Collusion in Educational Peer Assessment: How Much do We Need to Worry about It?

Yang Song, Zhewei Hu, Edward F. Gehringer
Department of Computer Science
NC State University
Raleigh, U.S.
{ysong8, zhu6, efg}@ncsu.edu

***Abstract*—Several decades of research have shown peer assessment to be an effective pedagogical approach. Researchers have shown that peer assessment has the potential to provide students more copious, timely and helpful feedback, and also helping reviewers to learn as well. In recent decades, peer assessments in educational settings have increasingly been facilitated by online tools. Some MOOC platforms also rely on aggregated peer-assessment scores to assign grades for each artifact. However, in peer assessment, students can potentially game this process, and thereby harm the validity and reliability of the aggregated scores. Most of instructors assume that the majority of the students' peer assessments are honest since most of the peer assessment is done in double-blind fashion. This assumption only holds in the absence of organized collusion – when no more than a small number of students game the peer assessment and give each other very high scores. This paper identifies two types of collusion that we have observed. They are small-circle collusion and pervasive collusion. Small-circle collusion refers to the behaviors of students who form small circles and give higher peer review grades to each other. Pervasive collusion refers to students assigning top grades to all the submissions they review. We also present our algorithms for detecting these two types of colluders. Our experiments are based on a peer-assessment dataset shared by multiple peer-assessment systems. By removing these colluders' peer assessments, we are able to estimate how much inflation is brought by colluders in educational peer assessment.**

Keywords—collusion, peer assessment, peer rating, peer ranking

I. INTRODUCTION

Fifty years of research has shown great potential for peer assessment as a pedagogical approach [1]. Compared with the assessments given by teaching staff, students receive more copious and timely assessments — usually before they move on to next task — from their peers. In addition, peer assessment helps the reviewers to reflect on their own work and understand their own strengths and weaknesses. Some research even suggests that this metacognitive process makes peer assessment more beneficial to assessors than assessees [2].

Though much peer assessment is still done face-to-face in classrooms, in recent decades, educational peer assessments have increasingly been facilitated by online systems. There are dozens of systems designed to facilitate peer assessment [3–5], some MOOC platforms have also integrated peer assessment [6, 7]. Online peer assessment is better suited for large courses and students from different locations. It also facilitates data collection on the peer-assessment process.

In the peer-assessment process, the key to success is to ensure fair and accurate assessment. Some researchers have attempted to calculate reviewers' reputations based on their credibility, then use reviewers' reputations to weight their reviews [8–10]. Another approach which has also proven to be successful is calibration [11]. The calibration can be done on a few artifacts (either on sample artifacts or real students' artifacts) with representative mistakes that students may make. Researchers have found that common understanding on rating standards can be better established after calibration [8].

Those attempts on improving peer-assessment quality work only under the assumption that students are honest and treat the peer assessment seriously. Most researchers simply assume that this is true without paying enough attention the possibility of collusion. This is partially because that there is no formal definition of collusion in peer assessment [12]. Some researchers treat some types of collusion (e.g. pervasive collusion discussed in the next section) as parts of reliability measures, such as "spread" [13], but we argue that there is a difference between unreliability and collusion: collusion refers to the intended cooperation in peer assessment, either to help some students unfairly gain higher aggregated peer-assessment scores, or to avoid the need to spend serious amounts of time on peer assessments.

Another reason for the dearth of research on students' collusion in peer assessment is that there is no tool designed to help users of peer-assessment systems, especially teaching staff, to identify potential colluders. As a result, they may have to use "spreadsheet and visual checks" [14] to detect collusion patterns, which is time-consuming and subjective.

In this paper, we present our work on defining and investigating students' collusion in peer assessment. We found that the collusions inflated the average peer-assessment scores by 1–2%, in addition, instructors should be vigilant against a particular type of collusion. We also developed a tool which can help instructors identify potential colluders in peer-reviewed assignments. This tool has been made available as a

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

web service which can be plugged to different online peer assessment tools.

The rest of this paper is organized as follows: Section II introduces the different types of collusion we investigated; Section III introduces our research methodology including our research questions, the algorithms, and the dataset; Section IV presents our experiment results; and Section V provides further discussion on student collusion in peer assessment.

II. Different Types of Collusion

Collusion among students is a threat to peer assessment validity and fairness that cannot be mitigated by training or calibration. Unfortunately, this issue has not drawn enough attention from researchers in this area. A few researchers have claimed that they have observed this phenomenon, but no formal definition has yet been given on students' collusion in peer assessment. To prevent collusion, some instructors modify the design of their peer assessment, e.g., by preserving anonymity, having reviewers review different artifacts in the second rounds [15]. In this section, we define and discuss two types of collusion we investigated in this research. Related research and observations are also reviewed and discussed.

