The Advancing MOOCs for Development Initiative (AMDI):

CourseTalk Website Data Analytics Report

Released: December 2015

Authors:

Lucas Koepke
Maria Garrido
Technology & Social Change Group
University of Washington Information School
Seattle, WA

Scott Andersen
IREX
Washington, DC.

Contact:

Maria Garrido

migarrid@uw.edu

TECHNOLOGY &
SOCIAL CHANGE GROUP

UNIVERSITY of WASHINGTON
Information School

TECHNOLOGY & SOCIAL CHANGE GROUP

The Technology & Social Change Group (TASCHA) at the University of Washington Information School explores the design, use, and effects of information and communication technologies in communities facing social and economic challenges. With experience in 50 countries, TASCHA brings together a multidisciplinary network of social scientists, engineers, and development practitioners to conduct research, advance knowledge, create public resources, and improve policy and program design. Our purpose? To spark innovation and opportunities for those who need it most.

CONTACT

Technology & Social Change Group
University of Washington Information School
Box 352115
Seattle, WA 98195

Telephone: +1.206.616.9101
Email: tascha@uw.edu
Web: *tascha.uw.edu*

ABOUT THE AUTHORS

Maria Garrido is a Research Assistant Professor at the University of Washington Information School.

Lucas Koepke is a Data Analyst at the Technology & Social Change Group.

ACKNOWLEDGEMENTS

COPYRIGHT, LICENSING, DISCLAIMER

Table of Contents

Table of Figures

Executive Summary

In January of 2015, through funding from the United States Agency for International Development (USAID) and CourseTalk via the Advancing MOOCs for Development Initiative (AMDI), IREX and the University of Washington's TASCHA program (Technology & Social Change Group) agreed to create a report for CourseTalk, the world's largest source for reviews of online courses and MOOCs, offering access to more than 100,000 reviews of 40,000 courses from 65 MOOC providers.

The aim of this report, hereafter referred to as *The CourseTalk Website Data Analytics Report*, was to provide a baseline analysis of the CourseTalk user ecosystem. The report affords a comprehensive look at the demographic composition of CourseTalk's registered users as gleaned from website analytics, narratives they have shared through course reviews, and the behavior they exhibit when browsing on the CourseTalk website.

The objectives of *The CourseTalk Website Data Analytics Report* were later amended at the request of the United States Agency for International Development (USAID) to include key search terms related to workforce development issues in emerging economies. This would be accomplished by mapping the geographic and demographic sphere of CourseTalk users, researching the CourseTalk database of nearly 90,000 total course reviews (as of May 2015), conducting an analysis of keyword searches, a synopsis of reviews by subject matter, information gleaned from datasets based on user search terms, and a succinct list of potential recommendations to improve website analytic data. Among the key findings were that priorities should be given to registering new users allowing them greater opportunities to stay active, and that a distinct focus should be placed on profile completion to allow better insights into core MOOC user group motivation and behavior.

Based on these findings and other identified trends, the *CourseTalk Website Data Analytics Report* concludes with a list of recommendations to improve CourseTalk's ability to reach MOOC users in developing economies. For example, that CourseTalk should prioritize reviewer engagement to foster a sense of community, that increased integration should be made with social media, and that CourseTalk should enhance its data analytics capabilities to assess user engagement data across core activities by key user groups.

The *CourseTalk Website Data Analytics Report* also aimed to achieve a greater understanding of both 'who' was using MOOCs and 'why' they were being used in the three target countries of the *Advancing MOOCs for Development Initiative*; specifically, the Philippines, Colombia and South Africa. Unfortunately, there were only a small sample of self-identified reviews from these countries; a total of 4 from Colombia, 5 from the Philippines, and 5 from South Africa, making any wholesale data allusions next to impossible for the purposes of the Website Analytics Report. However; the findings and recommendations presented by TASCHA in this report suggest potential improvements that could be made in both targeting and tracking those MOOC users in developing countries who benefit by using the CourseTalk website ecosystem. The findings likewise afford select insights to assist MOOC related

marketing campaigns within the three countries affiliated with the *Advancing MOOCs for Development Initiative*.

Finally, it is the sincere hope of both IREX and the University of Washington's TASCHA program that the *CourseTalk Website Analytics Report* can be used in combination with the soon to be released *Advancing MOOCs for Development Initiative's 'MOOC Baseline Needs Assessment Report'*, that is presently being conducted in the Philippines, Colombia and South Africa. The MOOC Baseline User Research, (anticipated in January 2016), comprises MOOC user and nonuser survey data sets, key informant interviews and focus groups to improve both internal CourseTalk data collection of user demographics, reveal motivations for taking MOOCs, and to improve the marketing and promotion of MOOCs in the developing world.

Scott Andersen

Director, The Advancing MOOCs for Development Initiative

1 Introduction

The Advancing MOOCs for Development Initiative aims to generate data on the use of Massive Open Online Courses (MOOCs) in developing countries to better understand their potential for improving employment opportunities for men and women between 18 and 35.

This report (The CourseTalk Website Data Analytics Report) is the first installment in a series of research products designed to meet the needs of the Advancing MOOCs for Development Initiative. It will be followed by the MOOCs Baseline Needs Assessment Report (January 2016), a wide-ranging research process conducted through surveys of MOOC users and non-users, key informant interviews with thought leaders in the domains of technology for workforce development, and focus groups in the Philippines, Colombia and South Africa. The final objective of this research process is to advance the understanding of MOOC enrollment and completion rates, and increase the utility of MOOCs in workforce development issues for young adults in the developing world.

The analysis presented in this report takes a comprehensive look at the demographic composition of CourseTalk registered users, the narratives they share through their course reviews, and the behavior they exhibit when browsing and searching for topics they are interested in on the CourseTalk website.

Originally, this research was designed to serve as a baseline for research activities that would follow throughout the course of the Advancing MOOCs for Development Initiative. As originally conceived, an analysis of CourseTalk's user ecosystem would be employed to classify the entire user base into distinct data clusters, creating a comprehensive picture of the different types of MOOC users, that would better allow further contrast among the categories of users based on their level of enrollment in courses, i.e., their keyword search and browsing behavior, and the differences among various geographical regions.

The goal of this proposed user cluster analysis was to help describe the CourseTalk user base, and to serve as a foundation for designing the MOOC user and potential-user surveys. However, given the limited nature of the data available with regards to respondents in the three target countries (the Philippines, Colombia and South Africa) on the CourseTalk website, it was determined that this approach would not be feasible, and instead, the analytical framework was redesigned to accommodate the data available for overall MOOC usage and course reviews.

With this redesigned research approach of focusing upon all MOOC users from all CourseTalk providers, the findings that follow provide a thorough overview of CourseTalk user demographics, an analysis of course reviews written by CourseTalk users, and an understanding of how CourseTalk users browse and search the site for topics of interest.

