Some Scaling Laws for MOOC Assessments

Nihar B. Shah
UC Berkeley
nihar@eecs.berkeley.edu

Joseph Bradley
UC Berkeley
joseph.kurata.bradley@gmail.com

Sivaraman Balakrishnan
UC Berkeley
sbalakri@berkeley.edu

Abhay Parekh
UC Berkeley
parekh@berkeley.edu

Kannan Ramchandran
UC Berkeley
kannanr@eecs.berkeley.edu

Martin J. Wainwright
UC Berkeley
wainwrig@stat.berkeley.edu

ABSTRACT

A problem that arises with the increasing numbers of students in Massive Open Online Courses (MOOCs) is that of student evaluation. The large number of students makes it infeasible for instructors or teaching assistants to grade all assignments, while current auto-grading technology is not feasible for many topics. As a result, there has recently been an increase in the use of peer-grading, where students grade each other; in this way the number of graders automatically scales with the number of students. However, in practice, peer-grading has been observed to have high error rates and has come under criticism.

In this paper, we take a statistical approach to assessing the feasibility of peer-grading for MOOCs. Our assessments lead to some good news and some bad news. Our study reveals that based on simple yet general models, peer-grading at scale is unfortunately unlikely to work well as a standalone option, i.e., as the number of students increases, the expected number of students mis-graded will grow proportionately. The good news however is that by considering a hybrid approach that combines peer-grading with auto-grading using dimensionality-reduction techniques, we can tackle the scaling challenge efficiently. Concretely, an automated approach is used for 'dimensionality reduction', a classical technique in statistics and machine learning, and peer-grading is used to evaluate this lower dimensional set of submissions.

Our contributions to the current MOOCs literature is two-fold. First, we provide a principled analytical approach to a problem area that is predominantly empirically driven. In this context, we approach the problem at a high level to provide fundamental laws on scaling of assessments in MOOCs based on simple but general models. Secondly, we provide a constructive approach to tackling the scaling problem using tools from statistics. Our proposed hybrid approach can potentially inform the way next-generation MOOC-assessment algorithms evolve.

INTRODUCTION

We discuss the scalability of grading, with special attention given to peer-grading in Massive Open Online Courses (MOOCs). For budget-constrained MOOCs, grading by the instructor(s) becomes infeasible as course sizes grow. Large courses require alternative approaches, such as auto-grading or peer-grading.

We focus our discussion on MOOCs for which auto-grading using pre-trained models is difficult. For subjective topics and complex problems, it is often difficult to design machine grading systems which are accurate [2]. In these cases, it is important to include humans in the grading loop.

Peer-grading is a system of grading where students taking a course are graded by fellow students in the same course. Peer-grading is a natural choice for MOOCs since the total number of graders in a peer-graded system self-scales in proportion to the number of students enrolled. For instance, Coursera employs peer-grading in its human-computer interaction (HCI) course. Since the students are not expert graders, in this peer-grading system, the submission provided by each student is graded by 3 to 5 students. The final grade of a student is computed as the median of these individual grades [11]. Alternative algorithms for aggregating peer-grades are proposed in [19, 9, 23, 7, 16, 12].

Research has shown (e.g., [11, 12]) that current auto-grading and peer-grading systems make many mistakes. Qualitative observations about the inaccuracy of MOOC assessment have led to criticism of auto-grading and peer-grading [2, 20]. For MOOC course credits to gain increased acceptance, it is critical that these errors be reduced.

In this paper, we view the problem of assessments in MOOCs through the lens of statistical analysis. Our approach is orthogonal to and nicely complements the largely empirical nature of works in this field. We study the *scaling* behavior of peer-grading, i.e., the behavior when the number of students gets large. Our analysis reveals that under reasonable assumptions, these systems will *incorrectly grade a constant fraction of the students* in expectation. This constant fraction is not a problem for small courses, where an instructor can handle complaints from students who feel they were mis-graded. However, this is not a scalable solution for large courses since student complaints will overwhelm instructors. Our analysis gives some insight into why current peer-grading systems mis-grade many students.

Importantly, our proposed framework is flexible: it allows for diverse grade collection systems, applies to any arbitrary choice of (adaptive) assignments of graders to submissions and any arbitrary choice of aggregating the peer-grades.

Current efforts to improve grading have examined many aspects: improving auto-grading models [14, 24, 10], improving algorithms for aggregating peer-grades [19, 9, 23, 7, 16, 12], combining auto-and-peer-grading [12, 3], and dimensionality reduction [4, 5, 18, 21, 8, 15]. These efforts have shown useful improvements in practice. However, we argue that these methods do not change the scaling behavior of grading; at most, they decrease the constant fraction of mis-graded students.

