Students, Teachers, Exams and MOOCs: Predicting and Optimizing Attainment in Web-Based Education Using a Probabilistic Graphical Model

Bar Shalem[1], Yoram Bachrach[2], John Guiver[2], and Christopher M. Bishop[2]

[1] Bar-Ilan University, Ramat Gan, Israel
[2] Microsoft Research, Cambridge, UK

Abstract. We propose a probabilistic graphical model for predicting student attainment in web-based education. We empirically evaluate our model on a crowdsourced dataset with students and teachers; Teachers prepared lessons on various topics. Students read lessons by various teachers and then solved a multiple choice exam. Our model gets input data regarding past interactions between students and teachers and past student attainment. It then estimates abilities of students, competence of teachers and difficulty of questions, and predicts future student outcomes. We show that our model's predictions are more accurate than heuristic approaches. We also show how demographic profiles and personality traits correlate with student performance in this task. Finally, given a limited pool of teachers, we propose an approach for using information from our model to maximize the number of students passing an exam of a given difficulty, by optimally assigning teachers to students. We evaluate the potential impact of our optimization approach using a simulation based on our dataset, showing an improvement in the overall performance.

1 Introduction

Recent years have marked an enormous leap in the use of the Internet and web-based technology. This technology had a huge impact on education, where web-based and online training are emerging as a new paradigm in learning [26]. Distant learning technology makes it easier to access educational resources, reduces costs and allows extending participation in education [28,2,40]. Intelligent online educational technologies enable a deep analysis of student solutions and allows automatic tailoring of content or the difficulty of exercises to the specific student [11]. One innovation that could affect higher education is massive open online courses (MOOCs), online training geared to allow large-scale participation by providing open access to resources [36,16]. MOOC providers offer a wide selection of courses, some already attracting many students. [1]

[1] See, for example the report on Peter Norvig and Sebastian Thrun's online artificial intelligence course, with its "100,000 student classroom", in `http://www.ted.com/talks/peter_norvig_the_100_000_student_classroom.html`.

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 82–97, 2014.

However, web-based education also brings with it new challenges. Students may become frustrated due to ambiguous instructions or lack of prompt feedback [25]. This triggers the need to manage the quality of online teaching material, and highlights the need for an objective system for measuring performance and for efficient resource allocation [10,43,47,36].

However, measuring the quality of teaching materials or predicting the attainment of students are challenging. Teachers who teach a similar subject are likely to have completely disjoint student cohorts, of different ability levels, backgrounds, and demographic traits. Further, students may solve different tasks or get different exams (with potentially some overlap in tasks or questions).

Many questions arise in such settings. How can we aggregate observations on outcomes in order to evaluate the abilities of students, the competence of teachers and the difficulty of exams? Can we systematically predict the attainment of students? Do demographic and personality traits correlate with performance? How can we optimize resource allocation, such as the assignment of teachers to students, so as to maximize performance?

Our Contribution: We propose a probabilistic graphical model for assessing teaching material quality and student ability, and for predicting student attainment in online education. Our model gets input data regarding past interactions between students, teachers and exams and past outcomes (whether a student succeeded in answering questions in the exam), and provides predictions regarding future interactions. We evaluate our model based on a dataset crowdsourced from Amazon's Mechanical Turk (AMT). We divided the AMT workers into "teachers" and "students". Each teacher prepared "lessons" on various topics, in the form of summaries of Wikipedia articles. For each topic we constructed a multiple choice "exam", and students were asked to solve it based on the lesson prepared by one of the teachers. We show that our model can predict outcomes in such settings and estimate the abilities of students, the competence of teachers and the difficulty of questions. We show that our model outperforms heuristic approaches for predicting outcomes. We also explore how demographic profiles and personality traits correlate with student performance in this task. Finally, given a limited pool of teachers, we propose an approach for using information from our model to optimize performance in our domain, such as the number of students passing a difficult exam. We do so by choosing the optimal assignment between teachers and students, based on our model's estimates, and evaluate the potential impact of this approach using a simulation based on our dataset.

2 Probabilistic Graphical Model for Predicting Attainment in Web-Based Education

We now describe our model for predicting performance in web-based education. Our domain consists of online *exams* given to *students* who studied various *topics* with the help of *lessons* prepared by *teachers*. We denote the student set as S, the teacher set as T, the topic set as M, and the set of questions comprising the exam on topic m as Q_m. We denote the exam on topic m as E_m. A student

$s \in S$ learns topic $m \in M$ based on the lesson prepared by teacher $t \in T$, then answers the exam E_m on topic $m \in M$. We say the *outcome* for this attempt was a success, denoted $r_{s,q} = 1$, if student s answers question q correctly, and otherwise we say it is a failure, denoted $r_{s,q} = 0$. The *raw score* of student s in the exam E_m is the number of questions she answers correctly. This raw score reflects not only the ability of the student, but also how well she was taught the topic by her teacher, and the difficulty level of the questions in the exam. Thus our dataset consists of observations of the form $z_i = (s, t, m, q, r_{s,q})$. Every student is taught each topic by a single teacher (though she may receive a different teacher for different topics).

Given our observations $Z = \{z_i\}_{i=1}^{w}$, we wish to predict future outcomes: how well is a student s likely to do in an exam E_m on topic m when she is taught by teacher t? We refer to our problem as the *attainment problem*. The full input data to the attainment problem potentially includes an entry for the outcome on every question for every student, so its size is $|S| \cdot |Q|$. Typically, however, the input data only includes a smaller set of observations: for example, a student may only have been taught some of the topics, or was only tested using some of the questions on a topic. Given the input data, our goal is to predict the outcomes on the missing entries, so a *query* is a tuple $u_j = (s, t, m, q)$. A query is similar to the input entries, except it is missing the outcome r, to be interpreted as requesting the model to predict whether student s would answer the question q regarding the topic m correctly when taught by teacher t.

