Beyond the MOOC platform: Gaining Insights about Learners from the Social Web

Guanliang Chen∗
Delft University of Technology
Delft, the Netherlands
guanliang.chen@tudelft.nl

Dan Davis†
Delft University of Technology
Delft, the Netherlands
d.j.davis@tudelft.nl

Jun Lin
Eindhoven University of Technology
Eindhoven, the Netherlands
j.lin@student.tue.nl

Claudia Hauff
Delft University of Technology
Delft, the Netherlands
c.hauff@tudelft.nl

Geert-Jan Houben
Delft University of Technology
Delft, the Netherlands
g.j.p.m.houben@tudelft.nl

ABSTRACT

Massive Open Online Courses (MOOCs) have enabled millions of learners across the globe to increase their levels of expertise in a wide variety of subjects. Research efforts surrounding MOOCs are typically focused on improving the learning experience, as the current retention rates (less than 7% of registered learners complete a MOOC) show a large gap between vision and reality in MOOC learning.

Current data-driven approaches to MOOC adaptations rely on data traces learners generate within a MOOC platform such as edX or Coursera. As a MOOC typically lasts between five and eight weeks and with many MOOC learners being rather passive consumers of the learning material, this exclusive use of MOOC platform data traces limits the insights that can be gained from them.

The Social Web potentially offers a rich source of data to supplement the MOOC platform data traces, as many learners are also likely to be active on one or more Social Web platforms. In this work, we present a first exploratory analysis of the Social Web platforms MOOC learners are active on — we consider more than 320,000 learners that registered for 18 MOOCs on the edX platform and explore their user profiles and activities on StackExchange, GitHub, Twitter and LinkedIn.

Categories and Subject Descriptors

•Human-centered computing → User models; Social networks; •Applied computing → Learning management systems;

∗The author’s research is supported by the Extension School of the Delft University of Technology.
†The author’s research is supported by the Leiden-Delft-Erasmus Centre for Education and Learning.

1. INTRODUCTION

Online education recently entered a new era of large-scale, free and open-access which has revolutionised existing practices. This new era dates from 2011, when the University of Stanford released its initial three MOOCs. Today, a wide range of courses in the humanities, business and natural sciences are offered for free with millions of learners taking advantage of them.

At the same time, however, the initial predictions of the “MOOC revolution” (universities will become obsolete) have not come to pass. On the contrary, MOOCs today generally suffer from a lack of retention [22, 20] — many learners sign up, but on average less than 7% complete a course.

Examining the current nature of MOOCs reveals an important clue as to why they, as yet, fail to realize their full potential. Although the “MOOC revolution” changed online education with respect to scale and openness, it did not involve any truly novel pedagogical approaches or education technologies. Currently, many MOOCs revolve around a set of videos, a set of quizzes and little else (the so-called “xMOOCs”). Instead, new approaches are necessary that support learning under the unique conditions of MOOCs: (i) the extreme diversity among learners (who come from diverse cultural, educational and socio-economic backgrounds [13]), and, (ii) the enormous learner-staff ratio, which often exceeds 20,000:1.

In order to improve the learning experience and retention, MOOC data traces (i.e. learners’ clicks, views, assignment submissions and forum entries) are being employed to investigate various aspects of MOOC learning, such as the effect of lecture video types on learner engagement [12], the introduction of gamification [7], the impact of instructor involvement [29] and the significance of peer learning [8].

Few data-driven research works go beyond the data learners generate within a MOOC platform. We argue that we can potentially learn much more about MOOC learners if we move beyond this limitation and explore the learners’ traces on the wider Web, in particular the Social Web, to gain a deeper understanding of learner behavior in a distributed learning ecosystem. Hundreds of millions of users are active on the larger Social Web platforms such as Twitter and existing research has shown that detailed user profiles can be built from those traces, covering dimensions such as
age [24], interests [1], personality [3], location [14] and occupation [25].

While MOOC learners are usually invited to participate in pre-course surveys that include inquiries about their demographics and motivations, not all of them do (and those who do may fill in non-credible or false information), with return rates hovering around 10%\(^1\). In addition, these surveys can only provide a very limited view of the learners as the return rates drop with every question that is added to the questionnaire and, finally, questionnaires offer us only a snapshot-based perspective as learners cannot be polled continuously across a period of time.

We hypothesize that the Social Web can provide us with a source of diverse, fine-grained and longitudinal learner traces we can exploit in order to (i) derive more extensive learner profiles for a larger learner population than is possible through pre/post-MOOC surveys, and, (ii) investigate questions that cannot be investigated solely based on the traces learners leave within MOOC environments (e.g. the uptake of learned concepts in practice).