A second contribution of this paper is to show that, on the upside, this scaling behavior may be improved via a *combination* of current methods used in MOOC assessments. Specifically, we present an algorithm for assessment that employs a combination of auto-and-peer-grading with 'dimensionality reduction'. We prove that under very general error models, our algorithm has the potential to create vanishing error rates, i.e., to ensure that the expected fraction of students mis-graded goes to zero as MOOC course sizes grow. This result provides a statistically sound, constructive path towards solving the scaling problem for assessments in MOOCs.

Any study involving MOOCs by definition must address scale. It is thus paramount when studying questions like grading for MOOCs that we tackle the most important "high order bit" first. This motivates us therefore to address peer-grading for MOOCs with a view to first understanding the fundamental scaling behavior, namely to study scaling laws with respect to simple yet informative models. This will help inform more detailed subsequent studies involving specific algorithms and practical solutions needed to make this a reality.

SCALING ANALYSIS OF PEER-GRADING

We are interested in MOOCs which are *massive* and *open*. In other words, the courses are very large, and are offered for free or at low costs. This cost constraint prevents the number of instructors from scaling proportionally to the number of students. Thus, we may assume that instructors cannot hand-grade every student's submission.[1] Instead, grading is done via peer-grading, where students evaluate their peers' submissions, and/or auto-grading, where a computer software evaluates the students' submissions.

Since grading requires considerable effort, we expect the average student to grade at most a few peers. We assume that most student graders perform better than random guessing but are imperfect, i.e., frequently make errors in the grading; empirical studies have shown this to be true (e.g., [11]). For our analysis, we assume that there is a "true" grade for each students' submission. The grades provided by expert graders can be assumed to match the true grades, but students (i.e., peer-graders) provide only noisy measurements of the true grade.

Theorem 1 below analyses the scaling behavior of typical peer-grading systems described above. For very general settings, we show that in expectation, the grades of a constant fraction of students will be in error. In order to keep our claims independent of the model and inference algorithm, we consider a generic grading scenario and show that under this scenario, a constant fraction of students having a specific true grade are *statistically indistinguishable* from those for students having a different true grade.

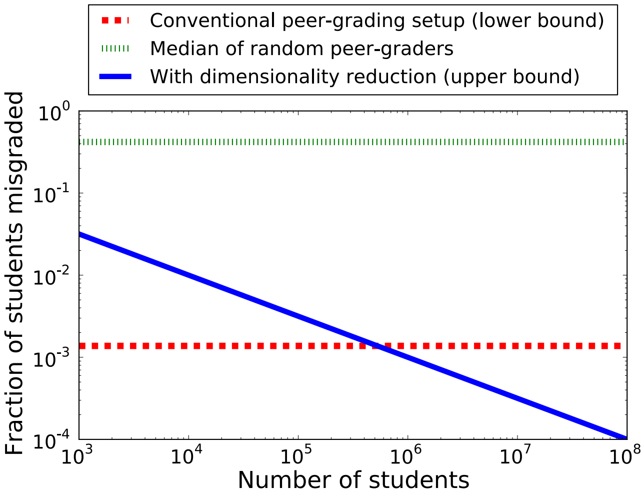

Figure 1. An illustration of the scaling of the expected fraction of students mis-graded in three settings: (a) lower bounds on error under conventional peer-grading setups, (b) error in the peer-grading algorithm that selects peer-graders uniformly at random and sets the final grade as the median of the peer-grades, and (c) an upper bound on the error under peer-grading with dimensionality reduction. The plot considers a setting with $q = 0.25$, $k = 3$, and no perfect graders or experts. In (c), the clustering algorithm is assumed to output $\frac{d}{2 \log d}$ clusters and wrongly clusters $\sqrt{d}$ non-representative submissions. Observe how the fraction of students mis-graded remains constant in the conventional setups (a) and (b), but falls to zero with dimensionality reduction (c).

