ABSTRACT
We develop a random forest classifier that helps assign academic credit for student class forum participation. Our twelve predictors are quantities that are available from typical forum facilities, such as the number of endorsed contributions, and the number of answers and views. We add page rank and centrality to these base measures. The classification target are the four classes created by student rank quartiles. Course content experts provided ground truth by ranking a limited number of post pairs. We expand this labeled set via data augmentation. We compute the relative importance of the predictors, and compare performance in matching the human expert rankings as we vary the number of predictors used in training. We reach multiclass AUC measures between 0.64 and 0.94, depending on the number of deployed predictors. This performance is compared with the simple formulas that are currently used by some instructors for estimating the amount of credit to apportion for forum activity in their classes. To test generality and scalability we trained a classifier on the archive of the Web economics StackExchange reputation data. We used this classifier to predict the quartile assignments by human judges of forum posts from a university artificial intelligence course. Our first attempt reaches an average AUC of 0.69.

Keywords
Forum, MOOCs, residential courses, FAQ, online learning

1. INTRODUCTION
Massively Open Online Courses (MOOCs) have in past years, provided courses to populations outside traditional venues of higher education. For these settings, online forum facilities, that are built into the course delivery platforms, such as Coursera and Open edX are the primary means of communication among learning peers, and for interacting with instructors.

Beyond the practical needs for coordinating logistics in geographically distributed settings, online discussion forums can serve pedagogical goals as well. The process of discussions is a critical dimension of the learning process. Online asynchronous discussion forums provide the basis for collaborative learning. Students answering each others’ questions can be helpful for all parties [11]. This support function is particularly useful in Science and Engineering courses. But as discussion centric humanities courses embrace distance learning, discussions on forums will likely gain even more prominence.

However, it is not only in the context of distance learning that forum facilities have found uses. Even when in person class time is available, many residential college courses have adopted the tool. The need for students to ask questions, voice concerns, or to point out errors in course material are as salient in residential settings as they are in less traditional situations, such as distance learning [4]. Figure 1 shows the rapid increase in contributions to just one of the several available online forum tools, over the past years, in a large private university.

Given the importance of the discussion in the learning process at both the theoretical and empirical level, instructors in many Universities are assigning between 1% and 25% of their course grading component to online forum contribution. Despite office hours held by numerous course assistants, students helping each other online is helpful. Not only does student activity make more answers available for future use, but responses are provided very quickly in many cases, as students in residential settings follow similar working hours.

Three challenges arise when apportioning course credit to reward student’s forum contribution. First, students can attempt to game the system. On surveying the instructors, we learnt about instances of students copying a peer’s forum posts, adding spaces or other innocuous characters to fool automated contribution counters.

A second, more complicated problem is that of apportioning fair credit to students at the end of the quarter. Contributions take many forms. Asking an insightful or intriguing question can contribute as much to the course as providing answers. Taking the time to upvote another student’s effort is a contribution as well. Yet for courses with hundreds of students, manual assessment is not feasible. Instructors instead develop ad-hoc formulas over the statistics of their
forums, hoping to capture the right signals. This practice leads to non uniform grading across courses that is based on diverging intuitions.

The third challenge is not so much a hindrance as a missed opportunity. As courses are offered repeatedly over the years, a treasure of course knowledge accumulates in forum archives. Many times this information moves offline at the end of each term, and is therefore unavailable for consultation by instructors or future students. While it is likely of value to have future students discover their own questions, the availability of previous relevant questions could help in many circumstances. It can help the instructors by providing them a snapshot of previous forum activity, which can draw their attention to topics that students struggled most with. Thus, the instructors can tailor their lectures accordingly. For the students, it can serve as a ready-reckoner, that can improve their productivity significantly, if they do not need to wait for the instructors’ response indefinitely. Interfaces for exploring such archives effectively must also be developed.

In an effort to address these three challenges we are developing a coherent system for supporting the use of forums. Figure 2 shows a block diagram of our proposed system. In this paper, we focus on QuanTyler, the module responsible for helping with forum credit apportioning. This component is highlighted in the Figure. The plan is to have its operation be customizable. For example, instructors will be able to decide the granularity of partitioning the class into their quantiles of choice.

