Robustness of Scanner Exams with TUMexam

Simon Y. Kassahun, Stephan Günther*
*Chair of Network Architectures and Services, Department of Informatics
Technical University of Munich, Germany
Email: simon.kassahun@tum.de, guenther@tum.de

***Abstract*—Analysing the optimization possibilities for exams**

This paper reports on the opportunities to facilitate the conventional exam evaluation process by using scanner-based evaluation software such as TUMexam. It closely analyses the time cost and error probability for evaluating a regular exam to estimate the potential improvements of automated systems based on empirical data from a previously written exam.

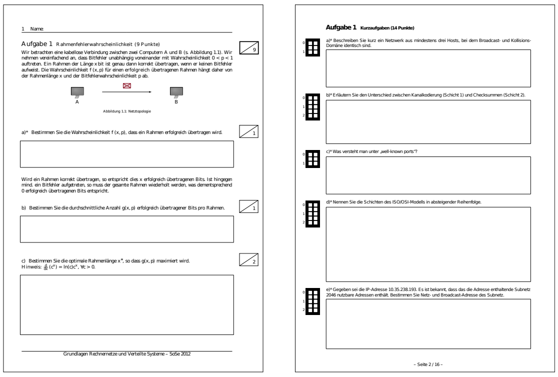
***Index Terms*—TUMexam, exam evaluation**

1. Introduction

Exams are essential to most educational institution but are often very time consuming and messy. Many institutions rely on the conventional method of manually creating and evaluating exams. These require a lot of time from the faculty to be reviewed and also leaves students waiting for potentially weeks until they can find out their result. Manually calculating the test scores also carries the risk of making mistakes, which are not only annoying for the reviewers but can also potentially negatively impact a student's credit score. There is now a variety of commercial evaluation products available with many different approaches and scopes such as for example EvaExam[1] or eSystem[2]. With the intention to address the aforementioned problems, the Chair of Network Architectures and Services from the Department of Informatics at the Technical University of Munich began developing their own software called TUMexam[3]. TUMexam has been developed since 2015 with the aim to provide solutions ranging from templates and attendance records to facilitating the preparation for the evaluation and correction. Exams are currently still being manually corrected but each problem has boxes representing the amount of points awarded for a correct answer. Instead of writing down a number, the examiner ticks the corresponding checkboxes. After that step, all exams are scanned and then digitally analysed to count the collective score as well as calculate the final grade. The software is also being offered to other chairs. However, the decisive factor for most potential users is whether a switch to the new evaluation system brings a significant improvement to the processing time. This will be the main focus of this paper.

2. Methodology

To evaluate the robustness of scanner exams, it is compared to the conventional method, specifically to how often mistakes are made in a conventional exam evaluation and how much time could potentially be saved. One example for a conventional exam and one which uses TUMexam can be seen in Figure 1. The final exam from 2012 uses one box with two sectors for each (sub)problem. The two sectors are needed for the two correction passes. In contrast the final exam from 2017 which uses TUMexam has several boxes forming a table. The two columns are used for the two correction passes and each line represents 0.5 credits.



1 Name:

Aufgabe 1 Rahmenfehlerwahrscheinlichkeit (9 Punkte)

Wir betrachten eine kabellose Verbindung zwischen zwei Computern A und B (s. Abbildung 1.1). Wir nehmen vereinfachend an, dass Bitfehler unabhängig voneinander mit Wahrscheinlichkeit 0 < p < 1 auftreten. Ein Rahmen der Länge x bit ist genau dann korrekt übertragen, wenn er keinen Bitfehler aufweist. Die Wahrscheinlichkeit f (x, p) für einen erfolgreich übertragenen Rahmen hängt daher von der Rahmenlänge x und der Bitfehlerwahrscheinlichkeit p ab.

Abbildung 1.1: Netztopologie

a)* Bestimmen Sie die Wahrscheinlichkeit f (x, p), dass ein Rahmen erfolgreich übertragen wird.

Wird ein Rahmen korrekt übertragen, so entspricht dies x erfolgreich übertragenen Bits. Ist hingegen mind. ein Bitfehler aufgetreten, so muss der gesamte Rahmen wiederholt werden, was dementsprechend 0 erfolgreich übertragenen Bits entspricht.

b) Bestimmen Sie die durchschnittliche Anzahl g(x, p) erfolgreich übertragener Bits pro Rahmen.

c) Bestimmen Sie die optimale Rahmenlänge x^*, so dass g(x, p) maximiert wird.
Hinweis: $\frac{d}{dx}(c^x) = \ln(c)c^x, \forall c > 0$.