THEOREM 1. *Consider the following setting. There are d students enrolled in the course. Each student submits one submission, and these d submissions must be graded. Each student has a true underlying grade which is equally likely to be either 'pass' or 'fail', and the goal is to infer these true grades. Each (peer-)grader grades k submissions. Each student is an imperfect grader with probability $\gamma \in (0, 1]$, and is a perfect grader with probability $(1 - \gamma)$. When asked to grade a submission, an imperfect grader grades incorrectly with probability q, for some $q \in (0, 0.5)$, while a perfect grader always grades correctly. A correlation between grading and answering skills may also be assumed: a randomly selected student with a true grade of 'pass' is a perfect grader with a probability $\rho \in [0, 1)$. A constant fraction $\alpha \in [0, 1)$ of submissions may also be graded by the instructors who grade perfectly. The parameters k, γ, q, ρ and α are all independent of the total number of enrolled students d. The values of these parameters may be known to the peer-grading algorithm.*

A peer-grading algorithm must first assign (peer or instructor) graders to each submission. We allow this assignment to be executed in a completely adaptive manner, i.e., the algorithm is allowed to wait for the output of one grader before making the next submission-grader assignment. There is no constraint on the number of people who may grade any individual submission. Upon receiving all the (peer and instruc-

[1] The term 'submission' will be used generically to refer to solutions to examinations, homeworks, or any other material submitted for evaluation by the students.

tor) grades, the algorithm must then assign final grades to each submission.

Under any peer-grading algorithm, in expectation, the final grades of a constant fraction of students will be in error.

The scaling results derived here imply that if the quality of peer-graders does not improve much with an increase in the number of enrolled students, then as the number of students in the course increases, the number of mis-graded students will increase proportionately. The student experience will suffer in such a scenario, and instructors may be faced with an overwhelming number of student complaints. This scaling analysis is illustrated in Figure 1 along with a comparison to the dimensionality-reduction based peer-grading algorithms to be presented in the next section.

The remainder of this section discusses the assumptions and some extensions of Theorem 1. The assumption of having the parameters γ, ρ and q independent of the total number of enrolled students d means that the average abilities of the students do not vary (significantly) when the number of students enrolled in the course increases. We assume that the number of submissions αd graded by the instructors increases at most linearly with the total students d: a reasonable assumption for freely offered courses. Finally, we assume k to be a constant (and not increasing in d) since grading is a task that requires significant time and effort on the part of the grader, and we must impose a limit on the number of submissions that any student needs to peer-grade. Note that this restriction does *not* restrict each submission to be graded by at most k peer-graders.

The result of Theorem 1 may be extended to cover many alternative settings. A similar argument holds for the case when the grades collected from the graders are on a finer scale (with more than two possible values of the grades), when the true underlying grade may not be uniformly distributed, or when the peer-grading is ordinal where students are asked to compare two or more submissions instead of assigning numeric scores to the submissions [22]. Peer-grading may be augmented to also include grading by people not taking the course, e.g., by people who have previously taken the course. If the number of such 'outside-graders' remains linear in the number of students, then arguments similar to those of Theorem 1 continue to apply. In order to ensure fairness, the use of extraneous information about the students' skills is generally avoided in the grading process [19], as assumed in the theorem. One could potentially obtain some information about the students' grading abilities from their performance in previous homeworks or tests, but this information will not change the scaling laws as long as the number of homeworks/tests conducted remains independent of the number of students.

BREAKING THE BARRIER VIA DIMENSIONALITY REDUCTION

In this section, we provide one possible means of breaking the barrier of a constant fraction of submissions being graded erroneously: *dimensionality reduction*. We discuss two methods for dimensionality reduction: clustering and featurization. Clustering is the more intuitive method, but featurization is more general. These methods combine auto-grading with peer-grading in a manner to be discussed in the sequel. The peer-grading interface remains the same as before, where students grade submissions submitted by their peers.

Clustering

For simplicity of exposition, the following discussion will assume a submission to be the answer to a single question. The algorithm may be applied separately to each of the questions. Suppose we use a computer program to cluster the collection of all d submissions provided by the d students with respect to the similarity of their content. The clustering algorithm is such that *ideally, within each cluster, all submissions have the same true grade*. Note that multiple clusters could have the same true grade. Given this assumption, grading any one submission in the cluster effectively grades all submissions in the cluster. The total number of submissions to grade is effectively reduced.

A cluster can be graded by collecting peer-grades for any submission within the cluster. These grades may be aggregated via any reasonable algorithm, for instance, taking a median of the received grades. Finally, the submission of any student is assigned the grade that its cluster receives. This approach is formalized in Algorithm 1, and illustrated pictorially in Figure 2.