Predicting Outcomes Using a Probabilistic Model: We propose a probabilistic graphical model for the attainment problem, called the Student-Teacher-Exam-Performance model — **STEP**. Given the input observations $Z = \{z_i\}_{i=1}^{w}$ and queries $U = \{u_j\}_{j=1}^{l}$, the model's output consists of predictions regarding the outcomes for the entries in the query set $R = (r_1, \ldots, r_l)$. STEP also outputs information regarding latent variables, such as the ability level of each student, the competence of each teacher and difficulty of each question. The outcomes in the query set U, as well as the abilities, competences and difficulties, are modeled as *unobserved random variables*. In contrast, the outcomes in the observation set Z are *observed variables*. The structure of our STEP model is governed by independence assumptions regarding the variables. Pearl discusses Bayesian Networks [42] (now referred to as directed graphical models), which represent conditional independence assumptions as a graph where each vertex corresponds to a variable and the edges capture dependencies between adjacent variables. We base STEP on a prominent extension of Bayesian Networks, called Factor Graphs (see [29]), which describes a factorial decomposition of an assumed joint probability distribution between the variables.

We first define the crux of the model in the form of a Factor Graph representation. We then set the observed variables in the graph to the values of the observations Z, consisting of the identities of the students, teachers, topics and questions, and most importantly the outcomes in our observation set. We then use approximate message passing algorithms [29] to infer marginal probability distributions of the target unknown variables: student abilities, teacher

competences, question difficulties, and of course the unobserved outcomes of the query set. We thus get a posterior distribution over these unobserved variables.

The Graphical Model: Recall that the variable $r_{s,q}$ indicates whether student s answered question q correctly ($r_{s,q} = 1$ indicates the answer was correct and $r_{s,q} = 0$ indicates it was incorrect). This variable is an observed variable for every entry $z_i = (s, t, m, q, r) \in Z$ (though it is unobserved in the query set U). We model the process which causes a student $s \in S$ to either answer a question correctly or incorrectly. We assume every student $s \in S$ has an inherent *ability* $a_s \in \mathbb{R}$ reflecting how easy she finds it to learn new topics and answer questions on them, and that every teacher $t \in T$ has an inherent *competence* $c_t \in \mathbb{R}$ reflecting her ability to teach students and provide them information on a topic. We assume every question $q \in Q$ has an inherent *difficulty* $d_q \in \mathbb{R}$ determining how likely it is that a student could answer it correctly.

Our model is a joint probabilistic model with a factor graph representation given in Figure 1. The model has two parts. The first part reflects the probability that student s actually knows the correct answer to a question q, denoted by the variable $k_{s,q}$, as determined by the student ability parameter a_s, the teacher competence parameter c_t (where t is the teacher who taught s the topic of that question), and the question difficulty parameter d_q. In Figure 1, this is shown to the left and above the vertex of $k_{s,q}$. The second part determines the observed outcome, depending on $k_{s,q}$ and is shown to the right of the vertex of $k_{s,q}$.

$k_{s,q}$ is a Boolean variable. A value of 1 indicates that the student s knows the correct answer to the question q, while a value of 0 indicates she does not know the answer (but may still give the right answer to the question by making

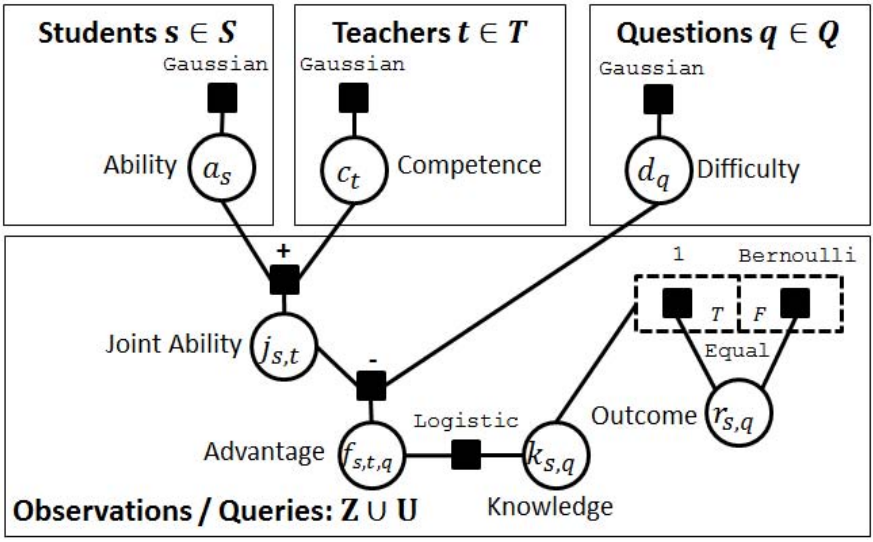

Fig. 1. Factor graph for the STEP model

a lucky guess). The probability of $k_{s,q}$ having the value 1 increases with the student ability and teacher competence and decreases with the difficulty of the question. By $f_{s,t,q,}$ we denote the difference between the "total joint ability" of the student and the teacher $(a_s + c_t)$ and the difficulty of the question (d_q), so $f_{s,t,q} = (a_s + c_t) - d_q$. [2] The variable $f_{s,t,q}$ reflects the "advantage" the student has over the question after she is taught the relevant topic by the teacher.

