Original scientific paper
Accepted 31. 12. 2015.

Blaženka Divjak
Marcel Maretić

Geometry for Learning Analytics

Geometry for Learning Analytics

ABSTRACT

Learning analytics is focused on the educational challenge of optimizing opportunities for meaningful learning.

Assessment deeply influences learning, but at the same time data about assessment are rarely considered and utilized by learning analytics.

Current approaches to analysis and reasoning about peer-assessment lack rigor and appropriate measures of reliability assessment. Our paper addresses these issues with a geometrical model based on the taxicab geometry and the use of the scoring rubrics.

We propose and justify measures for calculation of the final grade in peer-assessment and related inter-rater and intra-rater reliability measures. We present and discuss a geometrical model for two important peer-assessment scenarios.

Key words: taxicab geometry, metrics, learning analytics

MSC2010: 53A35, 91E45, 97B10

Geometrija za analitike učenja

SAŽETAK

Analitike učenja usredotočene su na obrazovne izazove vezane uz postizanje svrsishodnog učenja. Vrednovanje postizanja ishoda učenja izrazito utječe na učenje. Međutim, podaci o procesu vrednovanja vrlo rijetko se koriste u postojećim analitikama učenja. Nadalje, postojeće implementacije i analize procesa istorazinskog (vršnjačkog) vrednovanja nisu zadovoljavajuće. Ovaj rad predstavlja izradu i upotrebu matematičkog modela za opis i računanje vezano uz istorazinsko vrednovanje. Razvijeni model zasniva se na Manhattan (taxicab) metrici te korištenju rubrika za vrednovanje ishoda učenja. U radu su opisane i opravdane metode računanja konačne ocjene vršnjačkog vrednovanja, mjere pouzdanosti takvog vrednovanja kao i ocjene za pojedine vrednovatelje. Razvijeni geometrijski model razmatran je u kontekstu dva važna scenarija istorazinskog vrednovanja.

Ključne riječi: taxicab geometrija, metrika, analitika učenja

1 Introduction and motivation for research

1.1 Learning analytics and related challenges

Learning analytics (LA) belongs to interdisciplinary scientific fields connected to educational sciences and technology enhanced learning that has emerged rapidly in last five years. The most cited definition of LA is ”*Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimizing, learning and the environment in which it occurs*”. In [8] it is stated that the definition comes from the first international Conference on Learning Analytics and Knowledge (LAK 2011) and adopted by the *Society for Learning Analytics Research* (SoLAR). Further, Ferguson in [8] states that LA is focused on the educational challenge: *How can we optimize opportunities for online learning?* Even better, we should look for opportunities for *meaningful* learning.

Research methods and methodology in LA are still very much under development. There is a great opportunity for mathematicians to contribute to development of various kind of measures and the research of mathematical models. There is a vocal support for broadening the scope and usefulness of LA and special issue in LA is research in student assessment (cf. [5]).

1.2 Assessment and reliability measures

Assessment is of fundamental importance to students. It deeply influences learning. At the same time assessment data are rarely utilized by learning analytics. One of the possible reasons is that available data is not granular enough. Fundamental issues of peer-assessment are reliability and validity (cf. [7]). Research on indicators and metrics to be potentially used in the context of reliability and validity of assessment, peer-assessment and self-assessment, is (currently) very limited.

1.2.1 MOOC, online learning context

Completely new playground for learning analytics the so called *networked learning* [14], e.g. Massive Open Online Courses (MOOCs), social learning platforms, online learning and e-learning in general. In networked learning the number of participants rapidly increases as well as the interactions between learners in the form of discussions and mutual learning. Dealing with tens of thousands of learners in one MOOC it is very natural/appropriate to use self-assessment for tasks leading to a certificate. This approach generates huge amount of assessment data but also asks for sound metrics for calculation of final grade and for estimates on the reliability of assessment. For our work in this paper it brings forwards challenging scenario when we have inexperienced evaluators (scenario A). In this case we will demand more peer-evaluations per assignment to attain sufficient reliability. The second discussed scenario corresponds to a situation with expert (or experienced) evaluators assessing a complex problem solving task. Here we must take into account that experts' time is expensive and their judgments, but their evaluations can be trusted (scenario B). A number of assessments per assignment can be lower in scenario B. Further, in scenario B we can anticipate for situations that some evaluators are experts for only some of the assessment criteria. For example, an expert in project management and scheduling can skip assessment for criteria on financial regulation if he/she lacks the required expertise.

