Improving Peer Grading Reliability with Graph Mining Techniques

http://dx.doi.org/10.3991/ijet.v11i07.5878

Nicola Capuano[1], Santi Caballé[2] and Jorge Miguel[2]
[1] University of Salerno, Fisciano (SA), Italy
[2] Open University of Catalonia, Barcelona, Spain

***Abstract*—Peer grading is an approach increasingly adopted for assessing students in massive on-line courses, especially for complex assignments where automatic assessment is impossible and the ability of tutors to evaluate and provide feedback at scale is limited. Unfortunately, as students may have different expertise, peer grading often does not deliver accurate results compared to human tutors. In this paper, we describe and compare different methods, based on graph mining techniques, aimed at mitigating this issue by combining peer grades on the basis of the detected expertise of the assessor students. The possibility to improve these results through optimized techniques for assessors' assignment is also discussed. Experimental results with both synthetic and real data are presented and show better performance of our methods in comparison to other existing approaches.**

***Index Terms*—Peer Grading, Assessment, MOOCs, e-Learning, Graph Mining.**

I. INTRODUCTION

Massive Open Online Courses (MOOCs) are becoming increasingly popular with millions of students enrolled, thousands of courses offered and hundreds of educational institutions involved. According to [1], they represent *a continuation of the trend in innovation, experimentation and use of technology initiated by distance and on-line learning, to provide learning opportunities for large numbers of learners*.

Due to their scale, MOOCs introduce new technical and pedagogical challenges that require overcoming the traditional e-learning model based on tutor assistance to maintain a cheap and unrestricted access to high quality resources. Because of both the high numbers of students enrolled and the relatively small number of tutors, in fact, tutor involvement during delivery stages has to be limited to the most critical tasks.

In [2], the key challenges that MOOCs designers and providers are facing are analysed. Among these challenges, assessment is one of the most prominent. Given their discrepancy in number, it is not possible for the tutors to follow up with every student and review assignments individually. This also represents a major obstacle to the *credential programs* launched by MOOC players and targeted to people that want to achieve credits toward a degree or earn credentials to show to prospective employers.

A typical approach to overcome the assessment problem is to use close questions in exams and assignments so that grading can be done automatically [3]. Unfortunately, automated grading is limited, disappointing and insufficient, with no partial marks and, in some cases, with no detailed explanations of answers. It may result particularly unsatisfactory when applied to complex tasks like the evaluation of the students' ability of proving mathematical statements, expressing their critical thinking over an issue, demonstrating proficiency in skills like creative writing, etc. [4][5].

For these tasks, an approach that is gaining a growing consensus is *Peer Grading* where students are required to grade a small number of their peers' assignments as part of their own assignment. The final grade of each student is then obtained by combining information provided by peers. The positive aspect of this approach is its capability of easily scale to any size: the number of assessors in fact naturally grows with the number of students. Conversely the use of peer grading may be seen as unprofessional and unreliable given that it is based on grades assigned by unreliable graders (students) lacking the needed expertise, both didactical and on the specific subject to be assessed.

This paper aims at mitigating this issue by defining new alternative approaches, based on graph theory, to combine together peer grades coming from students, different from standard operators like median or mean. Described methods are capable of weighting grades provided by students according to assessors' proficiency in the subject matter. In this way the opinion of high-skilled students has a greater impact on the final grade than that of low-skilled ones.

Moreover, smart student-assessor assignments methods are presented. Differently from random assignment, such methods try to balance the number of reliable and unreliable assessors throughout the set of assignments. This way, situations where a student is evaluated only using grades proposed by unreliable assessors are avoided and the reliability is constant among all assessments.

The paper is organized as follows. The next section presents related work on peer grading as well as some existing aggregation approaches proposed by recent literature. Section 3 presents the approaches we defined both for the aggregation of peer grades and for the assessors-assessee assignment. Defined approaches are evaluated in Section 4 with synthetic data and, in Section 5, with data coming from an experiment with real students. Obtained results are compared with results coming from other methods discussed in Section 3. Eventually, Section 6 summarizes conclusions and outlines on-going work.

II. RELATED WORK

Peer grading has been used for many years as a tool to improve learning outcomes. The literature reports on many learning benefits for peer-assessors like the exposure to different approaches, the development of self-learning abilities, the enhancement of critical thinking, etc. [6].

Even if some studies suggest a good correlation between peer grading results and instructor ratings in conventional classrooms and online courses (at least for specific, high structured domains) [7], there is still a general concern on the use of peer grading as a reliable strategy to approximate instructor marking in massive contexts like MOOCs. Moreover, students themselves seem to distrust the results of peer grading.

In order to address the issue of accuracy of peer grading, several approaches, at various stages of development, have been proposed so far. For example, the *Calibrated Peer Review* (CPR) method [8] proposes a calibration step to be performed by students before starting to assess other students' assignments. During the calibration, each student rates the same small set of assignments that have been already rated by the instructor. The discrepancy between grades provided by a student and the instructor measures her accuracy in assessment and is then used to weight subsequent assessments provided by that student. The more accurate is an assessor the more weight is given to her judgment in the peer grading task.

CPR has been experimented in several contexts demonstrating to be an effective instructional tool. Despite that, it requires additional work from those students who are asked to take part in the calibration step. Moreover, this method does not take into account the progresses that students make over time until a new calibration step is performed. For this reason, additional approaches have been defined able to automatically tune peer grades based on different parameters.

In [9], three probabilistic models for tuning peer-provided grades are presented. Such models estimate the reliability of each assessor as well as her *bias* (i.e., a score reflecting the assessor's tendency to inflate or deflate her grade) based on the analysis of grading performance on special "ground truth" submissions that are evaluated either by the instructor or by a big number of peers (hypothesising that the mean of many grades should tend toward the correct grade). Reliability and bias of each student are then used to tune the provided grades to other (non ground-truth) submissions.

A similar approach has been applied in [10], where a Bayesian model has been used to calculate the bias of each peer assessor in general, on each item of an assessment rubric and as a function of the assessor grade assigned by the instructor. As in the previous case, obtained biases are used to tune the grades provided during peer grading. Differently from the previous method, bias calculation is based on the results of a whole round of assessment rather than on just few "ground truth" submissions so, in the calibration step, the instructor should rate all the submissions.