The CourseTalk Website Data Analytics report is structured as follows:

A.) The **first** part describes the demographic nature of CourseTalk users in terms of their gender, age, employment status, and geographical location when available. This section serves as the

foundation for understanding user behavior and participation in the analysis of course reviews presented in the rest of the report.

B.) The **second** part of the report elaborates on three techniques to gauge narratives of employability, workforce development, MOOC learning features, and CourseTalk mentions in almost 90,000 reviews that CourseTalk made available for this research. The analysis elaborates on the potential of each technique to understand CourseTalk user behavior through the lens of the review narrative and presents the different findings, as well as the pros and cons of applying each of the three techniques in review analysis.

C.) The **third** part of the report takes a closer look how users search by keyword, outlining the trends, main keywords searched in the timeframe of the available data, and the distribution of search keywords in different geographical regions.

D.) The **fourth** and final part presents a set of recommendations for CourseTalk and provides some guidelines for AMDI's future research and communications strategy.

2 Mapping the demographic sphere of CourseTalk users

The user profile data is the information provided by people who have registered an account on the CourseTalk website. The only required fields to register are a name and email address, while optional fields include birth date, country, employment status, gender, and more. Although data from these optional fields will have an unknown bias due to the high non-response rate, the information provides a baseline demographic profile with which to compare findings from the MOOC user survey and also from past research.

The data available includes all users who registered starting with CourseTalk's creation in October 2012 through May 11, 2015, when access to the database was granted. Since all this data comes from optional response fields during registration, there are unknown biases in who answers which questions, and thus, it cannot be considered a reliable description of the entire CourseTalk user population. Despite this issue, there are interesting pieces of information to consider for future direction of the services and products offered by the company.

First, 27% of the total registered user population) reported their gender (69.3% male, 30.7% female), and approximately 25% of users reported their birthdate. As of this writing, the median age for registered users was 33.3 years. This varies slightly by gender, with male users averaging 32.2 and female users 35.4 (Figure 1).

Occupational/User demographics afforded a substantially higher completion rate with 45% of all respondents (Table 1). The largest single response category was "professional" (45.6%), followed by "students," which combined account for 41.4% of the responses. Together, these categories account for 87% of the responses. The next largest category, "professor/instructor," is interesting with 6.4% of the total responses. This finding is particularly useful for the AMDI initiative, combined with the review data, by showing that users come from a variety of academic and professional backgrounds. In particular, the participation of professors and instructors in CourseTalk's MOOC sphere, which comprises almost 7% of the total population, is worth further exploration. Previous research shows that there is an increasing demand for MOOCs that cater more closely to the learning needs of this group. Courses on pedagogical innovations, multimedia use in classrooms, conversational learning, and techniques for teaching in a MOOC environment are increasingly populating the sphere of MOOCs. Some companies are starting to offer a variety of courses, pedagogical platforms, spaces for social learning among professionals in the education sector, and a variety of teaching tools geared specifically for this niche market. For example, Telefonica de Espana Educacion Digital, offers a variety of products and services for the education professionals in Spain and Latin America.

Table 1 Occupational/User type for registered CourseTalk users

	Percent
Professional	45.6
College Student	28.3
Professor/Instructor	6.4
K-12 Student	3.9
Parent	2.6
Total	**100**

Only 16% of all registered users entered a response in the "country" field, the number of responses for the top 10 countries is in Table 2. By far the largest number are in the United States, comprising 34.1% of the responses, and almost 3 times more users than the 2nd highest (India, with 11.8% of the responses). The rest of the list includes countries such as Brazil (5%), Egypt (3.1%), and China (1.5%). A total of 4 reviews from Colombia, 5 from the Philippines, and 5 from South Africa were available, while 3% of all reviews in which the authors profile had a foreign geographic location attached.

Table 2: Top ten countries for registered CourseTalk users

Country	**Percent**
United States	34.1
India	11.8
Brazil	5.0
United Kingdom	3.6
Egypt	3.1
Canada	2.8
Spain	2.0
Australia	1.8
China	1.5
France	1.4

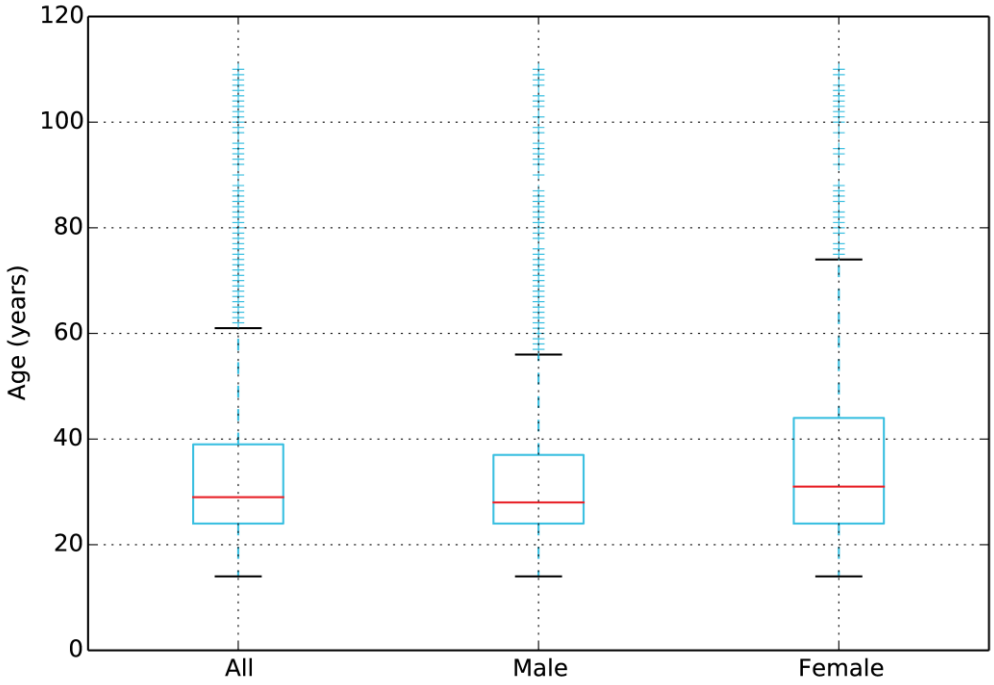

Figure 1: Age of registered CourseTalk users

Note: A handful of birthdates are very likely inaccurate resulting in unusually high calculated ages.

3 CourseTalk course reviews

The CourseTalk database contains a vast collection of course reviews, comprising 86,294 total records. Fields in these data include the profile ID for the writer (if registered), the date of the review, the course reviewed, the review text, the rating the reviewer assigned to the course, and the number of people who labeled that review as "helpful" or "not helpful." Since some reviews did not appear to be in English, a preliminary data cleaning pass using Python and the Natural Language Toolkit classified the language of each review. This left 79,196 reviews in English.

The size of this dataset means a purely qualitative approach is impractical, since reading even a fraction of these reviews would be prohibitively time-consuming. To mitigate this challenge, three key methods isolate important features in the body of reviews:

- The first *is a simple keyword search* to extract reviews for further reading using a single word or phrase.