1.2.2 Educational Rubrics

In order to increase transparency of assessment criteria and validity of assessment us of a scoring rubric is highly recommended (cf. [12]). Further, data from the rubric can be analyzed and utilized for estimation of reliability. Among several sets of assessment data, [5] mentions "*achievement mapped against explicit learning outcomes or assessment criteria (e.g. rubric results)*".

A widespread definition of the educational rubric describes it as *a scoring tool for qualitative rating of authentic or complex student work* [12]. A rubric consists of grading criteria and standards of attainment for those criteria (examples: [3, 4]).

Using rubrics provides several benefits such as increased consistency of assessment, attainment of the desired validity in assessment without sacrificing the need for reliability and promotion of learning [12].

Previous research claims that the use of rubrics in mathematics supports students' reflection and critical skills (deep learning) by clearly communicating what is asked from them [3].

Table 1 illustrates the scoring rubric for one criterium (whole scoring rubric is available in [3]). It refers to an assignment in a mathematics course where student had to relate a real world problem to the course material.

Rubrics are especially useful when more than one teacher/student is involved in the process of assessment. Grading can then be implemented as a combination of teacher's grading and automated grading. Rubrics are also vital in the case of a complex task assessment including problem-based learning, group work or peer-assessment that are authentic to the skills being tested (cf. [3]). Peer-assessment is a process where students grade assignments or tests of their peers based on teacher's benchmarks (cf. [16]).

Peer-assessment has several advantages over traditional (teacher) assessment and a few very strong disadvantages – comprehensively described and systematized in [4]. One known disadvantage is the so called "*reliability risk*" introduced by the fact that students are assessing their own peers – some of whom may be their friends. The teacher must be aware of the included risks and anonymize assessment tasks whenever possible (see [4]). Influential papers [12, 16] claim that the measurement of reliability is a problem for both peer-assessment *and* the use of scoring rubrics.

Table 1: *An example: Grading the "problem description"-criterium with a rubric (only one row of the rubric is shown)*

	0 points	1 points	2 points	3 points
problem description	*poor description, irrelevant context*	*problem is described but has no connection to the prescribed context*	*description of the problem is presented in a clear and interesting fashion but lacks the relevant context*	*problem is described in a clear and interesting fashion and is positioned/placed in a relevant real context*
⋮	…	…	…	…

We have to be aware that at this moment most teachers just use the available software (like Moodle Workshop) oblivious whether of the fact that the embedded metrics are not well justified or even missing (cf. [17]). Our paper addresses these problems with a geometrical model based on the taxicab geometry.

1.3 Research questions

We pose three research questions.

RQ1: How to model and implement the grading process for peer-assessment?

RQ2: How to calculate the grade in peer-assessment?

RQ3: What are the appropriate inter-rater (agreement among graders) and intra-rater (accuracy of a single grader for his several grading efforts) measures for peer-assessment?

In the following sections we are going to answer abovementioned research questions.

2 RQ1 – Problem description and modeling

Let us assume that students' assignments are graded with the help of a scoring rubric with n criteria. Students participate in an activity for which they are graded by their peers. Each participant is asked to grade several (i.e. 3) assignments, and consequently each student should receive several gradings for his own assignment.
The set of participating students is enumerated and we speak of student k, or assignment k (instead of the assignment of student k).

A particular grading is represented as a point in an n-dimensional vector space. Let $\mathcal{S}_k = \{S_k^1, \ldots, S_k^m\}$ denote a set of gradings for assignment k where

$$\begin{aligned} S_k^1 &= (c_{k,1}^{(1)}, \ldots, c_{k,n}^{(1)}) \\ S_k^2 &= (c_{k,1}^{(2)}, \ldots, c_{k,n}^{(2)}) \\ &\vdots \\ S_k^m &= (c_{k,1}^{(m)}, \ldots, c_{k,n}^{(m)}). \end{aligned}$$

Optionally, some assignments receive teacher's grading

$$T_k = (c_{k,1}^T, \ldots, c_{k,n}^T).$$

It is expected to have the teachers grade only a selection of assignments. If present, teacher's grading T_k is taken as a proper (final) grade for assignment k. The intent is to have teachers intervene (providing T_k) only in cases where received peer-assessments for a task k are indicated/detected as unreliable.

