LAB 3 PeerGrader simulator - implementation

Scorca Francesco s288876

1. Introduction

The purpose of the simulation is to observe the behavior of the PeerGrading system under different scenarios, characterized by different input parameters presented in Table 1. The reasons behind the choice of the set of values are the following.

- **Number of homeworks.** Supposing homeworks to require a week of work, it is unlikely that a class could perform more than 20 of these.
- **Number of students.** An exponential increase of this value is chosen to test a wider range of scenarios[1].
- **Number of evaluators.** In the real system this value must be strictly lower than the number of students, since the evaluators are sampled from the same class. Here we break this constraint to better observe the effects of this parameter.
- **Homework quality std.** The values from 0.01 to 0.2 seem to be reasonable to represent students' capability to perform according to their qualities. Two degenerated cases are tested to better observe the effects of this parameter on the system, namely 0, implying the same mark on every homework (the student's quality), and 10^{10}, implying random marks not related to the student's quality.
- **Evaluation std.** The choice of the set can be argued as the previous. Here the value 0 will imply perfect evaluations, 10^{10}, evaluations not related to the homeworks quality.

INPUT PARAMETER	VALUES TESTED
Number of homeworks	$\{2, 5, 10, 15, 20\}$
Number of students	$\{2, 4, 8, 16, 32, 64\}$
Number of evaluators	$\{2, 4, 8, 16, 32, 64\}$
Homework quality std	$\{0, 0.01, 0.05, 0.1, 0.2, 10^{10}\}$
Evaluation std	$\{0, 0.01, 0.05, 0.1, 0.2, 10^{10}\}$
Seed	$\{2, 3, 4, 5, 6, 7, 8\}$

Table 1: Grid of the input parameters for the different runs.

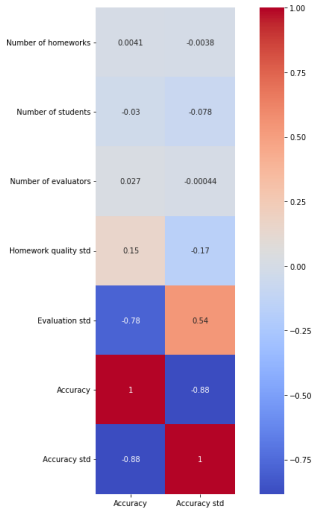

Figure 1: Correlation coefficients

2. Results

The presented results are averaged on the different seeds, to address the intrinsic randomness of the simulator, but we still refer to this average as "accuracy" for brevity. Moreover, since we have null values of sample standard deviations for the accuracy[2], some confidence intervals diverge. Thus, we present the sample standard deviations of accuracy instead of them. Finally, the points tested are marked in the figures, the curves used to connect them conjecture their trends.

2.1. Evaluation standard deviation effects

The most relevant phenomenon is the negative effect of the evaluation standard deviation increasing. It is possible to observe it from figure 1, showing a strong negative correlation with accuracy. This is even more noticeable in figure 2, showing the values of accuracy, with respect to this parameter: when it is null we get the maximum value, when it tends to infinite, we have really unstable results, negative in the majority of the cases. This instability is also noticeable from figure 3, showing that the increase of this value makes the system more and more sensitive to the values sampled.

[1] To test the same range with a linear increase we would have to perform much more simulations.

[2] In the case of null Evaluation standard deviation.

However, the key role of this parameter is quite intuitive, since the PeerGrading system intrinsically relies on the students' capability of evaluation.

2.2. Homework-quality standard deviation effects

A less intuitive effect is the increase of the accuracy with the increase of homework-quality standard deviation: a worsening in the students' capability of reflecting their qualities results in an higher quality of the system. In fact, in figure 1 we there is a positive correlation coefficient (0.15) with the accuracy, and figure 2 shows that as the value of this parameter increases accuracy gets higher and less disperse. This phenomenon can be explained looking at the denominator of the average relative error:

$$\frac{\sum_{h=1}^{H} Q_{hs}}{H} \tag{1}$$

Let us consider the degenerated cases, since the others will present a behavior in the middle:

- Homework-quality std = 10^{10}. Q_s is a Uniform in [0,1], then when averaged on the different values sampled, it will tend to 0.5 (mean value).
- Homework-quality std = 0. Q_s will assume always the same value (X_s). Since averaging makes no effect, for X_s near to 0 the error diverges. Consequently, we can argue that the system gets particularly sensitive to the values sampled, and this is noticeable also from figure 3, and from the positive correlation coefficient with the accuracy std in figure 1.