At the heart of our contribution are three experiments whose outcomes are required to inform the development of the QuanTyler module. In a first step we describe how we established ground truth of what a ‘good’ and credit worthy forum contribution looks like. Any automated algorithm requires ground truth for measuring success, and for training the models.

Second, we explore in the first of three experiments how contributions to a forum can be classified into quantiles, such that the implied ranking matches ground truth. We show the hyperparameters needed to make our chosen classifier technology—Random Forest—work well in support of the post evaluation task. We reached a near perfect AUC measure in this task.

However, the expense of obtaining ground truth limits our ability to obtain human judgments of forum posts at scale. At the same time, this requirement for human judgment limits us to examining just one course in depth at a time. To break out of this confinement, we examine how a much larger source of labels for a forum-like enterprise might be used for training, and to test generalizability.

To this purpose we used Stack Exchange, [2] which is an online Q & A platform. Constructive activity on the sites earns users reputation, which we can use as a surrogate for contribution quality. The underlying measures for computing reputation are similar to the measures available in academic forum facilities: number of questions asked and answered, number of upvotes, among others. The two facilities are thus quite analogous. But Stack Exchange has amassed more publicly available, high quality data than is available for academic course forums.

Stack Exchange is partitioned into sites for varying disciplines. We chose the Economics archive [1] from the sites and used it as a source for attempting transfer learning. In a second experiment we trained a random forest model on Stack Exchange reputation data, and tried predicting on the set of human rated forum posts in an Artificial Intelligence (AI) class. While not as good a classifier as the one trained on the forum data itself, our first attempt at transfer learning reached an $AUC = 0.69$, which we hope to improve further, going forward. We will show that data from Stack Exchange cannot be used in its raw form, and will explain what needs to be done to make the data suitable.

As a final, third experiment we will demonstrate that (at least one of) the ad hoc formulas currently deployed at our university does not at all reflect our human experts’ judg-
2. RELATED WORK

Online discussion forums empower students and instructors to engage one another in ways that promote critical thinking, collaborative problem solving and knowledge construction [16] [13]. Research has shown that linking some form of assessment to forum participation is a crucial element in promoting and enhancing online interactivity. [12] [22].

Quantitative methods for content analysis are most widely used in assessing effective forum participation. [7] presents an overview of 15 different content analysis instruments used in the CSCL studies.

One of the most cited works, used as a starting point in many Computer Supported Collaborative Learning (CSCL) studies is the model proposed by [9]. In [9], the author presented a framework and analytical model to evaluate computer-mediated communication (CMC). The analytical model was developed to highlight five key dimensions of the learning process exteriorized in messages: participation, interaction, social, cognitive and metacognitive dimensions. Although this model provides an initial framework for coding CMC discussions, it lacks detailed criteria for systematic and robust classification of electronic discourse [10].

Many researchers have strongly endorsed Social Network Analysis as a key technique in assessing the effectiveness of forum interactions [6] [23]. Social Network Analysis is a research methodology that seeks to identify underlying patterns of social relations based on the way actors are connected with each other [20] [18].

In [14], the authors discuss a conceptual framework for assessing quality in online discussion forums. Drawing from the previous works of [9], [15], and [8], the authors propose a framework and analytical model to evaluate computer-mediated communication (CMC). The analytical model was developed to highlight five key dimensions of the learning process exteriorized in messages: participation, interaction, social, cognitive and metacognitive dimensions. Although this model provides an initial framework for coding CMC discussions, it lacks detailed criteria for systematic and robust classification of electronic discourse [10].

In [19], the author proposes the use of the following metrics to assess forum participation: initiative, effectiveness-depth, effectiveness-breadth, value, timeliness, participation, scholarship, style, instructor points. Our system directly or indirectly includes most of these measures and also adds the crucial element of social network analysis to assess forum participation.

