Improving Peer Assessment with Graph Neural Networks

Alireza A. Namanloo
Ontario Tech University
Oshawa, Canada
ali@ali-naman.com

Julie Thorpe
Ontario Tech University
Oshawa, Canada
julie.thorpe@uoit.ca

Amirali Salehi-Abari
Ontario Tech University
Oshawa, Canada
abari@uoit.ca

ABSTRACT

Peer assessment systems are emerging in many settings, such as peer grading in large (online) classes, peer review in conferences, peer art evaluation, etc. However, peer assessments might not be as accurate as expert evaluations, thus rendering these systems unreliable. The reliability of peer assessment systems is influenced by various factors such as assessment ability of peers, their strategic assessment behaviors, and the peer assessment setup (e.g., peer evaluating group work or individual work of others). In this work, we first model peer assessment as multi-relational weighted networks that can express a variety of peer assessment setups, and can also capture conflicts of interest and strategic behaviors. Leveraging our peer assessment network model, we introduce a graph neural network which can learn assessment patterns and user behaviors to more accurately predict expert evaluations. Our extensive experiments on real and synthetic datasets demonstrate the efficacy of our approach, which outperforms a variety of peer assessment methods.

Keywords

Peer Assessment, Graph Neural Network.

1. INTRODUCTION

Peer assessment systems have emerged as a cost-effective and scalable evaluation mechanism in many settings such as peer grading in large (online) classes and peer review in conferences. In these systems, peers assess each others' work (e.g., assignments, papers, etc.) in lieu of a set of pre-appointed experts responsible for evaluation (e.g., instructors, teaching assistants, program committee members, etc.). These peer assessment systems not only make the evaluation of thousands of contributions plausible, but also help to deepen peers' understanding [22], and facilitate peers providing feedback to each other [20]. However, the reliability of peer assessment systems is directly impacted by the accuracy of peers in their assessments. Peers might lack knowledge or motivation to accurately evaluate others, or they might be strategic in their assessments for their own gain [1, 12, 11, 2, 15, 34, 25, 21].

Related work. Two classes of approaches are taken to address reliability challenges. One primarily focuses on designing strategy-proof peer assessment mechanisms, which incentivize peers to accurately assess each other [6, 24, 9, 16, 11, 35, 32]. The other class of approaches—most relevant to our work—emphasizes learning peer aggregation mechanisms, which aggregate noisy peer assessments for an item (e.g., assignment or paper) as an estimate of its ground-truth valuation (or expert evaluation) [19, 27, 7, 30, 4]. The learning methods for peer assessment aggregation fall into unsupervised [27, 19, 5] and semi-supervised [7, 30] approaches based on whether or not a subset of ground-truth labels are used for training in addition to peer assessment data. These models usually possess particular inductive biases such as peer's assessment accuracy being correlated with his/her item's ground-truth valuations (e.g., the grade of his/her assignment) [19, 27, 7]; or peer's accuracy in an assessment depending on the extent of its agreement with others' assessments or ground-truth valuations [30]. However, these machine learning methods are empirically shown to be only as effective as simple aggregation mechanisms such as averaging [23]. Moreover, these approaches are not flexible and general enough to accommodate a wide variety of peer assessment modes (e.g., when an individual assesses the group contribution of others or self assessments). Our focus in this paper is to develop a semi-supervised aggregation mechanism without any specific or restrictive inductive bias, accommodating various modes of peer assessments.

Contribution. We first introduce our graph representation model of peer assessment, which we call *social-ownership-assessment network (SOAN)*.[1] Our SOAN model can express a wide variety of peer assessment setups (e.g., self-assessment and peer assessment for both individual or group contributions) and represent conflict-of-interest relations between peers using auxiliary information, such as social networks. Leveraging our SOAN model, we then introduce a semi-supervised graph convolutional network (GCN) approach, called *GCN-SOAN*, which can learn assessment patterns and behaviors of peers, without any restrictive inductive bias, to predict ground-truth valuations. We run extensive experiments on real-world and synthetic datasets to evaluate the efficacy of GCN-SOAN. Our GCN-SOAN outperforms a wide variety of baseline methods (including sim-

[1] SOAN can read as "swan."

ple heuristics, semi-supervised, and unsupervised approaches) on the same real-world dataset [23], which was shown to be challenging for machine learning approaches. Our GCN-SOAN also outperforms others on a wide range of synthetic data, which captures strategic assessment behavior between users, follows the assumptions of competitor baselines, or considers strict and generous graders. GCN-SOAN can be a stand-alone approach or possibly be integrated with some incentivizing mechanisms (e.g., [32, 5, 35]).

2. PROPOSED APPROACH

Our goal is to predict the ground-truth assessments (e.g., expert evaluations of educational or professional work) from noisy peer assessments. We first discuss our proposed graph representation model, *social-ownership-assessment network (SOAN)*, for capturing the peer grading behavior. We then present a modified graph convolutional network (GCN), which leverages our SOAN model, to predict the ground-truth assessments. We call this approach *GCN-SOAN*.