2.3. Other attributes effects

It is not easy to understand the effect of the single attributes on the system, indeed correlation coefficients are near to 0, and neither the boxplots provide any meaningful information. To deepen the study, we can fix the values of the standard deviations, and observe accuracy varying the remaining parameters, as shown[3] in figure 4. Here, the increase of the number of evaluators improves the results, reflecting the Law of Large Numbers, since $\hat{Q}_{hs}$ is estimator for Q_{hs} obtained by averaging on this parameter. However, it is important to notice that improvements get smaller and smaller in every curve, so that increasing this value from 4 to 8 provides improvements always inferior than 0.05 in accuracy: a number of evaluators in this range is enough to have satisfying results. Regarding the number of students, we can see that for lower values the curves tend to overlap, then an increasing of this tends to provide more predictable results when varying the other parameters. Finally, an increase of the number of homeworks provide always very little improvement, and only when the other parameters are high enough. In particular, it is enough to perform 5 homeworks to get satisfying results, since increasing this value never brings an improvement in accuracy greater than 0.025.

[3] Both standard deviations are fixed at 0.1, a condition in the middle.

3. Proposal: corruption and integrity phenomenons

3.1. Setting

Corruption is a behavior often characterizing social interactions with benefits involved, in this case a mutual evaluation. To simulate this we can set the standard inputs in order to have a stable environment, basing on the previous observations, but at the same time quite realistic: both standard deviations are set to 0.1, with 32 students, 8 evaluators, and 5 homeworks to perform. The new input parameters are the following.

- Corruption rate in $\{0.2, 0.4, 0.6, 0.8, 1\}$. This parameter reflects the degree of corruption in society, indeed it represents the percentage of time a student will try to corrupt an evaluator.
- Integrity rate in $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. This parameter the percentage of time an evaluator is willing to refuse to be corrupted. Consequently, an integrity rate equal to 1 sets the upper bound of the performance of our simulator.

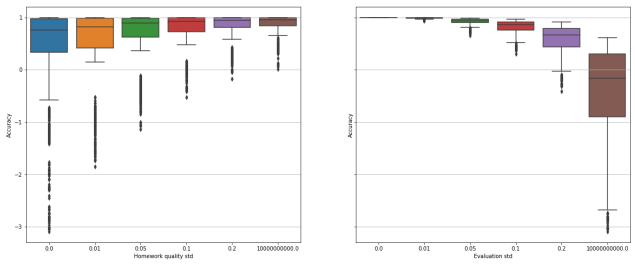

Figure 2: Accuracy standard deviation boxplots for Homework-quality std and Evaluation std values.

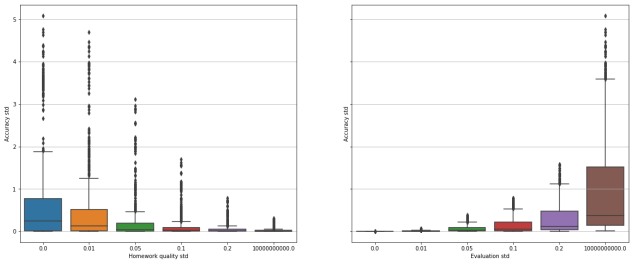

Figure 3: Accuracy deviation boxplots for Homework-quality std and Evaluation std values.

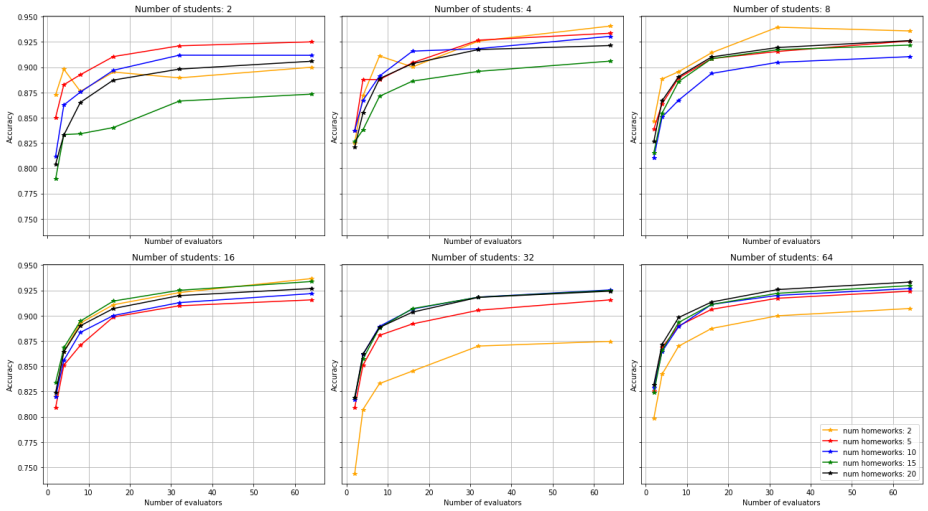

Figure 4: Curves for different numbers of homeworks, evaluators, students. Homework quality std and Evaluation std are fixed at 0.1.