Without the loss of generality we assume nonnegative grades $c_{i,j}^{(k)} \geq 0$. Ranges of points for criteria C_i are determined by the scoring rubric. We encode this as coordinates of a range vector

$$\mathbf{r} = (r_1, r_2, \ldots, r_n). \tag{1}$$

Values r_i communicate the relative weights of criteria C_i and must be carefully determined in advance during the design of the scoring rubric.

We need to calculate:

- the final grade for the assignment
- the measure for (inter-rater) reliability of gradings given for an assignment k (as deviation/divergence of the gradings)
- the assessment of the quality of gradings of a particular grader – (intra-rater, for “grading the grader”).

Inter-rater reliability measures agreement among graders for grading the same assignment. Intra-rater reliability tells how good of a grader some is – it measures how k's performed gradings agree with other (final) grades for these assignments.

Remark 1. *A 3-dimensional array is needed for storing $c_{i,j}^{(k)}$ data. For example, data of* 2000 *records is needed if $m = 4$ (number of desired peer gradings), $p = 100$ (class size) and $n = 5$ (number of rubrics criteria).*
In a MOOC setting a reality is to have data of millions of record.

2.1 Taxicab geometry

Taxicab geometry is one of non-Euclidean geometries introduced by Hermann Minkowski (1864 – 1909) at the turn of the 20-th century.
H. Minkowski described a set of metrics that can be used to measure distance and which satisfies the axioms of the metric space. These metrics are induced by the so called p-norms defined for every $p \in \mathbb{R}$, $p \geq 1$ as real scalar functions in n-dimensional space.
Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$.
p-norm for $p \geq 1$ is defined by

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i| \right)^{\frac{1}{p}}, \tag{2}$$

which induces the associated p-metric

$$d_p(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|_p. \tag{3}$$

Specifically, for $p = 1$ the induced metric is

$$d_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \ ,$$

better known as taxicab or Manhattan distance.

Taxicab geometry was named by Karl Menger in 1952. at the exhibit at the Museum of Science and Industry of Chicago. The justification for the name is straightforward – it measures distances traveled by a car in a city whose streets are laid out in a rectangular grid. Note that the shortest path in taxicab geometry is not unique (see Fig. 1).

For $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, $1 < r < p$, the following inequality holds

$$\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_r \leq n^{\frac{1}{r}-\frac{1}{p}} \|\mathbf{x}\|_p \, , \tag{4}$$

meaning that p-norms are equivalent. Specifically, this gives

$$d_2(\mathbf{x},\mathbf{y}) \leq d_1(\mathbf{x},\mathbf{y}) \leq \sqrt{n} \cdot d_2(\mathbf{x},\mathbf{y}). \tag{5}$$

Euclidean distance to a taxicab driver represents air distance, e.g. it provides a lower bound for the length of a minimal trip between A and B. On the other hand, taxicab distance *is the exact* distance that needs to be traveled from A to B in a grid.

The taxicab metric d_1 will be denoted as d from now on.

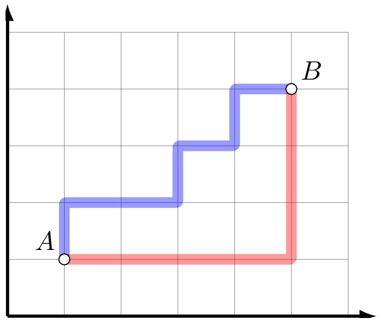

Figure 1: 2*-dimensional illustration of shortest taxicab paths between A, B*

Hypersphere in the n-dimensional taxicab space with center $C(c_1, c_2, \ldots, c_n)$ and a radius r is a locus of points satisfying

$$\sum_{i=1}^{n} |x_i - c_i| = r \ . \tag{6}$$

A 2-dimensional taxicab circle is shown inscribed in the Euclidean circle of the same radius in the Fig. 2.

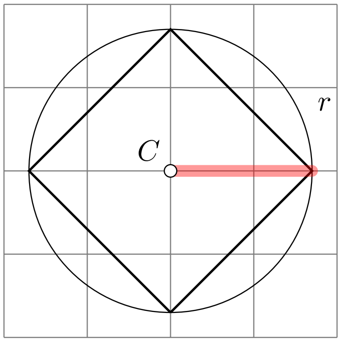

Figure 2: *Euclidean and taxicab* 2*-dimensional sphere with center C and radius r*

We argue that taxicab metric is adequate as a foundation for our model of LA of peer-assessment as possible rubric gradings are points laid out in hyper-rectangular grid. Taxicab metric is (topologically) equivalent to (but simpler than) Euclidean metrics. It is linear, e.g. distance on the criteria level contributes exactly the same amount to the total taxicab distance.