2.1 Social-Ownership-Assessment Model

We assume that a set of n users $\mathcal{U}$ (e.g., students or scholars) can assess a set of m items $\mathcal{I}$ (e.g., a set of educational, professional, or intellectual work). The examples cover various applications ranging from peer grading in classrooms to peer reviewing scientific papers, professional work, or research grant applications. We also consider each item $i \in \mathcal{I}$ possesses a (possibly unknown, but verifiable) ground-truth value $v_i \in \mathbb{R}^+$ (e.g., staff grade for a course work, or expert evaluation of intellectual or professional work).

The user-item assessments can be represented by *assessment matrix* $\mathbf{A} = [A_{ui}]$, where A_{ui} is the assessment (e.g., grade or rating) of user $u \in \mathcal{U}$ for item $i \in \mathcal{I}$. We let $A_{ui} = 0$ when the user u's assessment for item i is missing; otherwise $A_{ui} \in \mathbb{R}^+$. As the assessment matrix $\mathbf{A}$ is sparse, we equivalently represent it by an undirected weighted bipartite graph, consisting of two different node types of users $\mathcal{U}$ and items $\mathcal{I}$, and weighted assessment edges between them (see Figure 1a as an example).

We introduce a *social-ownership-assessment network (SOAN)*, an undirected weighted multigraph, consisting of three types of social, ownership, and assessment relationships on two node types of users and items. In addition to the assessment matrix $\mathbf{A}$, this network consists of two other adjacency matrices: *social matrix* $\mathbf{S} = [S_{uv}] \in \mathbb{R}^{n \times n}$ and *ownership matrix* $\mathbf{O} = [O_{ui}] \in \mathbb{R}^{n \times m}$. The social matrix $\mathbf{S}$, by capturing the friendship and foe relationships between users $\mathcal{U}$, can accommodate "conflict of interest" information. The ownership matrix $\mathbf{O}$, by capturing which users to what extent own or contributed to an item, not only completes conflict of interest information but also provides flexibility of modeling group contributions, self-evaluation, etc. We let $G = (\mathbf{S}, \mathbf{O}, \mathbf{A})$ denote the tuple of all three networks of SOAN. Figure 1 demonstrates some instantiations of our models for various settings. SOAN offers various advantages over the existing peer assessment models (e.g., [27, 19, 7, 4]):

Expressiveness. Our model is more *expressive* as it facilitates the representation of many various peer assessment settings that could not be accommodated in the existing models. Its expressive power can be realized in the settings such as self assessments (Figure 1b), peer assessments for both solo and group work (Figures 1e and 1f), and the mixtures of peer and self assessments for solo and group work (Figures 1c and 1d). For all of these settings, our SOAN model can also express conflict of interest (which is neglected in other models) through a social network (see Figure 1g).

Less Assumptions. Dissimilar to some existing models (e.g., [27, 19, 7, 30]), our model avoids making explicit or implicit assumptions about the relationships between ground-truth values (or grades) and the quality of peer assessments. It is still flexible enough to learn such correlations from assessment data if it exists. Our experiments below have shown that our model outperforms other models with restrictive assumptions regardless of whether their assumptions are present in the data or not.

2.2 Graph Convolutional Networks

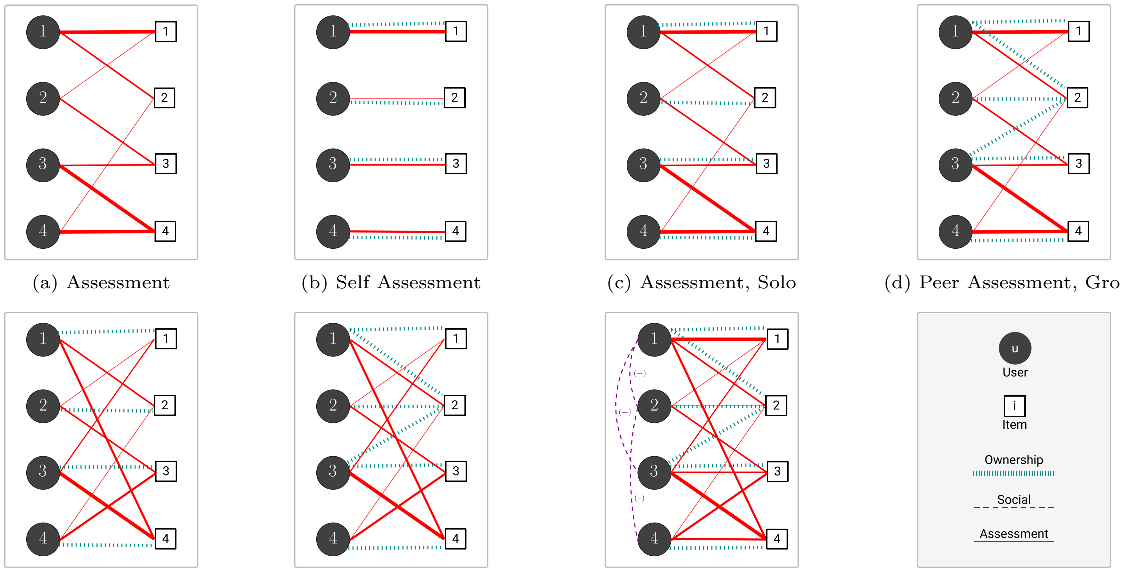
Our learning task is semi-supervised. Given a social-ownership-assessment network $G = (\mathbf{S}, \mathbf{O}, \mathbf{A})$ and a set of ground-truth valuations $V_\mathcal{D} = \{v_j | j \in \mathcal{D}\}$ for a subset of items $\mathcal{D} \subset \mathcal{I}$, we aim to predict v_i for $i \notin \mathcal{D}$. More specifically, we aim to learn the function $f(i|\boldsymbol{\theta}, G)$ for predicting the ground-truth valuation v_i by $\hat{v}_i = f(i|\boldsymbol{\theta}, G)$. The model parameters $\boldsymbol{\theta}$ are learned from both the observed ground-truth valuations $V_\mathcal{D} = \{v_j | j \in \mathcal{D}\}$ and social-ownership-assessment network G. We formulate the function f by a modified graph convolution network (GCN) with a logistic head:

$$f(i|\boldsymbol{\theta}, G) = \sigma\left(\mathbf{w}^{(o)}\mathbf{z}_i + b^{(o)}\right), \tag{1}$$

where $\sigma(.)$ is the sigmoid function for converting the linear transformation of the node i's embedding $\mathbf{z}_i$ into its predicted valuations. Here, $\mathbf{w}_o$ and b_o are the weight vector and the bias parameter for the output layer. The node (i.e., item) embedding $\mathbf{z}_i$ is computed with K layers of graph convolution network. Let $\mathbf{H}^{(l)}$ be the $(n+m) \times d$ matrix of d-dimensional node embeddings at layer l for all users $\mathcal{U}$ and items $\mathcal{I}$ such that user u and item i's vector embeddings are located at the u-th and $(m+i)$-th rows, respectively. In Eq. 1, the item i's embedding $\mathbf{z}_i$ is the $(m+i)$-th row of $\mathbf{H}^{(K)}$ with the updating rule of

$$\mathbf{H}^{(l+1)} = g^{(l)}\left(\mathbf{D}^{-1}\mathbf{M}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right). \tag{2}$$

The matrix $\mathbf{M}$ is constructed from the graph $G = (\mathbf{S}, \mathbf{O}, \mathbf{A})$ by $\mathbf{M} = \begin{pmatrix} \mathbf{S} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{0}_m \end{pmatrix} + \mathbf{I}$, where $\mathbf{P} = \mathbf{O} + \mathbf{A}$, $\top$ is the transpose operator, $\mathbf{0}_m$ is $m \times m$ zero matrix, and $\mathbf{I}$ is the identity matrix. In Eq. 2, $\mathbf{D}$ is the diagonal matrix with $D_{ii} = \sum_j \mathbb{1}[M_{ij} \neq 0]$ with $\mathbb{1}[.]$ as the indicator function. The core idea in Eq. 2 is to update the node embeddings at layer $l+1$, denoted by $\mathbf{H}^{(l+1)}$, from layer l's node embeddings $\mathbf{H}^{(l)}$. This update includes the multiplication of the layer l's embeddings $\mathbf{H}^{(l)}$ by the normalized matrix $\mathbf{D}^{-1}\mathbf{M}$, then linear transformation by learned weight matrix $\mathbf{W}^{(l)}$ at layer l, and finally passing through a non-linear activation function $g^{(l)}$. The initial embedding matrix $\mathbf{H}^{(0)}$ can be node-level features (e.g., textual features for items, user profiles for users, etc.). When the node-level features are absent, the common practice is to initialize the embeddings with one-hot indicators [14, 28, 10]. As our GCN is built upon SOAN, we refer to this combination as *GCN-SOAN*.



(a) Assessment (b) Self Assessment (c) Assessment, Solo (d) Peer Assessment, Group

(e) Peer Assessment, Solo (f) Peer Assessment, Group (g) Assessment, Group, Social (h) Legends

Figure 1: Different instantiations of social-ownership-assessment network (SOAN): (a) assessments provided by users to items as weighted edges; thicker lines for higher weights; (b) self assessments of users for their own items; (c) combination of self and peer assessments of solo contributions; (d) self and peer assessments of both solo and group contributions; (e) peer assessments of solo contributions; and (g) self and peer assessments of group contributions with the social networks between users for capturing conflict of interest.

The updating rule in GCN-SOAN (see Eq. 2) benefits from row normalization of the adjacency matrix similar to many other graph neural networks [31, 26, 28, 29]. As the choice of an effective normalization technique is an application-specific question [10, 3], we have decided to normalize our weighted SOAN model by taking an unweighted average, which has been suggested as a solution to address the sensitivity to node degrees for neighborhood normalization [10].