A. Pervasive Collusion (PC)

Topping suggested that some reviewers may "submit average scores, leading to lack of differentiation" [16]. However, more often than not, we found students in online peer assessment systems tend to submit high peer-review scores (sometimes even top scores available) [17]. We define this kind of behavior as *pervasive collusion*. This behavior makes it harder to identify the top artifacts (an artifact of medium quality may still get as good a score as a top artifact because of grade inflation) [18]. From the student's perspective, the main reason for this type of collusion is to reduce the workload and to avoid receiving arguments from the authors.

Even though students may not consciously conspire, teaching staffs have observed that "a nucleus of students starts to assign top grades to all the submissions they review, other initially honest students see what is happening, and join the colluders" [19]. Though quantitative analysis was not provided, this observation itself provides an important reason to investigate this type of collusion. In the worst case, if a higher percentage of students join this activity and give/receive (free) high peer-review scores, the honest reviewers could be penalized as outliers, since they disagree with their "colluding peers."

B. Small-circle Collusion (SCC)

Sometimes as we monitored students' behavior in online peer assessment systems, we found that some students gained an unfair advantage by giving higher peer-review scores to students they know. The reason is usually that students value personal qualities such as friendliness and trust more than mandates on academic conduct [20]. We define this kind of behavior as *small-circle collusion*. This is, to some extent, deliberate cheating. Another alternative is to consider small-circle collusion as "friendship bias" [21]—friends, if somehow able to figure out which is their friends' work, can give their friends higher scores even in blind review.

Though the average peer-review scores may not be used as part of students' final scores for their artifacts, they may still influence the teaching staff when deciding the final scores. From students' perspective, the main reason for small-circle collusion is to help friends. Similarly, "to help friends" is also known as one of the principal reason for students dishonest behaviors including plagiarism, cheating, and falsification [22].

III. Methodology

A. Research Questions

In this research, we mainly focused on two research questions related to collusion in peer assessment in education.

RQ 1: How much inflation can be created by collusion? Since some educators may aggregate peer-review scores into final scores for each artifact, the influence of peer assessments by colluders should be removed or at least mitigated. Unfortunately, the functionality of identifying and removing colluders' peer assessments is usually unavailable in online peer-assessment systems. In this case, if the teaching staff simply use the average or median of the peer assessment scores as the final scores, there could be a potential inflation of students' scores.

Even when the teaching staff does grading completely independently of the peer-assessment process (teaching staff grade all the artifact by themselves without even referring to the peer assessments), there is still another inflationary effect from the authors' perceptions — high peer-assessment scores may give them the illusion that their artifacts are already of good quality, and therefore they could be less willing to make modifications. They might even ignore the honest (yet helpful) advice.

RQ 2: If, to some extent, there is collusion and inflation in peer assessment, which approach, ranking or rating, can help lower the effect of collusion?

Peer ranking and peer rating have been long recognized as two main approaches for peer assessment [23]. Rating-based peer assessment systems usually ask assessors (mostly students, sometimes teaching staff as well [8, 22]) to rate artifacts. The assessors usually review one artifact at a time; therefore, detailed review rubrics can easily be applied to peer rating because the reviewed artifact holds the assessor's attention well.

Ranking-based peer assessment systems usually ask the reviewers to review a fixed number of artifacts. Instead of giving numerical scores to each of the artifacts, assessors need to rank them in order, from strong to weak. Researchers argue that ranking is a more reliable approach to peer assessment because the quantitative feedback (ranks in this case) is given by the comparisons between one artifact and others [5].

Some comparisons between ranking-based and rating-based systems have been done, e.g., on review reliability [25]. However, there has been no comparison of collusion in peer-rating and peer-ranking settings. Since collusion among students is a major concern of some educators [18], it is

important to know how to design the peer assessment to minimize the potential collusion.

B. Dataset

This study was conducted with the PeerLogic[1] dataset. The data is derived from multiple educational peer-assessment systems used by college students in a variety of disciplines including computer science, mechanical engineering, electrical engineering, statistics, art and English [26]. The shared data have been transformed and combined to facilitate further research [27]. There are 45,991 assessments (assessments give either a rank or comprehensive rating, comprising responses to all criteria in a single rubric) from 309 assignments used in this experiment. Table I provides more details about our dataset. Due to different class settings, the courses which used rating-based peer assessment required students to do fewer assessments (2 on average) than the courses which used ranking-based assessment (4 on average).

TABLE I. DETAILS ABOUT THE DATASET USED IN THE EXPERIMENT

	Rating-based	Ranking-based
Num. of tasks	168	141
Num. of participants	1519	1366
Num. of assessments	21343	24648
Avg. participants per assignment	50.6	46.9

All the peer assessments in our dataset were done in double-blind fashion.