Finally, the total amount of points awarded for an assignment with a final grade $G = (g_1, \ldots, g_n)$ is exactly the 1-norm of G

$$|G| = g_1 + g_2 + \cdots + g_n \, , \qquad g_i \geq 0. \tag{7}$$

3 RQ2 – How to calculate the final grade in peer-assessment?

We propose and analyze two approaches for the calculation of the final grade. Let $\mathcal{S} = \{S_k^1, \ldots, S_k^m\}$ denote a set of peer gradings for assignment k.

3.1 Mean value final grade

Traditionally the final grade is calculated as an arithmetic mean of available gradings:

$$M(\mathcal{S}) = (a_1^f, \ldots, a_n^f), \quad \text{where} \quad a_i^f = \frac{1}{m}\left(\sum_{j=1}^{m} c_{k,i}^{(j)}\right). \tag{8}$$

$M(\mathcal{S})$ is a center of mass of the set $\mathcal{S}$. Mean value final grade is sensitive to all available data (gradings). We can say that it is mostly sensitive to quantity, and less sensitive to extremes (outliers). Mean value final grade ”*respects the decision of the majority*”. More data produces better results. We consider it appropriate for scenario A.

3.2 Optimal final grade

We define $W(\mathcal{S})$ and $B(\mathcal{S})$

$$W(\mathcal{S}) = (w_1,\ldots,w_n), \qquad w_i = \min_j c_{k,i}^{(j)},$$

$$B(\mathcal{S}) = (b_1,\ldots,b_n), \qquad b_i = \max_j c_{k,i}^{(j)},$$

as amalgamation of the worst received grades (W) and best received grades (B) respectively. We define the **optimal final grade**

$$O(\mathcal{S}) = (o_1^f,\ldots,o_n^f), \quad \text{where} \quad o_i^f = \frac{1}{2}\left(W(\mathcal{S})+B(\mathcal{S})\right) \ . \tag{9}$$

Optimal final grade takes into consideration (is sensitive to) extremes: e.g. additional gradings within an axis-aligned hyperrectangle (box, see Fig. 3) encompassing the set $\mathcal{S}$ have no effect on a_k^f.

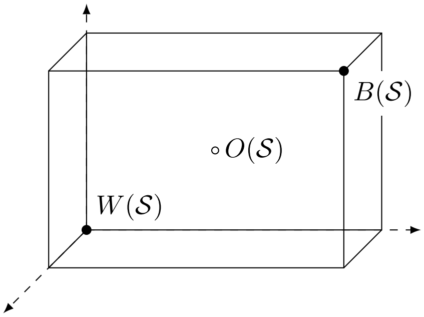

Figure 3: *A sketch of a 3-dimensional axis-aligned hyperrectangle*

Note that $W(\mathcal{S})$ and $B(\mathcal{S})$ as two juxtaposed vertices uniquely determine the hyperrectangle encompassing $\mathcal{S}$.

This approach is inspired by the TOPSIS method (*Technique for Order of Preference by Similarity to Ideal Solution*) in multi-criteria decision making (cf. [10]).

We find the optimal final grade approach adequate in situations where grading is performed by experts (i.e. several teachers), of whom abnormal/wild gradings are not expected (scenario B). After only a few expert gradings additional gradings should have little to no effect on the final grade. Expert's costs rise linearly, but benefits wane quickly. For this reason a balance must be struck to avoid overloading the experts with a workload that will have no effect.

3.2.1 Relative position of $M(\mathcal{S})$ and $O(\mathcal{S})$

$M(\mathcal{S})$ is positioned within the axis-aligned hyperrectangle encompassing $W(\mathcal{S}), B(\mathcal{S})$. Therefore

$$|M(\mathcal{S})-O(\mathcal{S})| \le \frac{1}{2}|B(\mathcal{S})-W(\mathcal{S})| = \frac{1}{2}\left(|B(\mathcal{S})|-|W(\mathcal{S})|\right) \tag{10}$$

Relatively large $|M(\mathcal{S})-O(\mathcal{S})|$ indicates a skewed data $\mathcal{S}$ with majority of data standing opposite to an outlier point.