The *Vancouver* algorithm [11] measures the grading accuracy of a student by comparing the grades given by her to each assignment with the average grade for that assignment. Differently from the other approaches, the assessor accuracy is used as a modifier of the assessor's grade rather than of the assesee's in order for the student's grade to reflect not only the quality of her homework but also the quality of her work as a reviewer.

The *PeerRank* method [12] builds a grade for a given student by weighting the grades proposed by her assessors on the basis of the grades received by assessors themselves. In other words, the grade received by a student is considered a measure of her ability to correctly rate other students. Given that students' grades recursively depend on other student's grades, an iterative graph-mining algorithm based on an equation similar to that used in *PageRank* [13] is proposed for their calculation.

Differently from other methods, *PeerRank* does not require any instructor's intervention. Indeed, there is no need to have a ground truth of professionally graded assignments. The same author has also proposed an improvement to the basic method that includes, as a component of the final grade of a student, the accuracy of proposed evaluations with respect to the average grades proposed by other peers. Such component is seen as an incentive for students to grade correctly.

Given the promising results shown in [12], we have implemented the *PeerRank* algorithm and have used it as the starting point for this study. Preliminary results have been already published in [14]. In the next sections, we discuss the methods that we have defined for improving peer grading results as well as obtained experimental results.

III. THE DEFINED METHODS

In a typical peer grading scenario an *assignment* is given to n different students. Each student elaborates her own solution (e.g. an essay, a set of answers to open-ended questions, etc.) generating a *submission*. Each student has then to grade m different submissions (with $m < n$) coming from other students (maybe based on an assessment rubric).

The assignment of submissions to assessor students is performed in accordance to an *assessment grid* i.e. a Boolean $n \times n$ matrix A where $A_{i,j} = 1$ iff student j has to grade the submission of the student i. The matrix A has the following properties:

1. the sum of the elements in each row and column is equal to m (i.e. each student grades and is graded by m other students);
2. the sum of the elements in the main diagonal is equal to 0 (i.e. none evaluates himself).

The assessment grid can be seen as the adjacency matrix of an m-regular directed graph where each node represents a student and each arc represents an assessment to be performed by a student on another one.

The easiest way to build the assessment grid is by filling it at random with an algorithm that preserves the properties above. A feasible algorithm starts with a null matrix and initialises its elements according to the following equation:

$$A_{mod(i+j-1,n)+1,i} = 1 \qquad (1)$$

for each $1 \leq i \leq n$ and $1 \leq j \leq m$, where *mod* indicates the remainder after division of the first term by the second one. The obtained matrix is then randomly shuffled in several iterations by selecting a couple of rows (or columns) i and j so that $A_{i,j} = A_{j,i} = 0$ and swapping them.

The grades proposed by students are then collected in the *grades matrix* G where $G_{i,j}$ is the grade proposed by the student j for the student i and $0 \leq G_{i,j} \leq 10$. In an ideal peer grading setting, every student performs the grading task so, the *final grade* g_i of each student i is obtained starting from the matrix G, by averaging all the grades obtained by peers (a matrix row) with the following equation:

$$g_i = \frac{1}{m}\sum_{j=1}^{n} G_{i,j} \quad \forall\, 1 \leq i \leq n. \qquad (2)$$

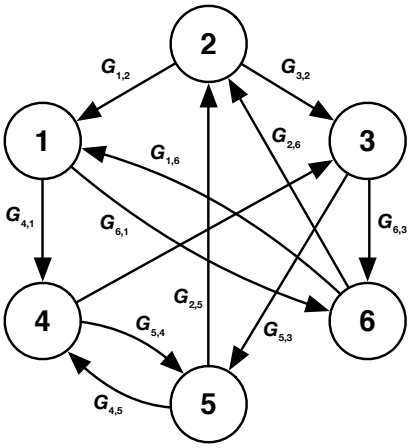

Figure 1. Graph interpretation of peer grading. Each node represents a student, each arc connects an assessor with an assessee (according to the assessment grid), the weight on each arc is a proposed grade (according to the grades matrix).

The same equation can be applied to non-ideal settings (when some students skip the grading task) by averaging on the total number $m'_i < m$ of grades proposed for *i*. Some authors propose to average all obtained grades apart the best and the worst [11]. Other authors use the median in place of the average. Despite that, the average is the most used aggregator and is the baseline against which we compare other aggregators proposed in next subsections.

It is worth noting that the grades matrix can be seen as a weighted *m*-regular directed graph where each node represents a student and each arc is the grade proposed by a student for another student. Figure 1 illustrates a sample peer grading setting with 6 students ($n = 6$) and 2 submissions to be rated by each ($m = 2$).

A. PeerRank

In [12] the author proposed to weight the grade that each assessor student gives to another student by her own grade i.e. to use the grade of a student as a measure of her ability to grade correctly. In other words, the grade g_i of a student *i* is so that:

$$g_i = \frac{\sum_{j \to i} G_{i,j} \cdot g_j}{\sum_{j \to i} g_j} \tag{3}$$

where both summations are performed over all students *j* having evaluated *i* (indicated with $j \to i$) i.e. so that $A_{i,j} = 1$.

Given that the grades of all assessor students are themselves weighted averages of grades obtained by their own assessors, an iterative process, named *PeerRank*, was proposed to calculate the final grade of each student. Let be g_i^t the grade of the student *i* at the *t*-th iteration, the grade of *i* at the iteration $t + 1$ is defined as follows:

$$g_i^{t+1} = (1 - \alpha) g_i^t + \alpha \frac{\sum_{j \to i} G_{i,j} \cdot g_j^t}{\sum_{j \to i} g_j^t} \tag{4}$$

where $0 \leq \alpha \leq 1$ is a constant affecting the convergence speed and g_i^0 is initialised according to Equation 2.

Equation 4 takes into account that each student only evaluates *m* peers according to the *assessment grid*. This is a more realistic setting with respect to the one described in [12] where each student was assumed to evaluate any other student. The author has also demonstrated useful properties for the defined grade updating rule as well as that after a limited number of iterations, it converges to stable values.