In this work we provide a first exploratory analysis of more than 329,000 MOOC learners and the Social Web platforms they are active on, guided by the following three Research Questions:

- **RQ1**: On what Social Web platforms can a significant fraction of MOOC learners be identified?
- **RQ2**: Are learners who demonstrate specific sets of traits on the Social Web drawn to certain types of MOOCs?
- **RQ3**: To what extent do Social Web platforms enable us to observe (specific) user attributes that are highly relevant to the online learning experience?

Our contributions can be summarized as follows:

- We provide a methodology to reliably identify a subset of learners from a set of five Social Web platforms and eighteen MOOCs. Depending on the MOOC/platform combination, between 1% and 42% of the learners could reliably be identified.
- We show that it is indeed possible to derive valuable learner attributes from the Social Web which can be used to investigate learner experience in MOOCs.
- We show that the tracking of learners over time (in the case of GitHub we consider three years of data traces) enables us to investigate the impact of MOOCs in the long-term.

2. **SOCIAL WEB & MOOCS**

The wider Web is starting to be viewed as a source of useful information in MOOC learning analytics — the field concerned with the understanding and optimization of learning in massive open online courses. Existing works focus on the analysis of Social Web platforms during the running of a MOOC in order to learn more about the interactions and processes occurring within a MOOC. These analyses are not conducted on the individual learner level, but on the aggregated group level, without the explicit matching of MOOC learners to Social Web profiles.

Alario et al. [2] investigate the learners’ engagement with two built-in MOOC platform components (Q&A and forum) and three external Social Web portals (Facebook, Twitter and MentorMob) during the running of a single MOOC. Learners’ MOOC and Social Web identities are not matched directly, instead, learners are asked to join a specific Facebook group and use a course-specific Twitter hashtag. The authors find that despite the active encouragement of the platforms’ usage to exchange ideas and course materials, after the initial phase of excitement, participation quickly dropped off. Similarly, van Treeck & Ebner [31] also rely on Twitter hashtags to identify the microblog activities surrounding two MOOCs. They (qualitatively) analyse the tweet topics, their emotions and the extent of actual interactions among learners on Twitter and find a small group of MOOC participants (6%) to have generated more than half of all microblog content.

Garcia et al. [11] analysed the concurrent Twitter activities of students taking a “Social Networking and Learning” MOOC to track their engagement and discussion beyond the MOOC environment by designating and tracking hashtagged conversation threads. In the same MOOC, [9] presented a generalisable method to extend the MOOC ecosystem to the Social Web (in this case Google+ and Twitter) to both facilitate and track students’ collaborations and discussions outside of the immediate context of the course.

[18] tracked Twitter interactions among MOOC students to understand the dynamics of social capital within a connectivist [27] MOOC environment, which is inherently decentralised and distributed across platforms. This work was primarily concerned with learner-learner relationships in the context of a MOOC—not individual learner traits. And, more broadly, [19] explored the types and topics of conversations that happen in the Social Web concurrent to a MOOC.

Four observations can be made based on these studies: existing works (i) analyze one or two Social Web platforms only, (ii) are usually based on experiments within a single MOOC, (iii) do not require a learner identification step (as an intermediary such as a Twitter hashtag is employed), and (iv) focus on learner activities exhibited during the running of a MOOC that are topically related to the MOOC content (e.g. ensured through the use of moderated Facebook group).

In contrast, we present a first exploratory analysis across eighteen MOOCs and five Social Web platforms exploring the learners’ behaviours, activities and created content over a considerably longer period of time.

3. **APPROACH**

In this section, we first describe our three-step approach to locate a given set of MOOC learners on Social Web platforms, before going into more detail about the analyses performed on the learners’ Social Web traces.

3.1 **Locating Learners on the Social Web**

On the edX platform, a registered learner \(i\) is identified through a username, his or her full name and email address (as required by the platform), i.e. \(i = (login, name, email)\).

On a Social Web platform \(P_j\), the publicly available information about a user \(u\) usually consists of (a subset of) username, full name, email address and profile description.
The profile description is often semi-structured and may also contain links to user accounts on other Social Web Platforms \( P_2, \ldots, P_5 \). A common assumption in this case (that we employ as well) is that those accounts belong to the same user \( u \).

For each Social Web platform \( P_j \) we attempt to locate \( l_i \) through a three-step procedure:

- **Explicit** If \( P_j \) enables the discovery of users via their email address, we use email to determine \( l_i \)'s account \( u_i \) on \( P_j \). If available, we also crawl the profile description of \( u_i \), the profile image (i.e. the user avatar) and extract all user account links to other Social Web platforms under the assumption stated before.