Algorithm 1 A grading algorithm with rigorous guarantees: combining dimensionality reduction and peer-grading

```
• Cluster submissions into $\Delta$ clusters based on 'similarity'
• Select one representative submission for each cluster
• Choose $\frac{dk}{\Delta}$ peer-graders for each cluster uniformly at random, while ensuring that no peer-grader is chosen for more than $k$ clusters
• For each cluster, assign the representative submission to its corresponding chosen peer-graders
• Upon receipt of peer-grades, for each cluster, assign median of the peer-grades as the final grade of all submissions in that cluster.
```

The following theorem shows that given a good clustering algorithm, the grading algorithm proposed in Algorithm 1 can reduce the number of erroneously graded students to a vanishing fraction, i.e., ensure that the expected fraction of students mis-graded goes to zero. The theorem considers peer-grading conditions worse than that considered in Theorem 1, namely, does not assume the existence of any perfect graders or instructor graders. In what follows, $D(p_1||p_2)$ denotes the Kullback-Liebler divergence [13] between two Bernoulli distributions having parameters p_1 and p_2 respectively, i.e.,

$$D(p_1||p_2) = p_1 \log \frac{p_1}{p_2} + (1 - p_1) \log \frac{1 - p_1}{1 - p_2} .$$

We use the standard Bachmann-Landau [1] notations of $o(\cdot)$, $O(\cdot)$ and $\Theta(\cdot)$ to represent the limiting behavior of functions. Informally, $f(x) = o(g(x))$ means that f is dominated by

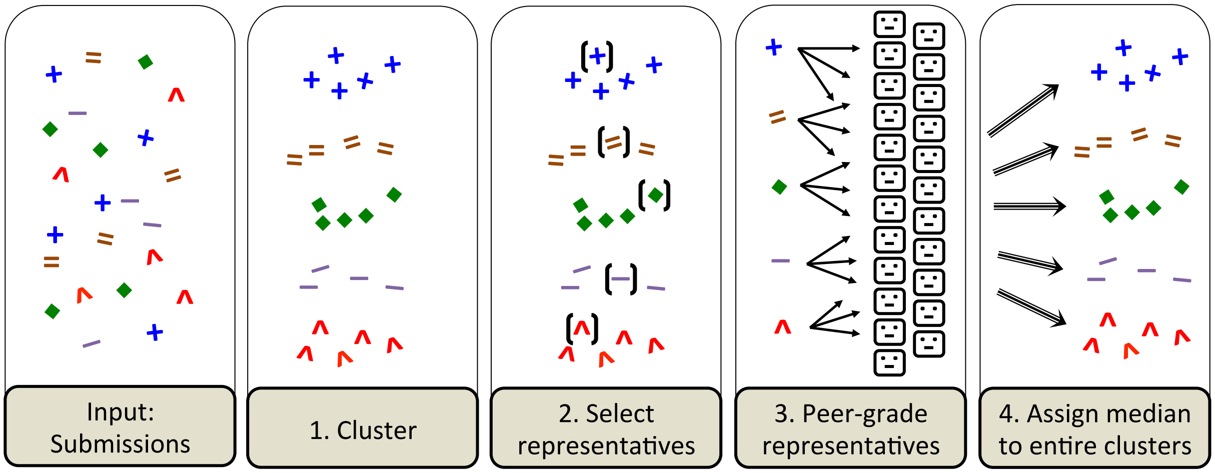
Input: Submissions
1. Cluster
2. Select representatives
3. Peer-grade representatives
4. Assign median to entire clusters

Figure 2. A block diagram of the dimensionality-reduction based assessment algorithm that is described in Algorithm 1 and mathematically proved to have a good scaling behavior in Theorem 2.

g asymptotically, $f(x) = O(g(x))$ means that f is upper bounded by g asymptotically, and $f(x) = \Theta(g(x))$ means that f is both upper and lower bounded by $g(x)$ with different constant pre-factors.

THEOREM 2. *Consider the following setting. There are d students enrolled in the course. Each student submits one submission, and these d submissions must be graded. Each student has a true underlying grade which is either 'pass' or 'fail', and the goal is to infer these true grades. Each (peer-)grader grades k submissions. When asked to grade a submission, a grader grades incorrectly with probability q, for some $q \in (0, 0.5)$. Here, k and q are independent of the number of students d.*

Suppose grading is performed as per Algorithm 1. Then, if the number of clusters satisfies $\Delta \leq \frac{d}{\log d} k D(0.5||1-q)$, and if the clustering algorithm employed in Algorithm 1 erroneously clusters at most $o(d)$ non-representative submissions and at most $o(\Delta)$ representative submissions at random, then the fraction of submissions that are mis-graded goes to zero as the total number of submissions d grows.

The theorem says that if the submissions can be clustered into clusters of average size of order $\log d$ or higher, then even if each student can grade only a constant number of submissions, the number of submissions graded incorrectly will be independent of the number of students d. The system can now scale to accommodate arbitrarily large numbers of students without the worry of the number of grading errors blowing up. This scaling analysis is illustrated in Figure 1.