Given that the proposed equation does not incentivize students to evaluate their peers accurately, the same author defined an updated version of Equation 4 as follows:

$$g_i^{t+1} = (1 - \alpha - \beta) g_i^t + \alpha \frac{\sum_{j \to i} G_{i,j} \cdot g_j^t}{\sum_{j \to i} g_j^t} + \beta \frac{\sum_{j \to i} 10 - \left| G_{j,i} - g_j^t \right|}{m} \tag{5}$$

where $0 \leq \beta \leq 1$ is a constant, so that $\alpha + \beta \leq 1$, that weights the reward given to a student according to the inverse normalised absolute error in the grades provided by her.

If $\beta = 0$ then Equation 5 degenerates to Equation 4. For $\beta > 0$, if $G_{j,i} = g_j^t$ for all *j*, then the grades assigned by *i* are all exact and the contribution of the third addendum is $10 \cdot \beta$. At the opposite, if $\left| G_{j,i} - g_j^t \right| = 10$ for all *j* then the grades assigned by *i* are completely wrong and the contribution of the third addendum is 0.

B. F-PeerRank

The *PeerRank* rule, described by Equation 5, prescribes that the influence of the grade of an assessor student on any grade she proposes is linear. For sake of simplicity we can decompose Equation 5 as the sum of tree different components as follows:

$$g_i^{t+1} = (1 - \alpha - \beta) g_i^t + \alpha \gamma_i^t + \beta \delta_i^t \tag{6}$$

where the constants α and β have the same meaning as in Equation 5, γ_i^t is the contribution coming from peer graders while δ_i^t is the incentive for accurate grading.

In order to improve the quality of the final grades, we propose an updated rule named *F-PeerRank* that applies a super-linear modifier to the grades proposed by peer assessors by modifying the γ_i^t component as follows:

$$\gamma_i^t = \frac{\sum_{j \to i} G_{i,j} \cdot f\left(g_j^t\right)}{\sum_{j \to i} f\left(g_j^t\right)}. \tag{7}$$

The function *f* that affects the contribution of rates proposed by other peer has the purpose of minimizing the contribution of low skilled student while maximising those of high skilled ones. In [14] we have proposed the *ExpPeerRank* rule where $f(x) = e^x$ with good results on synthetic data. As we will see in section 4, on real data an alternative rule (that we name *PowPeerRank*) where $f(x) = x^n$ (for some *n*) outperforms the *ExpPeerRank* rule.

C. BestPeer

Bringing this reasoning to the extreme, we can imagine to assign the maximum influence only to the best grader for each student and no influence at all to any other proposed grade. This is the case of another rule we propose, named *BestPeer*. It calculates the grade g_i for any student *i* with one of the previous methods and then assigns to each student the final grade g'_i according to the following rule:

$$g'_i = G_{i, \underset{j \to i}{argmax}\, g_j} \tag{8}$$

where the function *argmax* (argument of the maximum) returns the value *j* so that g_j is maximized.

This method is capable of performing particularly well when, for each student, at least one good grader is available. Unfortunately, this condition cannot be granted with the random assessor-assessee assignment proposed in Equation 1. So, in the next sub-section, we discuss an alternative assignment method that, under certain conditions, can overcome the limitations of the random assignment.

D. Smart Assignment

The randomized assessor-assessee assignment can generate settings in which some student is assessed by only unreliable graders (i.e. students with a low grade). In this case, even weighting the grades, the overall peer-assessment performance may be poor. Balancing reliable graders among students is a feasible approach to overcome this issue but, unfortunately, we have no information about the grades when the assessment grid is built.

To overcome this issue it is possible to initialize the assessment grid based on grades coming from previous assessments. To do that, a feasible algorithm starts with a null matrix and initialises its elements according to the following equation:

$$A_{mod(m(i-1)+j-1),n)+1,rank(i)} = 1 \quad (9)$$

for each $1 \leq j \leq n$ and $1 \leq i \leq m$ and where *rank*(*i*) is the position of the *i*-th student in the *student ranking* i.e. the list of the students ordered decreasingly on the average grade obtained in previous assessments.

Equation 9 does not ensure the fulfilment of the second property of assessment grids. For this reason an additional check is needed and, if $A_{i,i} = 1$ for some i: $1 \leq i \leq n$, then the closest column j so that $A_{i,j} = 0$ and:

$$\exists z \mid A_{z,i} = 0 \; and \; A_{z,j} = 1 \quad (10)$$

is selected and the values of $A_{i,i}$ and $A_{i,j}$ are swapped as well as values of $A_{z,i}$ and $A_{z,j}$. In other words, the assessor *i* does not assess himself but the student *z* assigned to the closest performer *j* that, in return, takes care of evaluating *i*.

A second option for optimizing the assessor-assessee assignment is to proceed incrementally (i.e., to perform the assessment session in *m* rounds). At the first round, each student is randomly assigned just one student to grade. At each subsequent round students are ranked in two lists:

1. list 1 orders students, decreasingly, on the average grade obtained in the preceding rounds (it so ranks the students basing on their quality as graders);
2. list 2 orders students, increasingly, based on the average grades obtained by their graders in the preceding rounds (it so ranks the students based on the quality of obtained grades).

Then, for the subsequent round, each student from list 1 has to grade the student from the list 2 with the same rank. This ensures that, in each step, the best graders are assigned to the students that, in the previous steps, have obtained grades from the worst ones. Some additional checks must be made to ensure that no student evaluates herself and that no student evaluates another student more than once.

This method has the advantage that it does not need any information about past assessments. Conversely, its incremental nature requires that every grade is assigned for a given round before starting the next one. This constraint can be very expensive, especially in massive contexts, when some student may be late in providing grades or may not provide grades at all. For these reasons we decided to focus our attention just on the first method.

IV. EVALUATION WITH SYNTHETIC DATA

In order to evaluate the performance of the defined methods, we made seven experiments with synthetic data. In all experiments 100 students are supposed to have submitted a solution to an assignment composed of 10 questions. For a correct answer a student gains 1 point while for a wrong answer she gains 0 points. The *real grade* of each student is then an integer belonging to [0, 10].

