a correct alignment. On top of that, the poor performance of a system on one single dataset would cancel out the fair performances over the remainder of datasets (and vice versa), thereby influencing the overall average performance. Further, one system is claimed to have superior performance over one another either the discrepancy between their averages is small or large. However, the slight difference between averages can be ignored and claiming the systems are significantly different might be wrong. Also, the sole comparison of averages is not substantiated by any evidence. In this paper, the appropriate statistical procedures are empirically and theoretically studied, which allow verifying the claim of significant difference among alignment systems. These methods also enable us to compare robustly the results of alignment systems which are obtained from multiple datasets, and to determine if one system is better than one another. In the case of comparing multiple systems, the chances are that they are declared insignificantly different; therefore, no single system might be the best as the result of the statistical analysis.

- *M. Mohammadi and Y.-H. Tan are with the Faculty of Technology, Policy, and Management, Delft University of Technology, Delft 2628, CD, The Netherlands. E-mail: {m.mohammadi, y.tan}@tudelft.nl.*
- *W. Hofman is with TNO, Delft NL-2600 AA, The Netherlands. E-mail: wout.hofman@tno.nl.*

Manuscript received 29 Oct. 2017; revised 23 May 2018; accepted 26 May 2018. Date of publication 30 May 2018; date of current version 5 Mar. 2019. (Corresponding author: Majid Mohammadi.)
Recommended for acceptance by L. Dong.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2842019

Careful readers might refute the claim of the superior performance of an alignment system based on the no free lunch theorem [5], [6]. According to the NFL theorem, there is no single system which performs well in all scenarios [5]. However, there are usually background knowledge available which can distinguish the performance of one system over the rest in one particular domain, e.g., one system performs better on biomedical ontologies and one on geoscience ontologies. Therefore, the outcome of this paper is not in contradiction to the no free lunch theorem as it is sought to find the superior system in a particular domain.

This article aims to leverage statistical tests that could be utilized for comparison of two or more systems upon multiple datasets. To this end, suppose that k systems are tested over N datasets (by datasets, we mean a pair of ontologies). Let P_i^j be the performance score on the ith dataset for the jth system. The goal is to decide if the systems are different from each other based on their performance scores P_i^j, which inherently indicates that one system is better. It has been considered that the content of this article should be read and understood independently from other resources; therefore, examples are presented in crucial sections.

Such an approach has been scrutinized in other areas of research [7], [8], [9], [10], [11], [12], [13]. Demšar [7] studied the statistical procedures for comparing two or more classifiers over multiple datasets. Garcia et al. [8], [9] extended the Demšar work and proposed more advanced non-parametric tests and their corresponding post-hoc procedures for comparison of multiple classifiers. Trawiski et al. [12] compared the regression learning algorithms and utilized various statistical tests to do so. Similar approaches are applied to other areas such as information retrieval [10] and evolutionary algorithms [13]. To the best of our knowledge, it is the first paper considering the statistical inference for comparison of two or more ontology matching systems.

The performance analysis of alignment systems is different from the areas of research which have been already considered the statistical inference. First and foremost, the number of datasets for matching, especially in the OAEI, is either large enough (roughly speaking more than 30 datasets) or very small (less than ten datasets.) In contrast, the number of datasets in other areas is usually assumed to be moderate, e.g., more than ten but less than 30. Such an assumption is valid due to either the lack of datasets or the difficulties of running the methods over a large number of datasets. From the statistical point of view, the moderate and small sample size put an obstacle in the way of checking the presumptions of the statistical tests and invalidate the results of parametric tests. Therefore, the current trend is to favor the non-parametric statistics for comparison. In ontology alignment, on the other hand, it is possible to check the presumption of parametric tests as there are enough datasets in several tracks such as benchmark and multifarm. We further investigate the case that a few datasets, e.g., less than ten, are available, and propose utilizing the McNemar's test for comparison. For the moderate number of datasets, the Wilcoxon signed-rank test is recommended as it is the case in other fields [7].

Another crucial point is the performance scores obtained from each dataset. In comparison of classifiers, for instance, the scores over a dataset are not independent of each other since the re-sampling methods (e.g., cross-validation) are usually exploited. In ontology alignment, however, there is no such a problem which facilitates the utilization of statistical tests. The thorough discussion about the usage of tests under various circumstances is presented in the experimental section.

This article is structured as follows. In Sections 2 and 3, the core concepts of ontology matching evaluation and statistical hypothesis testing are reviewed, respectively. The paired t-test, Wilcoxon signed-rank and McNemar's tests are studied in Section 4 for comparison of two systems. The Friedman and Quade tests are reviewed in Section 5 and followed by their post-hoc tests and the ways of p-value adjustment for dealing with *family-wise error rate*. The extensive experimental results are presented in Section 6, and the paper is concluded in Section 7.

2 Preliminaries

Ontologies are strong tools to model a domain formally. An ontology consists of a set of entities such as classes, object and data properties, and instances. The aim of ontology matching is to find the identical entities of two given ontologies.

A correspondence is the mapping of an entity from one ontology to one entity from the other. Correspondences are obtained by using several similarity metrics such as string, linguistic, and structural similarity measures [14]. The outcome of a matching system is a set of correspondences, called alignment.

The evaluation of an alignment is usually performed by three widely-used criteria precision, recall, and F-measure. When there are multiple datasets, these scores can be obtained by comparing the reference alignment and the alignment produced by a system. The *precision* measure is the ratio of the number of correctly discovered correspondence to the total number of correspondence found by an alignment system. Similarly, *recall* is the proportion of the number of successfully found correspondences to the total number of actual correspondences. Let A be the set of correspondences identified by a system and R be the reference alignment. Precision $P(A,R)$ and recall $R(A,R)$ are defined as

$$P(A,R) = \frac{|A \cap R|}{|A|}$$
$$R(A,R) = \frac{|A \cap R|}{|R|},$$

where $|.|$ is the cardinality operator. F-measure is the *harmonic* mean of precision and recall, i.e.,

$$F(A,R) = 2 * \frac{P(A,R) * R(A,R)}{P(A,R) + R(A,R)}.$$

The statistical tests proposed here require only single performance score. The performance score for statistical significance testing might be any of above-mentioned ones (or even a measure which has not mentioned here). However, it must be warned that employing different performance measure can lead to entirely different outcomes form statistical tests. As an instance, precision of a system might be one as all discovered correspondences by this system is correct. At the same time, this system could be unable to identify all

TABLE 1
The Tests for Comparison of Two Systems over N Datasets

Test	Presumptions	Applicability
Paired t	Normality of differences	N $>$ 30
Signed-rank	symmetry of differences w.r.t median	N $>$ 10
McNemar	-	N $<$ 10

Applicability is roughly the situation that test can be used and its results are valid and differences *refer to the differences in performance scores.*

the correspondences in the reference alignment so that its recall can be quite different from its precision. Therefore, the results of the statistical tests will be very different for two measures precision and recall. Also, the selection of the performance score must be justified by the expert: The performance score covers the needs of a problem, or it is an important yardstick in the particular domain.

The focus of this article is to compare the alignment systems when multiple datasets are available. Such a comparison is the case of various tracks in the OAEI such as benchmark and multifarm. However, there are several other tracks, i.e., the anatomy track, with only a pair of ontologies for alignment. The comparison over one mapping task has been the topic of the recent study [15]. As a complementary study, the methodologies in this paper are suitable for comparison over multiple alignment tasks.

The experiments of tests revised here are applied to benchmark and multifarm tracks. In the benchmark track, a test generator based on a *seed* ontology is utilized [16]. This generator creates various ontologies by changing the seed while it keeps the actual alignment between the seed and the generated ontologies. This track aims to verify the advantages and pitfalls of systems in distinct circumstances.

The multifarm track [17] concerns with the alignment of ontologies of different natural languages. Originally, it included seven different ontologies in eight different languages. Recently, more ontologies in other languages are also added so that ontologies in 10 different languages participated in the OAEI 2015. In the OAEI 2015, two types of alignment tasks were performed for this track: (a) The alignment of one ontology in different languages; (b) The alignment of different ontologies from different languages. The good results obtained for the first case does not indicate the decent performance in dealing with cross-lingual ontologies since the structures of both ontologies are the same. The latter case where two different ontologies of various languages are matched would indicate the real performance of systems in coping with ontologies in various languages.

3 Statistical Significant Testing

The hypothesis testing is of the essence in the realm of statistical inference. Here, we aim at utilizing this technique to compare alignment systems and to identify the systems with superior performances.

To leverage the hypothesis testing, a null hypothesis is required. The null hypothesis (shown by H_0) states that there is no significant difference between two or more populations according to available samples. The alternative hypothesis (shown by H_a), on the other hand, is the opposite of the null hypothesis and states that there is a meaningful difference between two or more populations based on available samples. Thus, it is desirable to reject the null hypothesis and instead, accept the alternative.

In the ontology matching case, especially in the OAEI, it is usually the case that the performance of various systems over a range of datasets are available and it is sought to verify which system is better than the others.

To compare k systems, the null and the alternative hypotheses are

$$\begin{aligned} H_0 &: \hat{P}^1 = \hat{P}^2 = \cdots = \hat{P}^k \\ H_a &: \text{at least one } \hat{P}^i \text{ differs,} \end{aligned} \tag{1}$$

where $\hat{P}^i$ is the average performances of the ith system. This paper reviews relevant tests to find the probability of occurring the performances given H_0 is correct (this probability is called p-value.) If the p-value is less than the nominal significance level α, which must be determined before performing the test, the null hypothesis is rejected, and it is drawn that systems are significantly different. Otherwise, it fails to reject the null hypothesis. The first test proposed in comparison of two systems is the paired t-test, but it could be statistically unsafe due to its strong presumptions. Therefore, the non-parametric tests, the Wilcoxon signed-rank [18] and McNemar's [19] tests, are proposed to be utilized since they have fewer and easy-to-satisfy presumptions.

The comparison of multiple systems is more challenging. The null hypothesis, in this case, is that all systems perform equally, and if it is rejected, it is drawn that there is at least one system with different performance. However, it cannot be determined what systems are significantly different. A post-hoc procedure is then applied to indicate where exactly the difference among performance scores are. The former test is called the *omnibus* test, and the latter is referred to as the *post-hoc* test. The repeated measures ANOVA [20], Friedman [21] and Quade [22] tests and their corresponding post-hoc procedures are discussed in details. The family-wise error rate (FWER), which is a serious issue in multiple comparisons, is studied and the ways of preventing such an error are scrutinized.