The inequality

$$0 \le \frac{2|M(\mathcal{S})-O(\mathcal{S})|}{|B(\mathcal{S})|-|W(\mathcal{S})|} \le 1 \tag{11}$$

resulting from (10) can be utilized for a normalized measure of skewness of the set $\mathcal{S}$.

Remark 2. *When teacher grading T_k is present, T_k is taken as a proper grade for assignment k.*

Example 1. *Grading set $\mathcal{S} = \{S_k^1, S_k^2, S_k^3\}$ for assignment k is given in the following table:*

	C_1	C_2	C_3	C_4	Σ
S_k^1	3	0	2	2	7
S_k^2	2	1	3	3	9
S_k^3	2	1	3	2	8

We can calculate the final grade for assignment k with (8) *and* (9)*:*

$$M(\mathcal{S}) = \frac{1}{3}(7,2,8,7), \quad |M(\mathcal{S})| = 8, \tag{12}$$

$$O(\mathcal{S}) = \frac{1}{2}\left((2,0,2,2)+(3,1,3,3)\right) = \frac{1}{2}(5,1,5,5), \quad |O(\mathcal{S})| = 8. \tag{13}$$

Note that looking at the total grade these assessment match, but they are far from agreement on granulated grades. Summative difference is 2*. Obviously, final grades can differ if calculated with mean and optimal value final grade. But, it can happen, as illustrated by this example, that agreement among evaluators is low.*

Suppose a teacher intervenes with

	C_1	C_2	C_3	C_4	Σ
T_k	2	1	2	2	7

.

Now T_k is a proper (final) grade for assignment k.

Gradings closer to the final grade (which is T_k in this case) are considered to be of better quality, and respective graders should be rewarded with more points for grading well. Here, for example, S_k^3 is the closest to the final grade T_k with $d(S_k^3, T_k) = 1$. S_k^1 and S_k^2 both have taxicab distance from T_k of 2 points.

4 RQ3 – What are the appropriate inter-rater and intra-rater measures for peer-assessment?

Main objectives regarding RQ3 are:

i. detection of inadequate grading set by measuring the agreement within a grading set,

ii. grading (rewarding) the grader proportionally to the measure quality of his effort.

4.1 The need for higher granularity of assessment data

We illustrate this with an example of "bad" grading that is visible only when analyzed at the higher level of detail.

Example 2. *Let's consider example gradings S_1 and S_2:*

$$\begin{array}{c|cccc|c} & C_1 & C_2 & C_3 & C_4 & \sum \\ \hline S_1 & 3 & 0 & 2 & 2 & 7 \\ S_2 & 1 & 1 & 3 & 3 & 8 \end{array} \left.\vphantom{\begin{array}{c}7\\8\end{array}}\right\} \begin{array}{c}\text{summative}\\ \Delta = 1\end{array}$$

$$\underbrace{\qquad\qquad\qquad}_{\substack{\text{granular} \\ d(S_1,S_2) \;=\; 5}}$$

Difference of totals (summative difference) for S_1 and S_2 is only 1 point, but the taxicab distance (sum of differences) is 5 points. Although gradings S_1 and S_2 seem coherent at the summative level, these gradings indicate a low quality of assessment(s) when observed at greater level of detail (criteria level).

We will use a set diameter as a measure for detection of inconsistent gradings. Grading set with a large diameter suggests inconsistent and possibly unreliable peer-assessment. A **diameter** of a set of gradings $\mathcal{S} = \{S_1, \dots, S_n\}$ is defined as
$\mathtt{diam}\,\mathcal{S} = \max_{i,j} d(S_i, S_j)$.

$\mathtt{diam}\,\mathcal{S}$ is also a diameter of a sphere encompassing $\mathcal{S}$. Note that, unlike in the Euclidean geometry, the encompassing sphere of the set $\mathcal{S}$ is *not* unique (see Fig. 4).
Any sphere of diameter $d(A,B)$ within the lightly shaded region of Fig. 4 is an encompassing sphere of $\{A,B\}$. This region is the intersection of two taxicab hyperspheres of radius $d(A,B)$ with centers A and B.