Each student has then to evaluate the submissions of *m* other peers. In our model, we suppose that each student *i* with a real grade $\bar{g}_i$ has probability $\bar{g}_i/10$ of marking correctly each answer of a peer submission. So if the student *i* grades the submission of a student *j* (with real grade $\bar{g}_j$), then the *proposed grade* $G_{j,i}$ is a random variable so that:

$$G_{j,i} \sim B(\bar{g}_j, \bar{g}_i/10) + B(1 - \bar{g}_j, 1 - \bar{g}_i/10) \quad (11)$$

where $B(m, p)$ represents a binomial distribution of *m* trials with probability *p*.

Each experiment is made of several iterations. For each iteration, real grades are randomly assigned (with different probability distributions). Then, the *assessment grid* is built (according to different methods) and the *grades matrix* is randomly filled according to the probability distribution given in Equation 11. The *final grades* are then calculated (according to different methods) and compared to real grades by calculating the Root Mean Square Error (RMSE). The details and the results of each experiment are discussed in the next sub-sections.

A. Binomial distribution of real grades

In this experiment, the real grades are assigned according to a binomial distribution: each student, for each of the 10 questions of her assignment, has a probability *p* to answer correctly and 1 – *p* to answer wrongly. In other words, the real grade of a student *i* is assigned according to:

$$g_i \sim B\,(10, p). \quad (12)$$

In each step of the experiment a probability *p* is chosen and 1000 iterations are performed. For each iteration, the real grades are assigned as described above (with probability *p*). Then a 100×100 *assessment grid* is randomly generated according to Equation 1 so that each student evaluates 4 other peers (*m*=4). A *grades matrix*, including all proposed grades, is then randomly generated from the distribution given in Equation 11.

For each iteration, the final grade of each student is calculated as the *Average* of grades proposed by peers (Equation 2), with *PeerRank* (Equation 5), with *PowPeerRank* and *ExpPeerRank* rules (that are special cases of the *F-PeerRank* rule described by Equations 6 and 7) and with the *BestPeer* method (Equation 8). In particular, for *PowPeerRank* we have selected $f(x) = x^2$ in Equation 7 while we have used *ExpPeerRank* as the base method to obtain a first estimation of student grades in *BestPeer*.

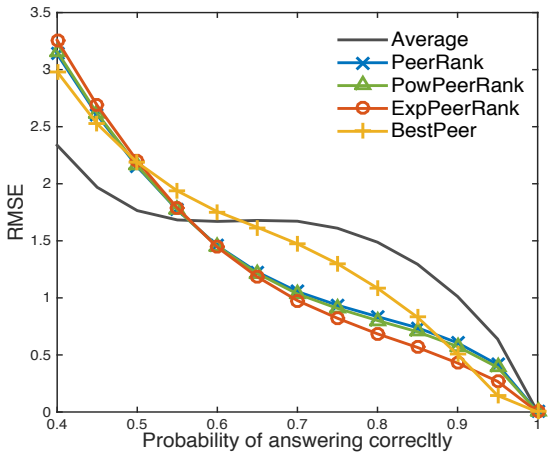

Figure 2. Experiment IV.A: performance in terms of RMSE of all methods on a binomial distribution of grades with different values for p (probability of answering correctly a question).

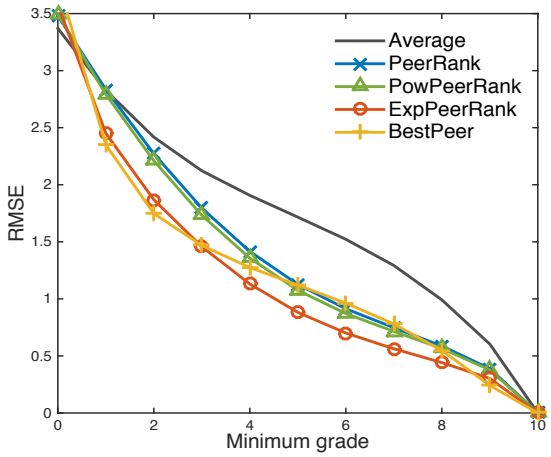

Figure 3. Experiment IV.B: performance in terms of RMSE of all methods on a uniform distribution of grades with different values for *min* (minimum grade for an assignment).

For each iteration, the RMSE between final and real grades is calculated over the 100 students. The obtained values are then mediated over all iterations.

Figure 2 plots the performance obtained by applying the five methods to the defined marking model in terms of mean RMSE against the probability p used to generate the real grades. As it can be seen both *PeerRank* and *ExpPeerRank* outperform the *Average* method for $p > 0.6$. Conversely, the performance of all methods is quite similar for $0.5 \leq p \leq 0.6$ while, for $p < 0.5$ the best method remains the *Average*.

Obtained results show that both *PeerRank* based methods need $p > 0.5$ to get any useful signal out of the data. It is worth noting that $p = 0.5$ means that students are answering (or marking) questions just as well by tossing a coin. This means that, in real contexts, assuming that $p > 0.5$ is not a so restrictive constraint. Moreover, as it can be seen, *PowPeerRank* performs a little better than *PeerRank* while *ExpPeerRank* outperforms both. Instead, *BestPeer* is better than other methods only for $p > 0.9$.

The best choice for this distribution of grades seems to be *ExpPeerRank* that ensures, in best cases, a decrease in RMSE of about 1 grade with respect to the baseline *Average* method. This means that, on average, each student will have a final grade closer to the real one of approximately 1 point over 10.

B. Uniform distribution of real grades

In this experiment, the real grades are assigned according to a uniform distribution rather than a binomial one: each student receives an integer random grade to the whole assignment from a minimum *min* (so that $0 \leq min \leq 10$) to a maximum of 10. Hence the real grade of a student i is assigned according to:

$$g_i \sim U\left(\{min, \ldots, 10\}\right) \tag{13}$$

where $U(S)$ defines a discrete uniform distribution over S.

Figure 3 plots the performance, in terms of mean RMSE against the minimum grade *min*, obtained by applying the five methods to the defined marking model with m=4 (m peers to be graded by each student) and $f(x) = x^2$ for the *PowPeerRank* rule.

Also in this case *ExpPeerRank* outperform the other methods in almost all conditions while *PowPeerRank* is a little more performant than *PeerRank*. Only for *min* = 0 the performance of all methods is quite the same.