C. Detection of Potential Colluders

The peer assessment in each task can be considered as a directed graph, in which each student (assessor or assessee) can be considered as a vertex and each review can be modeled as an edge from the assessor to the assesse. The weight of each edge can be considered as the numerical peer-review score (either a ranking or a rating score). To give a precise definition of the two types of collusion discussed in Section II, let $G = (V, E)$ be a directed graph for a peer assessment task, V be all the vertices in this graph (namely, all the students who have participated in the peer assessment); and E be all the edges (namely, all the peer assessments done by students). A *path* in G is a sequence of vertices $p_{vu} = (v = \mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_k = u)$ such that $(\mathrm{v}_i, \mathrm{v}_{i+1}) \in E$ for $1 \le i < k$. A *circuit* is a path where v and u are identical, $C_{vv} = (v = \mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_k, \mathrm{v}_1 = v)$.

We say that an edge is in a circuit, $e \in C_{vv}$, if $e \in \{(\mathrm{v}_k, \mathrm{v}_1) \cup (\mathrm{v}_i, \mathrm{v}_{i+1}), \text{for } 1 \le i < k\}$; a vertex is in a circuit if $u \in \{\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_k\}$.

We assume that there is no single-vertex circuit (one reviews him/herself, a.k.a self-review) in this directed graph. We also assume that there can be only 0 or 1 edge between any pair of vertices (one assessor can only give one aggregated numerical score to a particular assessee).

Let E_{in}^{v} be all the edges directed to vertex v, E_{out}^{v} be all the edges directed from vertex v, W_e be the weight of an edge e (in this case a weight is a peer-assessment score), α be the threshold of the weight to be considered as high enough to indicate potential collusion, and β be the threshold size of small-circle collusion circuits.

A potential small-circle collusion can be formally defined as:

- C_{vv} is a circuit in G,
- For each $e \in C_{vv}$, $W_e > \alpha$,
- $k \le \beta$,
- $\exists u \in C_{vv}$ that makes $\overline{W}^{u}_{within\_C} / \overline{W}^{u}_{\text{outside}\_C} \ge 1 + \varepsilon_1$, in which

$$\overline{W}^{u}_{within\_C} = \frac{\sum_{e \in \mathrm{E}_{in}^{u}, e \in C} W_e}{|e|_{\in \mathrm{E}_{in}^{u}, e \in C}}$$

$$\overline{W}^{u}_{\text{outside}\_C} = \frac{\sum_{e \in \mathrm{E}_{in}^{u}, e \notin C} W_e}{|e|_{\in \mathrm{E}_{in}^{u}, e \notin C}}$$

The first part of the formal definition of potential small-circle collusion is simply that the suspects need to be a circuit in the directed graph of the peer-assessment task. The second part indicates that the peer assessments in the potential small-circle collusions should have higher numerical scores than the threshold α. This will filter out the circuits of students who give each other medium or low scores. The third part of this definition adds a constraint on the number of vertices of the circuits – the circuits with more than β vertices are less likely to become a "small"-circle collusion. The last part of this definition emphasizes that there should be peer-assessment disagreement between the peer-assessment scores within and outside the circuits.

Fig. 1 shows a visualization of a potential small-circle collusion in a peer-assessment task. In this visualization, each vertex is a student in the task. Each directed edge represents a peer assessment from the assessor to the assessee. In this case, all the peer-assessment scores are higher than the threshold α (edges are in green) and the number of vertices (3) is smaller than β. However, we are still not confident that those three students are possibly colluders because they could be three of the strongest students in the class who happened to review each other.

Fig. 2 shows the peer assessment received by one of the three students. Each red edge represents a peer assessment with a score lower than α. From the visualization, we found that, though the student "a******1" received two favorable peer

[1] https://www.peerlogic.org/

assessments from the other two students who are in the potential small collusion circuit, (s)he overall received fewer favorable peer assessments than unfavorable ones. This indicates there is a disagreement between the reviews inside and outside the circuit. If the difference is greater than the threshold ε_1, our algorithm will consider this circuit as a case of potential small-circle collusion.