- **Direct** This step is only applied to the combination of learners and Social Web platforms \( (l_i, P_j) \) for which no match was found in the Explicit step. We now iterate over all extracted account links from the Direct step and consider \( l_i \)'s account on \( P_j \) to be found if it is in this list.

- **Fuzzy** Finally, for pairs \((l_i, P_j)\) not matched in the Direct step, we employ fuzzy matching: we rely on \( l_i \)'s name, \( \text{login} \) and search for those terms on \( P_j \). Based on the user (list) returned, we consider a user account a match for \( l_i \), if one of the following three conditions holds:
  
  1. The profile description of the user contains a hyperlink to a profile that was discovered in the Explicit or Direct step.
  2. The avatar picture of the user in \( P_j \) is highly similar to one of \( l_i \)'s avatar images discovered in the Explicit or Direct step (we measure the image similarity based on image hashing [28] and use a similarity threshold of 0.9), or,
  3. The username and the full name of the user on \( P_j \) and \( l_i \) are a perfect match.

### 3.2 Social Web Platforms

Our initial investigation focused on ten globally popular Social Web platforms, ranging from Facebook and Twitter to GitHub and WordPress. We eventually settled on five platforms, after having considered the feasibility of data gathering and the coverage of our learners among them. Concretely, we investigate the following platforms:

- **Gravatar**[^2] is a service for providing unique avatars to users that can be acquired across a wide range of sites. During our pilot investigation, we found Gravatar to be employed by quite a number of learners in our dataset. Given that Gravatar allows the discovery of users based on their email address, we employ it as one of our primary sources for Explicit matching. We crawled the data in November 2015. We were able to match 25,702 edX learners on Gravatar.

- **StackExchange**[^3] is a highly popular community-driven question & answering site covering a wide range of topics. The most popular sub-site on this platform is StackOverflow, a community for computer programming related questions. StackExchange regularly releases a full "data dump" of their content that can be employed for research purposes. We employed the data release from September 2015 for our experiments. We were able to match 15,135 edX learners on StackExchange.

- **LinkedIn**[^4] is a business-oriented social network users rely on to find jobs, advertise their skill set and create & maintain professional contacts. The public profiles of its users can be crawled, containing information about their education, professional lives, professional skills and (non-professional) interests. We crawled the data in November 2015. We were able to match 19,405 edX learners on LinkedIn.

- **Twitter**[^5] is one of the most popular microblogging portals to date, used by hundreds of millions of users across the globe. Twitter allows the crawling of the most recent 3,200 tweets per user. We crawled the data in December 2015 and January 2016. We were able to match 25,620 edX learners on Twitter.

- **GitHub**[^6] is one of the most popular social coding platforms, allowing users to create, maintain and collaborate on open-source software projects. The GitHub platform creates a large amount of data traces, which are captured and made available for research through two large initiatives: GitHub Archive[^7] and GHTorrent[^8]. For our work, we rely on all data traces published between January 1, 2013 and December 31, 2015. We were able to match 31,478 edX learners on GitHub.

In addition, we are interested in how many learners are observed across more one platform. The numbers of learners that can be matched across 2, 3, 4, 5 platforms are 14,824, 6,980, 3,129, 1,125, respectively.

### 3.3 Social Web Data Analysis

As our work is exploratory in nature, we employ a range of data analysis approaches that enable us to explore our gathered data traces from various angles.

**t-SNE.**

Many of our user profiles are high-dimensional: a LinkedIn user may be represented through a vector of his or her skills[^9] and a Twitter user profile may be encoded as a vector of the entities or hyperlinks mentioned in his or her tweets. If we are interested to what extent those user profiles are similar or dissimilar for users (learners) that are taking different kinds of MOOCs, we can visualize these similarities using t-SNE (t-Distributed Stochastic Neighbor Embedding [30]), a visualization approach for high-dimensional data that computes for each datapoint a location on a 2D (or 3D) map. t-SNE[^10] creates visualizations that reveal the structure of the high-dimensional data at different scales and has been shown to be superior to related non-parametric visualizations such as Isomaps [4].

[^2]: https://en.gravatar.com/
[^3]: http://stackexchange.com/
[^4]: https://www.linkedin.com/
[^5]: https://twitter.com/
[^6]: http://githubarchive.org/
[^7]: https://www.githubarchive.org/
[^8]: http://ghtorrent.org/
[^9]: The dimension of the vector space depends on the number of unique skills in the dataset, with a single skill being encoded in binary form.
[^10]: In this paper, we utilize t-SNE's scikit-learn implementation: http://scikit-learn.org/.
**Age and gender prediction.**

Predicting certain user attributes based on a user’s Social Web activities is an active area of research. It has been shown that attributes such as age [24], gender [5], personality [17], home location [23] and political sentiments [6] (to name just a few) can be predicted with high accuracy from Social Web data sources.