Our GCN-SOAN differentiates from vanilla GCN in various ways: (i) GCN-SOAN supports weighted graphs as opposed to GCN which solely is designed for unweighted graphs; (ii) it has asymmetric normalization as opposed symmetric normalization. These properties well-equip our GCN-SOAN to aggregate the information from multi-hop neighborhoods (e.g., neighbors, neighbors of neighbors, and so on) of SOAN, thus successfully capturing various assessment behaviors and patterns as evidenced in our experiments below.

Richer data as node-level features. To incorporate richer data for users (e.g., grader's profile, reviewer's expertise level, reviewer's interest, etc.) and items (e.g., textual or visual information) in peer assessment systems, our GCN-SOAN can readily accommodate those information in the form of their node-level features. For example, the expertise level of peers can be represented as the one-hot encoding for initial embeddings. These initial embeddings can be extended with any other type of peers' auxiliary information (e.g., education, age, sex, etc.). Similarly, initial embeddings of items can accommodate item features (e.g., the keywords for papers, textual features extracted from a paper, etc.).

Learning. Given the SOAN of G and a small training set of ground-truth valuations $V_{\mathcal{D}}$, we learn GCN-SOAN parameters by minimizing the *mean square error* of its predictions: $L(\boldsymbol{\theta}|G,\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\left(v_i - f\left(i|\boldsymbol{\theta},G\right)\right)^2$, where $|\mathcal{D}|$ is the number of items in the training dataset, and $f\left(i|\boldsymbol{\theta},G\right)$ is the estimated valuation of GCN-SOAN for item i. This loss function can be minimized by gradient-based optimization techniques (e.g., stochastic gradient descent, Adam, etc.).

As opposed to many existing peer assessment systems with unsupervised learning approaches (e.g., [19, 4, 27]), we deliberately have adopted a semi-supervised learning approach for predicting ground-truth assessment. This choice offers many advantages at some cost of access to a small training dataset. By learning from the training data, GCN-SOAN is well-equipped to mitigate the influence of strategic behaviors, assessment biases, and unreliable assessments in peer assessment systems. Of course, the extent of this mitigation depends on the size of training data.

3. EXPERIMENTS

We run extensive experiments on real-world and synthetic datasets to compare our GCN-SOAN model against other peer assessment methods. While the real-world datasets allow us to assess the practical efficacy of our approach, we generate various synthetic data to assess its robustness in various settings (e.g., strategic and biased assessments).[2]

Real-world dataset. The peer grading datasets of Sajjadi et al. [23] includes peer and self grades of 219 students for exercises (i.e., questions) of four assignments and their ground-truth grades.[3] For each specific assignment, the sub-

[2] See the longer version for additional experiments [17].

[3] The datasets can be found at `http://www.tml.cs.uni-tuebingen.de/team/luxburg/code_and_data/peer_grading_data_request_new.php`. The original datasets are for six assignments. However, two of the datasets have ordinal peer gradings, not applicable to our experiments.

Table 1: The summary statistics of real-world peer grading datasets.

	Average Grades			Number of					
Asst. ID	Ground-truth	Peer	Self	Exercises	Groups	Students	Items	Peer grades	Self grades
1	0.62 ± 0.27	0.70 ± 0.26	0.74 ± 0.22	3	75	183	225	965	469
2	0.71 ± 0.24	0.76 ± 0.23	0.80 ± 0.22	4	77	206	308	1620	755
3	0.69 ± 0.33	0.75 ± 0.31	0.82 ± 0.26	5	76	193	380	1889	890
4	0.59 ± 0.27	0.68 ± 0.29	0.76 ± 0.24	3	79	191	237	1133	531

missions are group work of 1–3 students, but each student individually has self and peer graded all exercises of two other submissions (in a double-blind setup). We treat all data associated with each assignment as a separate dataset, where all the submitted solutions to its exercises form the item set $\mathcal{A}$ and the user set $\mathcal{U}$ includes all students who have been part of a submission. We also have scaled peer, self, and ground-truth grades to be in the range of $[0, 1]$. Table 1 shows the statistics summary of our datasets.[4]

Synthetic datasets. We discuss different models used for the generation of our synthetic data.

Ground-truth valuation/grade generation. For all $i \in \mathcal{A}$, we sample the true valuation v_i from a mixture of two normal distributions $v_i \sim P(\mathrm{x}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{c=1}^{2} \pi_c \mathcal{N}(\mathrm{x}; \mu_c, \sigma_c)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$, with π_k, μ_k, and σ_k being the mixing probability, mean, and standard deviation of the component k.

Social network generation. We create social networks between users by *Erdős-Rényi (ER)* random graph $G(n, p)$ model: each pair of n users are connected to each other with the connection probability of p.

Ownership network generation. For all synthetic datasets, we randomly connect each user to just one item (i.e., one-to-one correspondence between users and items). This setup is in favor of existing peer assessment methods (e.g., [19, 27, 7]), which does not support group ownerships of items.