4 Comparison of Two Systems

This section is dedicated to comparing two systems over multiple datasets. The tests are summarized in Table 1.

4.1 Paired t-Test

A common way to detect the difference between two systems is to compute the paired t-test statistic. Let $d_i = P_i^1 - P_i^2$ be the difference between the performances of two alignment systems over the ith dataset. The t statistic is computed as

$$t = \hat{d}/\hat{\sigma}_d, \tag{2}$$

where $\hat{d}$ and $\hat{\sigma}_d$ are the average of differences d_i and standard deviation of samples, respectively. This statistic is distributed according to the Student's t-distribution with $N-1$ degrees of freedom where N is the number of datasets. After obtaining the probability of observing the performances given H_0 being true (p-value) according to the Student's t-distribution, H_0 can be rejected if p-value $\leq \alpha$.

TABLE 2
The F-Measure Scores, Their Differences, and Ranks over Each Dataset Obtained by the Wilcoxon Signed-Rank Test of Two Systems, Edna [24], GMap [25], over 20 Datasets from the *Benchmark* Track

	edna	GMap	d_i	rank
1	0.70	0.98	−0.28	13
1	0.70	0.98	−0.28	13
2	0.02	0.80	−0.78	20
3	0.62	0.95	−0.33	14
4	0.47	0.90	−0.43	17
5	0.31	0.86	−0.55	18
6	0.17	0.83	−0.66	19
7	0.01	0.00	0.01	1
8	0.62	0.87	−0.25	10
9	0.47	0.73	−0.26	12
10	0.31	0.56	−0.25	11
11	0.16	0.33	−0.17	5
12	0.78	0.98	−0.2	7
13	0.77	0.99	−0.22	9
14	0.78	0.98	−0.2	7
15	1.00	0.98	0.02	2.5
16	0.78	0.98	−0.2	7
17	0.55	0.96	−0.41	15.5
18	1.00	0.98	0.02	2.5
19	0.55	0.96	−0.41	15.5
20	1.00	0.96	0.04	4

The rejection of the null hypothesis indicates the superiority of the system with a higher average performance.

The major drawback of using the paired t-test is the imposed assumption on the performance differences d_i. According to this test, the performance differences must be normally distributed in order for the obtained results to be reliable. In the case of comparison among alignment systems, however, there is no provision on the normality of the performance differences. One way to overcome this is to provide the paired t-test with large enough samples (∼30 datasets) so that the normality can be assumed according to the *central limit theorem*. Another way is to check the normality of distribution using various tests. Ironically, these tests have less power on small samples; therefore, it is unlikely that such tests detect abnormalities.

Another pitfall of the paired t-test is the sensitivity to outliers. Outliers can skew the test statistic and increase the estimated standard error which adversely influences the power of the test. The existence of outliers can lower the power of the paired t-test as the averaging operator. In the case of normality violation, as a result, then non-parametric tests are considered due to their robustness and the fewer presumptions they impose on the sample distribution and robustness against outliers.

To verify the applicability of the paired t-test for the OAEI, we took pairs of systems from various tracks (e.g., benchmark, multifarm, etc.) and applied the normality test [23]. As there are large sample sizes in several tracks, such as benchmark and multifarm, the normality test might have a reliable outcome. Our investigation showed that in less than 7 percent of cases, the normality assumption holds. On top of that, it is usually the case that some systems fail to produce acceptable results for some particular task. Therefore, the existence of outliers seems to be inevitable.

4.2 Wilcoxon Signed-Rank Test

The non-parametric alternative to the paired t-test is Wilcoxon signed-rank test [18]. This method ranks the absolute values of performance differences between two systems. Then, it compares the average rank of positive and negative differences.

After computing the difference between two systems over the ith dataset, e.g., d_i, the differences are ranked based on the values of d_i, disregarding its sign. The number of $d_i = 0$ are evenly split between the sum of ranks. Let W^+ and W^- be

$$\begin{aligned} W^+ &= \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \\ W^- &= \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), \end{aligned} \tag{3}$$

and $T = min(W^+, W^-)$. If fewer than 25 datasets are available, then a table consisting critical values for T must be utilized [20]. If the number of datasets exceeds 25, then the statistics

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}, \tag{4}$$

follows the standard normal distribution and thereby calculating the p-value accordingly. If the p-value is less than α, then we reject the null hypothesis and accept that there is a significant difference between the performances of two systems. Consequently, it is drawn that the system with the higher sum of ranks is better.

An example elaborates the procedure of the test. Table 2 shows F-measure of two systems, edna [24] and GMap [25], over 20 tasks from the *benchmark* track along with the difference in their performance measures and the rank obtained by the Wilcoxon signed-rank test. According to this test, $T = min(200, 10) = 10$ and $N = 20$; therefore, the p-value is nearly zero and the null hypothesis is rejected with a high confidence. As a result, GMap is claimed to have outperformed edna.

This test assumes the symmetry of differences between the performances score concerning its median [26]. This assumption is not as strong as the normality assumption but can decrease the power of the test if not satisfied. The difference in performances is also considered in this test by assigning higher ranks to datasets over which the difference between two systems is bigger. In the next section, various McNemar's tests are proposed for comparison. The McNemar's test does not impose any presumptions for conducting the test. Further, the difference between performances are not taken into account and only the number of tasks which one outperformed the other matters.

4.3 McNemar's Test

The McNemar's test applies to a 2×2 contingency table. The test is usually applicable when there are two experiments over N samples. For such a test, the contingency table would be as Table 3.

Almost all versions of the McNemar's tests only consider the discordant pair, i.e., n_{01} and n_{10} [27]. Therefore, one

TABLE 3
A Simple Contingency Table

		Experiment 2		
		−	+	sum
Experiment 1	−	n_{00}	n_{01}	$n_{0.}$
	+	n_{10}	n_{11}	$n_{1.}$
	sum	$n_{.0}$	$n_{.1}$	N

drawback of these tests is that the accordant pair, i.e., n_{11} and n_{00}, is not taken into account while the bigger values of n_{00} and n_{11} indicate the proximity of systems. Ironically, such feature is in favor of comparison across multiple datasets because it is possible to easily find the discordant pair and then apply the test.

For comparison of two systems A and B over N datasets, n_{01} is the number of datasets over which the performance score of B is greater than A. By the same token, n_{10} is the number of datasets where the performance score of A is higher than B. As stated before, the cases of equal performances are not considered in this test.

The McNemar's asymptotic test statistic [19] is

$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}, \tag{5}$$

which is distributed according to the χ^2 distribution with one degree of freedom under the null hypothesis. The McNemar's asymptotic test is undefined when $n_{01} = n_{10} = 0$.

Edwards [28] modified the asymptotic test and proposed the following statistics

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}. \tag{6}$$

This statistic is arguably the most common one among other types of McNemar's test. However, it is pointed out that this test has higher type I and type II errors which makes it inappropriate [27]. This test is also undefined when $n_{01} = n_{10} = 0$.

According to the McNemar's *exact* test, n_{01} is distributed according to the binomial test with parameters $n = n_{01} + n_{10}$ and p = 0.5. Thus, one-sided p-value would be as

$$\text{one-sided p-value} = \sum_{x_{12}=0}^{min(n_{01},n_{10})} \binom{n}{x_{12}} \left(\frac{1}{2}\right)^2. \tag{7}$$

For the two-sided p-value, one can multiply the one-sided p-value by two. The McNemar's exact test never exceeds the nominal level; however, it is utterly conservative which results in generating large p-values and detecting fewer differences [27].

A mid-p-value is calculated by first subtracting half the point probability of the observed n_{01} from the exact one-sided p-value, then double it to obtain the two- sided mid-p-value [27], [29], e.g.,

$$\begin{aligned} \text{mid-p-value} &= 2\left[\text{one-sided p-value} - \frac{1}{2}\binom{n}{n_{01}}\left(\frac{1}{2}\right)^n\right] \\ &= \text{two-sided p-value} - \binom{n}{n_{01}}\left(\frac{1}{2}\right)^n. \end{aligned} \tag{8}$$

TABLE 4
The Tests for Comparison of Multiple Systems over N Datasets

Test	Presumptions	Applicability
ANOVA t	Sphericity	$N > 30$
Friedman	-	$N > 10$
Quade	-	$N < 10$

Applicability is roughly the situation that test results are valid.

If the null hypothesis is rejected, then it is concluded that the system which has won more dataset is better. The McNemar's asymptotic test is not considered in the rest of the paper due to its high type I and type II errors [27]. The McNemar's exact test is so conservative; therefore, it is unlikely to detect a difference among samples unless they are extremely different. As a result, the McNemar's asymptotic and mid-p tests are suitable for comparison of alignment systems. The similar conclusion will be drawn from the empirical evaluation of tests in further sections.

5 Comparison of Multiple Systems

In this section, the simultaneous comparison of multiple alignment systems is discussed. The null hypothesis here is that the performances of all systems are the same and the alternative one is that there is at least one systems behaves differently. In statistics, the comparison of multiple populations consists of two phases: The *omnibus* and *post-hoc* tests. The former test only detects if there is a significant difference among performances while the latter precisely indicates different groups. Table 4 summarizes the tests of this section.

5.1 Omnibus Tests

It is sometimes seen that omnibus tests are ignored, and post-hoc tests are only performed to detect the differences among various populations. However, it is statistically safer and recommended to carry out the omnibus test first. The three tests repeated measures ANOVA [20], Friedman [21] and Quade [22] tests are discussed in this section.

5.1.1 Repeated Measures ANOVA

The most well-known test for detecting the difference among more than two related samples is the repeated measures ANOVA. The null hypothesis is that all systems perform equally well. In the repeated measures ANOVA, the total variability is divided into variability between systems, variability between benchmarks and the residual error variability [7]. The between systems' variability is a measure between the variances of the means of the alignment systems [20]. The residual variability, on the other hand, is viewed as the variability by chance. The repeated measures ANOVA would reject the null hypothesis if the between-systems' variability was significantly larger than the residual variability.

As any parametric test, the repeated measures ANOVA is predicated on several assumptions whose violation can invalidate the obtained results. The first assumption is that the data are normally distributed. Although there is no guarantee that the data are normally distributed, statisticians do not ignore the ANOVA for abnormality of distribution unless the distribution is bi-modal [7], [30]. The most

important assumption of this test is *sphericity*. Sphericity refers to the conditions where the variances of the differences between each possible pair of groups are equal. This assumption is more likely to be violated as there is no guarantee for the parity of differences' variances. The violation of sphericity invalidates the obtained results and consequently influences the post-hoc test.