Let $e > 0$. A grading $\mathcal{S}$ is **acceptable** for an acceptable error e if the radius of the smallest encompassing sphere of $\mathcal{S}$ is smaller than e, i.e. if
$\mathtt{diam}\,\mathcal{S} < 2e$.

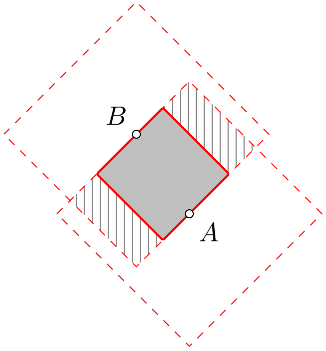

Figure 4: *Encompassing taxicab sphere for A, B*

4.2 Normalization

For the purpose of standardization (for comparison of results) and for easier application and interpretation of results by non-expert users (where acceptable error e can be set and recommended on a normalized $[0,1]$ scale) we introduce the normalization of the taxicab norm.
We define the norm $|\ |'$ for the points within the hyper-rectangle encompassing O and $\mathbf{r}$

$$|\mathbf{a}|' = \frac{1}{|\mathbf{r}|}\left(a_1, \dots, a_n\right) , \tag{14}$$

where $\mathbf{r}$ is the range vector (see equation (1)). Both $|\ |'$ and the induced metric $|\mathbf{x} - \mathbf{y}|'$ map to $[0,1]$ on the $(O,\mathbf{r})$-hyper-rectangle.

Since relative weights of criteria have already been taken into account in the design phase of the scoring rubric the normalization is simple. Any concerns about disparate sizes of r_i in some rubrics have to be addressed during the design of the scoring rubric.

Now we can use the relative acceptable error $e' = \frac{e}{|\mathbf{r}|}$ instead of e.

5 Implementation of the grading process

5.1 Simple grading process

Let $e > 0$ be acceptable error. Let $\mathcal{S}$ be a grading set for assignment k. Let g be a grading method (mean value final grade or optimal final grade).
If $\mathcal{S}$ is acceptable, a final grade $g(\mathcal{S})$ is assigned for assignment k. If $\mathcal{S}$ is not acceptable, we ask for teacher's grading.

5.1.1 Advantages and disadvantages of the simple grading process

In a situation with a cluster of gradings of poor quality, a single grading of good quality can be enough to demand

supervision (and rectify the situation). Also, simple grading is computationally very simple (of linear complexity). On the other hand, one outlying grading is enough to ”spoil” the grading set. Therefore, simple grading process demands supervision even when it would be easy to detect and eliminate an outlier.
Simple grading does not scale well. It can quickly become too work-intensive and overwhelming for the teacher even when the majority of gradings is performed by the peers and LMS. Simple grading is adequate for a face-to-face blended classroom and online classes of manageable size (scenario A).

5.2 Autonomous approach to grading

For scenario B we suggest a semi-autonomous approach – where we search for a maximal acceptable subset $\mathcal{S}'$ of $\mathcal{S}$ of minimal diameter. We must set the lower bound N (critical size) for $\#(\mathcal{S}')$ (number of assessments in $\mathcal{S}'$). Teacher's intervention will be asked for only if no such $\mathcal{S}'$ can be found.

This approach is described step-by-step in algorithm 1.

Algorithm 1: Semi-autonomous Grading Process

```
input  : Set of gradings $\mathcal{S} = \{S^{(1)}, \dots, S^{(m)}\}$,
         acceptable error $e \geq 0$
         grading calculation method $g$
         critical size $N$ (i.e. $N = 3$)
output:  Final grade or indicate gradings $\mathcal{S}$ as invalid

find a maximal $\mathcal{S}' \subseteq \mathcal{S}$ with acceptable error
if $\#(\mathcal{S}') \geq N$ then
    find $\mathcal{S}''$ of minimal diameter such that $\#(\mathcal{S}'') = \#(\mathcal{S}')$
    return $g(\mathcal{S}'')$ as a proper grade for assignment $k$
else
    Ask for teacher intervention (grading)
```

5.2.1 Advantages and disadvantages of the semi-autonomous grading process

One outlying grading is not enough to ”spoil” the computed grade. Discarding this grade can produce an acceptable set $\mathcal{S}'$ from which the final grade can be computed.
However, a cluster of bad gradings can prevail without demanding supervision, resulting in the wrong final grade being awarded. Additionally, as our intent is to reward good graders, in the case when good graders form a smaller cluster, they could even be unfairly ”penalized” for their effort.
Autonomous grading is somewhat computationally more intensive than the alternative method (`diam` for pairwise distances must be calculated for each subset tested).

