Learning in Teams: Peer Evaluation for Fair Assessment of Individual Contributions

Extended Abstract

Fedor Duzhin
Nanyang Technological University
Singapore, Singapore
fduzhin@ntu.edu.sg

ABSTRACT

We develop a game-theoretical model of a classroom scenario, where n students collaborate on a common task and the job of the course instructor is to grade the individual contribution of each student to teamwork. Our main result is a method of grading individual contributions based on the matrix of peer evaluations such that 1) the collective truth-telling is a strict Nash equilibrium and 2) the method of assessment is psychometrically reliable.

KEYWORDS

Peer Evaluation, Collaborative Learning, Cooperative Learning

ACM Reference Format:
Fedor Duzhin. 2023. Learning in Teams: Peer Evaluation for Fair Assessment of Individual Contributions: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Teamwork and report writing are taught at universities, but grading every individual student based on a team's report is a challenge. For instance, if all the team members get the same grade, then a free-rider problem may occur — see [11], [10], [4], or [1]. The most obvious solution to the free-rider problem is peer evaluation ([6]).

A system of peer evaluation is a procedure of calculating the "true" (at least, as it is perceived by team members) contribution of each of the team members to the common task based on mutual evaluations reported by team members. A system of peer evaluation may or may not have certain desired qualities. Among most important sought–after qualities of educational assessment are *validity* and *reliability* ([9], [13], [14]).

A valid assessment measures what it is supposed to measure. A reliable assessment yields the same results each time it is used in the same setting.

In this paper, we develop a mathematical model of peer evaluation for individual contribution to teamwork. We prove that the collective truth–telling in our model is a strict Nash equilibrium and argue that it means that the assessment system is valid. We also show that our assessment method is reliable.

2 RELATED WORKS

While we are constructing a mathematical model of peer evaluation in learning teams, there are a few superficially similar problems in game theory literature.

The first of them is peer grading or peer assessment — see, for example [18], [16], [8], and [17]. Despite an essentially the same name, the main scenario is completely different. In peer grading, a large group of students are required to submit their *individual* work (essays, assignments, reports etc.) to a common pool and then each student gets to evaluate a small number of peers' works according to criteria designed by the course instructor.

The second problem that is related to ours is peer nomination — see, for example, [2], [3], [12]. In peer nomination, the real-life motivation is a scenario where a number of researchers are competing for grants and they themselves select proposals that are worthy of funding. This set-up is somewhat similar to ours in that there assumed to be a ground truth — ranking, i.e., an order on the set of proposals. The key differences with our set-up are that in peer nomination the group is large, i.e., every agent only evaluates a fraction of other agents' proposals, and that the output is binary while the output in our scenario is numeric.

The third game–theoretical problem that somehow resembles ours is fair division — see, for example, [7] and [15]. In fair division, n agents, too, compete for some common resource. However, the similarity ends here. The main difference is that in fair division, each agent has their own opinion on the value of each resource and the objective is to distribute resources between the agents so that each agent's fraction of resource is at least $1/n$.

A rigorous mathematical theory of peer evaluation is outlined in [5]. However, their theory is very broad and does not provide a specific reliable method of grading (reliability is not even mentioned). Here, we are going to narrow the scope of the theory and, borrowing some ideas from [5], provide a practical method of grading.

3 MATHEMATICAL THEORY

3.1 Set-up

We assume that $n \geq 3$ students collaborate on a common task (such as a group project) and there exists the ground truth — individual contribution of each student to teamwork. If the true contribution of the i th student is t_i, then the ground truth is the vector

$$t = (t_1, t_2 \ldots, t_n) \in \Delta^{n-1} \subset \mathbb{R}^n,$$

where

$$\Delta^{n-1} = \{(t_1, \ldots, t_n) : t_1 \geq 0, \ldots, t_n \geq 0, \sum_{i=1}^{n} t_i = n\} \subset \mathbb{R}^n$$

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

is the $(n-1)$-simplex. Note that we require the mean individual contribution rather than the sum to equal 1.

The instructor observes the final product of teamwork (e.g., a report or a presentation) and evaluates the team with a score T. Intuitively, if the course instructor just wanted to give all students individual scores proportional to their effort, then the "fair" score given to student i should be $t_i \times T$.