It is interesting to note that *BestPeer* behaves better than in the experiment described in IV.A, with a RMSE lower or equal to *PeerRank*. The best performance is obtained when $min \leq 5$ (high variance of real grades) and with $min \geq 8$ (high average real grade).

This can be explained by the fact that, when there is a high variance in student levels, there is a high probability that a peer is evaluated also by unreliable graders that affect the quality of the final grade in all methods (at different levels) apart from *BestPeer* where only the best grade is selected. This advantage disappears when *min* increases because in that case, proposed grades increase their average quality.

C. Smart assignment and binomial grades distribution

This experiment replicates the Experiment IV.A with the difference that the *assessment grid* is generated according to Equations 9 and 10 rather than to Equation 1. In the model, we assume that the average grade obtained in previous assessments (needed to generate the *student ranking*) is equal to the assigned real grade. This is a simplification that supposes that students maintain a constant performance across several assignments. Given that, the results of this experiment can be considered as an upper bound of the results obtainable with smart assignment in real contexts.

Figure 4 plots the performance obtained by applying the defined methods to the marking model with random (dashed lines) and smart (plain lines) assignment methods. Given that the performance of *PowPeerRank* is quite similar to that offered by the standard *PeerRank* method, we have removed this method from the plot to maintain the readability higher.

As it can be seen, with a binomial distribution of real grades *Average*, *PeerRank* and *ExpPeerRank* are quite insensitive to smart assignment. Instead, as it might be supposed, *BestPeer* has a substantial improvement because the smart assignment ensures that each student is assessed by at least one good grader whose proposed grade is selected as the final one.

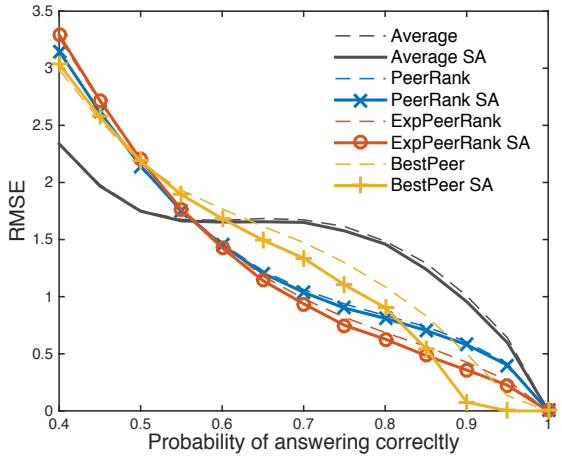

Figure 4. Experiment IV.C: performance in terms of RMSE of all methods on a binomial distribution of grades with different values for p and different assignment methods (SA stands for *Smart Assignment*).

D. Smart assignment and uniform grades distribution

This experiment replicates the Experiment IV.B with the difference that the *assessment grid* is generated according to Equations 9 and 10 rather than to Equation 1, with the same assumptions made in Experiment IV.C with respect to the average grade obtained in previous assessments.

Figure 5 plots the performance obtained by applying the four methods (also in this case we exclude *PowPeerRank* whose performance is similar to the standard *PeerRank*) to the defined marking model with random (dashed lines) and smart (plain lines) assignment methods in case of uniform distribution of real grades. In this case, while *Average* and *PeerRank* result again quite insensitive to smart assignment, *ExpPeerRank* and (to a greater extent) *BestPeer*, show a good improvement.

In particular, *BestPeer* outperforms all the other methods, especially in configurations with high grades variance ($min < 5$) and high average real grade ($min > 6$). Only for $min < 1$ its performance is comparable than that of other methods. Hence in this case, the best choice seems to be *BestPeer*, whose performance in contexts with a high variance of student levels is boosted by the smart assignment.

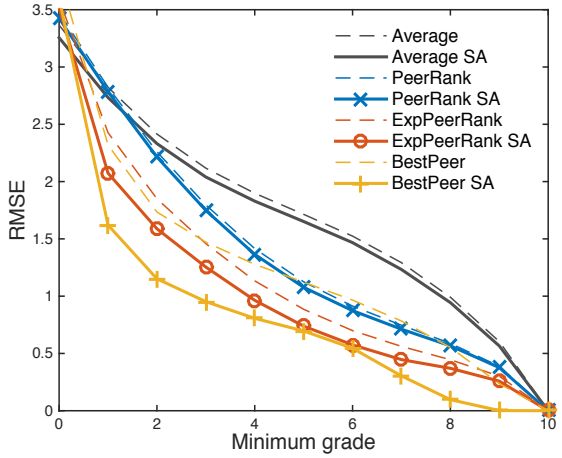

Figure 5. Experiment IV.D: performance in terms of RMSE of all methods on a uniform distribution of grades with different values for *min* and different assignment methods (SA stands for *Smart Assignment*).

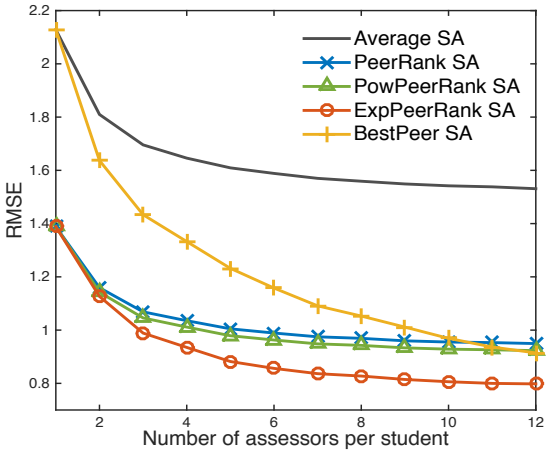

Figure 6. Experiment IV.E: performance in terms of RMSE of all methods on a binomial distribution of grades with different values for m (number of assessors per student) and *Smart Assignment*.

E. Variable assessors and binomial grades distribution

The number m of submissions that each student has to evaluate is one of the main parameters that must be defined to setup a peer grading session. On one hand, this number should be kept as small as possible to avoid overloading the students, with the risk that they do not respond adequately to the exercise providing rough, partial or void estimations. On the other hand, this number corresponds to the number of assessors for each submission. Taking this into consideration, m should be kept as big as possible to have sufficient information to estimate the final grades.