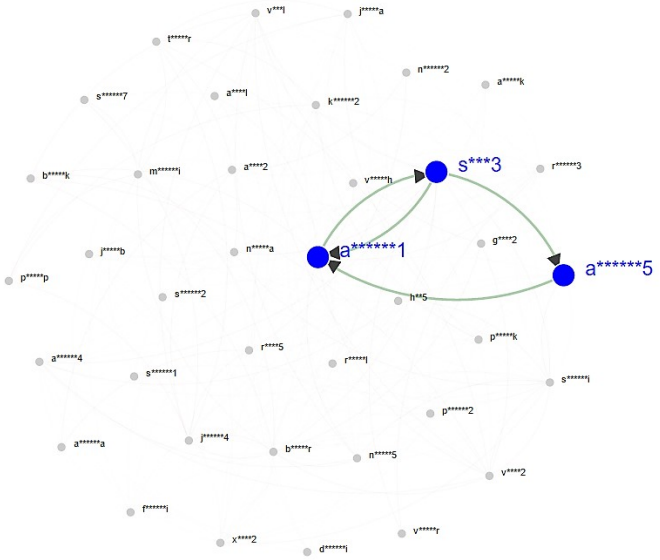

Fig. 1. An example of a small-circle collusion

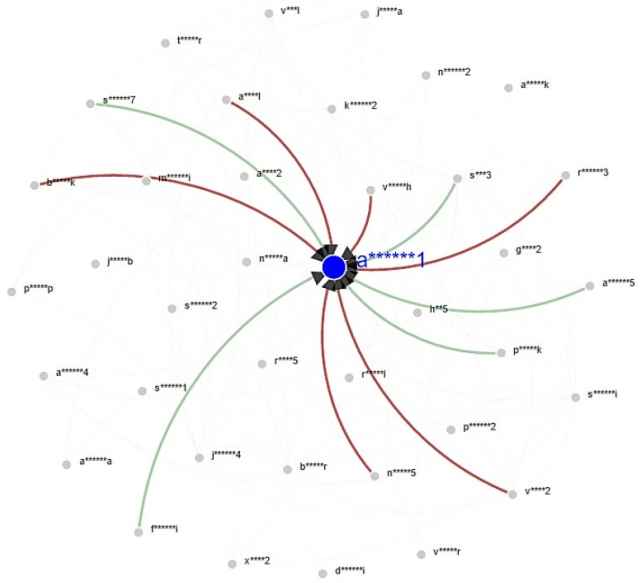

Fig. 2. Disagreement between peer-assessment scores within and outside the small-circle detected

A potential pervasive-colluder vertex $u \in \mathrm{V}$ can be formally defined as:

- $\mathrm{E}^{u}_{out} \geq Avg(\mathrm{E}^{v}_{out})_{v \in \mathrm{V}}$,
- $|\mathrm{e}|_{e \in \mathrm{E}^{u}_{out}, W_e > \alpha} / |\mathrm{E}^{u}_{out}| \geq 1 - \varepsilon_2$.

The first part of the definition makes sure that only the vertices with E^{v}_{out} greater than the average will be considered. This is to exclude reviewers who reviewed only a small number of artifacts that happened to be the stronger ones. The second part of the definition examines the percentage of peer-review scores lower than α given by this reviewer: if it is lower than ε_2 (in other words, more than 100%- ε_2 of the reviews given have peer-review scores higher than α), this reviewer will be considered to be a potential pervasive colluder.

Fig. 3 is another visualization of the same peer-assessment task. In this case, reviewer "f*******i" assessed 11 artifacts, which is higher than average. In addition, all the peer-assessment scores assigned were higher than α (green edges). The combination of the high number of reviews done and high percentage of scores higher than α made him/her a suspect for being a pervasive colluder.

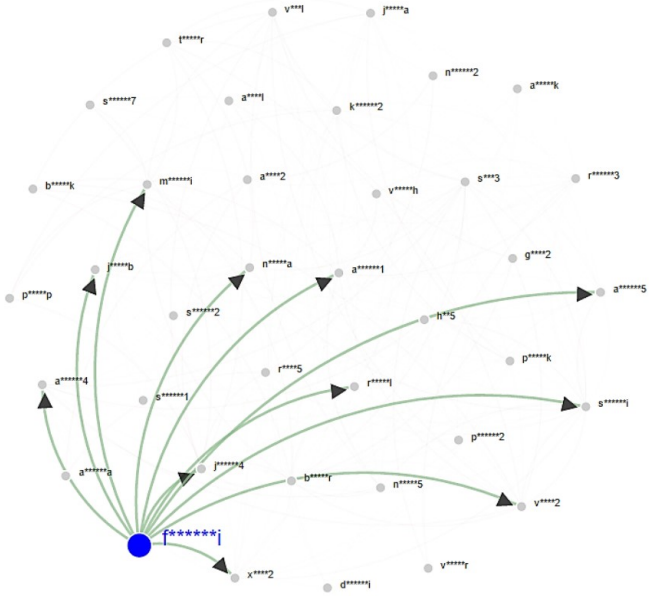

Fig. 3. An example of a pervasive colluder

Please note that our algorithm only detects the "suspect" colluders in our dataset. We do not have any further evidence since the dataset used in this research is anonymized and from different peer-assessment systems.

Our collusion detection for both types, and the visualization will be made available as a web service. This web service accepts data in PRML (Peer-Review Markup Language) format [27] and can detect and visualize potential collusions in any peer assessment task.