- Second, an *advanced keyword search* scores reviews based on a topic-level match, where each topic is a collection of keywords and phrases.
- Third, a *topic model* is used to discover hidden topics using the entire set of reviews. Before discussing these results, the following section provides descriptive statistics of the review data to give a general understanding of what the data is like. (For a detailed description of the methodology, please see Appendix 1)

3.1 Overview of CourseTalk course reviews

Of the 79,196 English reviews, 6,273 (7.9%) are associated with a registered user, while the remaining 72,923 (92.1%) are anonymous. The low number of reviews by registered users is a potential target metric for CourseTalk to increase, since 27% of reviews by registered users are labeled as "helpful" compared with just 0.7% of the anonymous reviews.

Another metric to gauge participation in the CourseTalk platform is the number of reviews submitted per registered user. Since these reviews are significantly more likely to be helpful, and thus arguably of higher quality than anonymous reviews, increasing that number is desirable as a business strategy. Of the registered users who wrote one or more reviews, the average number of reviews written was 1.7, with a minimum of 1 and a maximum of 93.

In terms of the number of different courses present in the data, a total of 9,158 unique courses are reviewed. With some overlap, 8,604 unique courses were reviewed anonymously and 959 were reviewed by registered users. Anonymous reviews again account for the bulk of the courses being reviewed, with the number increasing dramatically from 2013 to 2014 (Figure 2). The same is not true for registered reviews, with the number of unique courses reviewed declining after May 2014.

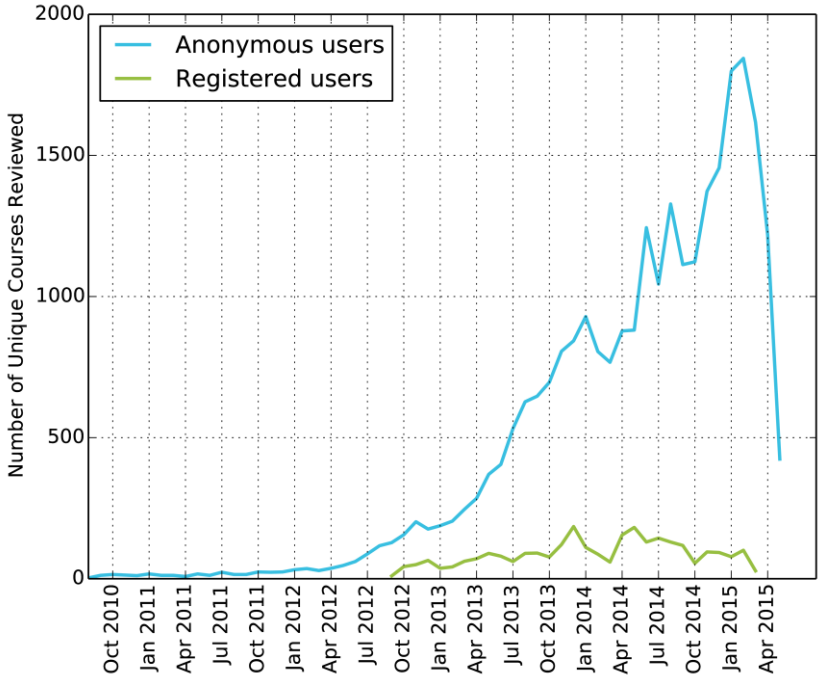

Figure 2: Number of unique courses reviewed per month, for anonymous and registered users

Digging deeper into the types of courses reviewed by registered users, 917 of the unique courses reviewed by registered users have an associated course name, rather than just an ID number. The top 20 in terms of the number of reviews are given in Table 3. Computer science courses are certainly common, for example *An Introduction to Interactive Programming in Python*, *Machine Learning*, and *Artificial Intelligence*. However the other courses cover a surprising variety of other topics. *A Beginner's Guide to Irrational Behavior* is second place with 242 reviews, followed by *Epidemics - the Dynamics of Infectious Diseases* with 179 reviews. Other top courses include *Greek and Roman Mythology*, *Comic Books and Graphic Novels*, and *Developing Innovative Ideas for New Companies: The First Step in Entrepreneurship*.

Table 3: Top 20 courses reviewed by registered users, sorted by number of reviews

Course name	# reviews	Percent
An Introduction to Interactive Programming in Python	476	7.6
A Beginner's Guide to Irrational Behavior	242	3.9
Epidemics - the Dynamics of Infectious Diseases	179	2.9
Modern & Contemporary American Poetry	152	2.4
Design: Creation of Artifacts in Society	151	2.4
Developing Innovative Ideas for New Companies: The First Step in Entrepreneurship	121	1.9
The Science of the Solar System	107	1.7
An Introduction to Operations Management	105	1.7
Learn to Program: The Fundamentals	85	1.4
Maps and the Geospatial Revolution	85	1.4
Greek and Roman Mythology	82	1.3
Machine Learning	79	1.1
Critical Thinking in Global Challenges	72	1.1
Gamification	66	1.1
Introduction to Engineering Mechanics	66	1.1
Think101x: The Science of Everyday Thinking	51	0.8

Comic Books and Graphic Novels	48	0.8
CS-191x: Quantum Mechanics and Quantum Computation	46	0.7
CS188.1x: Artificial Intelligence	45	0.7
Internet History, Technology, and Security	43	0.7

The following sections focus on the actual content of the reviews. This is challenging for several reasons but largely because the number of reviews requires some form of reduction before engaging in a qualitative analysis. The data reduction is achieved through two complimentary methods, the first extracting interesting reviews using a keyword search and the second discovering hidden topics in all the reviews using a topic model.

3.2 CourseTalk review analysis by keyword

Extracting reviews containing certain keywords progressed in two stages. The first was a simple exploratory approach, meant to capture and illuminate general themes surrounding MOOCs and the motivations for taking them, using a simple search using a single word or phrase to extract reviews and then reading the results. This is intended to build a baseline picture of what is in the review data, without blindly starting to read random reviews. The advanced keyword search will extend and refine this baseline picture. This algorithm uses a group of keywords to score the relevance of the reviews for each of three topics, and then display the highest scoring reviews for reading and analysis.

The baseline review summary used three words/phrases to extract reviews:

1. "CourseTalk" (13 results)
2. "first MOOC" (185 results)
3. "new job" (40 results)

These were chosen in three key areas. References to CourseTalk may help show if there are usability issues, or if the reviews on the CourseTalk site influenced the decision to take (or not take) a certain MOOC. Motivations for taking a MOOC are an integral part of this research project, and reviews containing the phrase "first MOOC" will hopefully capture a part of this landscape, such as the reasons for first time users to participate in MOOCs. Similarly for the phrase "new job," the goal is to start learning how people talk about employability and professional development in the context of MOOCs.