To determine how the selection of m impacts on the performance of defined methods, we have performed another experiment where the real grades are assigned according to a binomial distribution with probability $p = 0.7$ (a reasonable value in real contexts). In each step, the number m of assessors for each student is chosen from 1 to 12 and 1000 iterations are performed. For each iteration, real grades are assigned, then an *assessment grid* is generated according to Equations 9 and 10 (smart assignment) so that each student evaluates m peers. A *grades matrix*, including all proposed grades, is then randomly generated from the distribution given in Equation 11.

Figure 6 plots the performance obtained by applying the five methods to the defined marking model in terms of mean RMSE against the number of assessors m. As expected, the error decreases when the number of assessor increases but the decrease is smoother as m increases.

With *Average*, *PeerRank* and *PowPeerRank* algorithms, the increase in performance after the 4th assessor is negligible. *ExpPeerRank* offers good improvement until the 6th assessor while *BestPeer* has sensible improvements until the 10th assessor. Moreover, this latter becomes more performant of both *PeerRank* and *PowPeerRank* starting from the 12th assessor.

This fact can be explained by considering that the number of assignments evaluated by the best graders increase when more assessors are added. The impact on rules other than *BestPeer* is limited given that the resulting grades are obtained by considering also grades proposed by other assessors while the most positive impact is on *BestPeer* that only considers the grade assigned by the best grader.

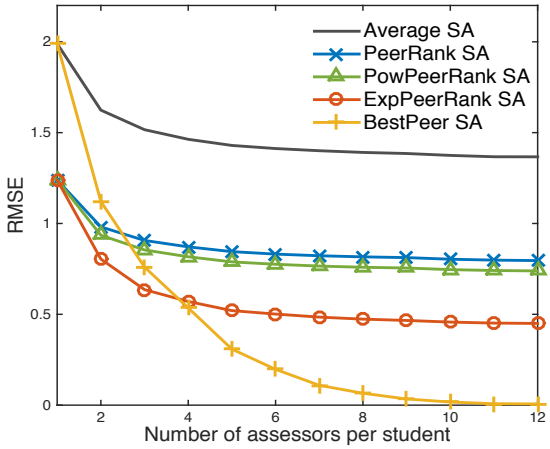

Figure 7. Experiment IV.F: performance in terms of RMSE of all methods on a uniform distribution of grades with different values for *m* (number of assessors per student) and *Smart Assignment*.

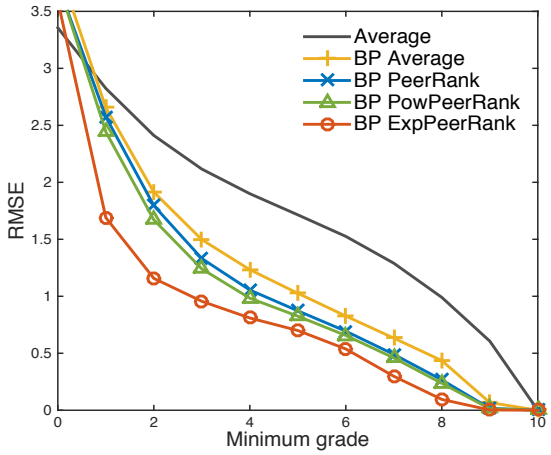

Figure 8. Experiment IV.G: performance of *BestPeer* (BP) with different support methods on a uniform distribution of grades with different values for *m* and *Smart Assignment* (against standard *Average*).

F. *Variable assessors and uniform grades distribution*

This experiment replicates the preceding one but the real grades are assigned according to a uniform distribution and each student receives an integer random grade to the whole assignment from a minimum of 6 to a maximum of 10 (i.e. $min = 6$ in Equation 13).

Figure 7 plots the performance obtained by applying the five methods to the defined marking model (with smart assignment) in terms of mean RMSE against the number of assessors m. Also in this case the error decreases when the number of assessors increases and the decrease is smoother as m increases.

It should be noted that the *BestPeer* method outperforms the other methods for $m \geq 4$. Moreover, for *BestPeer*, the RMSE asymptotically goes to 0 when the number of assessors increase. This is due to the same reasons already explained in the previous section and the effect is more evident with this distribution thanks to the high average level of the simulated class that results in a high number of reliable graders.

G. *Best Peers and support methods*

As described in Section III.B the *BestPeer* method calculates the final grade for any student with one of the other methods, then assigns to each student the grade coming from the assessor with the best final grade. In the previous experiments we used *ExpPeerRank* as support method for *BestPeer*. In this last experiment we wonder if *ExpPeerRank* is the best possible choice, at least in the configuration of Experiment IV.B.

We have so repeated Experiment IV.B only with *BestPeer* adopting different support methods. Obtained results are plot in Figure 8 against the standard *Average* method. As it might be supposed, *ExpPeerRank* (the method with the best performance in the majority of configurations) seems to be the best choice.

V. EVALUATION WITH REAL STUDENTS

To evaluate the effectiveness of the defined methods also with real users, we have applied them on peer grading data coming from a real on-line course held in Spring 2014 at the Open University of Catalonia (UOC) [15].

The on-line course had 58 students enrolled and was divided in 7 subsequent modules. After having completed the study of a module, each student received an invitation to answer three open questions. When the answers were collected, each student had to access each classmates' answers and evaluate it according to a 5-point scale (A, B, C+, C-, D) before approaching the subsequent module.

The core of the peer grading component was developed in Java and integrated in the UOC learning management system. It integrates two external Web survey applications: *Google Forms* to collect the answers to module questions and *LimeSurvey* to let students evaluate peers' answers to module questions. To exchange data between the two tools a *comma separated value* exchange model has been adopted and the *Super CSV* package has been selected to deal with such format in *Java*.

Table I shows the main statistics collected for each module. As it can be seen, the number of active students per module (students providing answers to module questions) has decreased about 70% over time: from 41 in module 1 to 12 in module 7 (on a total of 58 enrolled students). Despite it may seem discouraging, this result is in line with the problematic drop-out rate suffered by on-line courses (the mean drop-out ratio at UOC is about 50%).

Only a part of the active students also executed the peer grading task. The second row of Table I reports on the number of students that, for each module, succeeded in evaluating (at least some of) their peers. The remaining rows of Table I report the mean grade obtained by students for each question of each module normalized between 0 and 10. If we consider that the three questions are graded separately, data for 21 separate assignments is available.