The vector t is known to students but can't be observed by the instructor directly. What students report to the instructor is a matrix

$$A \in \mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0})$$

of evaluations of each student by each student, where $\mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0})$ is the set of $n \times n$ matrices with non-negative real entries.

Further, let the entry in row i, column j of matrix A, i.e., a_{ij}, be evaluation of student i by student j. If all students were truthful in their evaluations, then all columns of the matrix A would be proportional to t, i.e., A would have rank 1.

Definition 3.1. A *mechanism* (term adopted from [5]) is an algorithm of calculating the vector of individual grades, i.e., a function

$$f : \mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0}) \longrightarrow \Delta^{n-1},$$
$$A = (a_{ij})_{1\leq i\leq n, 1\leq j\leq n} \longmapsto f(A) = g = (g_1, g_2 \ldots, g_n).$$

Note that the output of the mechanism, i.e., the vector g of individual grades may or may not be equal to the vector t of true contributions.

3.2 Valid and reliable assessment

Definition 3.2. A mechanism is *incentive–compatible* if collective truth-telling is a strict Nash equilibrium, i.e., lying decreases one's own score given that others tell the truth.

A mechanism is *reliable* if, assuming that all students report the truth, then g_i is an increasing function of t_i. In particular, g_i does not depend on t_j for $j \neq i$.

Note that if collective truth-telling were not a strict Nash equilibrium, then the assessment method would measure the ability of manipulation with scores rather than honest teamwork, i.e., assessment would not be valid. To understand reliability, think of students A and B from different teams whose teams got the same score and whose true individual contribution is the same. For assessment to be reliable, students A and B should get the same final grade.

3.3 Relative contributions

We assume that at most one entry of the ground truth vector t is 0, i.e., that $t_i + t_j > 0$ and $a_{ik} + a_{jk} > 0$ whenever $i \neq j$. Also, we assume that $n \geq 5$. Note that our mechanism is reliable even for $n = 3$ and $n = 4$, but we do not know even if incentive–compatible mechanisms exist for $n = 3$ or $n = 4$.

The key ingredient of our mechanism is the *relative contribution*

$$r_{ij}^k = \frac{a_{ik}}{a_{ik} + a_{jk}}, \tag{1}$$

of student i to student j according to student k. Thus, for every i and every $j \neq i$, we have the vector

$$r_{ij} = \left(r_{ij}^1, \cdots, \widehat{r_{ij}^i}, \cdots, \widehat{r_{ij}^j}, \cdots, r_{ij}^n, \right) \tag{2}$$

of relative contributions of students i and j according to their teammates. The vector r_{ij} has $n-2$ entries. Note that $r_{ij} + r_{ji} = (1, \ldots, 1)$.

Now let

$$b_{ij} = \begin{cases} 1, & i = j, \\ \frac{\text{median}(r_{ij})}{\text{median}(r_{ji})}, & i \neq j. \end{cases} \tag{3}$$

Note that, according to our assumption, r_{ij} and r_{ji} cannot be both equal to 0, $b_{ij} \in \mathbb{R}_{\geq 0} \cup \{\infty\}$. We will call $B = (b_{ij})_{1\leq i,j\leq n}$ the *auxiliary matrix* for the raw peer evaluation matrix A.

3.4 Perceived contributions

Given a matrix of peer evaluations A, we first construct the auxiliary matrix B. If all students were truthful in their evaluations, then $b_{ij} = t_i/t_j$ would hold, i.e., columns of B would be proportional to the ground truth t with the coefficient of proportionality chosen so that $b_{ii} = 1$ for all i.

Now, in order to extract the vector s of *perceived contributions* from B, we divide each column of B by the mean of its entries, then take the vector of row medians, and then divide the result by its mean.

The vector s has the following important property. Let $n \geq 5$ and suppose that $n-1$ out of n students report evaluations that are perfectly consistent, i.e., $n-1$ out of n columns of the matrix A are proportional to each other. Then s is independent of the remaining column of A, i.e., evaluations reported by the last student don't affect the vector s of perceived contributions.