The theorem assumes a good performance of the clustering algorithm. This may seem to be at odds with our earlier discussion on the performance of auto-grading algorithms suggesting that they do not perform very well for topics that are subjective in nature. To this end, we note that while auto-grading requires the algorithm to understand the semantics of each submission in order to grade, *clustering only requires understanding the features on which similarity testing must be performed*. The job of clustering is thus a subset of the harder task of auto-grading. Indeed, designing algorithms for clustering submissions for educational assessment is an area of active research in the community, e.g., [4, 5, 18, 21, 8, 15]. Moreover, research has shown that students tend to answer in similar ways [15, 4] allowing the number of clusters to be small.

Let us now discuss the assumption on the performance of the clustering algorithm from a statistical perspective. We have assumed that the number of submissions clustered erroneously grows sub-linearly in d. In other words, we assume that the fraction of submissions clustered incorrectly reduces with an increase in the total number of submissions. An intuitive justification of this assumption is that clustering is often performed by comparing submissions with each other, and as the total number of submissions d grows, the number of submissions (in each cluster) available for comparison also grows, thereby allowing the clustering algorithm to improve its performance as d increases. Using literature on statistical guarantees for clustering problems (e.g., [6]), we make this intuition mathematically rigorous.

THEOREM 3. *(Adaption of [6, Theorem 2.2]) Consider a clustering algorithm that operates in the following manner. The algorithm has a black-box comparator that, given any pair of submissions, correctly identifies whether they belong to the same cluster or not with a probability at least $\frac{1}{2} + \epsilon$ (independent of all other pairs), for some $\epsilon > 0$. The value of ϵ is fixed, independent of d, and is unknown to the algorithm. Suppose there are $\Delta \leq c\epsilon^2 \frac{d}{\log d}$ clusters, of equal size, and the value of Δ is known to the algorithm. Here, c is a specific universal constant. Then there exists a clustering algorithm such that the expected number of submissions clustered incorrectly is upper bounded by a universal constant (independent of d).*

REMARK 1. *The setting of [6, Theorem 2.2] differs from the setting of peer-assessment considered in this section in some respects. We assume the average cluster-size to be at least $\tilde{c} \log d$ (for some constant $\tilde{c} > 0$) and that the clustering* *algorithm does not know the sizes of these clusters, while [6, Theorem 2.2] assumes that the sizes of all clusters are identical, lower bounded by $\tilde{c}\log d$, and the size of each cluster is known to the algorithm. On the other hand, the results of [6, Theorem 2.2] provide a very strong guarantee in that only $O(1)$ submissions are in wrongly clustered in expectation, whereas for our requirements, a weaker guarantee of $o(d)$ errors would suffice.*

The scaling analysis presented in this section suggests that reducing the dimension of the submissions by a logarithmic factor may suffice for designing a scalable grading system. Previous works on clustering which we reference used clustering to aid instructors in grading. This approach can lessen the burden on instructors, who can assign grades to groups of submissions. Our analysis motivates as well as provides a theoretical justification for the use of clustering tools [4, 5, 18, 21, 8, 15] for massive open online courses, where in conjunction with peer-grading, we show that they can help achieve scalability in the grading process.

Featurization

We briefly discuss a more general type of dimensionality reduction in this section. The clustering method discussed earlier assumes that many submissions are similar enough to be declared equivalent by a clustering algorithm. One could generalize to assume that *parts* or *aspects* of many submissions are similar and can be compared or clustered; we describe this as *featurization*, where the content of a submission is summarized by a set of features.

Suppose that submissions may be described by Δ features. Assume that the grade of each submission may be computed as a function of these features; e.g., a simple such function for pass/fail grades would be thresholding a weighted sum of the features. Then we have reduced the problem of grading to a traditional regression setting: submissions are examples, features are computed algorithmically for each example, and peer-grades provide noisy labels for the examples.

The regression model is simply another aggregation method for peer-grades. Assume for now that we use a generalized linear model $y^{(i,t)} \sim f(w^T x^{(i)}) + \epsilon^{(i,t)}$, where $x^{(i)}$ is the feature vector for submission i, $y^{(i,t)}$ is the t^{th} peer-grade for submission i, $f(\cdot)$ is the inverse link function, and $\epsilon^{(i,t)}$ is the noise added by the peer-grader. This model generalizes simple peer-grading systems which treat all submissions independently: d boolean features are indicators corresponding to the d submissions, so the feature vector $x^{(i)}$ for submission i has a "1" for feature i and "0" elsewhere. The model also generalizes clustering: we have Δ boolean features corresponding to the Δ clusters.