The well-known test for checking sphericity is Mauchly's test [31]. We have conducted this test over the results of the OAEI in recent years, and the assumption of sphericity is unexceptionally repudiated with an extremely-significant p-value. Even if the sphericity assumption is not rejected, Mauchly's test is reprimanded as it is not able to detect the transgression against sphericity in small samples and falsely detect it in large samples. As a result, it is recommended to exploit the non-parametric tests for comparison.

5.1.2 Friedman Test

The Friedman test [21] is the non-parametric counterpart of the repeated measures ANOVA and is the extension of the binomial Sign test (or the McNemar's exact test with p=0.5). Instead of using the scores themselves for computing the statistic, it first ranks the scores and uses them in the calculation of the statistic. The ranking procedure is among the scores of different systems over one specific dataset in a way that the best performance score takes the rank of 1 and the worst takes the rank of k, where k is the number of methods. The average rank is assigned if the scores tie.

Let r_i^j be the rank of the jth system on the ith dataset. If two systems perform equally, it is expected that their average ranks across all datasets are the same. The Friedman statistic

$$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{9}$$

is χ^2 distributed with $k-1$ degrees of freedom. It is investigated that the type II error of Eq. (9) is undesirably high; therefore, a better statistic is derived by Iman-Davenport [32]

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{10}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. An example in the next section elaborates the procedure of finding the Friedman statistic.

5.1.3 Quade Test

The Friedman test is only predicated on the ranks of systems over one single dataset. The Quade test [33], on the other hand, takes into account the performance variation among datasets and it is suitable when the number of datasets is small (roughly less than ten datasets). The underlying assumption behind the Quade test is that if the scores' variation over a dataset is larger, then it is a more challenging one to be aligned. Thus, the success of a system over such datasets indicates much better performance.

To find the ranks of each method, the range of scores over one dataset is computed by subtracting the maximum score from the minimum one. Then, the minimum and the maximum range takes the rank 1 and n, respectively. Let $Q_1, Q_2, \ldots, Q_n$ be the rank of n datasets and r_i^j be the ranks obtained by the Friedman test for each score. The Quade rank of each score is obtained as $S_i^j = Q_i(r_i^j - \frac{k+1}{2})$. Finally, the test statistic is

$$F_{Quade} = \frac{(n-1)\sum_{j=1}^{k}(S^j)^2}{A - \frac{1}{n}\sum_{j=1}^{n}(S^j)^2}, \tag{11}$$

where

$$S^j = \sum_i S_i^j \qquad A = \frac{n^2(n+1)(2n+1)k(k+1)(k-1)}{72},$$

and F_{Quade} is distributed according to F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. The next section includes an example of the calculation of this statistic.

TABLE 5
The F-Measure Scores and the Friedman Ranks (in the Parenthesis) of the Four Methods over 20 Tasks of the OAEI *Benchmark* Track

	edna	GMap	**LogMap**	**XMap**
1	0.70 (**4**)	0.98(**2**)	0.95(**3**)	1 (**1**)
2	0.02 (**2**)	0.80 (**1**)	0.00(**3.5**)	0 (**3.5**)
3	0.62 (**4**)	0.95(**1**)	0.87 (**2**)	0.66 (**3**)
4	0.47 (**4**)	0.90 (**1**)	0.72 (**2**)	0.65 (**3**)
5	0.31 (**4**)	0.86 (**1**)	0.52 (**2**)	0.51 (**3**)
6	0.17 (**3**)	0.83 (**1**)	0.28 (**2**)	0.15 (**4**)
7	0.01 (**1**)	0.00 (**3**)	0.00 (**3**)	0.00 (**3**)
8	0.62 (**4**)	0.87(**1.5**)	0.87 (**1.5**)	0.65 (**3**)
9	0.47 (**4**)	0.73 (**1**)	0.71 (**2**)	0.65 (**3**)
10	0.31 (**4**)	0.56 (**1**)	0.50 (**2**)	0.42 (**3**)
11	0.16 (**4**)	0.33 (**1**)	0.31 (**2**)	0.19 (**3**)
12	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
13	0.77 (**3**)	0.99 (**1**)	0.00 (**4**)	0.8 (**2**)
14	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
15	1.00 (**1.5**)	0.98 (**3**)	0.94 (**4**)	1.00 (**1.5**)
16	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
17	0.55 (**4**)	0.96 (**2**)	0.92 (**3**)	1.00 (**1**)
18	1.00 (**1.5**)	0.98 (**3**)	0.95 (**4**)	1.00 (**1.5**)
19	0.55 (**4**)	0.96 (**2**)	0.92 (**3**)	1.00 (**1**)
20	1.00 (**1.5**)	0.96 (**3**)	0.92 (**4**)	1.00 (**1.5**)
R_j	**3.2750**	**1.7250**	**2.8000**	**2.2000**

Each row and each column correspond to a dataset and a system, respectively. The last column shows the average Friedman rank.

5.1.4 An Example

In this section, the procedure of Friedman and Quade tests are elaborated by an example. Table 5 tabulates the *precision* of five methods, namely edna [24], GMap [25], LogMap [34] and XMap [35] across the OAEI *benchmark* track. The numbers in the parenthesis are the Friedman ranks of each method over the corresponding dataset. Then, the Friedman statistic can be calculated as

$$(\textbf{Friedman}) \quad \chi_F^2 = \frac{12\times 20}{4\times 5}\left(3.275^2 + 1.725^2 + 2.8^2 + 2.2^2 - \frac{4\times 5^2}{4}\right) = 16.575$$

$$(\textbf{Iman Davenport}) \quad F_F = 7.25.$$

As the experiment consists of four methods over 20 datasets, χ_F^2 has χ^2 distribution with $4-1=3$ degrees of

TABLE 6
The F-Measure Scores and the Quade Ranks (in the Parenthesis) of the Four Systems over 20 Tasks of the OAEI *Benchmark* Track

	Range	Q_i	edna	GMap	LogMap	XMap
1	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95(**3.75**)	1(**-11.25**)
2	0.8	19	0.02 (**-9.5**)	0.8 (**-28.5**)	0 (**19**)	0 (**19**)
3	0.33	13	0.62 (**19.5**)	0.95 (**-19.5**)	0.87 (**-6.5**)	0.66 (**6.5**)
4	0.43	14	0.47 (**21**)	0.9 (**-21**)	0.72 (**-7**)	0.65 (**7**)
5	0.55	17	0.31 (**25.5**)	0.86 (**-25.5**)	0.52 (**-8.5**)	0.51 (**8.5**)
6	0.68	18	0.17 (**9**)	0.83 (**-27**)	0.28 (**-9**)	0.15 (**27**)
7	0.01	1	0.01 (**-1.5**)	0 (**0.5**)	0 (**0.5**)	0 (**0.5**)
8	0.25	10	0.62 (**15**)	0.87 (**-10**)	0.87 (**-10**)	0.65 (**5**)
9	0.26	12	0.47 (**18**)	0.73 (**-18**)	0.71 (**-6**)	0.65 (**6**)
10	0.25	11	0.31 (**16.5**)	0.56 (**-16.5**)	0.5 (**-5.5**)	0.42 (**5.5**)
11	0.17	5	0.16 (**7.5**)	0.33 (**-7.5**)	0.31 (**-2.5**)	0.19 (**2.5**)
12	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
13	0.99	20	0.77 (**10**)	0.99 (**-30**)	0 (**30**)	0.8 (**-10**)
14	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
15	0.06	3	1 (**-3**)	0.98 (**1.5**)	0.94 (**4.5**)	1 (**-3**)
16	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
17	0.45	15.5	0.55 (**23.25**)	0.96 (**-7.75**)	0.92 (**7.75**)	1 (**-23.25**)
18	0.05	2	1 (**-2**)	0.98 (**1**)	0.95 (**3**)	1 (**-2**)
19	0.45	15.5	0.55 (**23.25**)	0.96 (**-7.75**)	0.92 (**7.75**)	1 (**-23.25**)
20	0.08	4	1 (**-4**)	0.96 (**2**)	0.92 (**6**)	1(**-4**)

Each row corresponds to a dataset. The first column is the range and the second is Q_i of the Quade test. The rest columns are the systems under comparison.

freedom and F_F is distributed according to the F-distribution with $4-1=3$ and $(4-1)(20-1)=57$ degrees of freedom. The p-values calculated for the Friedman and Iman-Davenport tests are 8.65×10^{-4} and 3.33×10^{-4}, respectively. Thus, the null hypothesis is rejected in both cases.

We perform the Quade test on the scores in the above table. Table 6 displays the datasets' ranks and scores' ranks of the Quade test. The test statistic is

$$F_{Quade}=10.16,$$

which is distributed according to the F-distribution with $(3,57)$ degrees of freedom. The corresponding p-value is 1.84×10^{-5} which results in rejecting the null hypothesis.

5.2 Post-Hoc Analysis

If the null hypothesis in multiple comparisons is rejected, a post-hoc test will be employed to say where exactly the differences occurred among performances of systems. For each of the tests mentioned in the previous section, a post-hoc test exists.

The following statistics must be computed for each pair of systems (i,j)

$$\begin{aligned}\text{Friedman}\qquad & z=\frac{R_i-R_j}{\sqrt{\frac{k(k+1)}{6n}}}\\ \text{Quade}\qquad & z=\frac{T_i-T_j}{\sqrt{\frac{k(k+1)(2n+1)(k-1)}{18n(n+1)}}},\end{aligned}\tag{12}$$

where R_i is the average ranks in the Friedman test and $T_i=\frac{2\sum_{i,j}Q_i r_i^j}{n(n+1)}$ in the Quade test. The probability of systems i and j having the same performance can be calculated using above mentioned statistics which are distributed according to the standard normal distribution. Similar to comparison of two systems, one can reject the null hypothesis and conclude that the two systems are significantly different provided that the computed probability is less than α. If the null hypothesis is rejected, then the system with lower average rank, in both Friedman and Quade tests, is claimed to be better.

In multiple comparisons, however, the family-wise error rate would increase the type I error if the p-value was not adjusted. With the significance level α, the probability of making type I error for each comparison is $1-\alpha$ and $m=k(k-1)/2$ comparison must be performed when k systems are available. Thus, the probability of making at least one type I error in m comparisons is $1-(1-\alpha)^m$ which is way higher than the nominal significance level α. For example, for $\alpha=0.05$ and $k=5$ the probability of making type I error is 0.4, which is undesirably high.