TABLE I. MAIN STATISTICS OF THE PERFORMED EXPERIMENT

Modules	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Active students	41	28	23	20	21	18	12
Peer Assessors	30	24	15	14	16	11	11
Mean grade (question 1)	7.3	8.0	7.5	7.3	7.8	7.5	7.5
Mean grade (question 2)	7.0	7.3	7.5	7.5	7.3	7.5	7.3
Mean grade (question 3)	7.5	7.8	7.3	7.8	7.3	7.8	7.5

In the experiment, students were asked to grade all their peers. Conversely, in a MOOC peer grading setting, students would be asked to evaluate only a small subset of other students. In the absence of an assessment made by an expert tutor, this peculiarity allows us to calculate the *approximate real grade* $\bar{g}_i$ of a student i as the mean grade obtained by her over the whole population of assessors.

According to [9], we have assumed that the mean of many student grades should tend towards the correct peer grade, especially for the first two modules where each submission were graded by 30 (for module 1) and 24 (for module 2) peer assessors.

Starting from such data we have then performed two different experiments as detailed in the next subsections. Once the assignment is selected among the 21 available, each experiment is made of several iterations. Given an assignment, for each iteration we have supposed that just m grades were proposed (randomly selected among those available) for each active student. This allow us to simulate the real conditions of a MOOC peer grading task.

So, for each iteration, the *assessment grid* is built by randomly selecting m assessors for each active student and the *grades matrix* is filled with grades proposed by that students. The *final grades* are then calculated (with different methods) and compared to the *approximate real grade* (obtained as previously described) by calculating the RMSE.

The purpose of the experiments is to determine which of the defined methods can estimate with better accuracy the *approximate real grade* (obtained by averaging all available evaluations) using only a small number m of randomly selected evaluations per submission. Considering that the *approximate real grade* is, in turn, an estimation of the real grade, we are indirectly finding the best estimator of the real grade.

A. Fixed number of peer assessors

This experiment is made of 7 steps (one for each module) and 21 sub-steps (corresponding to the three questions for each module). For each sub-step, 1000 iterations are performed. In each iteration, 4 assessors are randomly selected for each submission (m = 4) and the *assessment grid* and the *grades matrix* are filled as previously explained. The dimension of such matrices is equal to the number of active students in the related module (from 41×41 in the first step to 12×12 in the seventh).

For each iteration, the final grade of each student is calculated as the *Average* of grades proposed by selected peers (Equation 2), with *PeerRank* (Equation 5), with *PowPeerRank* and *ExpPeerRank* rules (Equations 6 and 7) and with the *BestPeer* method (Equation 8). The RMSE between final and real grades is calculated for each iteration over the active students.

Table II summarizes the performance obtained by the defined methods on the experimental data. The reported RMSE values are mediated over all iterations for each sub-step and over all stub-steps for each step.

As it can be seen both *PeerRank* and *PowPeerRank* outperform the *Average* method in all conditions. They show a better accuracy in predicting the *approximate real grade* even with a small number of available evaluations for each student. Conversely, the performance of *ExpPeerRank* and especially those of *BestPeer* are worst.

TABLE II. PERFORMANCE OBTAINED ON EXPERIMENTAL DATA

Module	RMSE per method				
	Average	*Peer Rank*	*PowPeer Rank*	*ExpPeer Rank*	*BestPeer*
1	1.00	0.96	0.94	1.40	2.13
2	0.87	0.82	0.81	1.16	1.87
3	0.88	0.83	0.82	1.13	1.82
4	0.82	0.77	0.77	1.01	1.80
5	0.81	0.76	0.75	1.02	1.74
6	0.80	0.76	0.75	1.07	1.87
7	0.65	0.61	0.61	0.77	1.49
Mean	**0.83**	**0.79**	**0.78**	**1.08**	**1.81**

This latter result can be explained by the fact that, with both *ExpPeerRank* and *BestPeer*, the final grade of each student is extremely influenced by the grade proposed by one grader: the most reliable. This moves the final grade away from the *approximate real grade* obtained by mediating all available evaluations. In particular, *BestPeer* suffers from an approximation issue too. Indeed, by just considering the grade proposed by the best grader, the final grade results in an integer from 1 to 5 (a point from the 5-point scale) normalized in the interval [0,10].

It should be noted that, when the total number of active student decreases (as the progressive module number increases), the performance of all methods improves. This behaviour is explained by the fact that the number of evaluations used for prediction is fixed (m = 4) while the total number of evaluations (used to calculate the *approximate real grade*) decreases. Therefore, the ratio of available data over the whole set increases, resulting in better performance.

B. Variable number of peer assessors

In this experiment the attention is focused just on one assignment (i.e. the first question of the first module) but the number m of assessors for each submission in increased from a minimum of 2 to a maximum of 10. In each step, the number m of assessors for each student is chosen in this range and 1000 iterations are performed. For each iteration the *assessment grid* and the *grades matrix* have been generated as in the previous experiment and the final grades are calculated according to the defined methods.

Figure 9 plots the performance obtained by the five methods in terms of mean RMSE against the number of assessors m. As in experiments 5 and 6 (executed on synthetic data), the error decreases when the number of assessor increases and the decrease is smoother as m increases. An exception is *BestPeer* that has uniform performance regardless of the selected number of assessors. This can be explained through the same approximation issue pointed out in the preceding sub-section.

As it can be seen, both the *PeerRank* and *PowPeerRank* methods show better performance with respect to the average aggregation rule. Indeed, the performance gap between these methods decreases with the increase of the number of assessors i.e. when the quantity of information available to the methods becomes closer to the information used to calculate the *approximate real grade*.