Grundlagen Rechnernetze und Verteilte Systeme – SoSe 2012

Aufgabe 1 Kurzaufgaben (14 Punkte)

a)* Beschreiben Sie kurz ein Netzwerk aus mindestens drei Hosts, bei dem Broadcast- und Kollisions-Domäne identisch sind.

b)* Erläutern Sie den Unterschied zwischen Kanalkodierung (Schicht 1) und Checksummen (Schicht 2).

c)* Was versteht man unter „well-known ports"?

d)* Nennen Sie die Schichten des ISO/OSI-Modells in absteigender Reihenfolge.

e)* Gegeben sei die IP-Adresse 10.35.238.193. Es ist bekannt, dass das die Adresse enthaltende Subnetz 2046 nutzbare Adressen enthält. Bestimmen Sie Netz- und Broadcast-Adresse des Subnetz.

– Seite 2 / 16 –

Figure 1. A sample page of the final exam GRNVS 2012 (left) and the the final exam GRNVS 2017 (right) [1]

Exams are normally corrected in two separated sessions. During the first correction pass the various subproblems are evaluated and given a score. In the conventional way, all these scores are manually summed up, for the most part by mental arithmetic, to give each problem a score. Finally all scores for each problem are summed up to calculate the total score but a calculator is used at this stage. The final score is listed on the cover along with the results of the individual problems which determine it. Since the final score is calculated from these summed up values, any previous error from one of the subtotals also affects the final score. During the second review, all subproblems are re-evaluated and the sums recalculated. In the case that all subproblems were given the same score, all sums should stay the same as well. If theses do not match, one of the two sums must be incorrect.

The significance of the aforementioned factors are assessed by repeating the counting process for a previous exam. That means counting all credits for each problem

[1] EvaExam - https://www.evasys.de/evaexam.html
[2] eSystem - https://www.speedwellsoftware.com/exam-software/
[3] TUMexam https://www.tumexam.de/

doi: 10.2313/NET-2019-06-1_04

separately, once for the first correction pass and once for the second correction pass. The timer is started as soon as the correct page is opened to solely record the time spent on summing up credits. After each step the time is measured and it is noted whether an error occurred. During the first review the timer is stopped as soon as a result has been calculated. It is also recorded whether an error occurred or not but this is only done for evaluation purposes. The timespan for marking an error is not included in the noted time as it would normally be impossible to tell if the first correction pass was free of errors without a second correction pass. When such a deviation between the summed up credits and the score listed on the exam is found, it is noted which of these two numbers are incorrect and and by what margin. The second pass is very similar but the timer continues until it is clear which score is correct. Since a calculator is usually used to calculate the final score on the cover, a calculator is also used for that specific part.

There are two types of errors which are counted and evaluated in this paper. An exam error describes the case, where the score written on the original exam is incorrect. A counting error describes the case, where the calculated score is wrong.

TUMexam automates the process of counting scores and calculating the grade. Because exams are specifically designed for TUMexam, the software only needs to distinguish between a ticked box and an empty one. It also flags unclear marks for later review. Since no case is known so far where TUMexam calculated an incorrect score, number of counting errors is here assumed to be zero.

3. Implementation

The exam used for this test was a final exam with a time frame of 90 minutes. It is the final exam from the year 2011 of the course Introduction to Computer Networking and Distributed Systems. The exam has 5 different problems, each of which includes multiple subproblems making up the combined score for that problem. 192 individual exam sheets were reviewed for this paper.

3.1. Errors

Errors are to be expected and can be very troublesome for the correctors as well as the students. Since they can lead to much time being spent on finding the mistake up to potentially lowering a student's credits score when they go unnoticed, it is very desirable to keep the amount of errors and their impact as low as possible.

When calculating the credit score 2112 times (11 scores are calculated per exam), 81 individual errors were found in total. This group consists of 24 deviations in which the score written on the exam is false and 57 cases in which the deviation is a counting error. The difference between the wrong score and the actual score when an error occurred is very similar between the two type of errors. On average, errors on the exam differ from the correct score by 1.27 credits, while counting errors are off by 1.07 credits.

Also noteworthy is the overall distribution of the errors. As can be seen in Figure 2, almost no errors occurred when summing up credits in Section 4. This is explained by the fact that the fourth problem is the shortest one (see Figure 3), has the lowest attainable score, and consisted of only one double page.