Assessment network (or peer grades) generation. To generate peer assessment for each item $i \in \mathcal{A}$, we first randomly select a set of k users $N(i) \subset \mathcal{U}$ such that $|N(i)| = k$. Then, for each $u \in N(i)$, we set u's assessment for item i, denoted by A_{ui}, using one of these two models. The *strategic* model sets $A_{ui} = 1$, if the grader u is a friend of the user j who owns the item i (i.e., $s_{uj}.o_{ji} = 1$); otherwise A_{ui} comes from a normal distribution with the mean of v_i and standard deviation of σ_H. This implies that friends collude to peer grade each others with the highest grade, but would be relatively fair and reliable in assessing a "stranger." The *bias-reliability* model draws A_{ui} from the normal distribution $\mathcal{N}(\mathrm{x}; \hat{\mu}, \hat{\sigma})$ with the mean $\hat{\mu} = v_i + \alpha$ and $\hat{\sigma} = \sigma_{max}(1 - \beta v_l)$, where v_l is the true valuation of item l owned by user u, and σ_{max} is the maximum possible standard deviation (i.e., unreliability) for peer graders. Here, the *bias* parameter $\alpha \in [-1, 1]$ controls the degree of generosity (for $\alpha > 0$) or strictness (for $\alpha < 0$) of the peer grader. The *reliability* parameter β controls the extent that the reliability of the grader is correlated with his/her item's grade (i.e, the peer graders with higher grades are more reliable graders). The inductive bias of many peer assessment models (e.g., [27, 7, 19]) include the assumption that the grader's reliability is a function of his/her item's grade. Our bias-reliability model allows us to generate synthetic datasets with the presence of this assumption. So we can compare our less-restrictive GCN-SOAN with those models tailored to this specific assumption in such datasets.

Baselines. We compare the performance of our GCN-SOAN model with PeerRank [27], PG1 [19], RankwithTA [7], Vancouver [4], Average, and Median. Average and Median (resp.) outputs the average and median (resp.) of each item's peer grades as its predicted evaluation. As PeerRank, PG1, and RankwithTA treat users and items interchangeably, they can't be directly applied to our real-world data with individual assessments on group submissions. For these methods, we preprocess our real-world dataset by taking the average of the grades provided by a group's members for a particular submission as the group assessment for the submission. For the PeerRank and PG1, we have used the same parameter settings reported by the original papers. The parameters for RankwithTA and Vancouver are selected by grid search with multiple runs, since the optimal parameters either were not reported or result in non-competitive performance.[5]

Experimental setup. We implement GCN-SOAN based on PyTorch [18] and PyTorch Geometric [8].[6] For all experiments, we use two GCN-SOAN convolutional layers with an embedding dimension of 64 and ELU as activation functions of all hidden layers. We train the model for 800 epochs with Adam optimizer [13] and a learning rate of 0.02. We initialize the node embeddings with vectors of ones. We use Monte Carlo cross-validation [33] with the training-testing splitting ratio of 1:9 (in synthetic data) or 1:4 (in real-world data), implying that just 10% or 20% data is used for training and the rest for testing. To make our results even more robust, we run all tested methods (our model and baselines) on four random splits and report the average error over those splits. For each random split, we compute the root mean square error (RMSE) over testing data as the prediction error.

Results: Real-World Datasets. To assess the effectiveness of GCN-SOAN in predicting ground-truth valuations, we compare it against the baseline methods on eight real-world

[4] While GCN-SOAN can easily accommodate user/item features, we were not able to explore its full potential due to a lack of access to datasets with such features.

[5] We set the Vancouver's precision parameter to 0.1. For RankwithTA, we set 0.8 and 0.1 (respectively) for the parameters controlling the impact of working ability on grading ability and grading ability on the grade (respectively).

[6] The implementation of GCN-SOAN can be obtained from: `https://github.com/naman-ali/GCN-SOAN/`

Table 2: Root mean square error of various methods over two classes of real-world datasets. The first and second best are shown with dark and light gray backgrounds, respectively. ↑ and ↓ denote better and worse than Average. GCN-SOAN (ours) is the only method that consistently has outperformed Average for all datasets. Results are averaged over three runs.

	Peer evaluation				Peer and self evaluation			
Model	Asst. 1	Asst. 2	Asst. 3	Asst. 4	Asst. 1	Asst. 2	Asst. 3	Asst. 4
Average	0.1917	0.1712	0.1902	0.1989	0.1944	0.1681	0.2023	0.2117
Median	0.1991 ↓	0.1843 ↓	0.2047 ↓	0.2250 ↓	0.2111 ↓	0.1750 ↓	0.2333 ↓	0.2538 ↓
PeerRank	0.1913 ↑	0.1762 ↓	0.2235 ↓	0.2087 ↓	0.1888 ↑	0.1721 ↓	0.2203 ↓	0.2168 ↓
PG1	0.1919 ↓	0.1669 ↑	0.2110 ↓	0.2161 ↓	0.2009 ↓	0.1680 ↑	0.2111 ↓	0.2304 ↓
RankwithTA	0.1922 ↓	0.1903 ↓	0.2183 ↓	0.1740 ↑	0.1884 ↑	0.1845 ↓	0.2137 ↓	0.1792 ↑
Vancouver	0.1851 ↑	0.1688 ↑	0.1951 ↓	0.2071 ↓	0.1815 ↑	0.1672 ↑	0.1945 ↑	0.2101 ↑
GCN-SOAN (ours)	0.1795 ↑	0.1673 ↑	0.1869 ↑	0.1822 ↑	0.1778 ↑	0.1621 ↑	0.1840 ↑	0.1821 ↑