Remark 3. *Note that $\mathcal{S}$ declared acceptable by the simple approach is also acceptable by algorithm 1. Therefore, we may consider that the simple method reflects the cautious approach to the grade calculation, whereas algorithm 1 attempts to be as autonomous as possible.*

Remark 4. *$\mathcal{S}$ with a tight cluster of gradings of poor quality passes as acceptable by both approaches. A good practice would be to give the student the opportunity to contest the received final grade as a safety net for catching errors.*

6 RQ3 –How to model the evaluation for peer-assessment?

Let $G_1, \dots, G_m$ be gradings of different assignments performed by student k. Let F_i be a final grade for the task graded by G_i (F_i and G_i are grades for assignment i). Let $e > 0$ be radius of acceptable grading (acceptable error). Let $d_i = d(G_i, F_i)$.

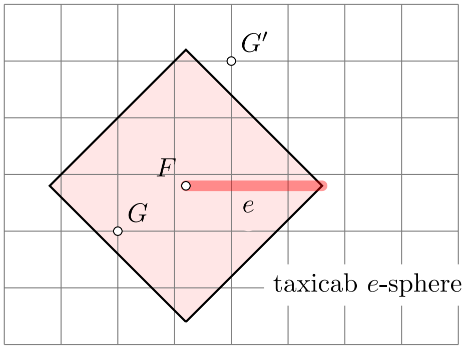

Figure 5

Fig. 5 shows an example of an e-sphere around F with two peer-assessments G i G'. The idea is to award the student who produced grading G that lies within the sphere proportional to $e - d(F,G)$. On the other hand, a student who produced grading G' will not receive points for his grading because G' lies outside of this sphere.

If we intend to award a maximum of A points for the task of peer-assessment grader k can be awarded A_i points for his grading G_i, where A_i is calculated by the following formula

$$A_i(d_i) := \begin{cases} \dfrac{A}{me}\left(e - d_i\right), & d_i < e \\ 0, & d_i \geq e \end{cases} .$$

If G_i is within the e-sphere around F_i, student k is rewarded an amount of points proportional to $(e - d_i)$ for each G_i for a maximum of A/m (see fig. 6).

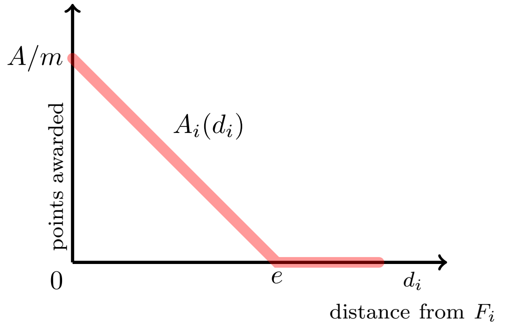

Figure 6: *Award A_i for grading G_i with $d_i = d(G_i, F)$*

Finally, grader k is awarded a total of $A(k)$ points for his effort with gradings $G_1, \ldots, G_m$ where $A(k) = \sum_{i=1}^{m} A_i(d_i)$.

6.1 Analysis of acceptable error

e must be set in advance (arbitrarily). But after all the data has been collected (after peer-assessment) we can analyze e to see how good was our aim for e. More precisely, we could analyze what would be the expected gain of ”lazy” peer-assessment.

A good practice in general for addressing this issue is to ask for mandatory written argumentation/explanation along each grade in peer-assessment.

Remark 5. *We have analyzed data gathered in a course UPC at FOI, University of Zagreb. We argue that grading at the summative level is not granular enough for making judgments about the quality of the grade, as was illustrated in example 2.*

Calculated Pearson correlation of summative and granular difference for data in our case study is 0.57. This indicates that a significant proportion of gradings may seem reliable at the summative level, while being inconsistent at the criteria level (the exact proportion is 12/62 for the UPC course).

7 Conclusion and future research

Peer-assessment and its analysis present an interesting challenge in LA. It is an important topic in LA because peer-assessment actively enhances the learning process and contributes to deeper learning. Peer-assessment is becoming a necessity in a MOOC setting. The obvious appeal of peer-assessment is the delegation of assessment workload from teacher to students. In a MOOC traditional forms of teacher's assessment quickly become impractical because of the vast number of participants. However, even with peer-assessment teachers remain heavily involved as now they must conduct and supervise the assessment process.