IV. Experimental Results

In our experiment, we set α to be the top 80% quantile of the peer-assessment scores for each task, namely, a peer-assessment is considered to be favorable if it is higher than

80% of the peer-assessments in a task. β was set to 4 and both ε_1 and ε_2 were set to be 5%.

A. Pervasive Collusion

We created an algorithm to detect pervasive colluders and calculate the percentage of reviews by pervasive colluders in each task. Fig. 4 shows the distribution of those percentages (*x*-axis shows the percentages of reviews done by pervasive colluders, the interval between each bin boundary is 5%; *y*-axis shows the number of tasks). We found that in more than 80% of tasks, less than 10% of reviews were submitted by pervasive colluders. Only very few tasks had 50% or more reviews done by pervasive colluders (will be discussed in the next section).

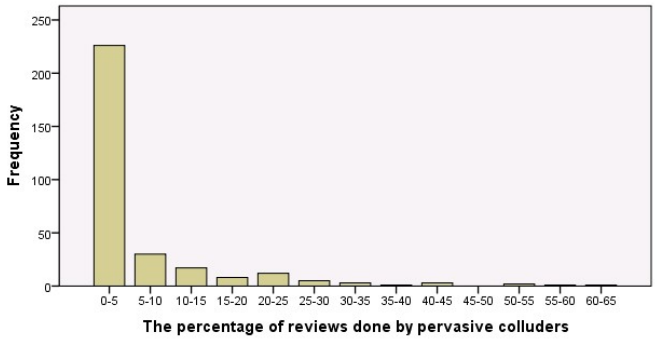

Fig. 4. Distribution of pervasive colluders' reviews percentages

B. Small-circle Collusion

We also computed the percentage of reviews done within small collusion circles for each task. Those small-circle collusions were detected with Hawick and James' algorithm [28] with necessary modifications, e.g. parameters such as α and β. Fig. 5 illustrates the distribution of those percentages (the *x*-axis shows the percentages of reviews done by small-circle colluders, the interval between each bin boundary is 5%; and the *y*-axis shows the number of tasks). We found that more than 70% of tasks contain 5% favorable reviews or less from small-circle collusions. Very few tasks had more than 15% of reviews from small collusion circles.

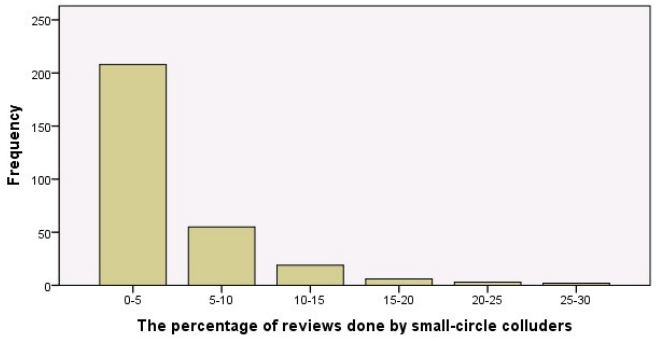

Fig. 5. Distribution of small-circle colluders' reviews percentages

C. Score Inflation Brough by Collusion

We removed all the reviews done by pervasive colluders and small-circle colluders and calculated the average peer-assessment scores for each task. We further compared those average scores with the original average scores (without removing colluders' reviews).

By comparing two average review scores for reach task, we can estimate the score inflation brought about by collusion. Fig. 6 is the distribution of score inflation for all the tasks in our dataset (the *x*-axis shows the score inflation based on 100 points, the interval between each bin boundary is one point, and the *y*-axis shows the number of tasks). This distribution indicated that for more than half of the review tasks, the inflation caused by colluders was less than 1 point. However, in the worst case, both type of collusion, together could bring almost 5 to 10 points inflation.