In our work we focus on the prediction of age and gender, as those two attributes can be inferred of Social Web users with high accuracy. We also have intuitions concerning the age and gender (in contrast to, for instance, their personalities) of the learners that take our MOOCs (e.g. a computer science MOOC is likely to have a larger pool of male participants), enabling us to judge the sensibility of the results.

The main challenge in this area of work is the collection of sufficient and high-quality training data (that is, Social Web users with known age, gender, location, etc.). Once sufficient training data has been obtained, standard machine learning approaches are usually employed for training and testing.

In our work, we make age and gender predictions based on tweets and employ the models provided by [26]11, who utilized the English language Facebook messages of more than 72,000 users (who collectively had written more than 300 million words) to create unigram-based age & gender predictors based on Ridge regression [16]. The age model \(M_{age}\) contains 10,797 terms and their weights \(w_t\). To estimate the age of a user \(u\), we extract all his English language tweets (excluding retweets), concatenate them to create a document \(D_u\) and then employ the following formulation:

\[
\text{age}_u = \omega_0 + \sum_{t \in M_{age}} w_t \times \frac{freq(t, D_u)}{|D_u|}.
\]

Here, \(|D_u|\) is the number of tokens in \(D_u\), \(\omega_0\) is the model intercept and \(freq(t, D_u)\) is the term frequency of \(t\) in \(D_u\). Only terms in \(D_u\) that appear in \(M_{age}\) have a direct effect on the age estimate. The model is intuitively understandable; the five terms with the largest positive weights (indicative of high age) are \{grandson, daughter, daughters, son, folks\}. Conversely, the five terms with the largest negative weights (indicative of a young user) are \{parents, exams, pregnant, youth, mommy\}.

The gender prediction is derived in an analogous fashion based on model \(M_{gender}\), which consists of 7,137 terms and their weights. In contrast to the age estimation (which provides us with a continuous estimate), we are interested in a binary outcome. Thus, after the regression stage, classification is performed: if the estimate is \(\geq 0\), the user is classified as female and otherwise as male. Once more, the model is intuitive; the largest negative weights (indicating maleness) are \{boxers, shaved, ha3ircut, shave, girlfriend\}.

**Learning Transfer.**

Existing investigations into student learning within MOOC environments are commonly based on pre- & post-course surveys and log traces generated within those environments by the individual learners [15]. With a crude, binary measure of learning, the success (pass/no-pass) of the learner could be labeled. While learning is an important success measure, we also believe that the amount of learning transfer [21] that is taking place should be considered: do learners actually utilize the newly gained knowledge in practice? Are learners expanding their knowledge in the area over time or do they eventually move back to their pre-course knowledge levels and behaviours? While most Social Web platforms do not offer us insights into this question, for MOOCs (partially) concerned with the teaching of programming languages (such as Functional Programming) we can rely on the GitHub platform to perform an initial exploration of this question.

GitHub provides extensive access to data traces associated with public coding repositories, i.e. repositories visible to everyone12. GitHub is built around the git distributed revision control system, which enables efficient distributed and collaborative code development. GitHub not only provides relevant repository metadata (including information on how popular a repository is, how many developers collaborate, etc.), but also the actual code that was altered. As the GitHub Archive13 makes all historic GitHub data traces easily accessible, we relied on it for data collection and extracted all GitHub data traces available between January 1, 2013 and June 30, 2015 (five months after the end of the programming MOOC in our dataset). We then filtered out all traces that were not created by the 31,478 learners we identified on the GitHub platform. Of the more than 20 GitHub event types14, we only consider the PushEvent as vital for our analysis.

Every time code is being updated ("pushed" to a repository), a PushEvent is triggered. Figure 1 contains an excerpt of the data contained in each PushEvent. The most important attributes of the event are the created_at timestamp (which allows us to classify events as before/during/after the running of ), the actor (the user doing the “push”) and the url, which contains the URL to the actual diff file. While the git protocol also allows a user to “push” changes by another user to a repository (which is not evident from inspecting the diff file alone), this is a rare occurrence among our learners: manually inspecting a random sample of 200 PushEvents showed 10 such cases.

```
{
  "__id__": ObjectId("55b6005de4b07ff432432dfe1"),
  "created_at": "2013-03-03T18:36:09-08:00",
  "url": "https://github.com/john/ RMS/compare/1c55c4cb04...420e112334",
  "actor": {"_id": "john",
    "actor_attributes": {
      "name": "John Doe",
      "email": "john@doe.com"
    }
  },
  "repository": {
    "id": 2.37202e+06,
    "name": "RMS",
    "forks": 0,
    "open_issues": 0,
    "created_at": "2011-09-12T08:28:27-07:00",
    "master_branch": "master"
  }
}
```

**Figure 1: Excerpt of a GitHub PushEvent log trace.**

A diff file shows the difference between the last version of the repository and the new one (after the push) in terms

12Data traces about private repositories are only available to the respective repository owner.
13https://www.githubarchive.org/
of added and deleted code. For each of the identified PushEvents by our learners, we crawled the corresponding diff file, as they allow us to conduct a more fine-grained code analysis. As a first step in this direction, we identified the number of additions and deletions a user conducts in each programming language based on the filename extensions found in the corresponding diff file.