In contrast with both the aforementioned works, each of which focus on certain specific aspects for assessing forum participation, our approach for assessing a student’s contribution uses a combination of qualitative measures, quantitative measures, engagement level measures and also measures from social network analysis, in order to provide a holistic view of the student’s contribution. We develop a system that the instructors can configure and use easily for assessing the forum participation for their class.

3. CURRENT PRACTICE

On surveying several instructors who use forum participation in their grading scheme, the authors found that most of the instructors rely on the basic quantitative statistics that the forum machinery readily offers. The following were some of the grading schemes that are currently used by instructors at the University for awarding forum participation grades:

\textbf{Scheme 1:} In this scheme, scores of each student were calculated using the following formula:

\[ \text{Score} = 1 \ast \left( \frac{\text{numberofquestionsasked}}{\text{numberofquestionsasked}} \right) + 4 \ast \left( \frac{\text{numberofquestionsanswered}}{\text{othercontributions}} \right) + 0.5 \ast \left( \text{othercontributions} \right) \] (1)

\textbf{Scheme 2:} In this scheme, scores of each student were calculated using the following formula:

\[ \text{Score} = 3 \ast \left( \frac{\text{numberofquestionsanswered}}{\text{numberoffollowups}} \right) + 1 \ast \left( \text{numberoffollowups} \right) \] (2)

Anyone above the 90th percentile received full credit, and all the other students received a score of 0.

\textbf{Scheme 3:} Award full credit if at least one forum contribution was made, and the student was online on the forum for at least x number of days, or viewed at least y posts. Here x and y were set by the instructor using intuition.

\textbf{Scheme 4:} Award full credit to a student if he/she made at least one contribution to the forum.

All the above methods rely solely on the basic statistics directly provided by Piazza[21]. Figure 3 shows a snapshot of a sample class on Piazza. The concern, however, is whether these methods accurately and meaningfully award credit to the deserving students. Broadly, following are two major limitations of using the current grading schemes:

- **Lack of qualitative measures:** All the schemes above overlook the quality of contributions. This exclusion negatively impacts the grades of the students who post few,
but very high quality contributions. More importantly, relying solely on the quantity of the contributions encourages posts that do not constitute meaningful forum participation. This behavior, in turn, can cause the forum’s quality to devolve.

- **Reward not proportionate to effort**: Most of these schemes do not award credit proportionate to the amount of effort and time the student invested. For instance, using the third and fourth scheme meant that two students with vastly varying quantity and quality of contributions would be awarded the same score. Concretely, let us consider two students A and B. Student A made only one forum contribution during the course by posting a “+1” to another student’s question. However, student B regularly made meaningful forum contributions throughout the quarter. Using grading scheme 4, both would receive equal credit. This lack of fairness can deter students from engaging in meaningful forum contributions.

Despite the above limitations, the instructors have no choice but to rely on grading schemes like the ones discussed above. The large volumes of forum posts that accumulate by the end of the term make it impossible for the course staff to manually go through them to apportion credit. However, even if hypothetically, one were to have the course staff manually go through each of the contributions, there is a significant amount of subjectivity in assessing forum contributions. Having TAs manually grade contributions would lead to a lack of grading uniformity. A trivial contribution according to one TA, might be a significant contribution to another. Thus, there is a need for an automated way to assess the forum participation of students using a holistic grading scheme. Automation can lead to a standardized approach across the entire class.

The next sections discuss how the authors developed a system to assess forum participation by each student in order to award them fair credit. We go beyond the ready at hand statistics, provided by the forum, and incorporate measures that provide insight into the dynamics of student’s interaction in the forum. These dynamics manifest in the social networks that are created by the online interactions. We briefly introduce the concepts, because they contribute to our set of predictors.

4. POTENTIAL PREDICTORS

Our goal being the ranking of forum contributions, we assembled a number of predictors, some of which are readily available from online forum facilities. Others reflect student behaviors computed from interactions. We next describe each predictor that was entered into consideration for the final system. The sources of the corresponding measures are the data sets generated by forum facility software during the length of an academic term. Each class generates a separate data set, such as the one we used from the AI class.