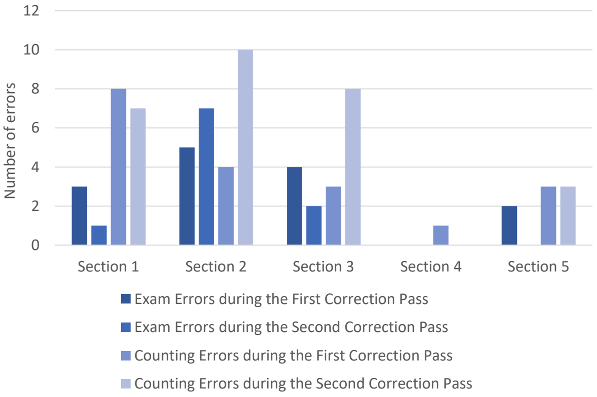


Figure 2. The number of errors per problem and type

3.1.1. Exam Errors.

Out of the 24 cases where an exam had an incorrect score listed on it, 10 errors were found on the second correction pass. This effectively means that 0.86% of all problems have an incorrect score. In the worst case scenario, as many as 5.2% of the exams could have an incorrect credit score. The other 14 errors were made during the first correction pass.

3.1.2. Counting Errors.

Counting errors did not occur consistently across the test. Only 19 errors were made during the first correction. Another 28 errors happened during the second correction and 10 when counting the combined scores on the cover.

3.2. Time

As the second key factor for an efficient and successful exam evaluation, a large time frame has probably a more noticeable impact and could potentially be greatly improved with automation. To find out how much time could be saved in the future, each part of a single exam sheet is measured individually and compared to the same part of the other correction pass and other exam sheets different. The results are shown in Figure 3. The dark blue graph represents the first pass and the light blue graph represents the second pass. The results are grouped together to improve visibility.

3.2.1. First correction pass.

During the first correction pass the time needed to sum up credits solely relies on how long it takes the corrector to calculate a result. Since there is not any comparable credit score to verify the result, there is no need to recalculate the result again and rather helps identify errors for the second correctors when credit scores do not match. This influencing factor of needing to recalculate a score can also be observed in the small standard deviation of 5.1 s in relation to the entire first correction pass.

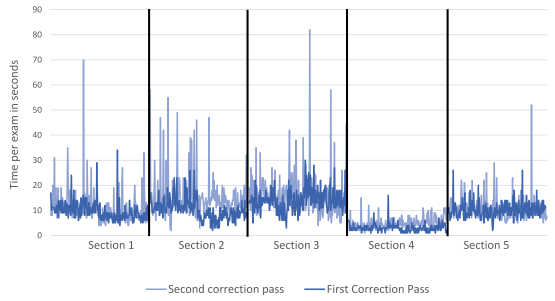


Figure 3. Comparison of the time spent on the two correction passes per section

Figure 4. Box plot of the time spent per problem during the first correction pass

3.2.2. Second correction pass. The key difference of the second correction pass from the previous one it that errors made when summing up credits are noticeable to the corrector. The time spent on noticing such errors, finding the cause, and deciding which score is ultimately correct are now included in the measured time. When comparing these new results, we can see a significant increase of the variance. The median itself has not changed much.

Looking at the box plots (Figure 4 and 5), the individual problems have a very similar pattern but there are far more extreme outliers.

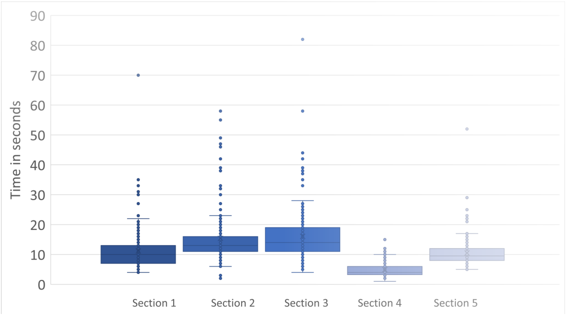


Figure 5. Box plot of the time spent per problem during the second correction pass

3.3. Estimations for an entire exam

The exam cover has two rows for both correction passes, however since they are distinct from each other, unlike the other parts of the exam, only the second correction pass was counted and then doubled when calculating the total time. Not taking any other factors, such as interruptions or navigating to the correct page, into account, a single exam takes on average 131.8 s. All 192 exams combined take 25 294 s, or approximately 7 h of pure counting time.