4. MOOC LEARNERS & THE SOCIAL WEB

As a starting point for our investigation we utilize eighteen MOOCs that have run between 2013 and 2015 on the edX platform — the largest MOOCs conducted by the Delft University of Technology (situated in the Netherlands) to date; the courses cover a range of subjects in the natural sciences, computer science and the humanities and were all taught in English. An overview of the MOOCs can be found in Table 1; we deemed the MOOC titles not to be self-explanatory, so we also added the MOOC’s “tag line”. Apart from the Pre-universis Calculus (specifically geared towards pre-university learners) and the Topology in Condensed Matter (aimed at MSc and PhD physics students) courses, the MOOCs were created with a wide variety of learners in mind. All courses follow the familiar MOOC recipe of weekly lecture videos in combination with quizzes and automatically (or peer-) graded assignments.

The MOOCs vary significantly in size. The largest MOOC (Solar Energy 2013) attracted nearly 70,000 learners, while the smallest one (Topology in Condensed Matter 2015) was conducted with approximately 4,200 learners. While the majority of learners register for a single MOOC only, a sizable minority of learners engage with several MOOCs and thus the overall number of unique learners included in our analysis is 329,200.

To answer RQ1, Table 1 summarizes to what extent we were able to identify learners across the five Social Web platforms, employing the three-step procedure described in Section 3.1. Note that the numbers reported treat each course independently, i.e. if a learner has registered to several courses, it will count towards the numbers of each course.

The percentage of learners we identify per platform varies widely across the courses between 4-24% (Gravatar), 1-22% (StackExchange), 3-42% (GitHub), 4-11% (LinkedIn) and 5-18% (Twitter) respectively. Functional Programming is the only MOOC we are able to identify more than 10% of the registered learners across all five Social Web platforms. While this finding by itself is not particularly surprising — two of the five Social Web platforms are highly popular with users interested in IT topics (i.e. GitHub and StackExchange) and those users also tend to be quite active on Social Web platforms overall — it can be considered as an upper bound to the fraction of learners that are active on those five platforms and identifiable through robust and highly accurate means.

Table 2: Overview of the percentage of MOOC learners (329,200 overall) identified through the different matching strategies on the five selected Social Web platforms. A dash (—) indicates that for this specific platform/strategy combination, no matching was performed.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Explicit</th>
<th>Direct</th>
<th>Fuzzy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravatar</td>
<td>7.81%</td>
<td>—</td>
<td>—</td>
<td>7.81%</td>
</tr>
<tr>
<td>StackExchange</td>
<td>4.32%</td>
<td>0.01%</td>
<td>0.25%</td>
<td>4.58%</td>
</tr>
<tr>
<td>GitHub</td>
<td>9.04%</td>
<td>0.02%</td>
<td>1.23%</td>
<td>10.29%</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>—</td>
<td>0.48%</td>
<td>5.41%</td>
<td>5.89%</td>
</tr>
<tr>
<td>Twitter</td>
<td>—</td>
<td>0.67%</td>
<td>7.12%</td>
<td>7.78%</td>
</tr>
</tbody>
</table>

5. RESULTS

In this section, we present an overview of our findings. As we collected different types of data (tweets vs. skills vs. source code) from different Social Web platforms, we describe the analysis conducted on each platform’s data traces independently in the following subsections.

5.1 Learners on Twitter

Our Twitter dataset consists of 25,620 unique users having written 12,314,067 tweets in more than 60 languages, which offers many insights into RQ2. The majority language is English (68.3% of all tweets), followed by Spanish (7.3%), Dutch (3.1%), Portuguese (3.1%) and Russian (2.2%)\textsuperscript{16}. The popularity of the Dutch language among our Twitter users can be explained by the fact that all MOOCs we consider in this analysis are offered by a Dutch university.