datasets. These datasets differentiate on (i) which assignment dataset is used and (ii) whether both peer and self-grades are used or only peer grades. For GCN-SOAN, we just create an assessment network, thus allowing us to measure how the assessment network alone can improve the predication accuracy. As shown in Table 2, our GCN-SOAN model outperforms others in five datasets and ranked second in remaining ones with a small margin. RankwithTA and PG1 are the only models that slightly outperform GCN-SOAN for those three datasets. Notably, GCN-SOAN is the only model which consistently outperformed the simple Average benchmark. This observation is consistent with Sajjadi et al.'s findings [23] on the same dataset that the existing machine learning methods (not including ours) could not improve results over simple baselines. However, the conclusion does not hold anymore as our machine learning GCN-SOAN approach could consistently improve over simple baselines. This improvement mainly arises from the expressive power and generalizability of GCN-SOAN (discussed in Section 2).

Results: Synthetic Data with Bias-Reliability. We run an extensive set of experiments with the bias-reliability peer grade generation model to assess our GCN-SOAN under various peer assessment settings. For these experiments, we define a *default* setting for all parameter of synthetic generation methods (e.g., bias parameter α, reliability parameter β, etc.). For each experiment, we fix all parameters except one; then, by varying that parameter, we aim to understand its impact on the performance of GCN-SOAN and other baselines. Our default setting includes a number of users $n = 500$ and number of items $m = 500$; random one-to-one ownership network; $\boldsymbol{\mu} = (0.3, 0.7)$, $\boldsymbol{\sigma} = (0.1, 0.1)$, and $\boldsymbol{\pi} = (0.2, 0.8)$ for the ground-truth generation method;[7] the number of peer grades $k = 3$,[8] $\sigma_{max} = 0.25$, bias parameter $\alpha = 0$, and reliability parameter $\beta = 0$ for assessment network generation; and no social network generation.[9]

Figure 2a shows how the prediction error changes with the number of peer graders k while the other parameters are fixed to the default setting. Unsurprisingly, the performance of all methods improves with k. GCN-SOAN not only outperforms others for any k, but also exhibits significant improvement over others for a relatively small k (e.g., $k \leq 4$). This superiority of GCN-SOAN with minimal number of peer graders is its strength to make peer assessment suitable and practical for different applications, as so many peer assessment requests will put unnecessary stress and burden on users, thus impeding the practicality of the system.

Figure 2b illustrates the errors for each model while changing the bias parameter α (and keeping other parameters fixed to default). GCN-SOAN performs significantly better than other models for any bias values, including generous ($\alpha > 0$) and strict graders ($\alpha < 0$). GCN-SOAN owes this success to its ability to learn students' grading behavior by leveraging a small portion of ground truth grades and assessment network structure. These experiments show that our model could be a great choice for those peer assessment settings where the peer grades are intentionally or unintentionally overestimated/underestimated.

Figure 2c reports the errors for various values of reliability parameters β. Recall that the β controls the extent that the accuracy of each peer in his/her assessments is correlated with his/her item's grade. Our results show that GCN-SOAN is very competitive to other models, even those built based on this correlation assumption (e.g., [27, 19]). We observe that only when $\beta > 0.8$, PeerRank outperform GCN-SOAN. One might argue that $\beta > 0.8$ is implausible scenario in practice. However, this result suggests that our model is still competitive choice for settings in which peer assessment accuracy is correlated with peer success.

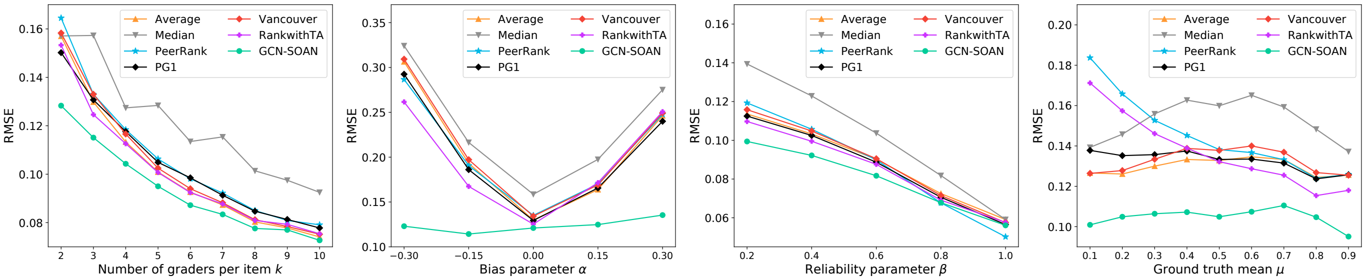
To study how various ground-truth generation distribution impacts the prediction error of various method, we first change the default biomodal mixture of normal distributions (for ground truth generation) to a normal distribution by setting $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2 = 0.15$. Then, we only vary the mean of distribution while other parameters are fixed to default. As shown in Figure 2d, GCN-SOAN consistently outperforms others regardless of the underlying ground-truth distribution. Notably, PeerRank and RankwithTA do not perform well when most users own items with low grades.