To adjust the p-values, suppose $p_1,\ldots,p_m$ are the probabilities of m hypotheses $H_1,\ldots,H_m$. There are various ways for the p-value adjustment in order to prevent FWER. The most straightforward one is the Nemenyi correction [36] which divides α by the number of comparisons. Dividing the p-value by the number of comparisons prevent the FWER. The adjusted p-value (APV) for each hypothesis i by the Nemenyi correction is: $APV_i=min\{m*p_i,1\}$. However, the Nemenyi correction is highly conservative and has high type II error. It means that there are several null hypotheses which must be rejected but they are retained if the Nemenyi APV is employed. Other than the Nemenyi correction which adjusts the value of α in one single step, there are other ways that adjust the p-values in a sequential manner.

The Holm procedure [37] takes the most significant p-value (let it be p_1) and compares it with $\frac{\alpha}{m-1}$ and $m=k(k-1)/2$. If $p_1<\frac{\alpha}{m-1}$ then it rejects the corresponding null hypothesis H_1 and compare the next most significant p-value, p_2, with $\frac{\alpha}{m-2}$ and so forth. This procedure is terminated when a certain null hypothesis cannot be rejected. In other words, let $p_1\le p_2\le p_3\ldots\le p_m$ be the ordered p-values and $H_1,H_2,\ldots,H_m$ be the corresponding hypotheses. The Holm procedure rejects the hypotheses $H_1,\ldots,H_{i-1}$ if $p_i>\frac{\alpha}{m-1}$ for the smallest i. As the Holm procedure starts with most significant p-value, it is called a 'step-down' method.

Akin to the Holm procedure, Shaffer [38] is a sequential method for the p-value adjustment. However, this method instead uses the logical relationship among the family of hypotheses. In Shaffer procedure at stage j, the null hypothesis H_j is rejected if the corresponding p-value is less than α/t_j, where t_j is the maximum number of hypotheses which are possible to be retained given that $j-1$ hypotheses are false. Shaffer proposed to find the maximum number of possibly correct hypotheses. The possible number of true hypotheses can be recursively obtained as

$$S(k)=\bigcup_{j=1}^{k}\left\{\binom{2}{j}+x : x\in S(k-j)\right\},\tag{13}$$

where $S(k)$ is the set of possible numbers of true hypotheses with k systems being compared. t_j is then calculated using $S(k)$.

Similar to Shaffer correction, Bergmann and Hommel [39] proposed a method which finds the maximum number

of possibly correct hypotheses dynamically. They defined the exhaustive set to formulate their procedure. An index set of hypotheses $I \subseteq \{1, \ldots, m\}$ is called *exhaustive* if exactly all $H_j, j \in I$ could be true.

Any hypothesis H_j is rejected if $j \notin A$, where A is the acceptance set defined as

$$A = \bigcup \{\text{I: I exhaustive, } \min\{P_i : i \in I\} > \alpha/|I|\}, \quad (14)$$

containing all the retained hypotheses. Theoretically speaking, this is the most powerful method for adjusting p-values for all pairwise comparisons. Our experiments also confirm that this method would detect more significant differences compared to the three methods discussed above. However, Bergmann method is not time-wise efficient especially when there are more than nine systems for comparison.

6 Experimental Study

In this section, the experiments regarding the statistical tests are discussed. First, the measures for power and replicability of tests are reviewed based on which various tests are compared. Then, the comparison of multiple systems are applied to the OAEI 2015 *benchmark* and *multifarm* tracks, and the corresponding results are reported.

6.1 Comparison of Two Systems

The power of statistical tests is formally defined as the probability of rejecting false null hypotheses. In reality, however, it is impossible to say if the null hypothesis is wrong beforehand; therefore, it is not possible to gauge the power of statistical tests from the formal definition. Instead, there are two ways to compare statistical tests concerning their power. First, the number of rejected null hypotheses in one thousand experiments are counted with a nominal significant level α. Another way is the average p-value in one thousand experiments; the lower the average is, the better the test will be.

For each way of the power estimation, there is a corresponding *replicability* measure. Bouckaert [40] defined the replicability as the probability that two experiments with the same pair of algorithms produce the same results. He estimated this probability as (in n experiments)

$$R(e) = \sum_{1 \le i \le j \le n} \frac{I(e_i = e_j)}{n(n-1)/2}, \quad (15)$$

where I is the indicator function, and e_i is the outcome of the ith experiment (0 if the null hypothesis in the ith experiment is rejected, and 1 otherwise.) If the hypothesis is accepted in p and rejected in q experiments, $R(e)$ can be easily computed as

$$R(e) = \frac{p(p-1) + q(q-1)}{n(n-1)}. \quad (16)$$

Instead of using the number of rejected or retained hypotheses, Demšar [7] proposed a robust estimator based on the p-value obtained in each experiment. Demšar defined the replicability $R(p)$ as

$$R(p) = 1 - 2var(p) = 1 - 2\frac{\sum_i (p - \hat{p})^2}{n-1}, \quad (17)$$

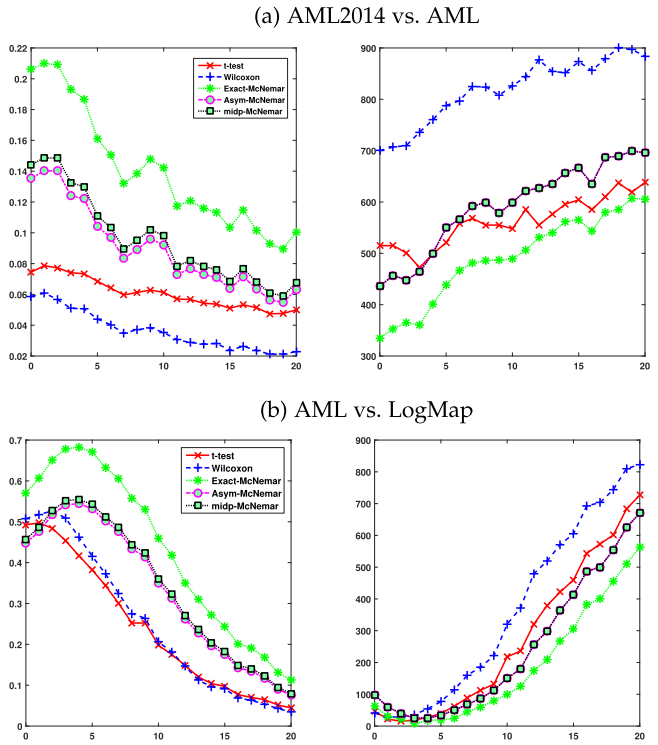


Fig. 1. Comparison of the paired t-test, Wilcoxon signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests from the power perspective in 1,000 experiments. The x-axis is k, and the y-axis is: (a) Left plot: The average p-value. (b) Right plot: The number of rejected null hypotheses.

where $\hat{p}$ is the mean of the p-values and p_i is the p-value of the ith experiment.

Since no single ontology matching system performs better than others in all scenarios [5], [6], it is usually the case that researchers would like to show the superiority of a system in one specific domain. In this case, there are some systems which perform better than others. To show this in simulation, some datasets are randomly selected from the OAEI 2015 *benchmark* track so that the probability of selecting the ith dataset is proportional to $1/(1 + e^{-kd_i})$, where d_i is the difference between the performances and k is the bias [7]. For $k = 0$, the probability of selecting all datasets are the same. With higher values of k, it is more likely to pick the sets in favor of one system. This procedure is only considered for the simulation study because doing such experiments with datasets chosen in favor of one system is, in one way or another, cheating.

For the above procedure, three different situations with 5, 20 and the whole datasets are considered. In each of this situation, the suitable tests are recommended for utilization.

First of all, 20 datasets are selected from the OAEI 2015 benchmark track with the procedure mentioned above. The comparison is between top two systems with two systems with mediocre performances so that the various numbers of k will effectively change the selected datasets. Fig. 1 plots the power estimation defined by the average p-value (left-hand side) and the number of rejected null hypotheses (right-hand side) in thousand experiments on five statistical tests studied in this paper. The x-axis in all plots is k as defined above, and the y-axis is the average p-value for the left plot and the number of rejected hypotheses for the right one. The McNemar's test with continuity correction is

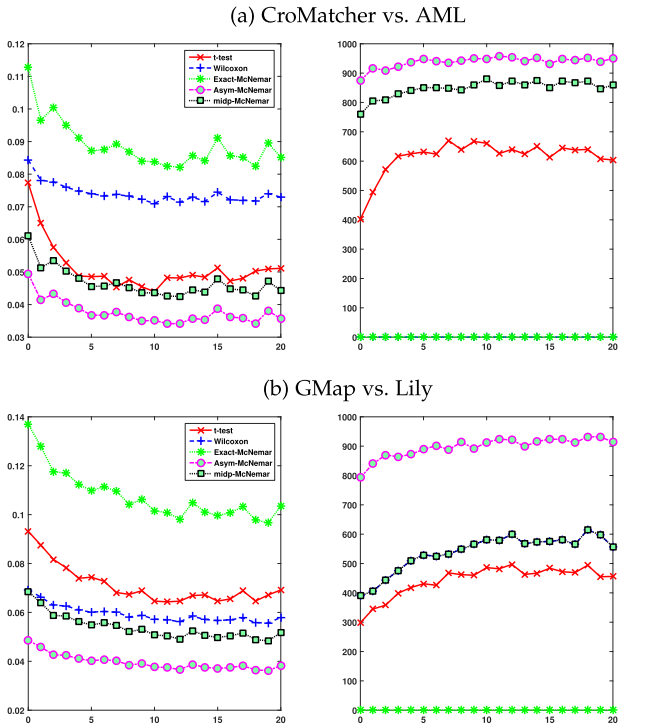


Fig. 2. Comparison of the paired t-test, Wilcoxon signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests from the power perspective in 1,000 experiments. The x-axis is k, and the y-axis is: (a) Left plot: The average p-value. (b) Right plot: The number of rejected null hypotheses.