For each Twitter user with at least 100 English language tweets we estimated their age according to the approach described in Section 3.3. The results for our Twitter user set overall and three exemplary MOOCs (that is, we only consider users that participated in a particular MOOC) are shown in Figure 2: we have binned the estimations into six age brackets\textsuperscript{17}. The average MOOC learner is between 20 and 30 years of age, though we do observe that different types of courses attract slightly different audiences: In the Functional Programming MOOC, the 20-40 year old learners are overrepresented (compared to the “Overall” user set — computed across all eighteen MOOCs), while Framing and Retrieval for email matching and the September 2015 data dump for our content analysis.

\textsuperscript{15}Note that StackExchange stopped the release of MD5 hashes in September 2013, thus we use the 2013 data dump for email matching and the September 2015 data dump for our content analysis.

\textsuperscript{16}We generated these numbers based on Twitter’s language auto-detect feature.

\textsuperscript{17}Based on the ground truth data provided by 20,311 edX learners, the prediction precision is 36.5%.
<table>
<thead>
<tr>
<th>MOOC</th>
<th>Year</th>
<th>#Learners</th>
<th>Gravatar</th>
<th>Stack-Exchange</th>
<th>GitHub</th>
<th>LinkedIn</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solar Energy</td>
<td>2013</td>
<td>67,143</td>
<td>3,510</td>
<td>1,570</td>
<td>3,677</td>
<td>2,997</td>
<td>3,828</td>
</tr>
<tr>
<td>Solar Energy</td>
<td>2014</td>
<td>34,524</td>
<td>1,923</td>
<td>874</td>
<td>2,229</td>
<td>1,625</td>
<td>2,152</td>
</tr>
<tr>
<td>Solar Energy</td>
<td>2015</td>
<td>26,178</td>
<td>1,147</td>
<td>435</td>
<td>1,184</td>
<td>1,181</td>
<td>1,557</td>
</tr>
<tr>
<td>Introduction to Water Treatment</td>
<td>2013</td>
<td>34,897</td>
<td>1,559</td>
<td>508</td>
<td>1,198</td>
<td>1,362</td>
<td>1,741</td>
</tr>
<tr>
<td>Introduction to Drinking Water Treatment</td>
<td>2014</td>
<td>10,458</td>
<td>457</td>
<td>129</td>
<td>430</td>
<td>427</td>
<td>548</td>
</tr>
<tr>
<td>Introduction to Water and Climate</td>
<td>2014</td>
<td>9,267</td>
<td>545</td>
<td>149</td>
<td>510</td>
<td>452</td>
<td>558</td>
</tr>
<tr>
<td>Technology for Biobased Products</td>
<td>2014</td>
<td>9,811</td>
<td>1,438</td>
<td>583</td>
<td>1,451</td>
<td>1,155</td>
<td>1,447</td>
</tr>
<tr>
<td>Next Generation Infrastructures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introduction to Water Treatment</td>
<td>2014</td>
<td>20,531</td>
<td>1,438</td>
<td>583</td>
<td>1,451</td>
<td>1,155</td>
<td>1,447</td>
</tr>
<tr>
<td>Functional Programming</td>
<td>2014</td>
<td>38,682</td>
<td>9,087</td>
<td>8,477</td>
<td>16,220</td>
<td>4,274</td>
<td>6,801</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>2015</td>
<td>33,547</td>
<td>2,392</td>
<td>1,165</td>
<td>4,432</td>
<td>2,469</td>
<td>2,800</td>
</tr>
<tr>
<td>Pre-university Calculus</td>
<td>2015</td>
<td>28,015</td>
<td>1,928</td>
<td>960</td>
<td>2,477</td>
<td>1,406</td>
<td>2,064</td>
</tr>
<tr>
<td>Introduction to Aeronautical Engineering</td>
<td>2014</td>
<td>20,481</td>
<td>1,134</td>
<td>605</td>
<td>1,373</td>
<td>921</td>
<td>1,192</td>
</tr>
<tr>
<td>Framing</td>
<td>2015</td>
<td>20,481</td>
<td>1,134</td>
<td>605</td>
<td>1,373</td>
<td>921</td>
<td>1,192</td>
</tr>
<tr>
<td>Topology in Condensed Matter</td>
<td>2015</td>
<td>4,231</td>
<td>277</td>
<td>129</td>
<td>600</td>
<td>201</td>
<td>302</td>
</tr>
<tr>
<td>Solving Complex Problems</td>
<td>2014</td>
<td>32,673</td>
<td>2,803</td>
<td>1,620</td>
<td>3,928</td>
<td>1,085</td>
<td>1,124</td>
</tr>
<tr>
<td>Delft Design Approach</td>
<td>2014</td>
<td>13,543</td>
<td>1,319</td>
<td>514</td>
<td>1,376</td>
<td>1,085</td>
<td>1,124</td>
</tr>
<tr>
<td>Responsible Innovation</td>
<td>2014</td>
<td>10,735</td>
<td>877</td>
<td>274</td>
<td>800</td>
<td>713</td>
<td>753</td>
</tr>
<tr>
<td>Unique Users</td>
<td>2014</td>
<td>20,351</td>
<td>1,438</td>
<td>583</td>
<td>1,451</td>
<td>1,155</td>
<td>1,447</td>
</tr>
</tbody>
</table>