We assume that $k_{s,q}$ depends on the advantage $f_{s,t,q}$ as follows:

$$P(k_{s,q} = 1 | f_{s,t,q}, \tau_q) := \int_{x=-\infty}^{x=\infty} \phi(\sqrt{\tau_q}(x - f_{s,t,q}))\theta(x)\, dx$$
$$= \Phi\left(\sqrt{\tau_q} f_{s,t,q}\right) . \quad (1)$$

Where ϕ is the Gaussian density: $\phi(x) := \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$, Φ is the sigmoidal cumulative Gaussian distribution: $\Phi(t) := \int_{x=-\infty}^{t} \phi(x)\, dx$, and $\theta(\cdot)$ is the Heaviside step function. The integral presentation allows for the following interpretation of this probability: this is a binary process which results from evaluating the step function θ over a variable f which is added a Gaussian noise of variance $\frac{1}{\tau}$. Another way to view this is that the data is assumed to come from a probabilistic generative process: the student's ability, teacher's competence and question's difficulty are sampled from random Gaussian distributions which reflect the distribution of those properties in the population. A random "performance noise" for each entry in the observation set Z, which may be either positive or negative, is added to the total joint ability (the sum of the student ability and teacher competence); If this number is greater than the difficulty of the question, then the student knows the correct answer so $k_{s,q} = 1$, otherwise $k_{s,q} = 0$.

The outcome variable $r_{s,q}$ is a mixture of two distributions. If student s knows the answer to question q, i.e. $k_{s,q} = 1$, she answers correctly with probability 1, so $r_{s,q}$ is constrained to be a point-mass distribution. If she does not know the correct answer, i.e. $k_{s,q} = 0$, we assume s guesses an answer uniformly at random; Question q is a multiple choice question with b possible answers, so the outcome variable $r_{s,q}$ is assumed to have a Bernoulli distribution, with success probability $\frac{1}{b}$. The mixture is expressed in Figure 1 using a *gate*, marked by a dashed pair of boxes, that switch the factor connecting to $r_{s,q}$, depending on the state of the variable $k_{s,q}$. Gates were introduced in [38] as a powerful representation for mixture models in factor graphs. Such gates represent conditional independence relations based on the context of a "switching variable".

Probabilistic Inference: We now explain how to infer the outcomes in the query set and the unobserved variables. Given the data in the observation set $\mathbf{Z} = \{z_i = (s_i, t_i, m_i, q_i, r_i^Z)\}_{i=1}^{w}$ and the query set $\mathbf{U} = \{u_j = (s_j, t_j, m_j, q_j)\}_{j=1}^{l}$

[2] Other operators can be used to aggregate the student ability and teacher competence into the total joint ability. For example, a max operator $\max(a_s, c_t)$ can indicate that either a strong student or a competent teach allow the student to determine the correct answer, while a min operator $\min(a_s, c_t)$ can indicate that both a strong student and a competent teacher are required. The complexity of performing the inference in such alternative graphical models depends on the operator used.

we are interested in predicting the missing outcomes for the query set $\{r_j\}_{j=1}^{l}$. We do so by simultaneously inferring several approximate posterior (marginal) distributions: the Gaussian density of the ability of each student, competence of each teacher and difficulty of each question, the Bernoulli distribution indicating whether each student knew the answer to each question for all such entries in the observation set and query set, and the Bernoulli distribution indicating whether each student gave a correct response to the question asked for all entries in the query set. The posterior distributions $\{p(r_j|\mathbf{Z},\mathbf{U})\}_{j=1}^{l}$ can be interpreted as the probability that the outcome in the j'th entry in the query set would be a success (i.e. the probability that the student would answer the question correctly). This posterior distribution is a Bernoulli distribution, so we can simply denote the probability of a successful outcome as $p_{r_j} = p(r_j = 1|\mathbf{Z},\mathbf{U})$. When requested to make a binary prediction rather than estimate the probability of a successful outcome, we use the *mode* of that distribution: if $p_{r_j} > \frac{1}{2}$ we predict a success, and otherwise predict a failure.

To perform the inference and compute the posterior distribution in STEP, we use Expectation-Propagation approximate message passing (see [39,29]), using Infer.NET [37], a framework for probabilistic modeling. [3]

3 Model Evaluation

We evaluated STEP using a dataset crowdsourced from Amazon's Mechanical Turk (AMT). AMT is a crowdsourcing marketplace bringing together workers interested in performing jobs remotely, and requesters interested in obtaining human labor for tasks. We constructed tasks for a remote learning experience, both on the teacher's side and the student' side. We first selected 10 Wikipedia articles covering various topics such as Chad, Saffron and DNA. We composed an "exam" on each of those topics, consisting of 5 multiple choice questions (50 questions total). We divided the worker set to two groups: "teachers" and "students". Each teacher was required to write a short (1500 character) "lesson" on each of the topics. The teachers were notified which issues to focus on when preparing the students for the exam (for example the history of Chad or the chemical structure of DNA). However, they did *not* know which specific questions were in the exam. Each student was asked to study the topic using the lesson provided by a teacher we chose, then solve the exam on that topic. The time given to solve the exam was limited to 3 minutes per topic, making it difficult (though not impossible) for students to consult external resources other than the teacher's lesson.

Data Collection: Our dataset consists of observations regarding the questions solved by students, in the form discussed in the previous section: student, teacher,

[3] STEP's factor graph is loopy, as we have multiple participants who respond to the same question set and share the same teacher set. Thus EP computes the posteriors by iterating until convergence. The number of iterations used in Infer.NET is constant, so the procedure runs in time linear in the input, i.e. in $O(|S| \cdot |Q|)$.

topic, question, and correctness. We sourced 237 workers for the task from AMT. We used 10 of them as teachers, and 227 as students. Each teacher had prepared a lesson on each of the 10 topics. Lessons were allocated to students as follows. Each student got 10 lessons by 10 different teachers. For each student, the teacher permutation was modified by cyclic shift, i.e. student s got the lesson by teacher $(s+m) \mod |T|$ on topic m, where $|T|$ is the number of teachers. Each student answered all 5 questions on each topic resulting in a total of 11,350 entries in our dataset.