Of the 13 reviews that specifically mention the phrase "CourseTalk," a few come from users linking to a different course page or review on the CourseTalk website, and are thus not relevant here. The others, however, offer information relating both to the usability of the website and the benefits from the reviews. For example, here are solid recommendations for the reviews on CourseTalk as a resource:

"…this was my first MOOC but based on this experience and the feedback I have found on coursetalk.org, I have already signed up for 3 more courses over the next 6 months." (Introduction to Data Science, 2013-06-23)

And

"...since there was no resource like Coursetalk when I completed the course last June, I thought I'd now add my praise to the mix. MITx 6.002 is truly the gold standard to which all MOOCs should aspire." (6.002x: Circuits and Electronics, 2013-04-29)

This search also surfaced a couple of issues with the CourseTalk website:

"I quite wish coursetalk would enable a "comment" function so people would be able to discuss others' reviews." (Design of Computer Programs, 2013-12-08)

And

"Note: I would say I expended 10-15 hours a week but coursetalk, keeps whining and asking for a whole number." (CS-191x: Quantum Mechanics and Quantum Computation, 2013-04-22)

The second keyword phrase, "first MOOC," was chosen to shed light on the reasons and motivations to begin taking MOOCs. The 185 reviews containing this phrase cover a variety of courses, from epidemiology to portfolio management to thermodynamics. There are numerous examples in this set of a positive first experience with MOOCs. For example:

"I am a recently retired medical librarian. I wish I could have had an opportunity like this while I was working. It helped to appreciate the broader picture of epidemics and to see where the many information requests from library patrons might fit within the infectious disease landscape. This is my first MOOC and based on the experience I will do others. Loved the AUA (ask us anything) videos!" (Epidemics - the Dynamics of Infectious Diseases, 2013-12-04)

And

"I am a retired RN and this is my first MOOC. I really enjoyed the course and felt I learned a lot of new information and reviewed information that I had learned long ago. I consider myself a life long learner and am interested in the sciences. I have introduced many of my friends to the MOOC platform of learning. As this was my first course, I cannot compare it to any others but felt it was well thought out for people in the medical field or people interested in disease." (Epidemics - the Dynamics of Infectious Diseases, 2013-12-11)

And

"This was my first MOOC and I started it with little experience and knowledge of the subject matter. I chose it because it looked interesting, I wanted to return to learning and MOOC's seemed to be an obvious way to do that." (Comic Books and Graphic Novels, 2013-11-16)

One feature that emerged as playing a positive role in users' MOOC experience, in some cases unexpectedly so, is the online community created alongside the lectures and other formal course material. The following excerpts highlight comments about this.

"This was my first MOOC, and one aspect of the course that I did not expect but really loved was the role of the diverse student community in discussing, evaluating each others' work, and encouraging each other throughout the class. Really great! Not sure if that's typical for MOOCs but it was a huge part of the experience for me." (Design: Creation of Artifacts in Society, 2013-06-20)

"This course was my first MOOC. I found it extremely engaging as an adult learner. I signed up for the class for professional development, to learn about digital learning tools, and to create a personal learning network (PLN). All of those goals and more were met through this course. … I found the constructivist methodology extremely life-giving to me. The level of engagement from other students through the forums that were suggested by the instructors (FB, G+, Twitter) was beyond my expectations. I was able to connect with other students and to build relationships." (E-learning and Digital Cultures, 2013-03-11)

Some reviews bring up negative points about the MOOC, which provide important information for the research. In some cases, these negative opinions helped influence questions in the survey instruments, particularly around potential barriers or challenges facing other potential MOOC users. These challenges in particular are language, poor quality, and negative perceptions around the efficacy of online learning compared to traditional methods.

"This is my first MOOC - its years since I last took a course, and English is my second language, so the Course in Gamification is hard for a foreigner like me. I like the Course because it is a thorough review of the Gamification phenomenon. - The multi answer questions is hard to get right - partly, I think because it goes on in a foreign language, for me." (Gamification, 2014-03-09)

"This is the first MOOC that I have taken and I had some background on Machine Learning (PhD on automatic speech recognition) before I took this course. … Overall I feel that I would have learned more using the time on studying by myself. However, being my first MOOC it is possible my expectations were too high." (Web Intelligence and Big Data, 2013-05-25)

The following review does not find fault with the MOOC itself, but rather states that a MOOC "isn't real teaching," despite praising the professor:

"Dr. Rogers is obviously a top professor, knows the subject cold, etc. But, this being my first MOOC experience, I wouldn't recommend it--not for the course material or the professor, but for the fact that this isn't real teaching, which requires a real, two-way engagement, not the professor talking to a camera, and student writing some blog post or answering general questions." (HIST229x: Was Alexander Great? The Life, Leadership, and Legacies of History's Greatest Warrior, 2014-06-30)

Overall, this collection of quotes shows that although this is a small subset of the whole set of reviews, MOOCs do play an important role in learning. These examples also highlight some interesting motivations for starting a MOOC, such as "to return to learning." Other reviewers talk about being a "life long learner" or "adult learner."

The search for the phrase "new job" extracted 40 reviews for reading. Many of these are from courses focused on job hunting, salary negotiations, LinkedIn proficiency, and other areas inherently associated

with searching for jobs. However, some of the reviews directly touch on skill building for a new job or career, relevant from the perspective of this initiative. A sample of these reviews is shown below.

"In my job I used to develop reports with SSRS but the last time I did that was 2 years ago. In order to prepare for a new job opportunity this has been a good refresher course for me to do and as a student you are guided every step of the way." (Learn SQL Reporting Services Beginning Report Training, 2013-07-02)

"I started a new job and needed to learn python. With this class i feel i have the basics down so I can continue on my own to start reading the python code that my job requires." (An Introduction to Interactive Programming in Python, 2013-06-17)

"I'm so pumped to switch careers I quit my full time job to go through this course and switch careers. Looking forward to finishing the course and starting a new job soon, and looking forward to more courses by Victor!" (Become a Web Developer from Scratch, 2014-09-25)

"As someone who was really not familiar with InDesign at all prior to being thrown into editing a document for a new job, this class was really helpful to me to get my bearings. I did not upload a project, to be honest I'm not a designer, but did find all the tips and information VERY helpful." (Basic InDesign: Layouts, Type, and Images, 2014-08-13)

"I'm starting a new job that requires a lot of work in Linux. This course has helped me regain much of the knowledge I forgot years ago. It has also helped me catch up with some new tools. The teacher did an excellent job delivering the lectures. I can't recommend this course enough." (Learn Linux in 5 Days and Level Up Your Career, 2014-12-19)

These excerpts highlight how for some people, MOOCs have played a critical role in their jobs and advancing their career goals. Whether re-learning forgotten skills or learning new ones, MOOCs are helping these users succeed. Do note that "success" may not necessarily equate with "completion;" it is presumed the reviewer for the InDesign class did not complete the course by uploading a project, but still achieved their goal of applying their new skill to a task at their job. This theme will be further explored in the analysis of the Advancing MOOCs for Development Initiative's Baseline Needs Assessment Report user and potential-user survey data and complemented with the qualitative data collected through the interviews to key informants in each of the three target countries.