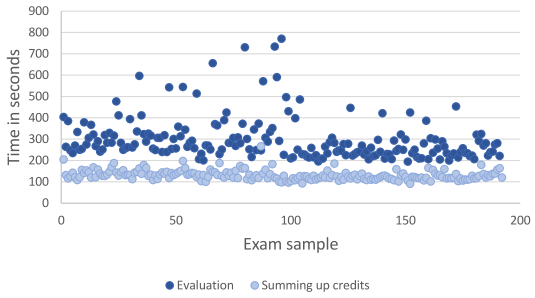


Figure 6. The time spent per exam

To put the calulated time for summing up all scores into perspective, it is compared to the time spent evaluating the same exam sheets. Since all records were saved along with their time stamps, it is possible to loosely reconstruct a more realistic time frame for procedures where exams need to be opened first and also takes other short interruptions into account. To make sure no major events affect this result, all breaks of longer than 10 min have been excluded. With these measurements an entire exam can take vaguely between 14 h and 15 h, slightly more than double the time it took to sum up credits. Figure 6 shows how the times compare for an invidiual sample.

3.4. Impact of errors on the time

To further analyse how big the impact of counting errors is, they are compared to the majority of samples where no error was made. Since errors do not have any affect on the first correction pass timewise, they are not included in this part. However, all errors which happened during the second correction pass are relevant but regardless of the type of error because both require more time. Both error types are therefore included. Figure 7 shows all cases where a errors was made and how much time they consumed relative to an average case of the same section when no errors was made. It shows that a case where an error occurs will on average take 246 % the time

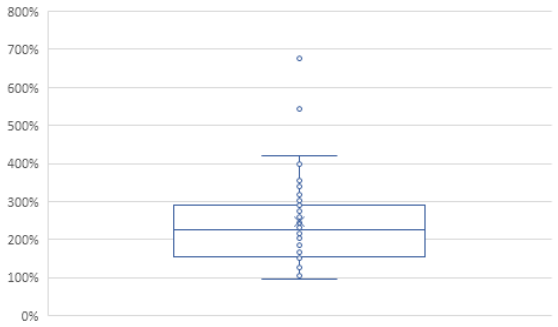


Figure 7. Box plot of the time spent per error relative to the average time spent on a sample of that problem where no error occurred

it would have taken if no error occured. The median is 224 %. Using the average time for a case without an error as a baseline, it is possible to calculate an estimate for the time all problems (without the cover) would require if the second correction pass was without any errors. The entire second pass consumed 3.1 h. Without and errors it would be around 2.9 h, saving approximately 12 min.

4. Conclusion and future work

An average 90 min exam with 192 participants and 5 problems takes under ideal circumstances require 7 hours to sum up all credits and calculate a final score. This number can be considered a rather low estimate and is very likely to increase a lot when including other factors. In a more realistic scenario you also have to factor in breaks, distraction and other interruptions. Depending on the workflow the time needed can increase to double the length or more. In this test only the time it takes to open the exam, note the test results and ordinary interruptions were factored in. Even though no breaks longer than 10 minutes were included, it raised the time required for the same exam evaluation significantly, to around 14 to 15 hours. While this is not an insignificant amount of time, it does also heavily depend on the number of participants as well as the scope of the exam.

0.86% of all problems had an incorrect score listed for the second correction pass. While this number might not seem significant, it is when put into the context of the entire exam. If each error occurred on a different exam, this would result in 5.2% of all exams having an incorrect score. None of the scores on the cover were calculated incorrectly (which does not include subsequent errors caused by incorrect scores from one of the problems). This is probably explained by the fact that a calculator is usually used for this part. One possible conclusion from this fact could be that using a calculator will likely decrease the number of errors. However using a calculator is also not risk free, as seen by the 10 counting errors which occurred when recounting the final score. Assuming all exam problems have a comparable credit score, a calculator also will not reduce the time by a significant margin as seen in Section 3.4

It is also worth mentioning that these evaluations are limited to the time consumed solely by summing up the credits. Streamlining other parts of the evaluation process, for instance by using software to automatically evaluate exercises and assist with the preparation of an exam, can also have huge advantages.

Conclusively, since TUMexam produces virtually no errors when summing up credits, there can be significant benefits to optimizing a conventional exam evaluation process in both time and error probability. It is likely that much smaller exams do not benefit from automation in a significant amount and even larger exams can expect to see much better results. However, this is out of the scope of this paper and could be evaluated in the future. Another possible direction could be to analyse other parts of the exam preparation and evaluation process, for example creating the exam problems, and how much time could then be saved by utilizing software.

References

[1] Archive of previous exams for the course "Introduction to Computer Networking and Distributed Systems at TUM", https://grnvs.net.in.tum.de/altklausuren/