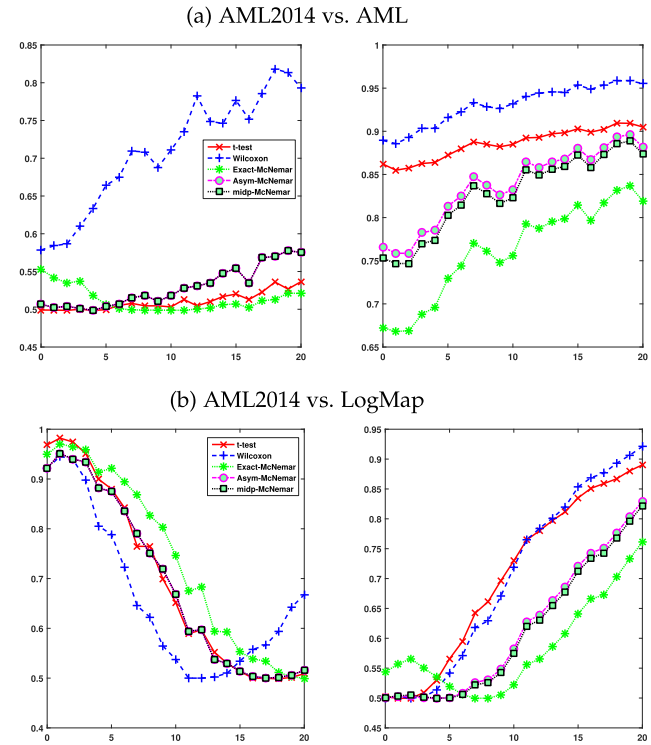


Fig. 3. Comparison of the paired t-test, Wilcoxon signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests via the replicability point of view. The x-axis is k and the y-axis is: (a) Left plot: The replicability estimation R(p). Right plot: The replicability estimation R(e).

dismissed because there is no guarantee that its type I error be below the nominal significance level [27]. The average p-value of the Wilcoxon signed-rank test is lower than or competitive with the paired t-test. This is probably because of the number of selected datasets is relatively high and presumptions of the paired t-test are likely to be satisfied through the *central limit theorem*. However, the number of rejected null-hypotheses in the Wilcoxon signed-rank test is higher than the paired t-test in both cases. Therefore, we suggest using the Wilcoxon signed-rank test when the comparison of two alignment systems is desired under this circumstance. It can also be readily seen that the McNemar's exact test (or the Sign test) is the most conservative one; thus, it should not be considered as means of comparison. Another interesting point is that the McNemar's mid-p and asymptotic tests are slightly different regarding the average p-values but almost the same with respect to the number of rejected null hypotheses. Further, these two tests are competitive with the paired t-test especially in terms of the number of rejected null hypotheses. As McNemar's tests are non-parametric, their utilization is recommended as an alternative to the Wilcoxon signed-rank test.

For the second scenario, five datasets are selected according to the above procedure. Fig. 2 shows the power estimations when five datasets are selected while the horizontal and vertical axes are the same as Fig. 1. Interestingly, the power of the Wilcoxon signed-rank test is less than McNemar's asymptotic and mid-p tests. The McNemar's asymptotic test shows high power, especially from the view of rejected hypotheses. When few datasets are available, McNemar's asymptotic and mid-p tests are preferred.

In addition to the power comparison, the statistical tests are compared concerning the replicability. Fig. 3 shows $R(e)$ on the right-hand side and R(p) on the left-hand side when 20 datasets are selected. Interestingly, the results of two measures are in contradiction. The Wilcoxon signed-rank test is (slightly) better than other methods regarding R (p). In terms of R(e), on the other hand, it is the least reliable one. However, the shape of this graph and Fig. 1 show that the test is less reliable in terms of R(e) when the p-value is in the proximity of α (here $\alpha = 0.05$). Thus, it can be drawn that the Wilcoxon is unreliable with respect to $R(e)$ because of its higher power.

For the case of selecting five datasets, the McNemar's asymptotic test indicates the better replicability regarding both perspectives while the Wilcoxon signed-rank test shows less replicability concerning both measures as shown in Fig. 4. Another interesting point is the paradoxical replicability of the Wilcoxon signed-rank and McNemar's exact tests. These tests could not reject any null hypothesis as can be observed from Fig. 2; therefore, the corresponding R(e) is one in all scenarios. Regarding R(p), on the other hand, the average p-values in thousand experiments shows their unreliability in comparison to others.

The final scenario is the case when the number of datasets is large enough. There are seemingly enough datasets so that the presumption of the paired t-test must be met. We paired various systems together from benchmark and multifarm tracks and performed Jarque-Bera test [23] to check the normality assumption required for the paired t-test. Ironically, the normality assumption is held in less than 7 percent; therefore, it is safer to conduct the Wilcoxon signed-rank test if all datasets are selected for comparison.

Table 7 tabulates the comparison of all pairs of systems with $k = 15$. The below diagonal numbers indicate the

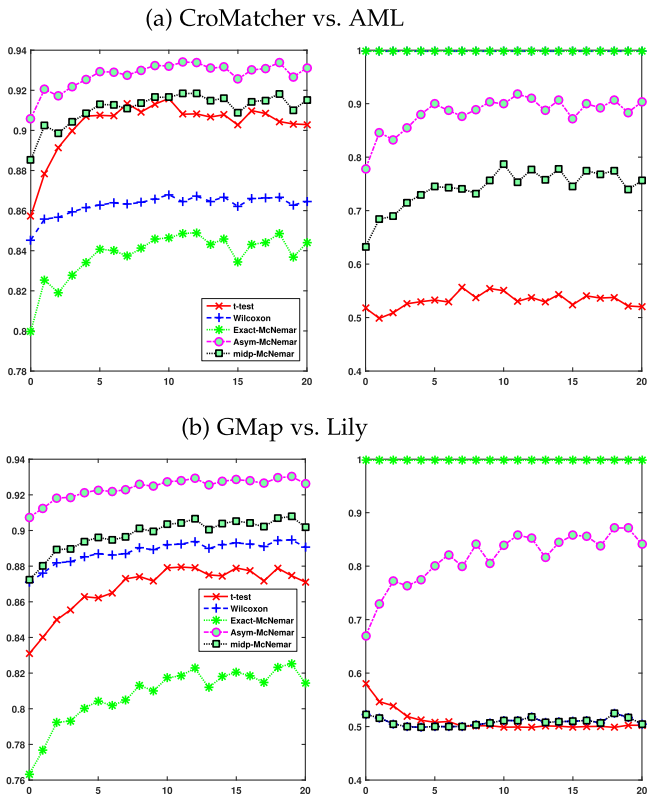


Fig. 4. Comparison of the paired t-test, Wilcoxon signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests via the replicability point of view. The x-axis is k, and the y-axis is: (a) Left plot: The replicability estimation R(p). (b) Right plot: The replicability estimation R(e).

average p-value and the corresponding replicabiliy measure R(p), and the above diagonal shows the number of rejected null hypotheses and the corresponding replicability measure R(e). The average p-value of the Wilcoxon signed-rank test is much lower than other methods in almost all cases. It is also recommendable by replicability measure R(p), but R (e) prefers other tests with the p-value higher than the critical value 0.05.

6.2 Comparison of Multiple Systems

In this section, the experiments across multiple alignment systems are studied. First, the power of various post-hoc procedures is reviewed and then the aforementioned multiple comparisons are applied to the OAEI 2015 *benchmark* and *multifarm* tracks and the corresponding results are reported.

Fig. 5 shows the results over the *benchmark* track of five methods by the Friedman test and various post-hoc procedures. The x-axis in this figure is the parameter k and the y-axis is the overall number of the rejected hypotheses with respect to a correction method. The Bergmann correction performs better than other methods as its number of rejected null hypothesis is consistently outweigh the number of rejected hypotheses of other methods. At the other extreme, the Nemenyi correction is the weakest method and must be ignored. Further, Holm and Shaffer methods are competitive with each other.

6.2.1 Benchmark *Track*

The *benchmark* track consists of artificially constructed ontologies based on a *seed* ontology [16]. In the OAEI 2015, two

TABLE 7
Comparison of the Paired t-Test, Wilcoxon Signed Rank and McNemar's (Asymptotic, Mid-p) Tests with k=15

(a) Wilcoxon signed rank test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		621/0.53	1000/1	80/0.85	873/0.78	334/0.57	450/0.51	171/0.74
AML14	0.08/0.84		1000/1	283/0.59	1000/1	143/0.75	1000/1	885/0.80
CMtch	0.01/0.98	0.00/0.99		1000/1	47/0.91	1000/1	1000/1	1000/1
GMap	0.47/0.50	0.26/0.62	0.00/0.99		1000/1	941/0.89	1000/1	1000/1
Lily	0.02/0.95	0.00/0.99	0.48/0.50	0.00/0.99		1000/1	1000/1	1000/1
XMAP	0.19/0.69	0.36/0.54	0.00/0.99	0.01/0.97	0.00/0.99		998/0.99	945/0.90
LogMap	0.16/0.73	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.99		23/0.96
Mamba	0.34/0.55	0.02/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.01/0.99	0.53/0.50	

(b) Paired t-test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		213/0.66	934/0.88	36/0.93	589/0.52	272/0.60	442/0.51	155/0.74
AML14	0.16/0.72		987/0.97	132/0.77	1000/1	158/0.73	1000/1	911/0.84
CMtch	0.02/0.95	0.01/0.98		1000/1	106/0.81	1000/1	1000/1	1000/1
GMap	0.49/0.50	0.46/0.50	0.00 /0.99		1000/1	980/0.96	1000/1	1000/1
Lily	0.05/0.90	0.00/0.99	0.34/0.55	0.00/0.99		1000/1	1000/1	1000/1
XMA	0.25/0.62	0.38/0.53	0.00/0.99	0.01/0.98	0.00/0.99		1000/1	962/0.93
LogMap	0.14/0.76	0.00/0.99	0.00/1	0.00/0.99	0.00/1	0.00/0.99		27/0.95
Mamba	0.36/0.54	0.02/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.98	0.50/0.50	

(c) McNemar's mid-p test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		213/0.66	934/0.88	36/0.93	589/0.51	272/0.60	442/0.51	155/0.74
AML14	0.16/0.73		987/0.97	132/0.77	1000/1	158/0.73	1000/1	911/0.84
CMtch	0.02/0.95	0.00/0.98		1000/1	106/0.81	1000/1	1000/1	1000/1
GMap	0.49/0.50	0.46/0.50	0.00/0.99		1000/1	980/0.96	1000/1	1000/1
Lily	0.05/0.90	0.00/0.99	0.33/0.55	0.00/0.99		1000/1	1000/1	1000/1
XMAP	0.25/0.62	0.38/0.53	0.00/0.99	0.01/0.98	0.00/0.99		1000/1	962/0.93
LogMap	0.14/0.76	0.00/0.99	0.00/1	0.00/0.99	0.00/1	0.00/0.99		27/0.95
Mamba	0.36/0.54	0.01/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.98	0.50/0.50	