3.5 Relative error of reported evaluations

Consider a peer evaluation matrix A and the vector of perceived contributions s calculated as described in section 3.4. Consider normalized columns of A, i.e.,

$$v_{ij} = \frac{n a_{ij}}{\sum_{i=1}^n a_{ij}}$$

Then

$$v_j = (v_{1j}, v_{2j}, \ldots, v_{nj}) \in \Delta^{(n-1)}$$

is j th version of truth and $\left|v_{ij}/s_i - 1\right|$ is the relative error of evaluation of student i by student j. Let

$$E_j = \frac{1}{n} \sum_{i=1}^n \left| \frac{v_{ij} - s_i}{s_i} \right| \tag{4}$$

be the average relative error of student j's version of truth. Notice that $E_j = 0$ if and only if $v_j = s$, i.e., evaluations reported by j are perfectly consistent with perceived contributions s. It may happen that $s_i = 0$ for some i (we assume that at most one entry of s may be 0). In that case, our convention is that $0/0 = 0$ and $1/0 = n$.

3.6 Our mechanism

Theorem 3.3. *The mechanism defined by*

$$g_j = (1-\varepsilon)s_j + \varepsilon \max(1 - E_j, 0), \tag{5}$$

where $\varepsilon > 0$, is incentive-compatible and reliable. Here, s_j is the perceived contribution of student j calculated as in section 3.4 and E_j is the relative error of evaluations reported by student j calculated according to (4).

Here, $\varepsilon > 0$ is a small number. The author sets $\varepsilon = 0.05$ in his classroom, but the exact value of ε is not important.

REFERENCES

[1] Praveen Aggarwal and Connie L O'Brien. Social loafing on group projects: Structural antecedents and effect on student satisfaction. *Journal of Marketing Education*, 30(3):255–264, 2008.

[2] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110, 2011.

[3] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey Rosenschein, and Toby Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[4] Charles M Brooks and Janice L Ammons. Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments. *Journal of Education for Business*, 78(5):268–272, 2003.

[5] Arthur Carvalho and Kate Larson. Sharing rewards among strangers based on peer evaluations. *Decision Analysis*, 9(3):253–273, 2012.

[6] Robert Conway, David Kember, Atara Sivan, and May Wu. Peer assessment of an individual's contribution to a group project. *Assessment & Evaluation in Higher Education*, 18(1):45–56, 1993.

[7] Geoffroy De Clippel, Herve Moulin, and Nicolaus Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139(1):176–191, 2008.

[8] Fedor Duzhin and Amrita Sridhar Narayanan. Peer grading reduces instructor's workload without jeopardizing student learning in an undergraduate programming class. *New Directions in the Teaching of Physical Sciences*, (15), 2020.

[9] Alison Laver Fawcett. *Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application.* John Wiley & Sons, 2013.

[10] William B Joyce. On the free-rider problem in cooperative learning. *Journal of Education for Business*, 74(5):271–274, 1999.

[11] Jane H Leuthold. A free rider experiment for the large class. *The Journal of Economic Education*, 24(4):353–363, 1993.

[12] Omer Lev, Nicholas Mattei, Paolo Turrini, and Stanislav Zhydkov. Peernomination: A novel peer selection algorithm to handle strategic and noisy assessments. *Artificial Intelligence*, page 103843, 2022.

[13] John R McClure, Brian Sonak, and Hoi K Suen. Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 36(4):475–492, 1999.

[14] Barbara M Moskal and Jon A Leydens. Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1):10, 2000.

[15] David C Parkes, Ariel D Procaccia, and Nisarg Shah. Beyond dominant resource fairness: Extensions, limitations, and indivisibilities. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):1–22, 2015.

[16] Victor Shnayder and David Parkes. Practical peer prediction for peer assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pages 199–208, 2016.

[17] James R Wright, Chris Thornton, and Kevin Leyton-Brown. Mechanical ta: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, pages 96–101, 2015.

[18] Hedayat Zarkoob, Greg d'Eon, Lena Podina, and Kevin Leyton-Brown. Better peer grading through bayesian inference. *arXiv preprint arXiv:2209.01242*, 2022.