The students were given a base payment of $2 for performing the task, and a bonus of up to $3 depending on their performance, measured by the number of questions they answered correctly. The teachers received a base payment of $10, for writing the lessons, and engaged in a contest for an additional bonus of $10: each teacher was randomly paired up with another teacher; The teacher with better performing students was awarded a $10 bonus. [4] In addition to answering the questions, each student completed a demographics survey regarding their age, gender, income and education. They also completed a short personality questionnaire called TIPI [22]. TIPI follows the Five Factor Personality Model, [13,45], a generally accepted model representing the "basic structure" underlying human personality, whose ability to predict human behavior has been thoroughly investigated [15,9]. [5] The key five personality traits are Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN for short).

Model Performance: We examined the performance of our STEP model, evaluated by randomly partitioning the data into a training set and a test set. We compared our model to heuristic approaches using two error metrics. The first error metric is the *prediction error*, which is the mean absolute difference between the actual answers (0 for an incorrect answer and 1 for a correct answer) and the model estimated probability of a correct answer. The second metric is based on a binary outcome prediction. We round the estimated probabilities of answering a question correctly to get a binary classification. The *classification error* is the proportion of entries where the model mis-classified the outcome.

We compared the performance of STEP with two heuristics. Given a target student s, our *student heuristic* examines all the entries with that student in the training set, and measures the proportion of those where the outcome was a success (i.e. the proportion of the student's entries where she gave a correct answer). This proportion is then used as the estimated probability of a successful outcome on each of that student's entries in the test set. Similarly, given a teacher

[4] While there is a high variance in the performance of participants in AMT [30,5], such contests are known to have good properties in terms of incentivizing the participants to exert significant effort on the task [27,3,20,52] (so long as participants are anonymous and are not colluding [35])

[5] Further, it is possible to automatically infer personality traits from peoples' social network profiles [7,32,6] or website choices [33,31], allowing such publicly available information to be used to profile students and make predictions about their performance in educational settingts.

t our *teacher heuristic* examines all the entries with that teacher in the training set and measures the proportion of those entries where the outcome was a success (i.e. the proportion of the teacher's entries where her student gave the correct answer, no matter who that student was). This proportion is then used as the estimated probability of a successful outcome on each of that teacher's entries in the test set. The student heuristic ignores information regrading who the teacher was, and the teacher heuristic ignores information regarding who the student was, while STEP uses all the available full information.

Figure 2 compares the quality of our model with the student and teacher heuristics, in terms of the classification error and prediction error metrics. The x-axis in both plots is the number of observations available in the training set. For each point in the plot we randomly selected a subset of questions, whose size was determined by the location on the x-axis, and used their entries as the training set. The remaining entries were used as a test set, with an unobserved outcome. We repeated the sampling 500 times and averaged the resulting error metrics. Figure 2 shows that the STEP produces better predictions than the heuristics, as it has a lower error for both error metrics discussed above. For both the heuristics and STEP, the error decreases as more data is given as input, but the improvement diminishes in the size of the data.

In addition to the outcome predictions regarding queries in the test set, the STEP model also returns information regarding the abilities of students, competence of teachers and difficulty of questions, captured as posterior distribution for the model parameters. These parameters allow us to rank students, teachers and questions, by their abilities, competence levels and difficulties, correspondingly. The values of these parameters are shown in Figure 3.

Figure 3 indicates high variances of the parameters. STEP sums together student ability and teacher competence and compares the sum with the question difficulty. The variability on the y axis between student abilities is larger than the variability between teacher competences, indicating that the identity of the student had stronger impact on performance than the identity of the teacher.

One simple way to "score" student abilities is by the proportion of questions they answered correctly. Correspondingly, we can score teacher competence by the proportion of questions that their students answered correctly. Similarly, we can score question difficulty by the proportion of all students who managed to correctly answer that question (here a high score means an easy question). Unsurprisingly, there is strong positive correlation ($r = 0.997$) between a student overall score in the full exam and her inferred ability, and between the average score of a teacher's students and her inferred competence ($r = 0.999$). Similarly, there is a strong negative correlation ($r = -0.946$) between the proportion of students who managed to solve a question and its inferred difficulty.

Demographics and Student Success: STEP predicts student success based on observed outcomes in previous interactions. Other sources of information regarding a student, such as demographic traits or personality traits may also help predict student performance. Previous work has already examined the correlation between a student's demographic or personality traits and success in online

