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.*

WebSci '16, May 22-25, 2016, Hannover, Germany.

DOI: http://dx.doi.org/10.1145/2908131.2908145

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 rate drops 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 **R**esearch **Q**uestions:

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 l_i is identified through a username, his or her full name and email address (as required by the platform), i.e. $l_i = (login_i, name_i, email_i)$.

On a Social Web platform P_j, the publicly available information about a user u^j usually consists of (a subset of) username, full name, email address and profile description.

[1]An estimate we derived based on the MOOCs we consider in this work. This percentage drops to 1% or less when considering post-course surveys, i.e. questionnaires conducted at the end of a MOOC.

The profile description is often semi-structured and may also contain links to user accounts on other Social Web Platforms $P_x, ..P_z$. 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 $email_i$ to determine l_i's account u_i^j on P_j. If available, we also crawl the profile description of u_i^j, 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 $name_i$ & $login_i$ 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:

(i) the profile description of the user contains a hyperlink to a profile that was discovered in the **Explicit** or **Direct** step,

(ii) 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,

(iii) 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 employed 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 14824, 6980, 3129, 1125, 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://stackexchange.com/
[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_i. 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:

$$age_u = w_0 + \sum_{t \in M_{age}} w_t \times \frac{freq(t, D_u)}{|D_u|}. \qquad (1)$$

Here, $|D_u|$ is the number of tokens in D_u, w_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 estimation is ≥ 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 everyone[12]. `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 Archive*[13] 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 types*[14], 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" : "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

[11] The models are available at http://www.wwbp.org/data.html

[12] Data traces about private repositories are only available to the respective repository owner.

[13] https://www.githubarchive.org/

[14] https://developer.github.com/v3/activity/events/types/

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-universiy 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.

In Table 2 we split up the matches found according to the type of matching performed (Explicit, Direct or Fuzzy). On Gravatar, we relied exclusively on Explicit matching, while the vast majority of learners on GitHub and StackExchange were also identified in this manner, with Direct and Fuzzy matching contributing little. On these platforms, users' email addresses are either directly accessible (Gravatar and GitHub) or indirectly accessible (StackExchange provides the MD5 hash of its users' email addresses[15]). In contrast, the LinkedIn and Twitter platforms do not publish this type of user information and thus the majority of matches are fuzzy matches. Overall, the Direct approach has the least impact on the number of matches found.

To verify the quality of our matchings, for each platform, we sampled 50 users identified through any matching strategy and manually determined whether the correct linkage between the learner's edX profile and the Social Web platform was found (based on the inspection of user profile information and content). We found our matching to be robust: of the 100 samples, we correctly linked 93 (StackExchange), 87 (GitHub), 97 (Twitter) and 95 (LinkedIn) respectively.

	Explicit	Direct	Fuzzy	Overall
Gravatar	7.81%	—	—	7.81%
StackExchange	4.32%	0.01%	0.25%	4.58%
GitHub	9.04%	0.02%	1.23%	10.29%
LinkedIn	—	0.48%	5.41%	5.89%
Twitter	—	0.67%	7.12%	7.78%

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.

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%)[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[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 *Re-*

[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.

[16]We generated these numbers based on Twitter's language auto-detect feature.

[17]Based on the ground truth data provided by 20,311 edX learners, the prediction precision is 36.5%.

MOOC	Year	#Learners	Gravatar	Stack-Exchange	GitHub	LinkedIn	Twitter
Solar Energy	2013	67,143	†3,510	1,570	†3,677	2,997	†3,828
Solar Energy	2014	34,524	†1,923	874	†2,229	1,625	†2,152
Solar Energy	2015	26,178	1,147	435	1,184	1,181	†1,557
Introduction to Water Treatment	2013	34,897	1,559	508	1,198	1,362	1,741
Introduction to Drinking Water Treatment	2014	10,458	457	129	430	427	†548
Introduction to Water and Climate	2014	9,267	†561	154	†510	452	†558
Technology for Biobased Products	2014	9,811	†545	149	†511	452	†547
Next Generation Infrastructures Explores the challenges of global & local infrastructure (ICT, energy, water and transportation).	2014	20,531	†1,438	583	†1,451	†1,155	†1,447
Functional Programming Teaches the foundations of functional programming & how to apply them in practice.	2014	38,682	‡9,087	‡8,477	‡16,220	‡4,274	‡6,801
Data Analysis Teaches data analysis skills using spreadsheets and data visualization.	2015	33,547	†2,392	1,165	‡4,432	†2,469	†2,800
Pre-university Calculus	2015	28,015	†1,928	960	†2,477	†1,406	†2,064
Introduction to Aeronautical Engineering	2014	20,481	†1,134	605	†1,373	921	†1,192
Introduction to Aeronautical Engineering	2014	13,197	†699	318	†837	609	†788
Topology in Condensed Matter Provides an overview of topological insulators, Majoranas, and other topological phenomena.	2015	4,231	†277	†292	†600	201	†302
Framing Analyzes how politicians debate and what the underlying patterns are framing and reframing.	2015	34,018	†2,838	1,034	†2,597	†2,211	†2,657
Solving Complex Problems How to solve complex problems with analytics based decision-making & solution designs.	2014	32,673	†2,803	1,620	‡3,928	†1,934	†2,647
Delft Design Approach How to design meaningful products & services.	2014	13,543	†1,319	514	‡1,376	†1,085	†1,124
Responsible Innovation How to deal with risks and ethical questions raised by the development of new technologies.	2014	10,735	†877	274	†800	†713	†753
Unique Users		329,200	25,702	15,135	31,478	19,405	25,620