Results: Synthetic Data with Strategic Assessment. We study the performance of all peer assessment methods under the strategic model discussed above. For this set of experiments, we define this default setting: number of users $n = 500$ and

[7] This setting for ground-truth distribution is motivated by two-humped grade distribution in academic classes.

[8] This choice of 3 is motivated by the fact that most practical applications (e.g., conference review or peer grading in classrooms) do not require more than 3 peer assessment per item due to time-consuming nature of assessment processes.

[9] We have run some other experiments with different default settings. The results were qualitatively similar.



(a) various number of graders k (b) various bias α (c) various reliability β (d) various ground-truth mean μ

Figure 2: Root mean square errors of various methods, synthetic data with bias-reliability peer generation model, default setting for all parameters but varying (a) number of peer graders k, (b) grading bias α, (c) reliability parameter β, or (d) ground-truth mean μ. Average over four runs.

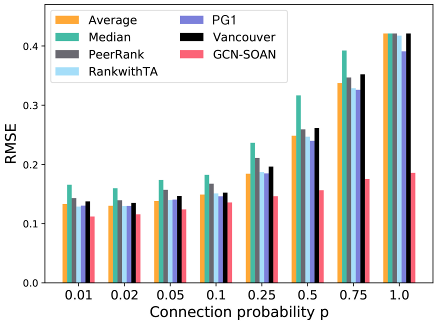


Figure 3: Average of root mean square errors, synthetic data with strategic peer grade generation model and random graph model.

number items $m = 500$; random one-to-one ownership network; $\boldsymbol{\mu} = (0.3, 0.7)$, $\boldsymbol{\sigma} = (0.1, 0.1)$, and $\boldsymbol{\pi} = (0.2, 0.8)$ for the ground-truth generation method; the number of peer grades $k = 3$ and $\sigma_{max} = 0.25$ for assessment network generation by the strategic model; and ER random graph model with $n = 500$ and $p = 0.05$ for social network generation. We only vary the connection probability p while keeping other parameters fixed to study how the connection density of colluding social networks impact the accuracy of peer assessment methods. Figure 3 show the outstanding performance of our model compare to other benchmarks and illustrate how our model is more resilient to colluding behaviors. This result suggest that GCN-SOAN is well-eqiupped to detect conflict-of-interest behaviors and mitigate the possible impact of any strategic behaviors.

Discussion. Our experiments show that GCN-SOAN learns very well various grading behaviors, even when graders have intentional or unintentional biases in their evaluations. We also observe that our GCN-SOAN can outperform other benchmarks even when their main inductive biases are strongly present in the dataset (e.g., when the grading ability of users are strongly correlated to the quality of their own work).

Our set of benchmarks, in spite of being very competitive, does not cover all ML-based peer assessment methods. We make a few remarks about this. PG3 [19] is missing in our experiments as its implementation was not publicly available and we could not properly implement it to gets its competitive performance. However, we expect that GCN-SOAN outperforms PG3 since the relative improvements of GCN-SOAN over PG1 is 10.27% (on average) compared to the average relative improvement of 1.76% for PG3 over PG1 (as reported in their original paper [19]). We also speculate that the other traditional ML-based methods might not outperform our GCN-SOAN for at least one reason: except graph neural networks, most ML methods assume that the data points (e.g., peer grades, ground-truth valuation, etc.) are identically and independently distributed (i.i.d.). However, as argued earlier, the peers' grading behaviors, ownerships, social connections, and valuations of their owned items are all dependent on each other. Ignoring these dependencies in machine learning methods with i.i.d. assumptions make them less competitive to our GCN-SOAN in which these dependencies are well-expressed by our proposed SOAN and well-exploited by our proposed graph neural network algorithm. This might explain why the literature has even been thin in successfully exploring other advanced machine learning models and Sajjadi et al. [23] concluded that machine learning methods cannot improve over simple heuristics (e.g., average).

4. CONCLUSION AND FUTURE WORK

We represent peer assessment data as a weighted multi-relational graph, which we call social-ownership-assessment network (SOAN). Our SOAN can easily express many different peer assessment setups (e.g., self assessment, peer assessment of group or individual work, etc.). Leveraging SOAN, we introduce a modified graph convolutional network approach, which learns peer assessment behaviors, to more accurately predict ground-truth valuations. Our extensive experiments demonstrate that GCN-SOAN outperforms state-of-the-art baselines in a variety of settings, including strategic behavior, grading biases, etc.

Our SOAN model provides a solid foundation for the broader investigation of graph neural network approaches for peer assessments. Our GCN-SOAN can be extended to mitigate the over-smoothing effect observed in our experiments, or to include a different set of network weights for each relation type of social, assessment, and ownership. Another promising direction is to assess the effectiveness of GCN-SOAN or its extensions on real-world assessment data, with the presence of social network data. Finally, it would be interesting to collect data with richer node features in order to evaluate how node features can further improve our GCN-SOAN.