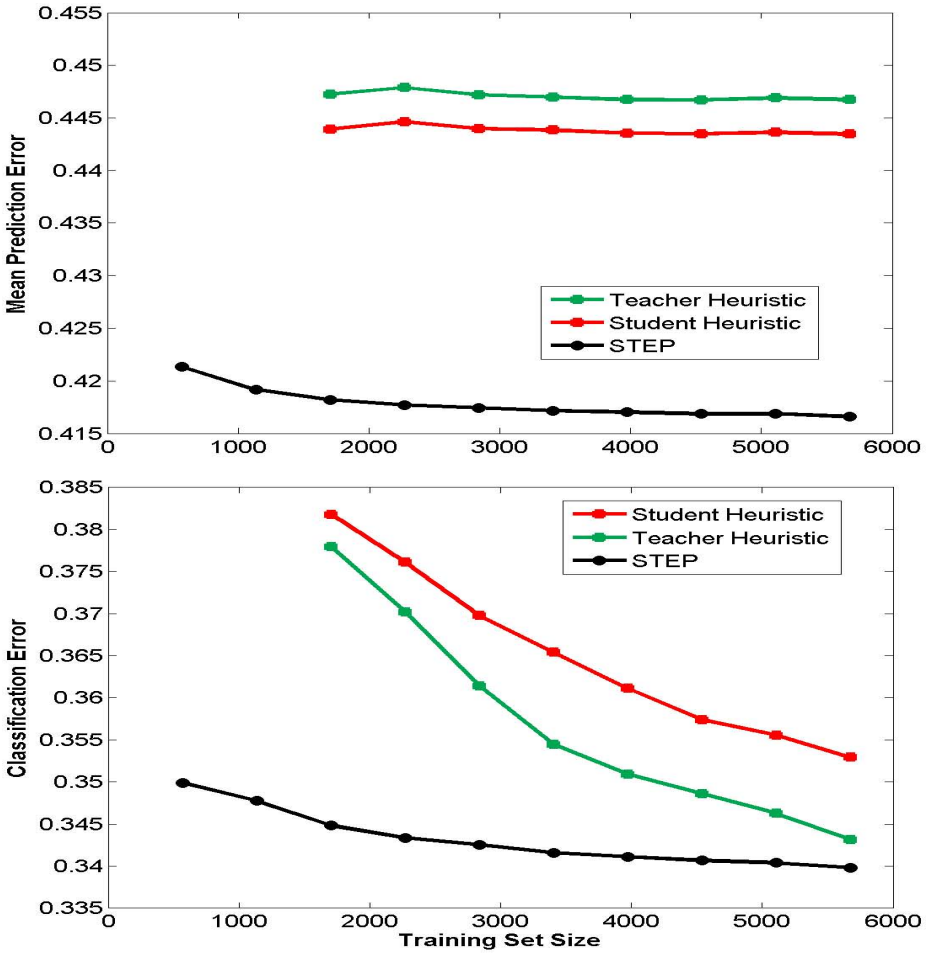

Fig. 2. Model quality - prediction and classification errors

tasks or in traditional educational settings [17,53,44,48,19]. We now examine such correlations in our web-based educational task. We measured a student's performance using the proportion of questions they answered correctly. We correlated this student performance score with other traits of the student, such as their age, level of education or personality. We found strong evidence for a positive correlation between a student's educational level and performance. We also found strong evidence of correlation between a student's personality and performance: both openness to experience and extroversion correlate positively with performance in our task; There is also some weak evidence for a positive correlation between a student's conscientiousness or agreeableness and performance. To test for the statistical significance we divided students into groups. For the educational level we used the questionnaire categories. For age and personality

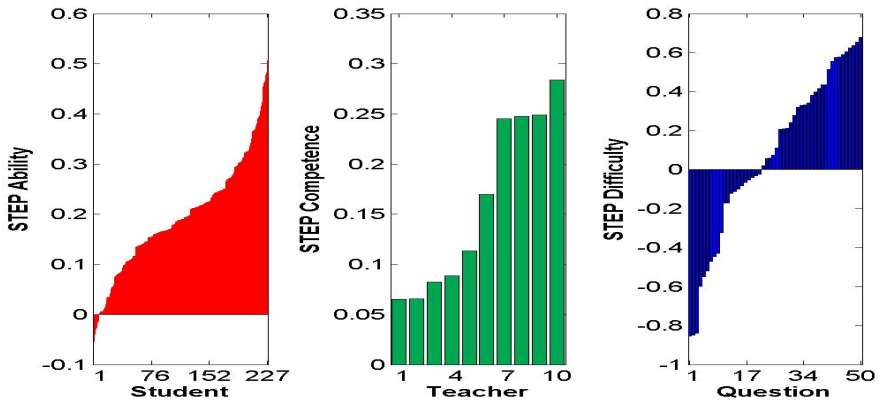

Fig. 3. STEP parameters - student ability, teacher competence and question difficulty

traits, we divided the student population into 3 equal size groups (low, medium and high) according to their responses in the questionnaire. We used a Mann-Whitney U-test (see [49]) to test the statistical significance of the differences between the low group and the high group. The statistically significant results (at a $p < 5\%$ level) are given in Table 1.

Table 1. Demographic/personality and performance

Property	Pearson Correlation	p value
Education	N/A	0.0001
Openness	0.2371	0.0001
Age	-0.1709	0.0032
Extroversion	0.2902	0.0068
Conscientiousness	0.1526	0.0405
Agreeableness	0.1867	0.0455

Table 1 shows that young or educated students had better performance. Further, those high in openess to experience or extroversion tended to do well in our task. Figure 4 visualizes these relations, showing the average performance for different groups (and showing the standard error).

Our results show a correlation between demographics or personality traits and performance in our task. Despite these correlations, there is a huge variability in performance even for workers with very similar demographic or personality profiles, highlighting the need to base predictions regarding attainment on observations regarding past performance, as done in the STEP model.