**Quantitative measures**: These measures take into account the volume of contributions made by an individual. Hypothetically, if all the contributions are assumed to be of identical quality, the time and effort invested by a student is directly proportional to the volume of contributions made by him/her. Hence, the forum participation score should be directly proportional to the number of forum contributions made by the student. These four features are: number of questions asked, number of questions answered, total number of contributions, and average post length by a student.

**Engagement level measures**: These measures are used to reward the passive engagement of the students, given that not everyone in the class might be comfortable actively posting on the forum. Some of the students are great listeners; they view or follow most of the posts and are regularly online on the forum. The two features used to apportion credit for the passive interaction are: total number of days a student was online on the forum, and the number of posts viewed by the student. In order to reward the students, who started important threads, which in turn, engaged many students, a third metric, average number of collaborators in the threads started by the student was added. A fourth metric, average number of views received by the questions posted by the student was added for similar reasons.

**Qualitative measures**: These measures are used to reward the students based on the quality of their contributions. These measures are the upvotes and endorsement counts available in forum datasets. Students can express appreciation for a post by adding an upvote to the contribution. Instructors can explicitly endorse answers provided by students, marking those answers as definitive. Upvotes and endorsements articulate human judgments, and can be thought of as crowdsourcing post quality assessment.

Another strength of qualitative measures is their robustness to student cheating by flooding the forum with meaningless threads to increase their contribution count. Our two qualitative features are: number of answers by the student that were endorsed, and total number of endorsements, including upvotes on the questions, answers and instructors’ endorsements.

**Social Network Analysis**: As discussed in the Related work section, Social Network Analysis (SNA) provides key insights to the student forum participation. A brief detour in...
the following section provides background for the measures we used for SNA.

In order to include the Social Network Analysis component in the credit apportioning system, the following networks were extracted from the class forum dataset. In the definitions below, nodes represent students and instructors. Typed edges represent interactions that are possible in the forum. Link weights encode the number of such interactions between the link’s nodes.

**Upvotes network:** An upvotes network is extracted, where an edge from student A to student B indicates that A upvoted B’s content at least once, and the weight of the edge encodes the number of times A upvoted B’s content.

**Endorsement network:** An endorsement network is extracted, where an edge from instructor A to student B indicates that A endorsed B’s content at least once, and the weight of the edge encodes the number of times A endorsed B’s content.

**Combined upvotes and endorsement network:** This is a union of the above two networks. An edge from A to B indicates that A either upvoted and/or endorsed B’s content at least once, and the weight of the edge encodes the sum of the upvotes and endorsements.

**Interaction network:** This graph models the interaction that happened on the forum over the duration of the course. In the interaction network, an edge from A to B indicates that B responded at least once to a question that A posted.

We use these networks for our final two measures: *degree centrality* in the interaction network, and *page rank* in the combined upvotes and endorsement network.

We calculate the degree centrality for every node in the interaction network. Degree centrality measures the number of links incident upon a node. Higher degree centrality of a student implies that the student answered questions or resolved doubts for a large number of students. On a high level, degree centrality in the interaction network translates to the "helpfulness" of the student.

Page rank in the combined upvotes and endorsement network was used in order to capture importance in both upvotes and endorsement information using a single metric. Page rank can additionally help uncover influential or important students in the network. Their reach extends beyond their immediate neighbors, and is therefore not captured by the earlier described upvote/endorsement measures. The higher the page rank in the combined network, the more "influential" the student.

### 5. GROUND TRUTH COLLECTION

In order to evaluate how effective each of the above predictors is in informing credit apportioning, human ground rule judgments are required. We obtained these by paying former students and teaching assistants of the AI or a related class to render judgments over a sample of posts. Given the high course enrollment of 700+, not all the posts could be evaluated. A survey instrument was used to collect the judgments, and participants were paid a $20 gift card. The number of posts sampled was limited by this cost, and time capacity of the 24 participants we could recruit.

Each item in the survey for the experts was a pair of two posts by different students. The experts were asked to indicate which of the two contributions was more helpful for the class as a whole. (See the precise instructions below). The task in preparing the survey was thus to find forum contribution pairs that would later help train an algorithm. The challenge was to select a set of posts that would cover a range of measures for all our candidate predictors, while being representative of the overall contributions. We describe here how this selection was accomplished.