(d) McNemar's asymptotic test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		619/0.53	1000/1	419/0.51	875/0.78	335/0.55	249/0.63	335/0.55
AML14	0.10/0.82		1000/1	523/0.50	1000/1	110/0.80	967/0.94	866/0.76
CMtch	0.00/0.98	0.00/0.99		1000/1	29/0.94	1000/1	1000/1	1000/1
GMap	0.17/0.71	0.11/0.80	0.00/1		1000/1	276/0.60	1000/1	1000/1
Lily	0.02/0.95	0.00/1	0.63/0.53	0.00/ 1		1000/1	1000/1	1000/1
XMAP	0.22/0.65	0.44/0.50	0.00/0.99	0.27/0.60	0.00/0.99		967/0.93	881/0.79
LogMap	0.29/0.58	0.01/0.98	0.00/1	0.00/0.99	0.00/1	0.01/0.98		10/0.98
Mamba	0.21/0.66	0.02/0.95	0.00/1	0.00/0.99	0.00/1	0.02/0.95	0.65/0.54	

Below diagonal: The average p-value and the corresponding $R(p)$. Above diagonal: The number of rejected null hypotheses in thousands experiments and the corresponding $R(e)$.

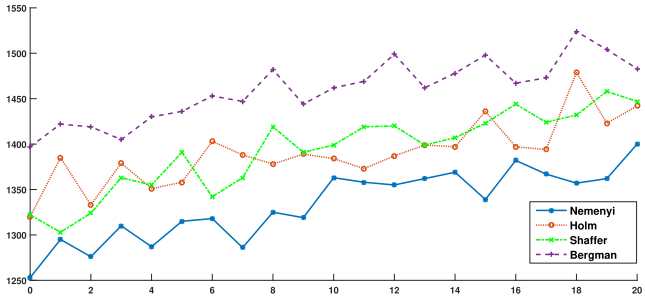
seed ontologies were employed: *biblio* and *energy*. We perform the statistical analysis of the results obtained from 94 datasets generated from the seed ontology *biblio*. The competition was among ten systems with their variation (e.g., AML and AML2014) in this track. Two of these systems did not produce any readable result so that they are not selected for comparison in this section. The remaining systems are edna [24], AML2014 [41], CroMatcher [42], GMap [25], Lily [43], XMap [35], LogMapLite [34] and Mamba [44]. The comparison is conducted based on the F-measure as it considers both undiscovered and falsely-discovered correspondences.

Table 8 tabulates the average ranks obtained by Friedman and Quade tests.

The Friedman statistic is 385.73 with 7 degrees of freedom; thus the corresponding p-value is 1.8×10^{-10}. The Quade statistic (with $(7, 651)$ degrees of freedom) and its p-value are 91.60 and 1.22×10^{-92}, respectively. The null hypothesis which is the equivalence of performances of systems is rejected by both tests.

Table 9 shows the adjusted p-values obtained by various correction procedures for Friedman and Quade tests for all pairs of systems. Based on this table, the rejected hypotheses can be simply discovered by the comparison of the adjusted p-values with the nominal significance level α while the FWER is inherently controlled. With $\alpha = 0.05$ and with the

(a) Comparison of multiple systems with selection of 10 datasets



(b) Comparison of multiple systems with selection of 40 datasets

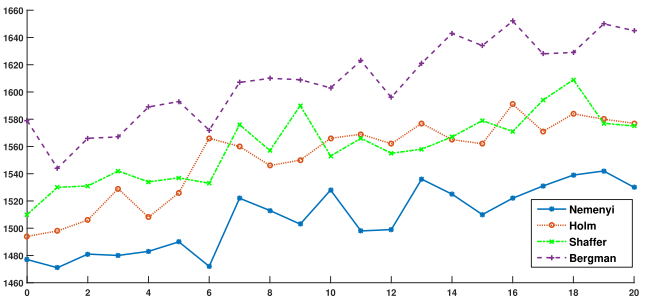


Fig. 5. The comparisons of correction methods for the Friedman test for various numbers of k in x-axis; Two different scenarios: (a) Selection of 10 datasets. (b) Selection of 40 datasets.

Friedman test, the first 18 hypotheses are rejected with the Nemenyi correction while 19 hypotheses are rejected with more advanced methods.

In the Quade test, on the other hands, the first 12 hypotheses are rejected with all correction methods. As mentioned above, the Quade test is more suitable when few datasets are available. In the benchmark track, which 94 pairs of ontologies exist, the Friedman test is expected to be more powerful, as can be readily drawn from Table 9. The sequential p-value adjustment methods reject the same number of hypotheses which means that they have the same power with respect to $R(e)$. From the $R(p)$ view, however, the Bergmann method is more powerful as it results in smaller adjusted p-values.

To better visualize and understand these results, Figs. 6 and 7 show the critical difference (CD) plot of the Friedman and Quade tests with various correction methods for $\alpha = 0.05$. The non-significant systems are connected to each other by a line. The results drawn from the table can be easily viewed from the CD diagrams as well. One difference between the Nemenyi and other sequential ways of p-value adjustment is the fixed critical difference in the former. It means that if the difference between any two methods is less than the critical difference shown at the top of the plot, then they are not significantly different. This is the reason we distinguish the plot of the Nemenyi correction with other methods.

The Quade test with four correction methods indicates that Lily and CroMatcher are together better than the remaining ones, and the rest are not significantly different (with $\alpha = .05$). The Friedman test also confirms the superiority of Lily and CroMatcher. With the Nemenyi correction, the Friedman test shows that Gmap, XMap, and AML2014 are not significantly different while GMap indicates better performance in comparison with AML2014 when other sequential-based correction methods are applied. Another difference between the Nemenyi correction and sequentially-corrected methods is the significant difference between AML2014 and LogMapLite: The Nemenyi correction cannot detect any difference between them whereas they are significantly different when Holm, Shaffer, or Bergmann correction is applied.

TABLE 8
The Average Ranks of All Systems Computed by Friedman and Quade Tests over the *Benchmark* Track

Algorithm	Friedman	Quade
Lily	1.51	1.37
CroMatcher	1.81	1.75
GMap	4.35	4.29
XMap	4.78	5.18
AML2014	5.37	5.56
Mamba	5.68	5.42
edna	6.09	6.24
LogMapLite	6.41	6.18

TABLE 9
The Adjusted p-Values by Four p-Value Adjustment Methods Across the OAEI 2015 *Benchmark* Track: (a) The Friedman Test and (b) the Quade Test

(a) Friedman test

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	Lily vs. LogMapLite	7.08×10^{-43}	1.98×10^{-41}	1.98×10^{-41}	1.98×10^{-41}	1.98×10^{-41}
2	CroMatcher vs. LogMapLite	7.30×10^{-38}	2.04×10^{-36}	1.97×10^{-36}	1.53×10^{-36}	1.53×10^{-36}
3	edna vs. Lily	8.85×10^{-38}	2.48×10^{-36}	2.30×10^{-36}	1.86×10^{-36}	1.86×10^{-36}
4	edna vs .CroMatcher	4.30×10^{-33}	1.20×10^{-31}	1.07×10^{-31}	9.02×10^{-32}	6.44×10^{-32}
5	Lily vs. Mamba	1.49×10^{-31}	4.18×10^{-30}	3.58×10^{-30}	3.14×10^{-30}	2.39×10^{-30}
6	CroMatcher vs. Mamba	2.68×10^{-27}	7.50×10^{-26}	6.16×10^{-26}	5.62×10^{-26}	2.94×10^{-26}
7	AML2014 vs. Lily	3.15×10^{-27}	8.81×10^{-26}	6.92×10^{-26}	6.61×10^{-26}	4.09×10^{-26}
8	AML2014 vs. CroMatcher	2.66×10^{-23}	7.45×10^{-22}	5.59×10^{-22}	5.59×10^{-22}	2.93×10^{-22}
9	Lily vs. XMap	5.40×10^{-20}	1.51×10^{-18}	1.08×10^{-18}	8.64×10^{-19}	7.02×10^{-19}
10	CroMatcher vs. XMap	1.11×10^{-16}	3.11×10^{-15}	2.11×10^{-15}	1.78×10^{-15}	1.22×10^{-15}
11	GMap vs. Lily	1.66×10^{-15}	4.64×10^{-14}	2.98×10^{-14}	2.65×10^{-14}	2.15×10^{-14}
12	CroMatcher vs. GMap	1.24×10^{-12}	3.46×10^{-11}	2.10×10^{-11}	1.98×10^{-11}	1.36×10^{-11}
13	GMap vs. LogMapLite	8.34×10^{-9}	2.34×10^{-7}	1.33×10^{-7}	1.33×10^{-7}	1.33×10^{-7}
14	edna vs. GMap	1.04×10^{-6}	2.93×10^{-5}	1.57×10^{-5}	1.57×10^{-5}	1.15×10^{-5}
15	XMap vs. LogMapLite	4.87×10^{-6}	1.37×10^{-4}	6.82×10^{-5}	6.33×10^{-5}	5.36×10^{-5}
16	GMap vs. Mamba	1.98×10^{-4}	0.0055	0.0025	0.0026	0.0016
17	edna vs. XMap	2.22×10^{-4}	0.0062	0.0027	0.0027	0.0016
18	AML2014 vs. LogMapLite	0.0035	0.099	0.039	0.039	0.028
19	AML2014 vs.GMap	0.0047	0.125	0.0446	0.0446	0.0312
20	XMap vs. Mamba	0.0114	0.318	0.102	0.102	0.056
21	LogMapLite vs. Mamba	0.041	1.159	0.331	0.331	0.206
22	edna vs. AML2014	0.041	1.159	0.331	0.331	0.207
23	AML2014 vs. XMap	0.098	2.756	0.590	0.590	0.295
24	GMap vs. XMap	0.233	1.00	1.00	1.00	0.93
25	edna vs. Mamba	0.245	1.00	1.00	1.00	0.934
26	edna vs. LogMapLite	0.38	1.00	1.00	1.00	1.00
27	AML2014 vs. Mamba	0.38	1.00	1.00	1.00	1.00
28	CroMatcher vs. Lily	0.388	1.00	1.00	1.00	1.00