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.

sponsible 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[18]. 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

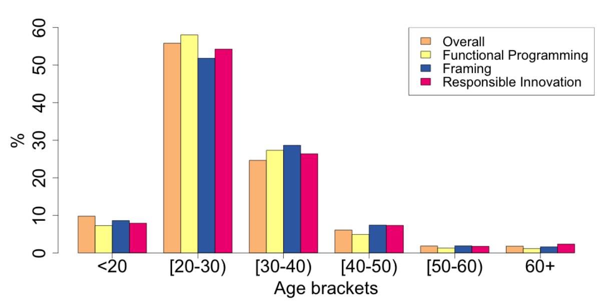

Figure 2: Percentage of our Twitter users across eight age brackets. The "Overall" user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific.

[18]The prediction precision is 78.3% based on the ground truth provided by 20,739 edX learners.

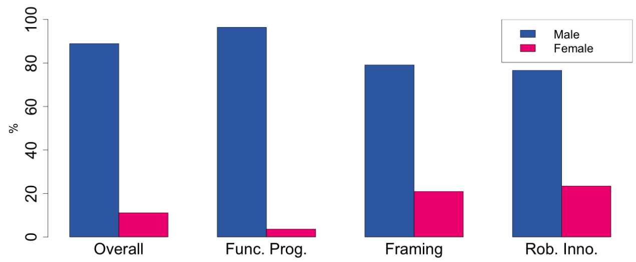

Figure 3: Percentage of our Twitter users of each gender. The "Overall" user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific.

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.

Figure 4: Overview of the most frequent job title bigrams among the learners of the *Data Analysis* (top), *Delft Design Approach* (middle), and *Responsible Innovation* (bottom) MOOCs.

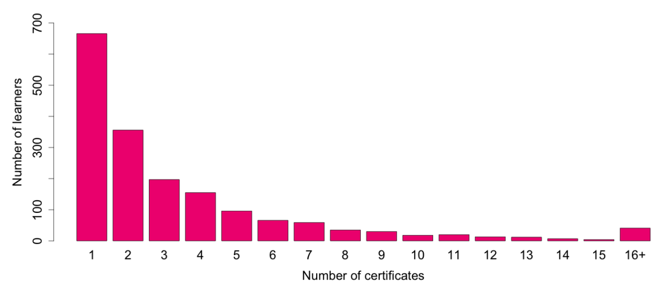

Figure 5: Fraction of learners displaying n numbers of MOOC certificate.

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 skill 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.

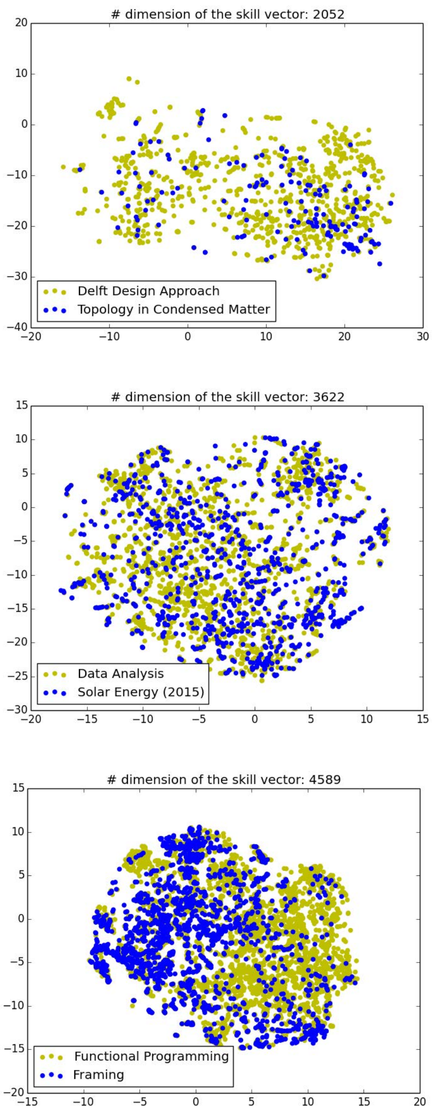

Figure 6: t-SNE based visualization of `LinkedIn` skill vectors for pairs of MOOCs. Each data point represents one skill vector (i.e. one user).