The assessment of complex problem-solving tasks is the recommended application of the discussed peer-assessment model. Mechanical grading (easily performed by a computer or LMS) is not applicable in the context of complex problem solving. We have presented a well founded LA model of peer-assessment. Issues and concerns that arise in peer-assessment are dealt with and measured with a mathematical model based on the use of scoring rubric. Finally, a framework for rewarding the graders for their peer-grading effort is proposed. Measures regarding reliability in this mathematical model are based on the taxicab geometry.

We look forward to opportunities for the testing of our model. Also, we are interested in applicability of this model in other settings: i.e., a similar use case is evaluation of project applications and applicability of taxicab geometry in decision making. An interesting feature to consider would be to allow partial gradings, because even experts are not experts for assessing all criteria.

Also, we are currently working on a implementation of our model of peer-assessment. We intend to release it as a plug-in for the Moodle LMS.

References

[1] B. DIVJAK, Notes on Taxicab Geometry, *KoG* **5** (2000), 5–9.

[2] B. DIVJAK, Implementation of Learning Outcomes in Mathematics for Non-Mathematics Major by Using E-Learning, in *Teaching Mathematics Online: Emergent Technologies and Methodologies*, A. A. JUAN, M. A. HUERTAS, S. TRENHOLM, C. STEEGMANN, Eds. IGI Global, 2012, 119–140.

[3] B. DIVJAK, Assessment of Complex, Non-Structured Mathematical Problems, *IMA International Conference on Barriers and Enablers to Learning Maths*, 2015.

[4] B. DIVJAK, M. MARETIĆ, Learning Analytics for e-Assessment: The State of the Art and One Case Study, *Central European Conference on Information and Intelligent Systems*, 2015.

[5] C. ELLIS, Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics, *British Journal of Educational Technology* **44**(4) (2013), 662–664.

[6] N. J. Entwistle, Approaches to studying and perceptions of university teaching-learning environments: concepts, inventory design and preliminary findings, *Powerful learning environments: Unravelling basic components*, 2003, 89–108.

[7] N. J. Entwistle, *Teaching for understanding at university: deep approaches and distinctive ways of thinking*, Basingstoke, Hampshire, Palgrave Macmillan, 2009.

[8] R. Ferguson *The state of learning analytics in 2012: a review and future challenges*, Technical Report KMI-12-01, vol. 4, March, 2012., 18.

[9] D. Gašević, S. Dawson, G. Siemens, Let's not forget: Learning analytics are about learning, *TechTrends* **59**(1) (2015), 64–71.

[10] C. L. Hwang, K. Yoon, *Multiple Attribute Decision Making and Applications*, New York, Springer Verlag, 1981.

[11] L. Johnson, S. Adams, V. Estrada, A. Freeman, *NMC Horizon Report: 2015 Higher Education Edition*, Austin, Texas, 2015.

[12] A. Jonnson, G. Svigby, *The use of scoring rubrics: Reliability, validity and educational consequences*, Educational Research Review, 2007.

[13] D. J. Nicol, D. Macfarlane-Dick, Formative assessment and selfregulated learning: a model and seven principles of good feedback practice, *Studies in Higher Education* **31**(2) (2006), 199–218.

[14] Z. Papamitsiou, A. A. Economides, Learning Analytics and Educational Data Mining in Practice, *A Systematic Literature Review of Empirical Evidence, Educational Technology & Society* **17**(5) (2014), 49–64.

[15] C. Redecker, Ø. Johannessen, Changing Assessment - Towards a New Assessment Paradigm Using ICT, *European Journal of Education*, **48**(1) (2013), 79–96.

[16] P. Sadler, E. Good, The impact of self-and peer grading on student learning, *Educational Assessment* **11**(1) (2006), 37–41.

[17] K. D. Strang, Effectiveness of peer-assessment in a professionalism course using an online workshop, *Journal of Information Technology Education: Innovations in Practice* **14** (2015), 1–16.

Blaženka Divjak
e-mail: blazenka.divjak@foi.hr

Marcel Maretić
e-mail: marcel.maretic@foi.hr

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, HR 42000 Varaždin, Croatia