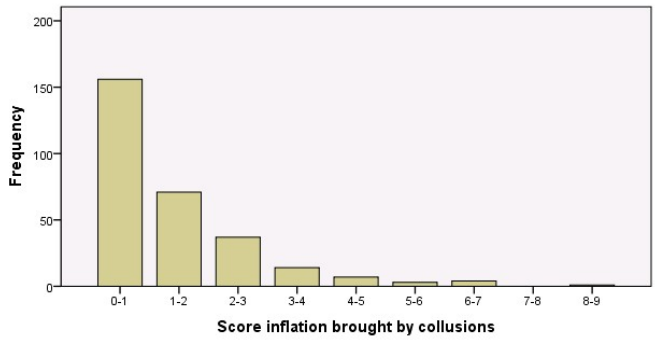

Fig. 6. Distribution of score inflations brought by collusions

D. Rating vs. Ranking

We also computed the collusion results and score inflation of data generated by ranking-based reviews and rating-based reviews separately (see Table II). We found that the average inflation in rating-based peer-assessment scores is slightly higher than the average inflation of ranking-based peer-assessment scores. There is not a big difference between the percentages of small-circle colluders' review in ranking (3.39%) and rating (3.74%). However, we did not find any pervasive colluders in ranking-based assessment, but we found 9.98% of the assessments done in the rating-based scenario were done by pervasive colluders. This indicates that in both ranking-based and rating-based assessments, students may return favors to each other. Small-circle collusion is the major collusion type for ranking-based peer assessment. However, in rating-based peer assessment, pervasive collusion is an even bigger challenge for educators: we found that on average almost 10% of the reviews in the rating-based system were done by pervasive colluders.

The *t*-ratio of score inflation on rating-based to ranking-based peer assessments is 2.39 with *p*-value of 0.0175 (< 0.05). This means that collusion is likely to cause more inflation in rating-based peer assessment. We hypothesize that the main reason is that the ranking-based peer assessment better facilitates comparisons. It is harder for a ranking-based reviewer to collude since all the artifacts are ranked against each other. Rating-based peer assessment relies less on

comparison. As a result, it is easier for colluders to give their friends high/top scores (small-circle collusion), or give each assigned artifact the same score (pervasive collusion).

TABLE II. COMPARISON BETWEEN RATING AND RANKING ON POTENTIAL COLLUSIONS

Assgt. type	Avg. inflation	Std. dev. inflation	Avg. SCC %	Avg. PC %
Rating	1.42	1.51	3.74%	9.98%
Ranking	1.03	1.26	3.39%	0.00%

V. DISCUSSION

Our experimental results indicate that collusion happens in peer-assessment tasks, and in some cases, the inflation cannot be ignored. In this section, we discuss additional questions related to our results.

A. Relationship between Pervasive Collusion and Small-Circle Collusion

We argue that pervasive collusion and small-circle collusion are not independent, but rather, if there is a large number of pervasive colluders in a peer-assessment task, it is also likely that our algorithm will find more cases of small-circle collusion. In this scenario, more pervasive colluders increased the numbers of blindly-assigned high scores (or even top scores). If two pervasive colluders happened to review each other, a small collusion circle is likely to be detected. In other words, a higher percentage of pervasive colluders is likely to increase the number of small collusion circles, though those students may instead be pervasive colluders.

An even worse result from a high pervasive collusion rate is that it makes the peer assessment more lenient, even for reviews from the students who are not detected as colluders. In our data set, the average peer-assessment score for non-colluding students on the 10 tasks with highest pervasive collusion rate is 95, yet the average peer-assessment score for non-colluding students is 85 for all tasks. This indicates that students, when receiving more favorable peer assessments, can also be more lenient when reviewing others. We hypothesize that this is caused by the "peer pressure" from the colluders in the peer-assessment task.

On the other hand, small-circle collusion does not induce "peer pressure" to encourage students to be lenient. We calculated the average peer-assessment scores from non-colluding students on 10 tasks with the highest percentage of small-circle collusions. Though there are many small collusion circles, the average peer assessment scores from honest students in those tasks was 86.4, which is almost the same as the average peer-assessment scores from all the honest students in the tasks. Hence, we suggest that teaching staff pay more attention to pervasive collusion in peer assessment because they brought more "hidden" inflation to peer-assessment scores (an increase in scores by honest peer reviewers rather than colluders), which is harder to estimate.

B. How does collusion influence reputation systems?

Reputation systems estimate the credibility of each reviewer in peer-assessment activities. Reputation systems may take various factors into account, but in this section, we only discuss reputation systems based on levels of agreement in peer-assessment scores [10] in this section. This type of reputation algorithm calculates reviewer reputations and aggregate scores for artifacts recursively untill they reach convergence [8, 9].

For small-circle collusion in which colluders give each other higher scores to each other within the circles, if the reviewers outside the circles rate/rank an artifact much lower, the reviewers in the circle are likely to get lower reputation scores. This means that the small-circle colluders tend to be punished by reputation system since their reviews do not agree with the majority of other reviewers on the same artifacts [10]. Therefore, teaching staff does not need to worry much about small-circle collusion if there is a reputation system in the peer-assessment system.

The pervasive colluders, if they are not the majority, may also be punished by reputation systems for the same reason. However, in the worst case (when there are more than 50% of the reviews are from pervasive colluders, as in 4 tasks from our dataset), the reputation system will fail to detect pervasive colluders since they are not the minority anymore. Instead, all the pervasive colluders will receive higher reputations than the honest reviewers, which could potentially hurt the honest reviewers and lead them to become pervasive colluders as well. This suggests that reputation systems can potentially punish the pervasive colluders by giving them low reputation scores only when they are not the majority.