Student Teacher Matching: Our experiment used teaching materials prepared by various teachers. Online education can allow high volumes of students to access training material though the Internet. However, direct student-teacher interaction, by a phone call or a chat, allows teaching more difficult material and

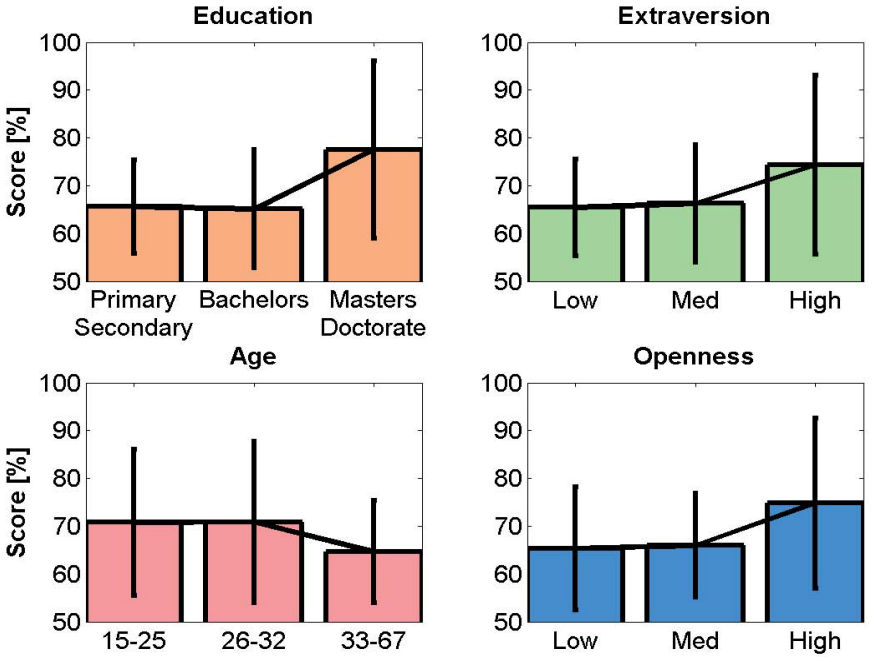

Fig. 4. Demographics and performance

achieves a higher rates of learning [41]. Such individual training requires having many teachers, as each teacher can only directly interact with few students. Nonetheless, one difference between traditional and online education systems is the flexibility in assigning teachers to students. Traditional education is constrained by physical limitations: a teacher who lives in one city cannot teach in another remote city. In online education a single student can be taught by many different teachers from across the globe, without leaving the comfort of their own home. We show that this allows us to optimize the assignment of teachers to students in order to improve the overall student performance. We use the model's estimate of a teacher t's competence in preparing teaching material as a *proxy* of how well they teach by *direct interaction*: though in our experiment the teaching materials prepared by a teacher can be used to train many students, we consider the case where a teacher can only interact with a single student.

STEP infers Gaussian posterior distributions for the competence of teachers and abilities of students. Given these parameters and a question (or exam) of a given difficulty, it infers p_c, the probability that student s would succeed in answering the question q if she is taught by the teacher t. Let $S \sim N(\mu_s, \sigma_s^2)$ be the inferred student s's ability, $T \sim N(\mu_t, \sigma_t^2)$ the inferred teacher t's competence, $D \sim N(\mu_d, \sigma_d^2)$ the question difficulty and $N \sim N(0, \sigma_n^2)$ the Gaussian noise used in the model. Let $p_c(s,t)$ be the probability that student s taught by teacher t knows the correct answer to a question of difficulty d (similar to the

Bernoulli variable $k_{s,q}$ in the previous section.) Under the assumptions of the STEP model, $p_c(s,t) = Pr(S + T + N > D)$, we can compute $p_c(s,t)$ for any student s and teacher t.

Consider a domain with an equal number n of teachers and students. Suppose every teacher has the capacity to teach a single student, and that we wish to maximize the number of students who pass an exam of difficulty d. How should we choose an assignment $A : S \to T$ between students and teachers, which respects the teacher capacity constraints (i.e. for any $t \in T$ there is only one student $s \in S$ such that $A(s) = t$), so as to maximize the expected number of passing students: $\arg\max_A \sum_{s \in S} p_c(s, A(S))$? The simplest way is a random assignment, which ignores the inferred abilities. However, when maximizing the number of passing students we only care if a student passes (rather than considering the exact score). If we have one good student and one bad student, and one good teacher and one bad teacher, we may be better off matching the good teacher to the bad student and the bad teacher to the good student, as the "returns on competence" can decrease with the ability of the student. [6] One heuristic is to sort students by increasing ability, and the teachers by decreasing competence and match them in that order. We call this the *inverse heuristic assignment*. Given an exam of difficulty d, matching a student s with teacher t has the expected return of $p_c(s,t)$. We can formulate maximizing the expected number of "passing" students as a Bipartite Maximum Weighted Matching (BMWM) problem [51]; We are given a bipartite graph of students on one side and the teachers on the other, and the edge between student s and teacher t has weight $w_{(s,t)} = p_c(s,t)$; The goal is find an assignment $A : S \to T$ matching each teacher to exactly one student so as to maximize the sum of weights of the matching. The BMWM output is the assignment A maximizing $\sum_{s \in S} w_{(s,A(s))}$. This *optimal assignment* (equivalently BMWM) can be found in polynomial time [51].

We compared the three matching algorithms (random assignment, inverse heuristic assignment and the optimal assignment) in terms of their performance, measured by the expected number of passing students. As the input data for the simulations we used the scaled output parameters of STEP on the real data discussed in the previous section. We only had 10 teachers in this dataset, so we randomly sampled a subset of 10 students many times, averaging the resulting performance under the three assignment methods. We performed the analysis on a range of question difficulty levels (matching the student abilities and teacher competences). The results are shown in Figure 5.

Figure 5 shows that for easy questions, the inverse heuristic outperforms random matching, and almost as good as the optimal assignment. However, as the difficulty increases, the inverse heuristic's performance degrades, until at some point it is even worse than random matching. For such moderate to difficult exams, there is a performance gain when switching to the optimal assignment. One possible reason for this is low ability students. If the exam is easy, such students are likely to pass when assigned a highly competent teacher, so the

[6] Such diminishing returns are prevalent in many resource allocation settings [12,18,8,2,40].