Table 1: Overview of the edX MOOCs under investigation, the number of learners registered to those MOOCs and the number of learners that could be matched (with either Explicit/Direct or Fuzzy matching) to our five Social Web platforms. Marked with † (‡) are those course/platform combinations where we were able to locate > 5% (> 10%) of the registered learners. The final row contains the unique number of users/learners (a learner may have taken several MOOCs) identified on each platform.

*Responsible Innovation* engage older learners to a larger than average degree.

We conduct an analogous analysis of our users’ gender distribution; the results are shown in Figure 3. The majority of MOOCs we investigate are anchored in engineering or the natural sciences, which traditionally attract a much larger percentage of male learners (in most parts of the world). This is reflected strongly in our Twitter sample: across all users with 100 or more English speaking tweets, 89% were identified as male. The MOOC with the highest skew in the distribution is *Functional Programming* with more than 96% of users identified as male. In contrast, the *Framing* and *Robust Innovation* exhibit the lowest amount of skewness: in both MOOCs, more than 20% of the users in our sample are classified as female.

The results we have presented provide us with confidence that microblog-based user profiling in the context of massive...
open online learning yields reliable outcomes. Future work will investigate the derivation of more complex and high-level attributes (such as personalities and learner type) from microblog data and their impact on online learning.

5.2 Learners on LinkedIn

LinkedIn user profiles are often publicly accessible, containing information about a user’s education, past and current jobs as well as their interests and skills. As shown in Table 1, for each of the MOOCs we were able to identify between 200 (Topology in Condensed Matter) and 2,997 (Solar Energy 2013) learners on the LinkedIn platform. To explore RQ2 we focus on two types of information in those profiles: job titles and skills. In our dataset, among the 19,405 collected LinkedIn profiles, 17,566 contain a job title (with on average 5.89 number of terms) and 16,934 contain one or more skills (37.42 skills on average).

In Figure 4, exemplary for three MOOCs (Data Analysis, Responsible Innovation, and Delft Design Approach), we present the most frequently occurring bigrams among the job titles of our learners. Interestingly, the Data Analysis MOOC attracts a large number of self-proclaimed “software engineers” and “business analysts,” despite the fact that it covers elementary material (it is an introduction to spreadsheet-based data analysis & Python) which we consider users in this area to be already familiar with. In contrast, the Delft Design Approach and Responsible Innovation job title bigram distributions are more in line with our expectations — the most frequent bigrams are “project manager” and “co founder” respectively, positions for which knowledge about the risks and ethical questions of new technologies (Responsible Innovation) and the design of new products (Delft Design Approach) are very relevant to.

As prior works [33] have indicated extrinsic factors such as recognition-by-others to play an important motivating role for MOOC learners, an explanation for the observed discrepancy between expected learners and actual MOOC participants, we also investigate to what extent our learners on LinkedIn present their MOOC achievements to the outside world. In Figure 5 we present a distribution of the number of MOOC certificates our users in the LinkedIn dataset list on their profile page. Each certificate represents a successfully completed MOOC. We limit our investigation to any certificate issued by the edX or Coursera platforms, as they offer a verifiable certificate interface to LinkedIn. We manually checked a random sample of 100 DelftX edX certificates listed by LinkedIn users to check whether each was actually issued to this specific user via edX. This was indeed the case for all sampled certificates. Overall, 9% of our users list one or more MOOC certificates on their public profile with the majority of users (57%) having achieved one or two certificates only. A small fraction of learners (2%) is highly active in the MOOC learning community, having collected more than 15 certificates over time. Future work will investigate the impact of MOOC certificates on professional development through the lense of LinkedIn.

Lastly, we investigate to what extent the users’ listed skills on their LinkedIn profiles can be considered indicative of their course preferences (to enable course recommendations for instance). A user can list up to 50 skills on his profile — skills are not restricted to a pre-defined set, any keyword or short phrase can be added as a skill. Across all LinkedIn
users in our dataset (19,405 users in total), the five most frequently mentioned skills are management (5,847 times), project management (4,894 times), java (4,087 times), microsoft office (4,073 times) and leadership (3,971 times). Thus, most of the users in our dataset present skills of themselves that are required for higher positions. We created a skill vocabulary by considering all skills mentioned at least once by a user in our dataset and then filtering out the fifty most frequent skills overall, leaving us with 28,816 unique skills. We create a user-skill matrix, where each cell represents the presence or absence of a skill in a user’s profile. We then applied truncated SVD [10] to reduce the dimensions of the matrix to 50 and then employed t-SNE (described in Section 3.3) to visualize the structure of the data in a two dimensional space.  