3.3 A closer look at reviews using an advance technique of keyword search

The method for scoring each review's relevance for a certain topic is discussed in detail in the methodology section (See Appendix 1), but essentially uses a list of keywords related to a topic to score each review based on the total number of keywords matched (a single keyword repeated multiple times in a review also contributes to the score). The idea is that a review that matches multiple keywords, or a single keyword multiple times, is potentially more interesting than a review that only matches one

keyword. The "score" can then be used as to separate reviews into different tiers of importance for a given topic.

For example, suppose the keywords for a topic are "change career," "new job," and "job application." The review *"I needed to change careers and find a new job, taking this MOOC really helped refresh my skills"* would score a 2 by matching two keywords (each matched once). Alternately, the review *"Learning Python helped me find a new job, and I recommend it to anyone looking for a new job as a programmer"* would also score a 2, this time by matching a single keyword twice.

This extends the simple word or phrase search, since a topic can't necessarily be condensed to just one keyword. The three topics used are listed in Table 4. Initially, single words such as "salary," "income," and "hired" were included in the employment topic. They were updated to phrases for the final list because there are a number of courses (such as on salary negotiation or filling out income tax forms) that use some of those basic keywords without really being relevant for professional development. The goal of the employment topic search was to focus more on courses being used for advancing employment opportunities through skill development, not learning skills related to employment such as salary negotiating.

Table 4: Topics and their associated keywords used to score the reviews using the advanced keyword search function

Topic name	Keywords
Employment	employment, promotion, new job, find job, change job, switch job, job application, career change, new career, switch career, change career, career switch
Learning/skill acquisition	earn certificate, certification, new skill, gain skill, increase skill, learn new, increase knowledge, new knowledge, gain knowledge, new method
Course features	professor, university, instructor, TA, forum, meetup, quiz, grade, video, assignment, free, duration, long

In practice, this search topic did not yield as many relevant reviews as expected, and only 14 reviews scored a 2 or higher on the set of employment keywords. Nonetheless, in this set there are certainly some reviews that specifically call out the MOOC when the reviewer changed jobs or needed new skills to succeed, for example this review:

"I changed jobs and thus was thrown into an environment that leaned on Excel far more than my prior place of employment. This has taken me from a bare bones novice to a pretty adept Excel user in about a months time. I have already advised many of my colleagues what a great program this is!" (Fast Track to Microsoft Excel Beginner + Advanced Training, 2015-01-04)

Even though very few reviews scored highly on the employment topic, this could partly be due to the specificity of the keywords used. It may be highly unlikely for a reviewer to use more than one of these phrases in their text. This could be an excellent area for further research, perhaps refining the keyword list to better capture the employment topic as related to skill development. Conversely, many more reviews scored highly in the learning/skills acquisition topic (70 reviews scored above a 1) and tangentially

touch on employment. The reason could be that changing jobs only receives a passing mention, but learning skills is mentioned frequently in the context of professional development and learning skills.

A sample of reviews that scored highly on the learning/skills acquisition topic are presented below. There are clear examples of people taking MOOCs to get back in to the workforce, switch careers, and even as an aid to assist a teacher in the classroom.

"I am looking to get back into the tech field after taking off about 8 years to pursue a career in teaching math. I have been reading about HTML, CSS, Bootstrap, Javascript, PHP and node.js as well as taking on my own small projects to learn new technologies and to get my skillset stronger." (Learn To Build Beautiful HTML5 And CSS3 Websites In 1 Month, 2015-02-15)

"I have a degree in marketing and before I became a stay at home mom I worked as a graphic designer. I wanted to use the time while out of 'corporate world work' and gain the knowledge in web design and development.... This course gives me the freedom to work at my own pace, gain the knowledge I was looking for, build up my portfolio and the price is great! Well worth it. Thanks Rob for putting this together." (The Complete Web Developer Course - Build 14 Websites, 2014-08-16)

"As a teacher of a recovery credit summer course, I used 'How to Learn Math' with a group of students who had faced math failure over many years. The curriculum was presented in easy to watch segments that engaged my students during our warm up. Many good discussions about everyone having the potential to learn new skills was a benefit." (How to Learn Math: For Students, 2015-02-03)

"My goal is to eventually quit my sales job and follow my passion of web development. ... I highly recommend this course if you are planning to start a new career or just want to learn some new skills that will help you in the future." (The Complete Web Developer Course - Build 14 Websites, 2014-08-17)

"I've been a graphic designer for 10 years and have always wanted to add web design to my skills set and I finally did through this course. It was the right timing in my life as I have started my own design business from home and have been taking courses on a regular basis through udemy to stay up to date with the latest design trends and add learn new skills!" (Learn Photoshop, Web Design & Profitable Freelancing, 2014-12-21)

Lastly, the course features topic easily returned the most reviews. Since in general the direct focus of the reviews is on the courses themselves, it is not surprising that over one quarter of the reviews (21,952) scored a 1 or higher. Of these, 7,212 scored a 2 or higher (the highest scored a 35). Many of these reviews are very detailed (and sometimes quite long), picking apart the courses in terms of production quality, course forums, grading, content, and teaching style. A few examples are shown here.

"In short, the things I had hoped to the MOOC experience would provide that a video/audio lecture series could not -- engaging, relevant assignments and an interactive learning community -- were sorely lacking in this course." (Social Psychology, 2013-10-02)

"As I found in the Maps Coursera course I took, the discussion forums are a great place to supplement your learning from the course. In a programming course such as this, they could be used to learn tricks and tips

and also get help on errors, but since so few students used them (or were likely aware of them at all), they didn't live up to their full promise." (Getting and Cleaning Data, 2014-06-15)

"I commend Professor Filreis and his very capable Teaching Assistants for their milestone success in online education and for bringing us a thrilling poetry class like no other. We are into week eight (The New York School) and it's fair to say that we have an enthusiastic international poetry salon going on. Our professor and TAs lead very interactive, inclusive and involved discussions via live webcasts, social media and designated threads on the insanely active online forums. During live webcasts students from around the world participate in discussions via telephone, Twitter, Facebook, YouTube boards and Coursera discussion forums." (Modern & Contemporary American Poetry, 2012-10-26)

The keyword search results for this topic affirmed that in general it is much easier to find course reviews that discuss aspects of the courses than the reviewer's job or how they will use their newly learned skills for employment. This analysis indicates that the reviews do not presently shed much additional light on using MOOCs for workforce development as designed. Certainly the keyword searches found examples where reviewers refreshed skills or learned new skills to switch careers or return to the workforce, but failed to unearth much in terms of replacing a university degree or learning skills to enter the workforce for the first time.

In conclusion, the keyword searches definitely extracted numerous interesting reviews. The number of reviews to search, and the diversity of topics in those reviews, reveals that there are many ways to refine and advance this method. Data from the surveys and interviews will further illuminate the intersection between MOOCs and employment.