Four algorithms were created using a weighted combination of the above explained candidate predictors.

- **Alg 1:** Using only qualitative measures and social network analysis measures.
- **Alg 2:** Using only quantitative measures and engagement level measures.
- **Alg 3:** Using all the measures but the weightage more skewed towards quantitative measures.
- **Alg 4:** Using all the measures but the weightage more skewed towards qualitative measures.

In addition, the current formula based grading scheme 1 that is used by some instructors at the University was also included as a variant:

\[
\text{Score} = 1 \cdot (\text{number of questions asked}) + 4 \cdot (\text{number of questions answered}) + 0.5 \cdot (\text{other contributions})
\]

Let us call this approach Alg 5. All five grading schemes were used to calculate each student’s score. Then, 10 new values are calculated, each of which are absolute values of ranking differences between one pair of rankings for the same student. Each algorithm pair is processed. Thus, we have Alg1vsAlg2, Alg1vsAlg3, Alg1vsAlg4, Alg1vsAlg5 and so on. For instance, if Student ID # 500 was ranked 30 by Alg 1, and 300 by Alg 3, then the Alg1vsAlg3 value for Student ID # 500 would be 270.

We then sort these ten rank differences in descending order. The top entry in the 10 columns gives us the students of interest. To sample student pairs, we compare these students of interest with the students immediately above and immediately below in the ranking by both the experiments under consideration. We clarify the procedure with the following example:

Let us assume that student ID #10 had the maximum absolute difference between ranking through Alg 1 and Alg 3. Also, using Alg 1, student ID #400 is directly above student ID #10 and student ID #5 is directly below student ID #10 in the ranking. Finally, let us assume that using Alg 3, student ID #20 is directly above student ID #10 and student ID #557 is directly below student ID #10 in the ranking.

Then the student ID pairs of interest that would be used are: (10, 400), (10, 5), (10, 20), (10, 557).
In addition to 40 such pairs of interest, additional student contributions were sampled at random and included in the survey. At most 4 question pairs and 4 answer pairs were sampled from all these selected student pairs and presented to the experts. A total of 89 question pairs or answer pairs were sampled from all these selected student pairs and presented to the experts. In order to avoid fatiguing the experts, the set was partitioned into two batches such that each question pair or answer pair was voted on by at least 12 experts.

The survey instructions were as follows:

Each of the following sections presents one pair of questions or answers that were posted to the course forum in the past. We ask that you indicate for each pair the contribution that might have been most helpful to the rest of the class.

One sample item from the survey is as follows:

Q1: I am very confused about alpha-beta pruning, as we do not have example code from lecture. When we say we prune certain leaf, what does it mean? Does it mean we do not store that choices?
Q2: To create our own label, must it be a binary label 1,-1? or it can be multi-categories with labels of any number? Is the feature still word counts or can be anything?

Which of the above two questions contributes more to the class community?

In order to learn the experts’ intuition about which predictors might be important in ranking student forum contributions, the following related question was introduced in the ranking survey. The question was presented to the survey takers once, at a random time in their processing the survey, such that they could drag the entries up and down to arrange the predictors in decreasing order of relevance.

Imagine you had the following statistics about forum contributions by all students at the end of the term. In your opinion, which statistics are important to evaluate the forum contributions of students to the class. Please drag the entries up and down to indicate their relative importance. The first entry would be the most important.

- Number of questions asked by the student
- Number of questions answered by the student
- Total number of posts viewed by the student
- Total number of days the student was online on the forum
- Total number of endorsements received by the student
- Total number of Forum contributions by the student (including questions, answers, notes, follow-ups, etc.)

Based on the majority vote for every rank, we arrive at a ranking order using the experts’ intuition. The results are summarized in Table 1. Rank 1 is the most important feature. The percentage of experts agreeing with each ranking is also included.

We compare these intuitions with our computationally determined predictor importance later on.