![t-SNE visualizations](image)

In Figure 6 we present the t-SNE based clustering of user skills exemplary for three pairs of MOOCs: Delft Design Approach vs. Topology of Condensed Matter, Data Analysis vs. Solar Energy 2015, and, Functional Programming vs. Framing. Recall, that a point in a plot represents a skill vector; t-SNE visually clusters data points together that are similar in the original (high-dimensional) skill space. The most distinct clustering can be observed for the final course pairing — users interested in functional programming are similar to each other, but different in their skill set from users interested in the analyses of political debates. This is a sensible result, which highlights the suitability of t-SNE for this type of data exploration. For the other two course pairings, the plots show less separation. In particular, for the Data Analysis vs. Solar Energy 2015 pairing, we observe a complete overlap between the two sets of users, i.e. there is no distinct set of skills that separates their interests. The pairing Delft Design Approach vs. Topology of Condensed Matter shows that the users of the design course have a larger spread of skills than those taking the physics MOOC. Still, the overlap in the skill set is considerable.

### 5.3 Learners on StackExchange

Our StackExchange dataset consists of 86,672 questions (1% of all StackExchange questions posted), 197,504 answers (1.2% of all answers) and 418,633 comments, which were contributed by the 31,478 unique users we identified as MOOC learners among our courses. Given that 51.5% of the identified users registered for the Functional Programming MOOC, we focus our attention on the StackOverflow site within StackExchange (the Q&A site for programming-related questions), where our learners contributed 71,344 questions, 177,780 answers and 358,521 comments.

Driven by RQ3, we first explored to what extent (if at all) MOOC learners change their question/answering behaviour during and after a MOOC. We restricted this analysis to the learners of the Functional Programming MOOC as those were by far the most active on StackOverflow. Among the 38,683 learners that registered for that MOOC, 8,068 could be matched to StackExchange. Of those users, 849 attempted to answer at least one question related to functional programming.

In Figure 7 (top) we plot month-by-month (starting in January 2014) the number of questions and answers by our learners that are tagged with “Haskell”, the functional language taught in the MOOC. Two observations can be made: (i) a subset of learners was already using Haskell before the start of the MOOC (which ran between 10/2014 and 12/2014), and, (ii) the number of Haskell questions posed by MOOC learners after the end of the MOOC decreased considerably (from an average of 32 questions per month before the MOOC to 19 per months afterwards), while the number of answers provided remained relatively stable. Figure 7 (bottom) shows that this trend is specific to the subset of MOOC learners: here we plot the frequency of “Haskell”-tagged questions and answers across all StackExchange users and observe no significant changes in the ratio between questions and answers. Finally, in Figure 7 (middle) we consider our learners’ uptake of functional programming in general, approximated by the frequency of questions and answers tagged with any of the nine major functional language names. We again find that over time, the ratio between questions & answers becomes more skewed (i.e. our learners turn more and more into answerers).

---

19 Scala, Haskell, Common Lisp, Scheme, Coljure, Racket, Erlang, Ocaml, F#
In this work, we have provided a first exploratory analysis of learners’ Social Web traces across eighteen MOOCs and five globally popular Social Web platforms. We argue that MOOC-based learning analytics has much to gain from looking beyond the MOOC platform and accounting for the fact that learning events frequently happen beyond the immediate course environment. This study embraces the data traces learners leave on various Social Web platforms as integral parts of the distributed, connected, and open online learning ecosystem.

Focusing on RQ1, we have found that on average 5% of learners can be identified on globally popular Social Web platforms. We observed a significant variance in the percentage of identified learners; in the most extreme positive case (Functional Programming/GitHub) we were able to match 42% of learners.

We also found that learners with specific traits prefer different types of MOOCs (RQ2) and we were able to present a first investigation into user behaviours (such as learning transfer over time) that are paramount in the push to make MOOCs more engaging and inclusive (RQ3).

In this work we were only able to explore the possible contributions of each Social Web platform to enhance massive open online learning on a broad level. In future work, we will zoom in on each of the identified platforms and explore...
in greater detail how learners’ behaviours and activities can be explored to positively impact our understanding of massive open online learning and improve the learning experience. For example, we plan to model learners’ motivation (e.g., career development needs) based on their professional profile provided in LinkedIn.

7. REFERENCES