C. Will collusions increase through each semester?

Although on average, collusion does not cause major inflation in peer-assessment scores, teaching staff may still worry that more students will learn to collude as the semester progresses [19]. From our dataset, we pick the courses which have at least two peer-assessment tasks and visualized their inflations against the day of the semester that the peer assessment is due (Fig. 7).

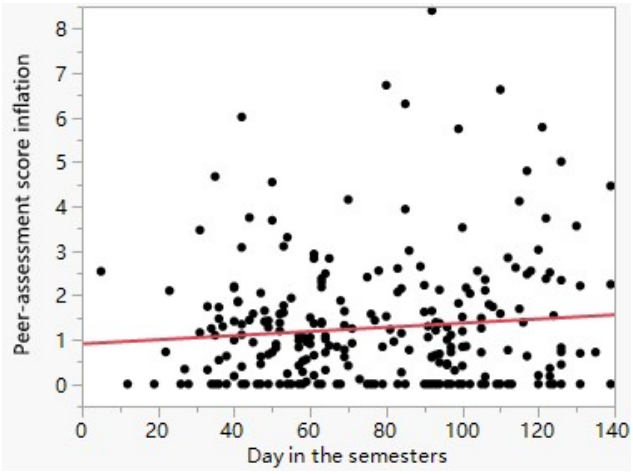

Fig. 7. Peer-assessment inflations based on their deadline of day of semesters

In Fig. 7, the *x*-axis shows on which day of the semester that the peer-assessment tasks were due, and the *y*-axis shows the inflation in the tasks. The slope of the fit line is 0.005. Though it is positive, the slope is relatively small considering the total score in our dataset is 100. In addition, the *R*-square value is 0.01 (the change of the *x* value only explains 1% of the changes in the *y* value), which is again, very small.

This finding, in general, disagrees with de Alfaro's observation [19]. Instead, we found that students' collusion changes little within a semester. However, we did find a few cases where the inflation kept increasing in a course through a semester, e.g. in the first peer-assessment task in a course (due on 28th day of the semester), there was 0 inflation but on the third task (due on 99th day of the semester) the inflation was 5.7 points.

VI. Conclusion

Educators apply peer assessment to provide authors timely and helpful reviews. Collusion, however, can undermine the value of peer assessment. In this paper, we identified two types of collusion, pervasive collusion and small-circle collusion, and gave formal definitions for both. We designed algorithms to detect potential colluders of both types with our PeerLogic dataset, and investigated the impact of collusion on peer-assessment scores. We found that on average the collusions inflated the average peer-assessment scores by 1–2%.

We also compared the collusion in rating-based and ranking-based systems. Generally speaking, rating-based peer assessment tends to have more collusion than ranking-based peer assessment. We hypothesize that the main reason is that the ranking-based peer assessment better facilitates comparison and thereby makes students less likely to collude, namely, rank a weak artifact higher than a stronger one. In rating-based peer assessment, since students have fewer chances to do comparisons, they are slightly more likely to collude. Our experimental results on the dataset from different peer-assessment systems show that in most cases, only a small portion of reviews are done by colluders. However, in a few extreme cases, more than 50% of reviews are done by colluders.

We also suggest that teaching staff pay more attention to pervasive collusion. First, pervasive colluders tend to influence other honest reviewers and make them more lenient than average. This makes it harder to differentiate stronger artifacts from weaker ones, and tends make authors feel overconfident of their work. This harms the peer-assessment process since authors do not get many honest/helpful reviews. Second, the pervasive colluders, if they become the majority, will deceive a reputation system and make the honest reviewers outliers. In this case, the opinions of the honest reviewers may be overwhelmed by the high scores given by colluders.

A. Future Work

We are working on collecting data from MOOC platforms. It will be interesting to investigate the peer-assessment data by MOOC students and learn the difference on the collusion patterns between MOOC students and university students.

Acknowledgment

This research is part of the PeerLogic project, which is funded by the National Science Foundation under grants 1432347.

References

[1] K. Topping and S. Ehly, *Peer-assisted Learning*. Routledge, 1998.

[2] K. Lundstrom and W. Baker, "To give is better than to receive: The benefits of peer review to the reviewer's own writing," *J. Second Lang. Writ.*, vol. 18, no. 1, pp. 30–43, Mar. 2009.

[3] E. Gehringer, "Expertiza: information management for collaborative learning," *Monit. Assess. Online Collab. Environ. Emergent Comput. Technol. E-Learn. Support*, pp. 143–159, 2009.

[4] "Mobius SLIP: UNCG develops a new online learning tool," *Research & Economic Development*, 06-Nov-2013. [Online]. Available: http://research.uncg.edu/spotlight/mobius-slip-uncg-develops-a-new-online-learning-tool/. [Accessed: 08-Jul-2016].