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 38,682 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 frquency 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[19]. 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#

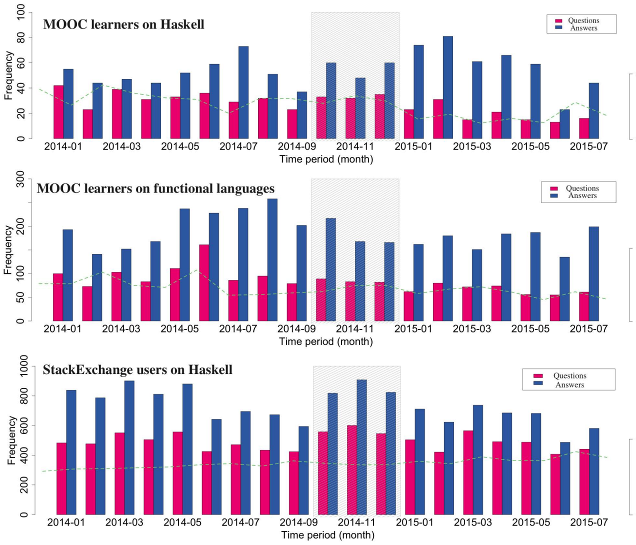

Figure 7: Overview of the number of StackOverflow questions and answers posted on a monthly basis between January 2014 and July 2015 by (i) our MOOC learners [top and middle], and (ii) all StackExchange users [bottom] for Haskell [top, bottom] and the nine major functional languages [middle]. Marked in gray is the time period of the *Functional Programming* MOOC. The dashed green line indicates the ratio of $\frac{Questions}{Answers}$ in each month.

Finally, we also explored whether our MOOC learners have a similar expertise-dispensing behaviour as the general StackOverflow user population. To this end, we make use of the two expertise use types proposed in [32]: *sparrows* and *owls*. In short, sparrows are highly active users that contribute a lot but do not necessarily increase the community's knowledge. Their answers, while relevant, might be of low quality or low utility as they are motivated by reputation scores, and gamification elements of the platform. Owls on the other hand are users that are motivated to increase the overall knowledge contained in the platform. Owls are experts in the discussed topic, and they prove their expertise by providing useful answers to important and difficult questions. [32] proposed the *mean expertise contribution* (MEC) metric to capture measure expertise, based on answering quality, question debatableness and user activeness. Based on this metric, they determined 10.0% of the StackOverflow users to be owls. We derived MEC for our set of *Functional Programming* MOOC learners that are active on StackOverflow and found 21.0% of them to be owls. Thus, the average MOOC learner is not only interested in gathering knowledge, but also in distributing knowledge to others, on a deeper level than the average StackExchange user.

5.4 Learners on GitHub

Finally, with respect to **RQ3**, we consider the concept of learning transfer, introduced in Section 3.3. As a social coding platform, GitHub is most suitable to explore programming-heavy MOOCs, thus we restrict our analysis (as in the previous section) to the *Functional Programming* MOOC. We are particularly interested in the extent of the learners' functional programming after the end of the MOOC — our MOOC learners ask fewer topic-related questions (on StackExchange) over time, but does it also mean they program less in the language? To his end, we explored th 6,371,518 PushEvents we extracted from our MOOC learners between January 1, 2013 and June 30, 2015. Figure 5.4 provides a first answer to this question. The amount of Haskell programming by our learners was increasing slowly over time even before the start of the MOOC. A spike is visible in November 2014 (weeks 3-6 of the *Functional Programming* MOOC) and immediately after the end of the MOOC the contributions increase. However, by March 2015, i.e. three months after the end of the MOOC, the contributions are beginning to decline again towards nearly pre-MOOC levels.

In contrast to Haskell, we observe a sharp rise in "Scala" (the main functional language in industry) activities after the end of the MOOC which peak in November 2015. These functional activities are not evenly spread across all users though, only 32% of the users we identified on GitHub exhibited any type of functional language activities after the end of the *Functional Language* MOOC.

In the future, we will not only consider the addition of lines of codes in a particular language, but also perform fine-grained code analyses to investigate which specific concepts the learners picked up on in the MOOC and later employed in their own works.

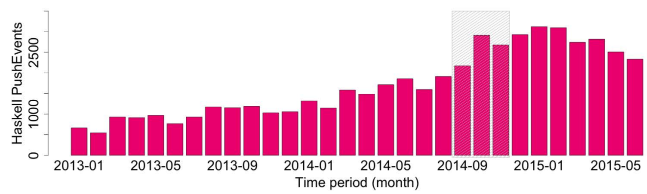

Figure 8: Month-by-month GitHub contributions in the Haskell language by the *Functional Programming* MOOC learners identified on GitHub.