6. EXPERIMENT PREPARATION

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>%support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td># of endorsements</td>
<td>60.7</td>
</tr>
<tr>
<td>2</td>
<td># of questions answered</td>
<td>57.1</td>
</tr>
<tr>
<td>3</td>
<td># of Forum contributions</td>
<td>46.4</td>
</tr>
<tr>
<td>4</td>
<td># of questions asked</td>
<td>46.4</td>
</tr>
<tr>
<td>5</td>
<td># of posts viewed</td>
<td>60.7</td>
</tr>
<tr>
<td>6</td>
<td># of days online</td>
<td>64.2</td>
</tr>
</tbody>
</table>

Figure 4: Distribution of predictor measures from 37 students.

Given the pairwise rankings of posts by the experts we needed to arrive at a ranking against which we could then train and test. We generated this ranking using the Copeland method [3]. The procedure counts the number times a student’s post was considered superior to the alternative post offered to an expert. The number of losses are then subtracted from these wins.

Copeland ranking ties can be broken by a second order Copeland approach [5]. However, we found that forcing a complete order did not lead to good classification, because the ties are a reflection of true similarity.

We included a number of posts from each sampled student in the survey. The final result was a list of 37 students for which we had rankings from twelve experts each. We collected this large number of rankings for each student because of the above mentioned subjectivity in evaluating posts. In addition to the rankings, we had for each of the 37 students the measures for all our 12 candidate predictors. Figure 4 shows the distribution of the measures after centering and scaling.

We examined the outliers and found that they are legitimate data points. This finding supports [11], which detected ‘super posters’ in MOOC forum activities.
Table 2: Accuracy and Kappa by number of predictors per tree

<table>
<thead>
<tr>
<th>mtry</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>7</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>12</td>
<td>0.88</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Rather than attempting a regression, we formulated the problem as one of classification into four classes: the rank quartiles. This decision was based on the application of apportioning credit. A granularity of four suffices, given that forum participation is not the only source of credit for a course. Partitioning a 10% course credit into 700 values is not meaningful.

Accordingly, we partitioned the ranked list of students into four roughly equal parts. The four classes could not be filled equally because of the ranking ties. Tied students should be in the same class, rather than being split across quartile boundaries. When such splitting occurred we moved all participants into one of the classes, such that the fewest moves were required. For example, if three of five students with rank seven were assigned to quartile two, and two were assigned to quartile three, all students ranked seven were moved to quartile two.

Given the sparsity of our human labeled set, we next augmented the 37 sets of data points as follows. We can think of the data as a table in which each row holds the data for one student, and each column holds the measurements of one predictor. The rows are sorted such that the students in the same quartile are contiguous. Data augmentation added rows to each quartile. Within each quartile $Q$, and for each predictor $P$ in $Q$’s rows we generated new values for $P$ within the range of $P$’s measured values. We thereby added 300 rows to each quartile.

Finally, we set aside 30% of the resulting set for testing. We call these sets forumTrainAug and forumTrainAug. The corresponding putative responses are forumTrainResp and forumTestResp. Our first exploration was to see whether we could construct a classifier that would use predictor measures to assign each student to one of the quartiles.

![Error vs. number of trees for model rf10K](image)

Figure 5: Classification errors by number of trees.

Table 3: Model RF8K: 300 samples augmented in each quartile

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Class</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>219</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>227</td>
<td>1</td>
<td>5</td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>223</td>
<td>1</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>0</td>
<td>205</td>
<td></td>
<td>0.10</td>
</tr>
</tbody>
</table>

The resulting model rf8K, trained on forumTrainAug with 10-fold cross validation repeated 3 times has the following characteristics:

Figure 6 shows the relative importance of our candidate predictors. The chart shows the amount of decrease in accuracy that is contributed by each of the predictors. The top three predictors are the number of student answers that were endorsed by an instructor, the total number of endorsements, and the days the student has spent online. Note that these predictors differ somewhat from those intuited by the experts, though there is overlap.