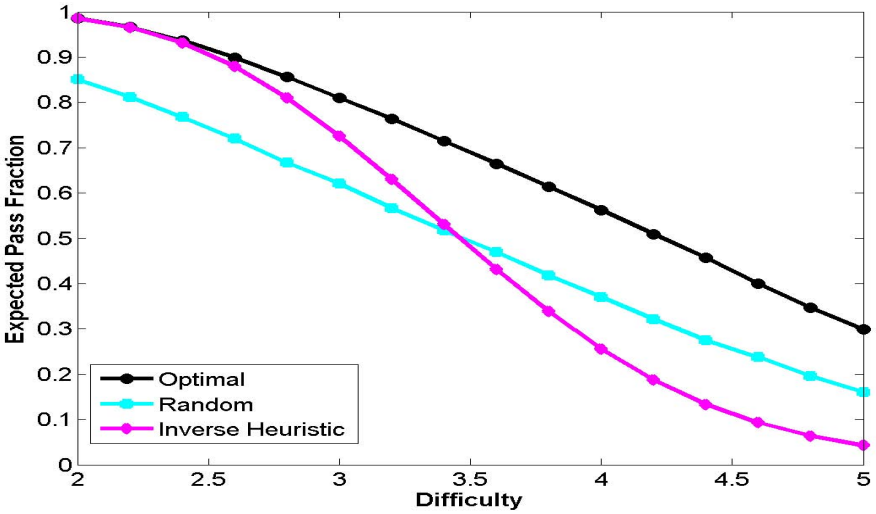

Fig. 5. Performance under assignment methods

inverse heuristic does well. However, if the exam is difficult, even a competent teacher cannot help such low ability students pass, so this heuristic "wastes" a very good teacher on a student that is very likely to fail nonetheless.

4 Related Work

Various models were proposed for assessing teacher competence [34,14,23]. To the best of our knowledge, we are the first to propose a probabilistic graphical models that simultaneously estimates student abilities, teacher competence and exam difficulties. The impact of demographics or personality on student attainment in *traditional* educational settings was studied in [17,53,44,48,19].

Our teachers' bonus was based on a competition. Such crowdsourcing contests were shown to allow the contest designer to elicit significant participant efforts [27,3].

Predicting attainment in cognitive tasks is a central topic in psychology. Psychometricians developed a framework called "test theory" to analyze outcomes in psychological testing, including intelligence and education [1]. One paradigm for designing such tests is "item-response theory" [24] (IRT for short), used to develop high-stakes adaptive tests such as the Graduate Management Admission Test (GMAT). Our STEP model relies on a probabilistic graphical model [29], and uses themes similar to the principles of IRT. A key difference is that we consider teacher competence as well, and tie the variables in the form of a factor graph. Frameworks using IRT principles and a probabilistic graphical model are [50,4,46]. However, the goal of these models is to *aggregate* multiple responses of participants to best determine the correct answers to questions, whereas our goal is to predict future performance of teachers and students in online education.

Our work ignored logical connections between questions. In many exams several questions rely on the same piece of knowledge, so a mistake regarding this information is likely to affect many responses. Frameworks such as Probabilistic Relational Models [21] combine a logical representation with probabilistic semantics, and can be used to express such structures.

5 Conclusion

We introduced the STEP model for estimating abilities and predicting outcomes in web-based education based on student abilities, teacher competences and question difficulties. We evaluated it on a crowdsourced dataset. We showed that STEP outperforms alternative approaches, and explored possible applications of this model. We have also analyzed the relation between attainment and demographics or personality traits. Finally, we have shown that the outputs of the STEP model regarding student abilities and teacher competences can be used to optimize the overall attainment of all the students by best matching teachers to students. This achieves an overall performance that is much better than a random or heuristic assignment.

Several directions remain open for future research. STEP was evaluated using data from a short experiment in AMT, which does not necessarily reflect a realistic online learning environment. Can a similar model predict outcomes in traditional education systems? Do our results generalize to real-world data from MOOCs? Can we build a dynamic model, that tracks fluctuations in student ability and teacher competence over time? How can we express dependency relations between tasks and areas of expertise?

References

1. Anastasi, A., Urbina, S., et al.: Psychological testing. Prentice Hall, Upper Saddle River (1997)
2. Anderson, T.: The theory and practice of online learning. Au Press (2008)
3. Archak, N.: Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder. com. In: Proceedings of the 19th International Conference on World Wide Web, pp. 21–30. ACM (2010)
4. Bachrach, Y., Graepel, T., Minka, T., Guiver, J.: How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In: ICML (2012)
5. Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., Van Gael, J.: Crowd iq: aggregating opinions to boost performance. In: AAMAS (2012)
6. Bachrach, Y., Graepel, T., Kohli, P., Kosinski, M., Stillwell, D.: Your digital image: factors behind demographic and psychometric predictions from social network profiles. In: AAMAS (2014)
7. Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D.: Personality and patterns of facebook usage. In: ACM WebSci (2012)
8. Bachrach, Y., Rosenschein, J.S.: Distributed multiagent resource allocation in diminishing marginal return domains. In: AAMAS (2008)