This setup could allow grading systems to draw on extensive research on feature engineering and modeling. The success of previous work on feature-based clustering of submissions in MOOCs indicates that useful features can be found. The fact that current aggregation methods for peer-grades can be generalized by simple regression models indicates that such models are reasonable.

Depending on the choice of the regression model, one can achieve scaling results similar to those for clustering. E.g., consider a logistic regression model (with a logit link function) to classify submissions as pass/fail. Then previous work [17] on logistic regression has shown that, with Δ features and $\Theta(d)$ samples (peer-grades), it suffices to have $\Delta = o(d)$.

Previous work has considered combining auto-grading with peer-grading [12, 16]. Both these works employ an adaptive process for grading, with a separate auto-grading processes used to supplement the peer-grading process. However, none of these works have rigorous guarantees on the performance. Moreover, these works separate peer-grading and auto-grading into multiple stages, and combine the auto-grades and peer-grades in a certain manner to compute the final grade. This approach leaves the scaling laws unchanged: either auto-grading makes a constant fraction of grading errors (in which case a constant fraction of errors remain after peer-grading as well), or auto-grading makes a vanishing fraction of errors (in which case peer-grading becomes unnecessary for the purpose of obtaining vanishing error rates).

CONCLUSIONS

In this paper, we gave a statistical analysis of the scaling properties of various grading mechanisms in MOOCs. We saw that under very simple and general models, the kinds of peer-grading systems employed today will not scale. We then showed that combining (auto-) dimensionality reduction and peer-grading has the potential to scale. Dimensionality reduction is already an active topic of research [4, 5, 18, 21, 8, 15], and the proposal of combining it with peer-grading falls under the more general paradigm of combining machine and human intelligence [12, 3].

While most current research on assessment in MOOCs is empirical, this paper provided a more analytical approach, helping understand the basic principles behind peer-assessment in current grading systems, and identifying a path for future research to overcome assessment errors. Accurate, reliable, and scalable assessment will help to pave the way for MOOCs to democratize education.

APPENDIX: PROOFS

This section presents proofs for the three theorems stated in the paper. The proofs focus on demonstrating the scaling laws and do not attempt to optimize the constants (The plots of Figure 1 have optimized constants). Since we are interested in only the scaling laws, without loss of generality, we assume d to be large enough to ignore any floor/ceiling effects of integers.

PROOF OF THEOREM 1.
Consider any values of the parameters $d, k, q, \alpha, \gamma, \rho$ as defined in the statement of the theorem. Consider any peer-grading algorithm $\mathcal{A}$ that minimizes the expected fraction of students mis-graded.

The high-level idea of the proof is as follows. We will select two students uniformly at random from the set of d students.

We will show that with a probability lower bounded by a positive value independent of d, the performance of the two students appear statistically indistinguishable to algorithm $\mathcal{A}$, irrespective of their true underlying grades. Conditioned on being statistically indistinguishable, algorithm $\mathcal{A}$ will mis-grade at least one of these two students in expectation. Aggregating over all students, we will show that the fraction of students mis-graded is lower bounded by a positive value that is independent of d. Given the formulation of Theorem 1 and this high-level idea, the rest of the math behind the proof is fairly simple.

Under the setting considered in the theorem, at least $(1-\alpha)d$ submissions will not be graded by instructors. These submissions must be evaluated based on the peer-grades assigned to them and the accuracy of these $(1-\alpha)d$ students in the peer-grading that they performed. The total number of peer-grades available to algorithm $\mathcal{A}$ is upper bounded by dk. Let μ denote the number of submissions from this collection of $(1-\alpha)d$ submissions that have received no more than $\frac{2k}{1-\alpha}$ peer-grades. The pigeonhole principle implies that

$$((1-\alpha)d-\mu)\left(\frac{2k}{1-\alpha}+1\right) \leq dk,$$

and some algebraic manipulations lead to

$$\mu \geq \frac{1-\alpha}{2}d.$$

Let us momentarily focus our attention only on these μ submissions. Suppose you choose two submissions uniformly at random from this set of μ submissions, say the submissions of students s_1 and s_2. We derive a lower bound on the probability that the performance of student s_2 is statistically indistinguishable from that of student s_1, irrespective of the true grades of students s_1 and s_2. Consider the peer-grading process of student s_2 under algorithm $\mathcal{A}$. In order to obtain a peer-grade for s_2's submission, Algorithm $\mathcal{A}$ may wish to enroll a perfect or an imperfect grader. The algorithm has the following information available at its disposal for inferring whether a certain peer-grader is perfect or imperfect: the performance of this peer-grader on the $(k-1)$ or fewer submissions that she has already graded, and the evaluation of this peer-grader's own submission coupled with the knowledge of the correlation between good graders and good grades.