5. REFERENCES

[1] H. Aziz, O. Lev, N. Mattei, J. Rosenschein, and T. Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-16)*, volume 30, 2016.

[2] H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*, 275:295–309, 2019.

[3] Y. Chen, X. Tang, X. Qi, C.-G. Li, and R. Xiao. Learning graph normalization for graph neural networks. *arXiv preprint arXiv:2009.11746*, 2020.

[4] L. de Alfaro and M. Shavlovsky. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *arXiv preprint arXiv:1308.5273*, 2013.

[5] L. De Alfaro and M. Shavlovsky. Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education (SIGCSE-14)*, pages 415–420, 2014.

[6] L. De Alfaro, M. Shavlovsky, and V. Polychronopoulos. Incentives for truthful peer grading. *arXiv preprint arXiv:1604.03178*, 2016.

[7] H. Fang, Y. Wang, Q. Jin, and J. Ma. Rankwithta: A robust and accurate peer grading mechanism for moocs. In *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE-17)*, pages 497–502, 2017.

[8] M. Fey and J. E. Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[9] X. A. Gao, J. R. Wright, and K. Leyton-Brown. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artif. Intell.*, 275:618–638, 2019.

[10] W. L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2020.

[11] S. Jecmen, H. Zhang, R. Liu, N. B. Shah, V. Conitzer, and F. Fang. Mitigating manipulation in peer review via randomized reviewer assignments. *arXiv preprint arXiv:2006.16437*, 2020.

[12] A. Kahng, Y. Kotturi, C. Kulkarni, D. Kurokawa, and A. D. Procaccia. Ranking wily people who rank each other. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1087–1094, 2018.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR-15)*, 2015.

[14] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the Fifth International Conference on Learning Representations (ICLR-17)*, 2017.

[15] Y. Kotturi, A. Kahng, A. Procaccia, and C. Kulkarni. Hirepeer: Impartial peer-assessed hiring at scale in expert crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 2577–2584, 2020.

[16] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[17] A. A. Namanloo, J. Thorpe, and A. Salehi-Abari. Improving peer assessment with graph convolutional networks, 2021.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[19] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y.-T. Ng, and D. Koller. Tuned models of peer assessment in moocs. In *Proceedings of The 6th International Conference on Educational Data Mining (EDM-13)*, 2013.

[20] C. Psenicka, W. Vendemia, and A. Kos. The impact of grade threat on the effectiveness of peer evaluations of business presentations: An empirical study. *International Journal of Management*, 30(1):168–175, 2013.

[21] K. Reily, P. L. Finnerty, and L. Terveen. Two peers are better than one: aggregating peer reviews for computing assignments is surprisingly accurate. In *Proceedings of the ACM 2009 international conference on Supporting group work (GROUP-09)*, pages 115–124, 2009.

[22] P. M. Sadler and E. Good. The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31, 2006.

[23] M. S. Sajjadi, M. Alamgir, and U. von Luxburg. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In *Proceedings of the Third (2016) ACM conference on Learning@Scale (L@S-16)*, pages 369–378, 2016.

[24] T. Staubitz, D. Petrick, M. Bauer, J. Renz, and C. Meinel. Improving the peer assessment experience on mooc platforms. In *Proceedings of the Third (2016) ACM conference on Learning@Scale (L@S-16)*, pages 389–398, 2016.

[25] I. Stelmakh, N. B. Shah, and A. Singh. Catch me if I can: Detecting strategic behaviour in peer assessment. In *ICML Workshop on Incentives in Machine Learning*, 2020.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR-18)*, 2018.

[27] T. Walsh. The peerrank method for peer assessment. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence (ECAI-14)*, page 909–914, 2014.

[28] H. Wang and J. Leskovec. Unifying graph convolutional neural networks and label propagation. *arXiv preprint arXiv:2002.06755*, 2020.

[29] X. Wang, M. Cheng, J. Eaton, C.-J. Hsieh, and F. Wu. Attack graph convolutional networks by adding fake nodes. *arXiv preprint arXiv:1810.10751*, 2018.

[30] Y. Wang, H. Fang, Q. Jin, and J. Ma. Sspa: an effective semi-supervised peer assessment method for large scale moocs. *Interactive Learning Environments*, pages 1–19, 2019.

[31] Y. Wang, Z. Hu, Y. Ye, and Y. Sun. Demystifying graph neural network via graph filter assessment. 2019.

[32] J. R. Wright, C. Thornton, and K. Leyton-Brown. Mechanical ta: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE-15)*, pages 96–101, 2015.

[33] Q.-S. Xu and Y.-Z. Liang. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

[34] Y. Xu, H. Zhao, X. Shi, and N. B. Shah. On strategyproof conference peer review. In *IJCAI*, pages 616–622. ijcai.org, 2019.

[35] H. Zarkoob, H. Fu, and K. Leyton-Brown. Report-sensitive spot-checking in peer-grading systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS-20)*, pages 1593–1601, 2020.