9. Barrick, M.R., Mount, M.K.: The big five personality dimensions and job performance: a meta-analysis. Personnel Psychology 44(1), 1–26 (2006)
10. Brabazon, T.: Digital hemlock: Internet education and the poisoning of teaching. University of New South Wales Press (2002)
11. Brusilovsky, P., et al.: Adaptive and intelligent technologies for web-based eduction. KI 13(4), 19–25 (1999)
12. Clearwater, S.H.: Market-based control: A paradigm for distributed resource allocation (1996)
13. Costa Jr., P.T., McCrae, R.R.: Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. Psychological Assessment Resources, Odessa (1992)
14. Darling-Hammond, L.: Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching (2010)
15. De Raad, B., Schouwenburg, H.C.: Personality in learning and education: A review. European Journal of Personality 10(5), 303–336 (1998)
16. Downes, S.: The rise of moocs. Stephens Web (2012)
17. Dumais, S.A.: Cultural capital, gender, and school success: The role of habitus. Sociology of Education, 44–68 (2002)
18. Eichler, H.-G., Kong, S.X., Gerth, W.C., Mavros, P., Jönsson, B.: Use of cost-effectiveness analysis in health-care resource allocation decision-making: How are cost-effectiveness thresholds expected to emerge? Value in health 7(5), 518–528 (2004)
19. Engle, J., Tinto, V.: Moving beyond access: College success for low-income, first-generation students. Pell Institute for the Study of Opportunity in Higher Education (2008)
20. Gao, X.A., Bachrach, Y., Key, P., Graepel, T.: Quality expectation-variance tradeoffs in crowdsourcing contests. In: AAAI (2012)
21. Getoor, L., Friedman, N., Koller, D., Pfeffer, A., Taskar, B.: 5 probabilistic relational models. Statistical Relational Learning, 129 (2007)
22. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality 37(6), 504–528 (2003)
23. Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., Whitehurst, G.: Evaluating teachers: The important role of value-added. Brookings Institution (2010)
24. Hambleton, R.K., Swaminathan, H., Rogers, H.J.: Fundamentals of item response theory, vol. 2 (1991)
25. Hara, N.: Student distress in a web-based distance education course. Information, Communication & Society 3(4), 557–579 (2000)
26. Harasim, L.: Shift happens: Online education as a new paradigm in learning. The Internet and Higher Education 3(1), 41–61 (2000)
27. Howe, J.: The rise of crowdsourcing. Wired Magazine 14(6), 1–4 (2006)
28. Khan, B.H.: Web-based instruction. Prentice Hall (1997)
29. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques (2009)
30. Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J., Graepel, T.: Crowd iq: Measuring the intelligence of crowdsourcing platforms. In: ACM WebSci (2012)
31. Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., Graepel, T.: Manifestations of user personality in website choice and behaviour on online social networks. Machine Learning 95(3), 357–380 (2014)
32. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. PNAS (2013)

33. Kosinski, M., Stillwell, D., Kohli, P., Bachrach, Y., Graepel, T.: Personality and website choice (2012)
34. Lavy, V.: Evaluating the effect of teachers group performance incentives on pupil achievement. Journal of Political Economy 110(6), 1286–1317 (2002)
35. Lev, O., Polukarov, M., Bachrach, Y., Rosenschein, J.S.: Mergers and collusion in all-pay auctions and crowdsourcing contests. In: AAMAS (2013)
36. Mackness, J., Mak, S., Williams, R.: The ideals and reality of participating in a MOOC. In: Networked Learing Conference (2010)
37. Minka, T., Winn, J.M., Guiver, J.P., Knowles, D.A.: Infer.NET 2.4 (2010)
38. Minka, T., Winn, J.: Gates. In: NIPS, vol. 21 (2008)
39. Minka, T.P.: A family of algorithms for approximate Bayesian inference. PhD thesis (2001)
40. Moore, M.G., Kearsley, G.: Distance education: A systems view of online learning. Wadsworth Publishing Company (2011)
41. Palloff, R.M., Pratt, K.: Lessons from the cyberspace classroom: The realities of online teaching. Wiley. Com (2002)
42. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference (1988)
43. Picciano, A.G.: Beyond student perceptions: Issues of interaction, presence, and performance in an online course. Journal of Asynchronous Learning Networks 6(1), 21–40 (2002)
44. Ridgell, S.D., Lounsbury, J.W.: Predicting academic success: General intelligence,big five personality traits, and work drive. College Student Journal 38(4), 607–618 (2004)
45. Russell, M.T., Karol, D.L.: Institute for Personality, and Ability Testing. In: The 16PF Fifth Edition Administrator's Manual, Institute for Personality and Ability Testing, Champaign (1994)
46. Salek, M., Bachrach, Y., Key, P.: Hotspotting – a probabilistic graphical model for image object localization through crowdsourcing. In: AAAI (2013)
47. Schochet, P.Z., Chiang, H.S.: Error rates in measuring teacher and school performance based on student test score gains. ncee 2010-4004. National Center for Education Evaluation and Regional Assistance (2010)
48. Scott, J.: Family, gender, and educational attainment in britain: A longitudinal study. Journal of Comparative Family Studies 35(4), 565–590 (2004)
49. Sprinthall, R.C., Fisk, S.T.: Basic statistical analysis. Prentice Hall, Englewood Cliffs (1990)
50. Welinder, P., Branson, S., Belongie, S., Perona, P.: The multidimensional wisdom of crowds. In: NIPS (2010)
51. West, D.B., et al.: Introduction to graph theory, vol. 2. Prentice Hall, Upper Saddle River (2001)
52. Witkowski, J., Bachrach, Y., Key, P., Parkes, D.C.: Dwelling on the negative: Incentivizing effort in peer prediction. In: HCOMP (2013)
53. Yorke, M., Thomas, L.: Improving the retention of students from lower socioeconomic groups. Journal of Higher Education Policy and Management 25(1), 63–74 (2003)