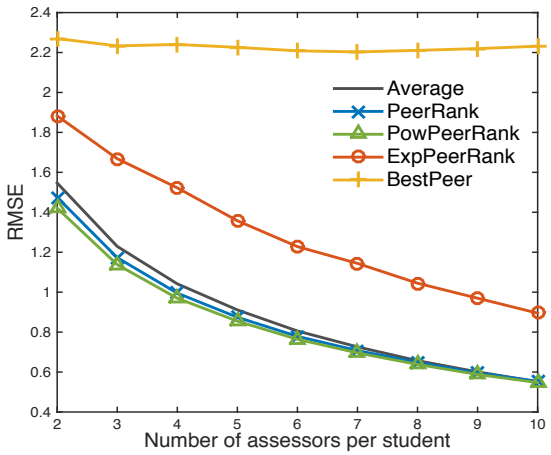

Figure 9. Performance in terms of RMSE of the defined methods on experimental data coming from the first assignment (module 1, question 1) with an increasing number of assessors per student.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed different aggregation methods based on graph mining techniques for peer-assessment as well as a smart assignment method aimed at balancing good graders among students. The assumption of this work, confirmed by other studies, is that the grade obtained by a student is not only a measure of proficiency in a given subject but also a measure of her ability to grade correctly. A limitation of this approach is that additional factors that may affect the student's ability to grade (e.g. the grader's tendency to inflate or deflate proposed grades, the general attitude to review other's work) are disregarded. Such parameters will be considered in a future work.

Experimental results with simulated data show that the *ExpPeerRank* method outperforms other methods in most configurations (with random or smart assignment) while, in particular circumstances (uniform distribution of grades) the *BestPeer* method with smart assignment performs better. Experimental results with real data, show instead a predominance of *PowPeerRank* over the other methods. Nevertheless, these latter results should be considered only preliminary given that they are calculated against approximated real grades rather than against grades assigned by expert tutors.

This simplification obviously advantages the average aggregation rule over the other methods. Taking this into consideration, the performance achieved by the defined methods in the experiment with real students can be considered as a lower bound to the performance obtainable in experiments specifically designed to assess the reliability of a peer grading task. A future experimentation will overcome these limitations by also involving expert tutors.

REFERENCES

[1] G. Siemens, "Massive Open Online Courses: Innovation in Education?", in Open Educational Resources: Innovation, Research and Practice, Commonwealth of Learning, pp. 5-15, 2013.

[2] T. Daradoumis, R. Bassi, F. Xhafa, S. Caballé, "A review on massive e-learning (MOOC) design, delivery and assessment", In proc. of the 8th Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 208-213., 2013.

[3] N. Capuano, M. Gaeta, P. Ritrovato, S. Salerno, "Elicitation of Latent Learning Needs Through Learning Goals Recommendation", in Computers in Human Behavior, Elsevier, vol. 30, pp. 663-673, 2014. http://dx.doi.org/10.1016/j.chb.2013.07.036.

[4] I. Caragiannis, A. Krimpas, A. A. Voudouris, "Aggregating partial rankings with applications to peer grading in massive online open courses", in proc. of the Int. Conf. on Autonomous Agents and Multiagent Systems, pp. 675-683, 2015.

[5] N. Capuano, R. King, "Knowledge-based assessment in serious games: an experience on emergency training", Journal of e-Learning and Knowledge Society, vol. 11, n. 3, pp. 117-132, 2015.

[6] D. Glance, M. Forsey, M. Riley, "The pedagogical foundation of massive online courses", First Monday, vol. 18, n. 5, 2013. http://dx.doi.org/10.5210/fm.v18i5.4350.

[7] L. Bouzidi, A. Jaillet, "Can online peer assessment be trusted?", in Educational Technology & Society, vol. 12, n. 4, pp. 257-268, 2009.

[8] P. A. Carlson, F. C. Berry, "Calibrated Peer Review™ and Assessing Learning Outcomes", in proc. of the 33rd Int. Conf. Frontiers in Education, 2003.

[9] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, D. Koller, "Tuned models of peer assessment in MOOCs", proc of the 6th Int. Conf. on Educational Data Mining, 2013.

[10] I. M. Goldin, "Accounting for peer reviewer bias with Bayesian models", proc. of the Intelligent Support for Learning Groups workshop, 11th Int. Conf. on Intelligent Tutoring Systems, 2012.

[11] L. de Alfaro, M. Shavlovsky, "Crowdgrader: Crowdsourcing the evalu-ation of homework assignments", in proc. of the 45th ACM Technical Symposium on Computer Science Education, pp. 415-420, 2014. http://dx.doi.org/10.1145/2538862.2538900.

[12] T. Walsh, "The PeerRank Method for Peer Assessment" in procs. of the 21st European Conference on Artificial Intelligence, 2014.

[13] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank citation ranking: Bringing order to the Web", Tech. Report 1999-66, Stanford InfoLab, 1999.

[14] N. Capuano, S Caballé, "Towards Adaptive Peer Assessment for MOOCs", proc. of the 10th Int. Conf. on P2P, Parallel, GRID, Cloud and Internet Computing, pp. 64-69, 2015. http://dx.doi.org/10.1109/3pgcic.2015.7.

[15] J. Miguel, S. Caballé, F. Xhafa, J. Prieto, "Security in online Web learning assessment", World Wide Web, vol. 18, n. 6, pp. 1655-1676, 2015. http://dx.doi.org/10.1007/s11280-014-0320-2.

AUTHORS

Nicola Capuano is research assistant at the University of Salerno. His main research interest is artificial intelligence and, among its applications, intelligent tutoring systems and knowledge representation. He works as a project manager and research consultant within several research and development projects. He is author of about 100 scientific papers. He is scientific referee and member of editorial boards for International journals and conferences (e-mail: ncapuano@unisa.it).

Santi Caballé is associate professor at the Open University of Catalonia in the area of software engineering and Web-applications for collaborative work and learning. He conducts research on e-learning, collaborative and mobile learning, distributed technologies and software engineering. He has been involved in the organization of several international conferences, conference tracks and workshops, and has published over 200 research contributions. He has also acted as an editor for books and special issues of international journals (e-mail: scaballe@uoc.edu).

Jorge Miguel is lecturer and assistant professor of Information Security at the San Jorge University. In 2015 he received the PhD degree in Network and Information Technologies from the Open University of Catalonia. His research interests are focused on e-Learning and Information Security and are supported by 20 research contributions to journals and international conferences as well as by active

participation in research projects at national and European level (e-mail: jmmoneo@uoc.edu).

This research was partially supported by the Spanish Government through the project: TIN2013-45303-P ICT-FLAG: Enhancing ICT education through Formative assessment, Learning Analytics and Gamification.