Recall that in the setting under consideration, a randomly selected student is an imperfect grader with probability γ and a perfect grader with proability $(1-\gamma)$. A randomly selected student has a 'pass' or a 'fail' grade on her own submission with a probability 0.5 each. Given a pass grade, the student is a perfect grader with a probability ρ. In order for the algorithm to confuse the imperfect grader for a perfect grader, the imperfect grader should have correctly graded all (at most $(k-1)$) submissions assigned to her, and should have a true underlying grade that is most common among perfect graders. Some algebraic manipulations lead to a lower bound on the probability of the algorithm mistaking an imperfect grader for a perfect grader as

$$\frac{q^{k-1}}{\max\left\{\frac{2(1-\gamma)-\rho}{1+\rho-2(1-\gamma)}, \frac{\rho}{1-\rho}\right\}+q^{k-1}}.$$

When an imperfect grader happens to be selected, there is a $q \in (0, 0.5)$ chance that she will mis-grade and a $(1-q)$ chance that she will grade correctly. It follows that for any arbitrary sequence of $\frac{2k}{1-\alpha}$ or fewer gradings for student s_2, the probability of observing this sequence is lower bounded by

$$\left(\frac{q^{k-1}}{\max\left\{\frac{2(1-\gamma)-\rho}{1+\rho-2(1-\gamma)}, \frac{\rho}{1-\rho}\right\}+q^{k-1}}q\right)^{\frac{2k}{1-\alpha}}.$$

The grades of the two students s_1 and s_2 may also depend on their performance as peer-graders. The probability that student s_2 peer-grades in a manner identical to s_1 is lower bounded by q^k. It follows that the probability that these two randomly chosen students appear statistically indistinguishable to algorithm $\mathcal{A}$ is lower bounded by

$$\left(\frac{q^{k-1}}{\max\left\{\frac{2(1-\gamma)-\rho}{1+\rho-2(1-\gamma)}, \frac{\rho}{1-\rho}\right\}+q^{k-1}}q\right)^{\frac{2k}{1-\alpha}} q^k.$$

Conditioned on students s_1 and s_2 appearing statistically indistinguishable to algorithm $\mathcal{A}$ irrespective of their true grades, the algorithm will mis-grade at least one of them in expectation.

We now return to the set of all d students. Aggregating all arguments from above, we get that the algorithm will mis-grade at least

$$\frac{d}{2}\frac{1-\alpha}{2}\left(\frac{q^{k-1}}{\max\left\{\frac{2(1-\gamma)-\rho}{1+\rho-2(1-\gamma)}, \frac{\rho}{1-\rho}\right\}+q^{k-1}}q\right)^{\frac{2k}{1-\alpha}} q^k.$$

students in expectation. This quantity is linear in d thereby proving our claim. □

PROOF OF THEOREM 2.
First consider the setting of a perfect clustering algorithm. The total number of grades obtained from the students is kd. Let $\beta_0(d)$ denote the expected number of students mis-graded when the total number of enrolled students is d. We show that in the setting described in the statement of the theorem, $\beta_0(d) = O(1)$. Since there are Δ clusters, each cluster receives $\frac{kd}{\Delta}$ grades. Each grade is correct with a probability $(1-q) \in (0.5, 1]$, and hence, applying the Chernoff bound, we get that the probability of mis-grading any individual cluster is upper bounded by

$$\exp\left(-D(0.5||1-q)\frac{kd}{\Delta}\right).$$

Thus, in expectation, the total number of mis-graded students is

$$\beta_0(d) = d \exp\left(-D(0.5||1-q)\frac{kd}{\Delta}\right).$$

Substituting $\Delta \leq \frac{d}{\log d} kD(0.5||1-q)$ as assumed in the statement of the theorem, we get that only $O(1)$ students are mis-graded in expectation.