(b) Quade test

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	edna vs. Lily	2.65×10^{-10}	7.42×10^{-9}	7.42×10^{-9}	7.42×10^{-9}	7.42×10^{-9}
2	Lily vs. LogMapLite	4.41×10^{-10}	1.23×10^{-8}	1.1×10^{-8}	9.27×10^{-9}	9.27×10^{-9}
3	edna vs. CroMatcher	5.64×10^{-9}	1.58×10^{-7}	1.47×10^{-7}	1.18×10^{-7}	1.18×10^{-7}
4	CroMatcher vs. LogMapLite	9.04×10^{-9}	2.53×10^{-7}	2.26×10^{-7}	1.89×10^{-7}	1.36×10^{-7}
5	AML2014 vs. Lily	5.25×10^{-8}	1.47×10^{-6}	1.26×10^{-6}	1.10×10^{-6}	8.40×10^{-7}
6	Lily vs. Mamba	1.45×10^{-7}	4.06×10^{-6}	3.33×10^{-6}	3.04×10^{-6}	1.89×10^{-6}
7	AML2014 vs. CroMatcher	7.36×10^{-7}	2.06×10^{-5}	1.62×10^{-5}	1.54×10^{-5}	8.09×10^{-6}
8	Lily vs. XMap	7.60×10^{-7}	2.13×10^{-5}	1.62×10^{-5}	1.60×10^{-5}	9.88×10^{-6}
9	CroMatcher vs. Mamba	1.86×10^{-6}	5.21×10^{-5}	3.72×10^{-5}	2.98×10^{-5}	2.05×10^{-5}
10	CroMatcher vs. XMap	8.41×10^{-6}	2.36×10^{-4}	1.60×10^{-4}	1.35×10^{-4}	9.26×10^{-5}
11	GMap vs. Lily	1.48×10^{-4}	0.0041	0.0026	0.0023	0.0019
12	CroMatcher vs. GMap	9.57×10^{-4}	0.0267	0.0163	0.0153	0.0105
13	edna vs. GMap	0.011	0.325	0.186	0.186	0.186
14	GMap vs. LogMapLite	0.0144	0.405	0.217	0.217	0.186
15	AML2014 vs. GMap	0.099	1.00	1.00	1.00	0.793
16	GMap vs. Mamba	0.142	1.00	1.00	1.00	1.00
17	edna vs. XMap	0.170	1.00	1.00	1.00	1.00
18	XMap vs. LogMapLite	0.195	1.00	1.00	1.00	1.00
19	GMap vs. XMap	0.249	1.00	1.00	1.00	1.00
20	edna vs. Mamba	0.290	1.00	1.00	1.00	1.00
21	LogMapLite vs. Mamba	0.327	1.00	1.00	1.00	1.00
22	edna vs. AML2014	0.381	1.00	1.00	1.00	1.00
23	AML2014 vs. LogMapLite	0.426	1.00	1.00	1.00	1.00
24	AML2014 vs. XMap	0.619	1.00	1.00	1.00	1.00
25	CroMatcher vs. Lily	0.623	1.00	1.00	1.00	1.00
26	XMap vs. Mamba	0.754	1.00	1.00	1.00	1.00
27	AML2014 vs. Mamba	0.854	1.00	1.00	1.00	1.00
28	edna vs. LogMapLite	0.937	1.00	1.00	1.00	1.00

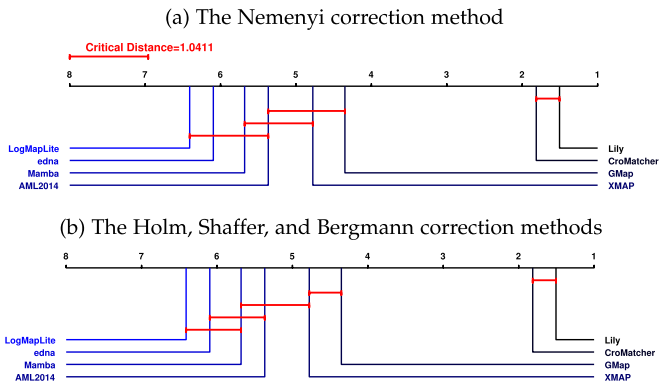


Fig. 6. The critical difference diagrams for the Friedman test with four p-value adjustment methods on the benchmark track: (a) The Nemenyi correction. (b) The Holm, Shaffer, and Bergmann correction. The x-axis is the average rank of each system obtained by the Friedman test.

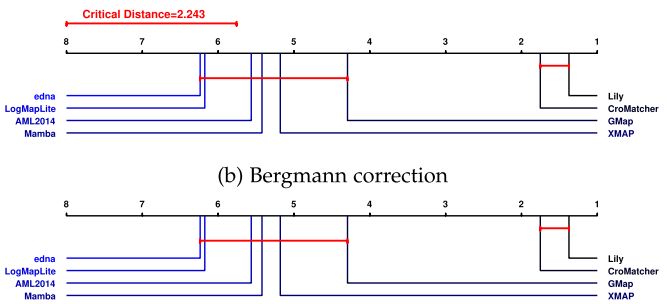


Fig. 7. The critical difference diagrams for the Quade test with four p-value adjustment methods on the benchmark track: (a) The Nemenyi correction. (b) The Holm, Shaffer, and Bergmann correction. The x-axis is the average rank of each system obtained by the Quade test.

TABLE 10
Average Rankings of Systems Computed by Friedman and Quade Tests

	AML	LogMap	CLONA	XMap
Friedman	1.07	2.48	2.68	3.77
Quade	1.05	2.51	2.56	3.88

The results of this track are in accordance with the theory. First, the Nemenyi correction is so conservative and detect fewer differences among alignment systems. Further, the Friedman test has more power than the Quade test when a sufficient number of datasets is available.

Last but not least, the results of this section is compared with the averaging. The average of Lily and CroMatcher systems, which are top two systems in the OAEI 2015 benchmark track is 0.90 and 0.88, respectively. These are indiscernibly the top systems from the statistical analysis point of view as well. At the other extreme, edna and LogMapLite are the worse ones with the average 0.41 and 0.46, respectively. Similarly, these systems are also the worst ones regarding the statistical analysis.

There are some small difference between the ranking of systems from averaging and the statistical analysis. For instance, AML2014 has a lower rank than Mamba from the statistical view while the latter system is claimed to have outperformed the other one with respect to averaging. However, the major difference between averaging and the statistical analysis is that several systems are declared insignificant. This seems rational since we cannot indicate the superiority of one system merely if its average is slightly higher than one another.

TABLE 11
The Adjusted p-Values by Four p-Value Adjustment Methods: (a) The p-Value Adjustment for the Friedman Test and (b) the p-Value Adjustment for the Quade Test

(a) Friedman test

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	AML vs. XMap	5.10×10^{-23}	3.06×10^{-22}	3.06×10^{-22}	3.06×10^{-22}	3.06×10^{-22}
2	AML vs. CLONA	4.13×10^{-9}	2.48×10^{-8}	2.05×10^{-8}	1.24×10^{-8}	1.24×10^{-8}
3	AML vs. LogMap	2.69×10^{-7}	1.61×10^{-6}	1.07×10^{-6}	8.07×10^{-7}	5.38×10^{-7}
4	LogMap vs. XMap	2.18×10^{-6}	1.31×10^{-5}	6.55×10^{-6}	6.55×10^{-6}	6.55×10^{-6}
5	CLONA vs. XMap	6.31×10^{-5}	3.79×10^{-4}	1.26×10^{-4}	1.26×10^{-4}	6.31×10^{-5}
6	CLONA vs. LogMap	0.462	1.00	0.462	0.462	0.462

(b) Quade test

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	AML vs. XMap	1.52×10^{-13}	9.14×10^{-13}	9.14×10^{-13}	9.14×10^{-13}	9.14×10^{-13}
2	AML vs. CLONA	8.04×10^{-5}	4.83×10^{-4}	4.02×10^{-4}	2.41×10^{-4}	2.41×10^{-4}
3	AML vs. LogMap	1.28×10^{-4}	7.67×10^{-4}	5.10×10^{-4}	3.83×10^{-4}	2.55×10^{-4}
4	LogMap vs. XMap	3.79×10^{-4}	0.0022	0.0011	0.0011	0.0011
5	CLONA vs. XMap	5.77×10^{-4}	0.0034	0.0011	0.0011	0.0011
6	CLONA vs. LogMap	0.91	1.00	1.00	1.00	1.00

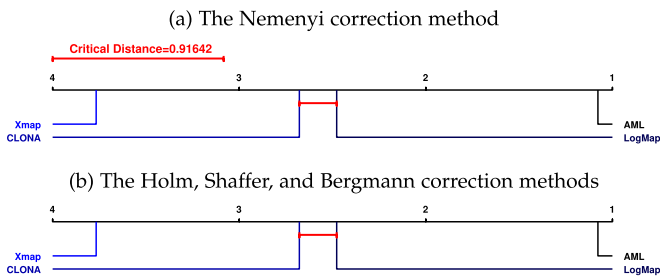


Fig. 8. The critical difference diagrams for the Friedman test with four p-value adjustment methods on the multifarm track: (a) The Nemenyi correction. (b) Holm, Shaffer, and Bergmann correction methods. The x-axis is the average rank of each system obtained by the Friedman test.

6.2.2 Multifarm *Track*

Another track in the OAEI which is considered here is *multifarm*. There are 47 pairs of ontologies which are matched by various systems. We take 4 of them (AML [45], CLONA [46], LogMap [34] and XMap [35]) which could produce acceptable mappings in the OAEI 2015. Then, we apply the statistical procedures over F-measure obtained for each dataset to determine the systems with improved performance.

The ranks computed by the Friedman and Quade tests are presented in Table 10.

The Friedman statistic (with 3 degrees of freedom) and its p-value are 98.80 and 5.80×10^{-11}, respectively. Similarly, the Quade statistic is computed as 138.30 with $(3, 46)$ degrees of freedom, and the corresponding p-value is approximately zero. Thus, both tests reject the null hypothesis, and it is concluded that there is a significant difference among the performances.

The post-hoc procedure is applied to F-measure of the aforementioned methods over the datasets in the *multifarm* track. The adjusted p-values of various post-hoc procedures are presented in Table 11. Based on this table, it can be easily understood what systems are significantly different from each other given the significance level α.