[5] D. Tinapple, L. Olson, and J. Sadauskas, "CritViz: Web-based software supporting peer critique in large creative classrooms," *Bull. Tech. Comm. Learn. Technol.*, vol. 15, no. 1, 2013.

[6] S. P. Balfour, "Assessing writing in MOOCS: Automated essay scoring and Calibrated Peer Review," *Res. Pract. Assess.*, vol. 8, no. 1, pp. 40–48, 2013.

[7] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned Models of Peer Assessment in MOOCs," *ArXiv13072579 Cs Stat*, Jul. 2013.

[8] J. Hamer, K. T. K. Ma, and H. H. F. Kwong, "A Method of Automatic Grade Calibration in Peer Assessment," in *Proceedings of the 7th Australasian Conference on Computing Education - Volume 42*, Darlinghurst, Australia, Australia, 2005, pp. 67–72.

[9] H. W. Lauw, E. Lim, and K. Wang, "Summarizing review scores of 'unequal' reviewers," in *In Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

[10] Y. Song, Z. Hu, and E. F. Gehringer, "Pluggable reputation systems for peer review: A web-service approach," in *IEEE Frontiers in Education Conference (FIE), 2015. 32614 2015*, 2015, pp. 1–5.

[11] A. . Russell, "Calibrated Peer Review™- A Writing and Critical-Thinking Instructional Tool," *Am. Biol. Teach.*, vol. 63, no. 7, pp. 474–480, Sep. 2001.

[12] R. Barrett* and A. L. Cox, "'At least they're learning something': the hazy line between collaboration and collusion," *Assess. Eval. High. Educ.*, vol. 30, no. 2, pp. 107–122, 2005.

[13] K. Cho, C. D. Schunn, and R. W. Wilson, "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives," *J. Educ. Psychol.*, vol. 98, no. 4, pp. 891–901, 2006.

[14] M. R. Fellenz, “Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions,” *J. Manag. Educ.*, vol. 30, no. 4, pp. 570–591, 2006.
[15] R. Ballantyne, K. Hughes, and A. Mylonas, “Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process,” *Assess. Eval. High. Educ.*, vol. 27, no. 5, pp. 427–441, Sep. 2002.
[16] K. J. Topping, “Peer assessment,” *Theory Pract.*, vol. 48, no. 1, pp. 20–27, 2009.
[17] Y. Song, Z. Hu, and E. F. Gehringer, “Closing the Circle: Use of Students’ Responses for Peer-Assessment Rubric Improvement,” in *Advances in Web-Based Learning -- ICWL 2015*, F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. H. Lau, Eds. Springer International Publishing, 2015, pp. 27–36.
[18] N. Falchikov, “Involving students in assessment,” *Psychol. Learn. Teach.*, vol. 3, no. 2, pp. 102–108, 2004.
[19] L. De Alfaro, M. Shavlovsky, and V. Polychronopoulos, “Incentives for Truthful Peer Grading,” *ArXiv160403178 Cs*, Apr. 2016.
[20] W. Sutherland-Smith and others, “Crossing the line: Collusion or collaboration in university group work?,” *Aust. Univ. Rev.*, vol. 55, no. 1, p. 51, 2013.
[21] K. G. Love, “Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction,” *J. Appl. Psychol.*, vol. 66, no. 4, pp. 451–457, 1981.
[22] A. Franklyn-Stokes and S. E. Newstead, “Undergraduate cheating: who does what and why?,” *Stud. High. Educ.*, vol. 20, no. 2, pp. 159–172, 1995.
[23] J. S. Kane and E. E. Lawler, “Methods of peer assessment,” *Psychol. Bull.*, vol. 85, no. 3, pp. 555–586, 1978.
[24] Y. Song, Z. Hu, Y. Guo, and E. F. Gehringer, “An experiment with separate formative and summative rubrics in educational peer assessment,” in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–7.
[25] Y. Song, G. Yifan, and G. Edward, “An Exploratory Study of Reliability of Ranking vs. Rating in Peer Assessment,” in *Accepted by ICALT 2017 : 19th International Conference on Advanced Learning Technologies*, Paris, France, 2017.
[26] F. Pramudianto, M. Aljeshi, Y. Song, and E. Gehringer, “Peer Review Data Warehouse: Insights From Different Systems,” in *Submitted to Computer-Supported Peer Review in Education Workshop at EDM 2016*, 2016.
[27] Y. Song, F. Pramudianto, and E. F. Gehringer, “A markup language for building a data warehouse for educational peer-assessment research,” in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–5.
[28] K. A. Hawick and H. A. James, “Enumerating Circuits and Loops in Graphs with Self-Arcs and Multiple-Arcs.,” in *FCS*, 2008, pp. 14–20.