6. CONCLUSIONS

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

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. 2011.

[2] C. Alario-Hoyos, M. Pérez-Sanagustín, C. Delgado-Kloos, M. Muñoz-Organero, A. Rodríguez-de-las Heras, et al. Analysing the impact of built-in and external social tools in a mooc on educational technologies. In *Scaling up learning for sustained impact*, pages 5–18. 2013.

[3] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. Personality and patterns of facebook usage. In *Web Science '12*, pages 24–32, 2012.

[4] M. Balasubramanian and E. L. Schwartz. The isomap algorithm and topological stability. *Science*, 295:7–7, 2002.

[5] D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[6] A. Bermingham and A. F. Smeaton. On using twitter to monitor political sentiment and predict election results. *IJCNLP 2011 Workshop*, pages 2–10, 2011.

[7] D. Coetzee, A. Fox, M. A. Hearst, and B. Hartmann. Should your mooc forum use a reputation system? In *CSCW '14*, pages 1176–1187, 2014.

[8] D. Coetzee, S. Lim, A. Fox, B. Hartmann, and M. A. Hearst. Structuring interactions for large-scale synchronous peer learning. In *CSCW '15*, pages 1139–1152, 2015.

[9] J. Cruz-Benito, O. Borrás-Gené, F. J. García-Peñalvo, Á. F. Blanco, and R. Therón. Extending mooc ecosystems using web services and software architectures. In *Interacción '15*, pages 52:1–52:7, 2015.

[10] L. Eldén. *Matrix methods in data mining and pattern recognition*, volume 4. SIAM, 2007.

[11] F. J. García-Peñalvo, J. Cruz-Benito, O. Borrás-Gené, and Á. F. Blanco. Evolution of the conversation and knowledge acquisition in social networks related to a mooc course. In *Learning and Collaboration Technologies*, pages 470–481. 2015.

[12] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of mooc videos. In *L@S '14*, pages 41–50, 2014.

[13] P. J. Guo and K. Reinecke. Demographic differences in how students navigate through moocs. In *L@S '14*, pages 21–30, 2014.

[14] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500, 2014.

[15] A. D. Ho, I. Chuang, J. Reich, and C. A. Coleman *et al.* Harvardx and mitx: Two years of open online courses fall 2012-summer 2014. *SSRN 2586847*, 2015.

[16] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[17] D. J. Hughes, M. Rowe, M. Batey, and A. Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[18] S. Joksimović, N. Dowell, O. Skrypnyk, V. Kovanović, D. Gašević, S. Dawson, and A. C. Graesser. How do you connect?: Analysis of social capital accumulation in connectivist moocs. In *LAK '15*, pages 64–68, 2015.

[19] S. Joksimović, V. Kovanović, J. Jovanović, A. Zouaq, D. Gašević, and M. Hatala. What do cmooc participants talk about in social media?: A topic analysis of discourse in a cmooc. In *LAK '15*, pages 156–165, 2015.

[20] K. Jordan. Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, 15(1):133–160, 2014.

[21] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012.

[22] D. Koller, A. Ng, C. Do, and Z. Chen. Retention and intention in massive open online courses. *Educause Review*, 48(3):62–63, 2013.

[23] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, 12:511–514, 2012.

[24] D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder. " how old do you think i am?" a study of language and age in twitter. *ICWSM '13*, pages 439–448, 2013.

[25] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras. An analysis of the user occupational class through twitter content. pages 1754–1764. The Association for Computational Linguistics, 2015.

[26] M. Sap, G. Park, J. C. Eichstaedt, M. L. Kern, D. Stillwell, M. Kosinski, L. H. Ungar, and H. A. Schwartz. Developing age and gender predictive lexica over social media. *EMNLP '14*, pages 1146–1151, 2014.

[27] G. Siemens. Connectivism: A learning theory for the digital age. 2014.

[28] G. Silvestri, J. Yang, A. Bozzon, and A. Tagarelli. Linking accounts across social networks: the case of stackoverflow, github and twitter. In *International Workshop on Knowledge Discovery on the WEB*, pages 41–52, 2015.

[29] J. H. Tomkin and D. Charlevoix. Do professors matter?: Using an a/b test to evaluate the impact of instructor involvement on mooc student outcomes. In *L@S '14*, pages 71–78, 2014.

[30] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[31] T. van Treeck and M. Ebner. How useful is twitter for learning in massive communities? an analysis of two moocs. *Twitter & Society*, pages 411–424, 2013.

[32] J. Yang, K. Tao, A. Bozzon, and G.-J. Houben. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *UMAP'14*, pages 266–277. 2014.

[33] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll. Understanding student motivation, behaviors and perceptions in moocs. In *CSCW '15*, pages 1882–1895, 2015.