3.4 CourseTalk reviews by topic

Latent Dirichlet Allocation, a generative probabilistic model, is a popular topic model for textual data analysis. This model posits that each observed document was generated as a mixture of (unobserved) topics, where in turn each topic is a collection of words. The goal of the LDA model is then to discover the optimal set of topics. Starting with a set of documents, the model forms topics, and the mixture of topics in each document, that maximizes the probability of generating the observed data. Technical details of the model specification are in Appendix 1 that includes more on the methodology used.

The LDA model makes the simplifying assumption to treat each document as an unordered "bag of words," dropping any syntactical structure or ordering of the words. Additionally, to focus the model on the most meaningful words in the reviews, all two letter words and stop words (for example "the" and "and") were dropped. To aid comparability between reviews and simplify the topics, all remaining words were lemmatized to their root form. This cleaned set of reviews is used to fit the LDA model.

Before actually fitting and interpreting the LDA model, the number of topics k needs to be set. This is a problem, because the topics are exactly what we are trying to determine from the data. Although this

may seem like a chicken and egg scenario, an empirical method for finding the optimal value of k was proposed by Arun et al. [1]. Their metric (detailed in Appendix 1) is computed over a range of k and the minimum value is the optimal number of topics. For the reviews, the results of the Arun measure are plotted in Figure 3, and shows that k = 53 is the appropriate choice.

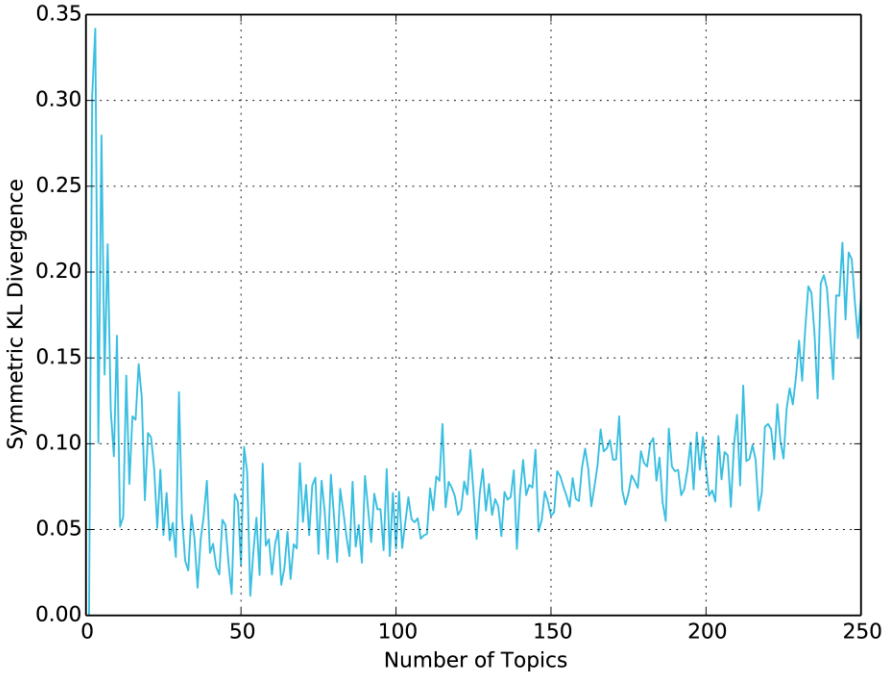

Figure 3: The Arun measure computed on the review corpus

Note: The minimum value of the symmetric KL divergence is the best number of topics. For the reviews the optimum value is k = 53 topics.

Using k = 53 as the input parameter to our final LDA model, the top 30 words within these 53 topics are shown fully in the table in Appendix 2, and a selection of 10 topics are presented here in Table 6 for discussion and interpretation. Note that the ordering of topics in the table is not random. Within a topic, each word has a certain probability of being chosen. Summing the probabilities for the top 30 words in each topic gives a measure of how focused that topic is on those words, higher values are more focused. The topics are sorted by this sum.

There are several key findings from this model that enhance the baseline results of the review texts from the keyword searches. The top 10 topics all focus on features and aspects of the courses, and do so in a highly positive light (see topics 1-4 in Table 5 as an example). Many of the reviews seem to be both highly positive and talk mostly about the MOOC being reviewed, which supports this finding.

Additionally, several topics focus on words from specific subject areas, such as Python programming and web development, likely due to courses in these areas having lots of reviews (Introduction to Interactive Programming in Python, for example, has 476 reviews). Topic 11 is quite clearly about web development, using words such as "web," "development," "html," "website," and more. Topic 12 focuses on using

Adobe products, such as Photoshop and Illustrator, and Topic 17 covers "apps" and "android" and also contains the word "entrepreneur." Finally, topic 26 covers learning programming, using words such as "programming," "language," "python," "cloud," and "Hadoop" (no idea why "daughter" is so highly ranked in this topic).

Table 5: A selection of interesting topics from the entire set of 53 (presented fully in Appendix 2)

Topic	P_sum	Top 30 words
1	0.583	easy, understand, follow, course, simple, great, learn, way, make, thanks, really, good, well, mobile, explain, find, along, start, use, thank, tutor, lot, like, technique, help, work, get, testing, beginner, instruction
2	0.555	course, recommend, highly, anyone, well, really, great, would, learn, enjoy, excellent, want, rob, instructor, make, way, take, content, easy, training, present, explain, interest, lot, engage, definitely, material, teaching, like, thank
3	0.528	clear, course, business, good, great, concise, explanation, information, well, excellent, instructor, example, lot, really, instruction, video, give, easy, presentation, content, plan, thanks, provide, lesson, teacher, point, job, manner, short, practical
4	0.502	look, course, forward, take, learn, ive, really, class, start, great, get, strategy, far, next, excite, make, recommend, like, anyone, never, future, new, see, way, want, try, good, one, best, enjoy
----		
11	0.435	course, web, development, learn, wine, html, website, great, cs, start, new, knowledge, learning, take, site, developer, need, launch, want, well, design, recommend, help, best, get, develop, skill, responsive, build, basic
12	0.415	design, photoshop, illustrator, class, course, designer, actionable, sketch, graphic, use, really, linkedin, great, learn, image, like, tool, art, create, take, draw, technique, process, would, skill, well, one, lot, make, video
---		
17	0.390	course, sell, build, apps, business, android, start, create, entrepreneur, brand, successful, building, mark, learn, way, get, new, develop, market, help, make, give, item, online, use, product, idea, great, io, need
---		
19	0.366	video, quality, course, lecture, good, production, audio, watch, time, slide, content, instructor, material, sound, engage, like, presentation, high, could, well, would, also, information, interview, much, make, listen, really, present, take
---		
26	0.348	course, programming, language, campaign, python, market, well, thumb, unity, problem, learn, concept, work, structure, programmer, help, object, solve, manageable, program, cloud, orient, good, skeptical, daughter, introductory, enter, hadoop, start, experience
---		
53	0.216	course, story, write, think, take, people, detailed, writer, profit, time, individual, idea, engaging, way, also, life, one, find, give, would, student, want,

		lead, street, new, act, realistic, well, interest, part

Note: The topics are sorted by the value of the "p_sum" column. This value can be thought of as a measure of the "compactness" of the topics: each word in a topic has a probability of being chosen, and p_sum is the sum of these probabilities for the top 30 words. The higher the value of p_sum, the more focused that topic is on these words.