Similar to the *benchmark* track, we visualize the results obtained over this track. The critical difference diagrams of

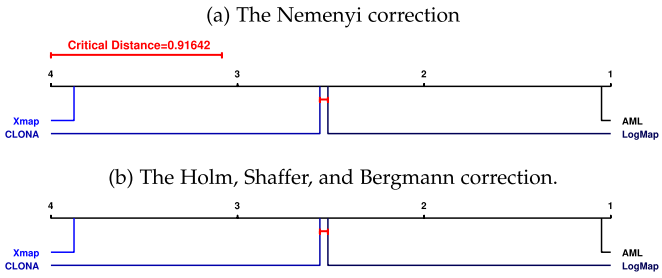


Fig. 9. The critical difference diagrams for the Quade test with four p-value adjustment methods on the multifarm track: (a) The Nemenyi correction. (b) The Holm, Shaffer, and Bergmann correction. The x-axis is the average rank of each system obtained by the Quade test.

statistical tests with correction methods are plotted in Figs. 8 and 9 where the x-axis indicates the rank of each system obtained by Friedman and Quade tests. In this plot, the methods which are not significantly different are connected to each other by a line. The results of various tests over this track are the same. The Friedman and Quade tests with each method of correction indicate that AML is the best and XMap is the worst system. Further, CLONA and LogMap are not significantly different, but they are better than XMap and worse than AML.

6.3 Summary

The recommendation for utilization of tests are summarized as the following

- For comparison of two systems and with large enough datasets (> 30 datasets), the normality test is first conducted to check the normality of differences. If the normality assumption holds, the paired t-test is the most appropriate statistic. Otherwise, the Wilcoxon signed-rank test is preferred.
- For comparison of two system with a moderate number of datasets (less than 30 but above 10), the test of normality is not reliable. Among the nonparametric tests, the Wilcoxon signed-rank test is preferred. In addition, if the number of datasets is less than ten, McNemar's asymptotic or mid-p tests are recommended.
- For the case of comparison among multiple systems, the repeated measures ANOVA is not recommended and its use must be prohibited. Instead, Friedman and Quade tests are recommended for the moderate or large (more than 10) and the small (less than 10) number of datasets, respectively.
- For controlling FWER, Bergmann correction is the most powerful one and is highly recommended. However, it takes a lot of time to conduct the comparison if there are more than ten systems. If there is any time restriction and there are more than ten systems, Shaffer correction is recommendable which is powerful and fast. The Nemenyi correction is too conservative, and its use should be prohibited.

7 Conclusion

The statistical methodologies for comparison of two or more alignment systems were studied in this paper. For comparison of two systems, three different situations related to the number of datasets were considered and an appropriate test was recommended for each of the case. For comparison of multiple systems, the use of ANOVA was prohibited due to its severe presumption *sphericity*. Instead, Friedman and Quade tests were proposed for comparison. For comparison of multiple systems, the family-wise error rate and the ways to prevent it are elaborated in details.

References

[1] A. Isaac, S. Wang, C. Zinn, H. Matthezing, L. Van der Meij, and S. Schlobach, "Evaluating thesaurus alignments for semantic interoperability in the library domain," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 76–86, Mar./Apr. 2009.
[2] P. P. Talukdar, Z. G. Ives, and F. Pereira, "Automatically incorporating new sources in keyword search-based data integration," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2010, pp. 387–398.
[3] M. Ba and G. Diallo, "Large-scale biomedical ontology matching with ServOMap," *Innovation Res. BioMed. Eng.*, vol. 34, no. 1, pp. 56–59, 2013.
[4] U. Bharambe, S. S. Durbha, and R. L. King, "Geospatial ontologies matching: An information theoretic approach," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 2918–2921.
[5] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.
[6] D. H. Wolpert, "What the no free lunch theorems really mean; how to improve search algorithms," in *Santa Fe Inst. Work. Paper*, 2012, Art. no. 12.
[7] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. Jan, pp. 1–30, 2006.
[8] S. Garcia and F. Herrera, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *J. Mach. Learn. Res.*, vol. 9, no. Dec, pp. 2677–2694, 2008.
[9] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.
[10] D. A. Hull, "Information retrieval using statistical classification," PhD thesis, Dept. Statistics, Citeseer, 1994.
[11] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.
[12] B. T. Nski, M. S. Etek, Z. Telec, and T. Lasota, "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," *Int. J. Appl. Math. Comput. Sci*, vol. 22, no. 4, pp. 867–881, 2012.
[13] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 3–18, 2011.
[14] J. Euzenat, P. Shvaiko et al., *Ontology Matching*, vol. 18. Berlin, Germany: Springer, 2007.
[15] M. Mohammadi, A. A. Atashin, W. Hofman, and Y. Tan, "Comparison of ontology alignment algorithms across single matching task via the McNemar test," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 4, Jun. 2018.
[16] J. Euzenat, M.-E. Roşoiu, and C. Trojahn, "Ontology matching benchmarks: Generation, stability, and discriminability," *Web Semantics: Sci. Serv. Agents World Wide Web*, vol. 21, pp. 30–48, 2013.
[17] C. Meilicke, R. GarcíA-Castro, F. Freitas, W. R. Van Hage, E. Montiel-Ponsoda, R. R. De Azevedo, H. Stuckenschmidt, O. SVáB-Zamazal, V. Svátek, A. Tamilin et al., "MultiFarm: A benchmark for multilingual ontology matching," *Web Semantics: Sci. Services Agents World Wide Web*, vol. 15, pp. 62–68, 2012.
[18] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
[19] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
[20] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2003.

[21] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statistical Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.
[22] D. Quade, "Using weighted rankings in the analysis of complete blocks with additive block effects," *J. Amer. Statistical Assoc.*, vol. 74, no. 367, pp. 680–683, 1979.
[23] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ. Lett.*, vol. 6, no. 3, pp. 255–259, 1980.
[24] J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaisé, C. Meilicke, A. Nikolov, J. Pane, M. Sabou et al., "Results of the ontology alignment evaluation initiative 2009," in *Proc. 4th Int. Conf. Ontology Matching*, 2009, pp. 73–126.
[25] W. Li and Q. Sun, "GMap: Results for OAEI 2015," *Ontology Matching*, vol. 1, 2015, Art. no. 150.
[26] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri, "A Bayesian Wilcoxon signed-rank test based on the Dirichlet process," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, 2014, pp. 1026–1034.
[27] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional," *BMC Med. Res. Methodology*, vol. 13, no. 1, 2013, Art. no. 1.
[28] A. L. Edwards, "Note on the correction for continuity in testing the significance of the difference between correlated proportions," *Psychometrika*, vol. 13, no. 3, pp. 185–187, 1948.
[29] H. Lancaster, "Significance tests in discrete distributions," *J. Amer. Statistical Assoc.*, vol. 56, no. 294, pp. 223–234, 1961.
[30] J. H. Drew, "Modern data analysis: A first course in applied statistics," *Technometrics*, vol. 33, no. 4, pp. 487–488, 1991.
[31] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Ann. Math. Statist.*, vol. 11, no. 2, pp. 204–209, 1940.
[32] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the fbietkan statistic," *Commun. Statist.-Theory Methods*, vol. 9, no. 6, pp. 571–595, 1980.
[33] J. Hodges, E. L. Lehmann et al., "Rank methods for combination of independent experiments in analysis of variance," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 482–497, 1962.
[34] E. Jiménez-Ruiz and B. C. Grau, "LogMap: Logic-based and scalable ontology matching," in *Proc. Int. Semantic Web Conf.*, 2011, pp. 273–288.
[35] W. E. Djeddi and M. T. Khadir, "XMAP: A novel structural approach for alignment of OWL-Full ontologies," in *Proc. Int. Conf. Mach. Web Intell.*, 2010, pp. 368–373.
[36] P. Nemenyi, "Distribution-free multiple comparisons," PhD thesis, Dept. Math., Princeton Univ., Princeton, NJ, USA, 1963.
[37] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian J. Statist.*, vol. 6, pp. 65–70, 1979.
[38] J. P. Shaffer, "Modified sequentially rejective multiple test procedures," *J. Amer. Statistical Assoc.*, vol. 81, no. 395, pp. 826–831, 1986.
[39] B. Bergmann and G. Hommel, "Improvements of general multiple test procedures for redundant systems of hypotheses," in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*. Berlin, Germany: Springer, 1988, pp. 100–115.
[40] R. R. Bouckaert, "Estimating replicability of classifier learning experiments," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, Art. no. 15.
[41] Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, P. Lambrix et al., "Results of the ontology alignment evaluation initiative 2014," in *Proc. 9th Int. Conf. Ontology Matching*, 2014, pp. 61–104.
[42] M. Gulic and B. Vrdoljak, "CroMatcher-results for OAEI 2013," in *Proc. 8th Int. Conf. Ontology Matching*, 2013, pp. 117–122.
[43] P. Wang and B. Xu, "Lily: Ontology alignment results for OAEI 2008," in *Proc. 3rd Int. Conf. Ontology Matching*, 2008, pp. 167–175.
[44] C. Meilicke, "MAMBA-results for the OAEI 2015," in *Proc. 10th Int. Workshop Ontology Matching—Ontology Alignment Evaluation Initiative*, 2015, Art. no. 181.
[45] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, "The agreementMakerLight ontology matching system," in *Proc. On Move Meaningful Internet Syst. Confederated Int. Conf.*, 2013, pp. 527–541.
[46] M. El Abdi, H. Souid, M. Kachroudi, and S. B. Yahia, "CLONA results for OAEI," in *Proc. 10th Int. Workshop Ontology Matching—Ontology Alignment Evaluation Initiative*, 2015.

Majid Mohammadi received the BSc and MSc degrees in software engineering and artificial intelligence, respectively. He is working toward the PhD degree in the Information and Communication Technology Group, Department of Technology, Policy and Management, Delft University of Technology. His main research interests include semantic interoperability, machine learning, and pattern recognition.

Wout Hofman is a senior research scientist with TNO, the Dutch organization for applied science, on the subject of interoperability with a specialization in government (e.g., customs) and business interoperability in logistics. He is responsible for coordinating semantic developments within the iCargo project.

Yao-Hua Tan is a professor of information and communication technology with the ICT Group, Department of Technology, Policy, and Management, Delft University of Technology and part-time professor of electronic business with the Department of Economics and Business Administration of the Vrije University Amsterdam.