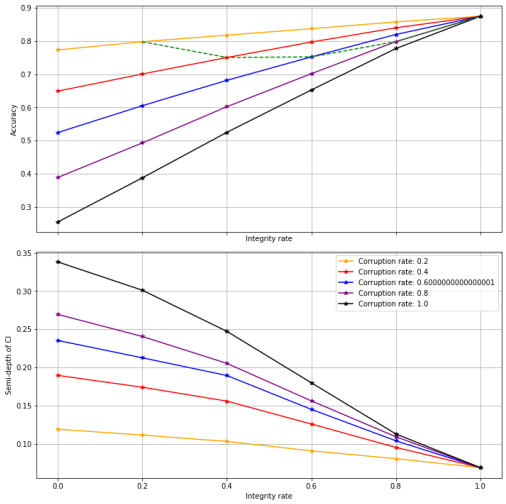

Figure 5: Curves for different values of corruption rate. The top curves present accuracy behavior relative to the integrity rate; the bottom ones present accuracy confidence intervals semi-depths with behavior relative to the integrity rate, with confidence level of 0.9.

In practise, we sample two values from a Uniform distribution over [0,1]. If the first value is lower than the corruption rate, the students will try to corrupt; if the second value sampled is lower than the integrity rate, the evaluator will refuse. Thus, only if the first condition is met and the second is not, corruption occurs and we increase the sampled evaluation by a bias value, in our case 0.2.[4]

[4]clipping the result to 1, maximum grade.

3.2. Results

From figure 5, we can suppose that fixing a corruption rate, the recovery provided by the increase of integrity rate is almost linear. However, except for integrity rate at 1, we never reach the optimum accuracy, even if the integrity rate is much higher than the corruption rate. In spite of this, considering the trend of the points where the integrity rate is equal to the corruption rate, it is possible to notice that, if we guarantee this condition, the drop of accuracy stays bounded. Indeed, the loss in accuracy has a symmetrical shape with maximum at 0.6, after which the benefits of an increase of integrity are superior to the losses of the increase of corruption. Focusing on the bottom graph we can notice that the rates have analogous effects also in term of stability. Indeed the confidence intervals get narrower with the increase of integrity rate in a loosely linear way, while increasing with corruption. Focusing only on corruption, a phenomenon worth mentioning is the following. Let us define the ”effective corruption” as the slope of a corruption line, normalized by its corruption rate. This index represents the relative effect of corruption and, as noticeable from figure 6, it is not constant, but increasing with the corruption rate. Let us consider an example to clarify this result: fixing an integrity rate, if we double the corruption rate, the loss in accuracy will be more than doubled.

4. Conclusions

Summarising, the proposed study led to the following conclusions:

- the system relies almost exclusively on the precision of the evaluators;
- the system performs better if the students are not good

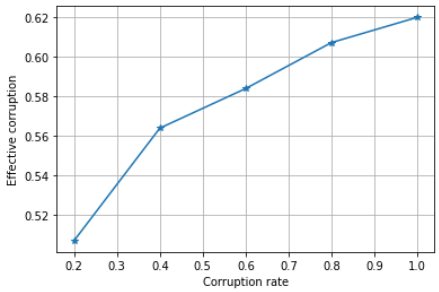

Figure 6: Effective corruption behavior relative to corruption rate

at reflecting their actual value, which highlights a problem in the quality metric adopted;

- a relative small number of homeworks and evaluators is enough to guarantee quite satisfying results[5].

Furthermore, it has been shown that introducing the possibility of corrupting the evaluators it is possible to draw the following considerations:

- the effects of corruption get harsher and harsher with its magnitude, both in absolute and relative terms, as shown by the effective corruption behaviour;
- if the percentage of population not willing to be corrupted is equal to the percentage of corruption of the system, the degradation of this is bounded to relatively small values;
- in spite of the previous consideration, even if of a small magnitude the loss is never null, showing how it is needed to introduce some means for eliminating corruption, for instance a punishment.

The last point can provide inspiration for future developments. Furthermore it would be appropriate to simulate[6] this scenario for different levels of evaluation variance, given its key role, and for different biases[7] in evaluations.

[5] Provided a small variance in evaluations.
[6] This was not possible for time constraint
[7] in the simulated case equal to 0.2