The handful of specific topics are not exclusively about those specific subject areas that have the most reviews. Topic 19 is not focused on general course features, but rather on the production quality of the MOOC in terms of audio and video. This topic is focused on words such as "video," "quality," "lecture," "audio," "slide," and "content."

Finally, Topic 53 is included in the shortened table of topics to show one of the many diffuse and unclear topics. Reading the list of words does not present a clear idea of what this topic is about. These diffuse topics seem to be the majority of the 53 topics, which suggests that there may be a lot of noise in the review data once the more specific topics have been accounted for.

3.4.1 Distribution of topics in reviews

In the topic model framework, each review can be considered as a mixture of topics (with differing weights for each topic component), and it may be useful to turn around and think about the topics as "threads" that run through the reviews. Some threads are more prevalent than others, while some only appear in a small subset of reviews. As a summary of the reviews, this will help show the important topics in two ways. First, topics that are a significant component of many reviews could be considered as universal themes. Secondly, topics that are focused on a small set of reviews may highlight themes in the reviews relevant to a certain user group or type of course.

Due to the large number of reviews and topics, the best way to see these trends is in a visualization. Figure 9 shows the weight of each topic in each review. To interpret this figure, focus on the distribution of each topic over the entire collection of reviews (that is, for each topic, scan vertically to see how densely it is present across the body of reviews). Specifically, note how the first 10 topics are broadly present in a large number of reviews, and recall that these topics focus on general praise for the MOOCs. On the other hand, topics 11 and 12 (specific to certain subjects) are present in a much smaller set of reviews.

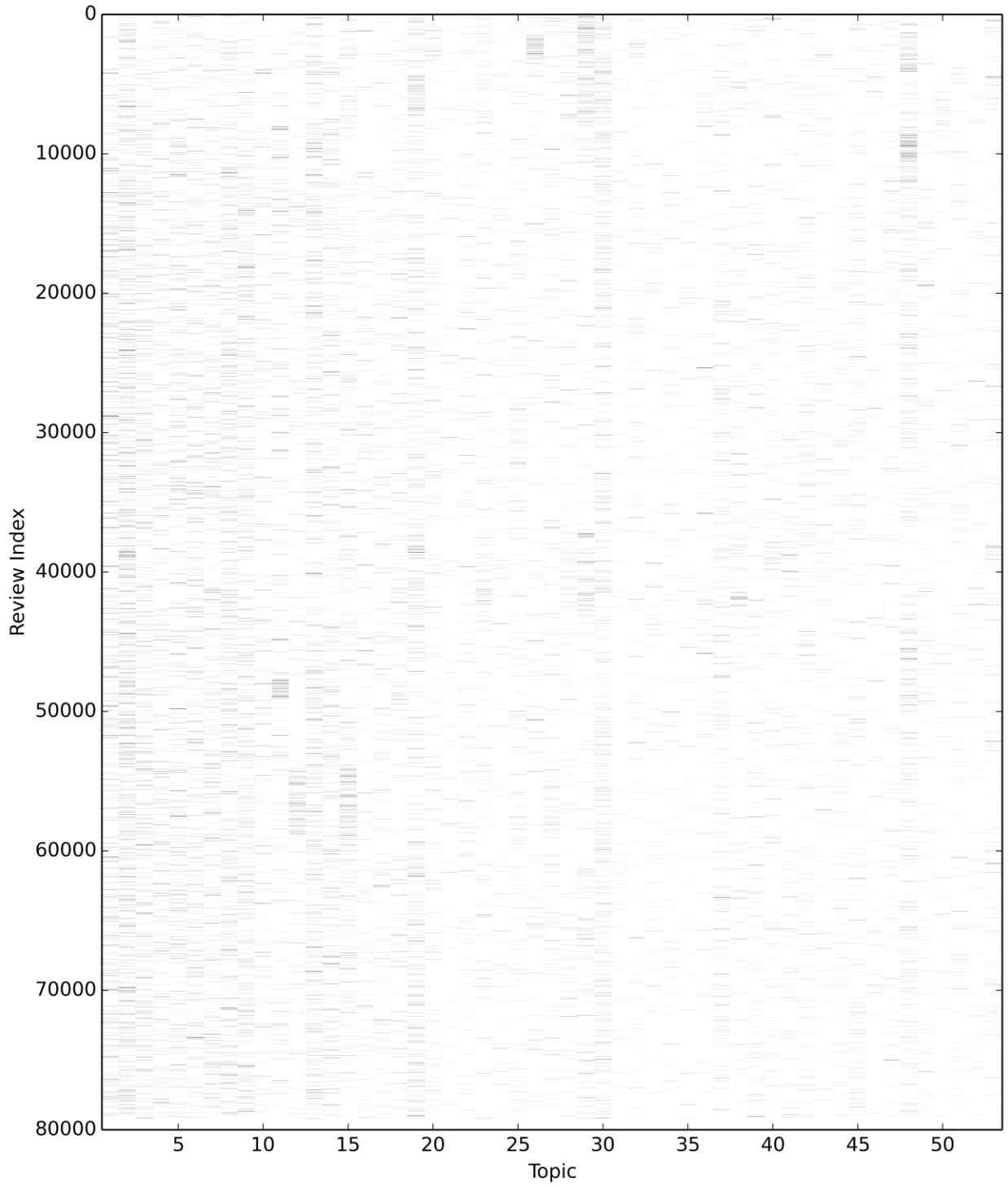

Figure 2: Distribution of topics in reviews.

3.4.2 Further topic analysis

Although these findings are a powerful summary of the topics present in the reviews, a key limitation is still the bag-of-words approach that is the basis of the LDA model. The top words in many of the topics suggest that clarity could be vastly improved with the addition of context or ordering for some of the

words. Such alternatives to the basic LDA model do exist, such as a topical n-gram model [11], syntactic topic models [5], and supervised topic models [8]. Exploring these methods would be one possible extension of the current analysis.

A second potential improvement would be to focus on some carefully selected subset of reviews. For the purposes of the Advancing MOOCs for Development Initiative perhaps filtering the IP addresses of users in the Philippines, Colombia and South Africa or encouraging residents of these countries to clearly identify themselves in some distinct manner.

There appears to be quite a bit of noise in the 53 topics, and more interpretable topics may arise from using a subset of the reviews. A suitable subset could be constructed, for example, from the set of "helpful" reviews, or the set of reviews written by registered users. However, this method may not provide additional information or different topics since the reviews are so heavily focused on positive aspects of the courses.