Now suppose the clustering algorithm incorrectly clusters $\beta_1(d)$ non-representative submissions and $\beta_2(d)$ representative submissions randomly, with $\beta_1(d) = o(d)$ and $\beta_2(d) = o(\Delta)$. Then the expected total number of mis-graded students is

$$\beta_0(d) + \beta_1(d) + \frac{d}{\Delta}\beta_2(d).$$

The expected fraction of students mis-graded is

$$\frac{\beta_0(d)}{d} + \frac{\beta_1(d)}{d} + \frac{\beta_2(d)}{\Delta} = \frac{O(1)}{d} + \frac{o(d)}{d} + \frac{o(\Delta)}{\Delta},$$

which goes to 0 as d gets large. □

PROOF OF THEOREM 3.
The parameters p, q, n and K in [6, Theorem 2.2], when translated to our setting, have the following relations with the parameters of our setting:

$$\begin{aligned} p &\geq \frac{1}{2} + \epsilon, \\ q &\leq \frac{1}{2} - \epsilon, \\ n &= d, \\ K &= \frac{d}{\Delta}. \end{aligned}$$

The assumption $\Delta \leq c\epsilon^2 \frac{d}{\log d}$, with c being large enough, ensures that the condition [6, Equation (7)] required by [6, Theorem 2.2] is satisfied. While the statement of [6, Theorem 2.2] guarantees correct clustering of all submissions with a probability at least $(1 - c_2 d^{-c_3})$ for some positive constants c_2 and c_3 that are independent of d, the proof of the theorem establishes the value of the constant c_3 as 1. It follows that the number of submissions clustered incorrectly is upper bounded by

$$d \times \left(c_2 d^{-c_3}\right) = c_2.$$

□

REFERENCES

1. Family of Bachmann-Landau notations. `http://en.wikipedia.org/wiki/Big_O_notation#Family_of_Bachmann.E2.80.93Landau_notations`. Retrieved October 29, 2014.
2. Professionals against machine scoring of student essays in high-stakes assessment. `http://humanreaders.org/petition/index.php`. Retrieved June 1, 2013.
3. Aggarwal, V., Minds, A., Srikant, S., and Shashidhar, V. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study. In *NIPS Workshop on Data Driven Education* (Dec. 2013).
4. Basu, S., Jacobs, C., and Vanderwende, L. Powergrading: a clustering approach to amplify human effort for short answer grading. *TACL 1* (2013), 391–402.
5. Brooks, M., Basu, S., Jacobs, C., and Vanderwende, L. Divide and correct: Using clusters to grade short answers at scale. In *Learning at Scale* (2014).
6. Chen, Y., and Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267* (2014).
7. Dıez, J., Luaces, O., Alonso-Betanzos, A., Troncoso, A., and Bahamonde, A. Peer assessment in MOOCs using preference learning via matrix factorization. In *NIPS Workshop on Data Driven Education* (2013).
8. Glassman, E. L., Scott, J., Singh, R., and Miller, R. C. Overcode: visualizing variation in student solutions to programming problems at scale. In *Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology*, 129–130.
9. Gutierrez, P., Osman, N., and Sierra, C. Collaborative assessment. In *EDM Workshop on Feedback from Multimodal Interactions in Learning Management Systems* (July 2014).
10. Kakkonen, T., Myller, N., Timonen, J., and Sutinen, E. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, Association for Computational Linguistics (2005), 29–36.
11. Kulkarni, C., Pang-Wei, K., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and self assessment in massive open online classes. *ACM Transactions on Computer-Human Interaction 9*, 4 (2013).
12. Kulkarni, C. E., Socher, R., Bernstein, M. S., and Klemmer, S. R. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@ scale conference*, ACM (2014), 99–108.
13. Kullback, S., and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics* (1951), 79–86.
14. Larkey, L. S. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (1998), 90–95.

15. Luxton-Reilly, A., Denny, P., Kirk, D., Tempero, E., and Yu, S.-Y. On the differences between correct student solutions. In *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*, ACM (2013), 177–182.

16. Mitros, P., Paruchuri, V., Rogosic, J., and Huang, D. An integrated framework for the grading of freeform responses, 2013.

17. Ng, A., and Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS* (2002).

18. Nguyen, A., Piech, C., Huang, J., and Guibas, L. Codewebs: scalable homework search for massive open online programming courses. In *Proceedings of the 23rd international conference on World wide web* (2014), 491–502.

19. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuned models of peer assessment in MOOCs. In *International Conference on Educational Data Mining* (2013).

20. Rees, J. Peer grading can't work. `http://www.insidehighered.com/views/2013/03/05/essays-flaws-peer-grading-moocs`. March 5, 2013.

21. Rogers, S., Garcia, D., Canny, J. F., Tang, S., and Kang, D. ACES: Automatic evaluation of coding style. Master's thesis, EECS Department, University of California, Berkeley, May 2014.

22. Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., and Ramchandran, K. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education* (Dec. 2013).

23. Walsh, T. The PeerRank method for peer assessment. *arXiv preprint arXiv:1405.7192* (2014).

24. Zhenming, Y., Liang, Z., and Guohua, Z. A novel web-based online examination system for computer science education. In *IEEE Frontiers in Education Conference*, vol. 3 (2003).