We experimented using three predictors only, but the degradation was noticeable. It is also advantageous to retain predictors that are less easy to spam than time online. For instance, the page rank predictor, while less important for

Table 4: Model RF8K: 100 samples augmented in each quartile

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Class</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>76</td>
<td>3</td>
<td>5</td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>75</td>
<td>1</td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>0</td>
<td>11</td>
<td>2</td>
<td>66</td>
<td></td>
<td>0.16</td>
</tr>
</tbody>
</table>
the classification, is more difficult to defraud.

Using rf8K we predicted forumTestResp. Figure 7 shows ROC curves for each quartile predictor.

The AUC values are near perfect with a mean of 0.99. This result is encouraging in that it signals inroads towards apportioning fair forum participation credit even for very large courses.

However, the result does not speak to generality. The model was trained on a science forum data set, and its human labels were few. We therefore added a second experiment to demonstrate how the approach behaves when training occurs on data of an unrelated domain, and the resulting classifier is then used to predict forum participation levels.

8. EXPERIMENT 2: STACK EXCHANGE TRANSFER LEARNING

We obtained the Stack Exchange (SE) records for the site dedicated to Economics [2]. The measures obtainable from this data set are very similar to those from typical forum facilities. However, the facility includes a natural quality target: user reputation. Stack Exchange contributors earn reputation by competently answering questions, and other constructive behaviors. That measure is therefore analogous to the quality ratings we needed to obtain for Experiment 1 from expensive, and limited human labor.

We began with the data from about 5300 SE contributors. In a first step we followed the same procedure as in Experiment 1 to obtain optimal mtry and forest size values, which were 2 and 4K respectively. After scaling, centering, and partitioning into quartiles we set aside a 30% test set (seTest) from the training set (seTrain). The respective reputation responses are seTrainResp, and seTestResp.

Since the forum training set was not involved in the SE training, we used the larger forumTrainResp as test target for the SE-trained forest. Figure 8 shows the problematic resulting AUC ROC curves. We addressed the lower triangle Q3 curve by reversing that classifier’s orientation. This step is an appropriate measure, because the curve lies consistently below the diagonal, indicating a true polarity issue. However, AUC values were low, and further investigation uncovered the reason (Figure 9). The Figure shows that quartile 2 is over-represented, while quartile 3 suffers from a scarcity of examples. We balanced the training set by sub-sampling the quartile 2 examples to 1200, and augmented quartile 3 examples analogously to our process in Experiment 1.

The resulting 4K tree model, again trained with 10-fold cross validation repeated three times yielded a training accuracy of 0.72, and a kappa of 0.63. Table 5 shows the model’s confusion matrix. When predicting seTest with this SE-trained classifier, a satisfactory mean AUC of 0.93 resulted, with classification behaviors shown in Figure 10.

Finally, with the SE model reasonably solid, we used this model to once again predict both forumTrainResp and forumTestResp. Table 6 shows results.

An important question remains: how do the ad hoc formulas devised by instructors perform? Are they sufficient? A
10. DISCUSSION
The AUC of 0.69 when using the Stack Exchange trained classifier on forum posts lags behind the classifier that is specialized on forum post evaluation. However, as a first step this result is encouraging. Forum assessment is gaining enough importance, and human judgments are expensive enough that training data from large, ready at hand, and similar enough facilities is extremely attractive for attempts in transfer learning.

Stack Exchange and maybe other reputation incentivized systems have accumulated enough labeled examples that alternatives to random forests, such as neural nets, which require large amounts of training data should be feasible as tools going forward.

11. CONCLUSION
Forum assessment is an active research area for good reason. A growing number of schools and companies are offering entire degree programs online. All require online commu-
cation among students and instructors. Demand for tools that help manage and assess forum activity is likely to rise as online education continues to capture market share.

The work described here has not yet added NLP to the mix of technologies that may prove helpful in the context of online forum management. Deploying natural language processing to enable topic browsing and other facilities is on our road map, as well as on that of other researchers.

As explained in the introduction, this work is part of a larger effort that fills modules into a forum centered architecture. The frequently asked questions module, as well as the spam detection facility will round out our efforts going forward.

12. REFERENCES