3.5 CourseTalk reviews by user search terms

This dataset is a log of terms input to the search box on the CourseTalk website. Each term is timestamped, and includes an originating IP address, as well as the number of course listings returned. In total, this dataset contains a large number of entries, with the monthly number of searches generally increasing over time One way of cleaning the search terms is to restrict to those that returned at least one course listing, as this implies that a match was made with the course listings. Using this measure, 96% of the searches returned at least one course listing. This seemed like a high proportion of the total, and digging deeper into this value showed that 72% of the searches are blank, implying that a blank search string still returns some default listing of courses.

Restricting to the searches that returned at least one course and had a valid (and not blank) search string, the top 10 search terms are in Table 6. These searches lean heavily towards computer science (indeed "computer science" is the single most frequent search term), followed by "python," "java," "web," "machine learning," and "android."

Table 6: Top 10 search terms

Search string	Percent
Computer science	2.9
Python	2.7
Java	2.0
Web	1.7
Machine learning	1.5
Statistics	1.5
Finance	1.4
Social media	1.2
Game	0.9
Sebastian Thrun	0.8
Android	0.8

Note: Top 10 search terms for the searches that were not a blank string and that returned one or more course listings.

Tabulating the top search terms by month largely shows variations on this overall theme (Table 7). The ordering of the terms differs by month, but in general the search terms are dominated by computer science and related fields.

Table 7: Top 5 search terms for each month.

Month	Top 5 search strings
Nov 2012	python, machine learning, computer science, chemistry, android
Dec 2012	python, design, English, computer science, algorithms
Jan 2013	statistics, machine learning, python, computer science, algorithms
Feb 2013	machine learning, statistics, data analysis, python, algorithms
Mar 2013	algorithms, data analysis, cryptography, nutrition, programming languages
Apr 2013	machine learning, statistics, algorithms, calculus, finance
May 2013	statistics, data science, artificial intelligence, algorithms, python
Jun 2013	statistics, machine learning, algorithms, data science, python
Jul 2013	
Aug 2013	
Sep 2013	statistics, python, finance, data analysis, machine learning
Oct 2013	python, data analysis, machine learning, java, algorithms
Nov 2013	computer science, social media, sebastian thrun, ariely, python
Dec 2013	computer science, social media, sebastian thrun, game, key word
Jan 2014	computer science, web, social media, sebastian thrun, game
Feb 2014	computer science, social media, sebastian thrun, python, game
Mar 2014	web, game, computer science, key word, sebastian thrun
Apr 2014	python, computer science, game, business, coursera
May 2014	web, python, computer science, coursera, sebastian thrun
Jun 2014	python, java, web, computer science, machine learning
Jul 2014	java, python, machine learning, linux, statistics
Aug 2014	java, python, machine learning, linux, android
Sep 2014	python, java, machine learning, statistics, computer science
Oct 2014	python, java, finance, computer science, statistics
Nov 2014	python, java, computer science, machine learning, finance
Dec 2014	python, java, computer science, machine learning, statistics
Jan 2015	python, java, computer science, machine learning, finance
Feb 2015	java, python, finance, computer science, machine learning
Mar 2015	python, machine learning, java, finance, computer science

Since each search term entry includes an IP address, we attempted to geo-locate the searches based on this information. Unfortunately, over 89% of the searches recorded an IP address from an internal CourseTalk server. However, of the remaining, 87% returned a valid location using the free Geoip2 database, and over 1/3 with a unique IP address. These locations are plotted on a world map in Figure 5, where the size of the dot is proportional to the number of searches originating from that location. The highest density is in the United States and Europe, however there are many hits logged from all over the world.

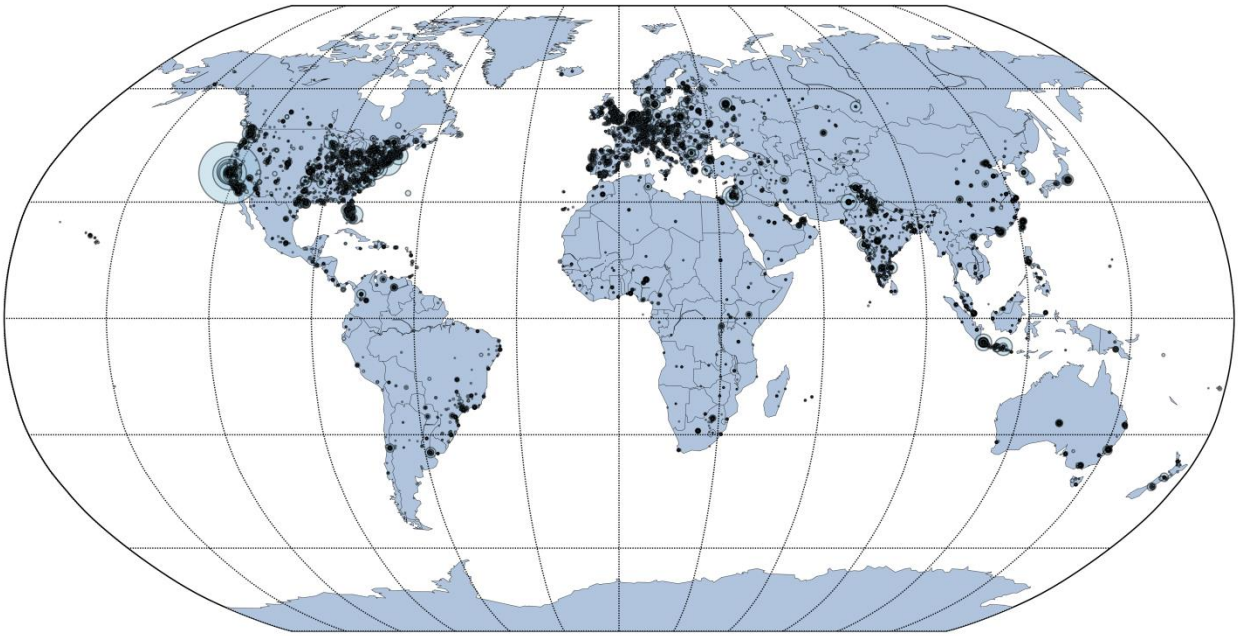

Figure 5: Approximate locations of the unique valid IP addresses associated with searches on the CourseTalk website

This high-level summary of the search terms largely shows that the majority of searches are focused on computer science and related fields. This affirms how prevalent interest in computer science related courses is in the MOOC sphere. The locations of the searches give a broader picture of access to the CourseTalk page than that provided by the registered users alone. However on the whole, the search term analysis does not add deep insight to the MOOC research.

Recommendations and Future Research

The findings below illuminate some areas where the data show potential for improvement:

1.) Activation & Retention: Priority should be given to registering new users and providing more opportunities for them to stay active. This analysis showed reviews by registered users are much more likely to be voted helpful (27% for registered vs. 0.7% for non-registered); however, only a fraction of the total reviews (~8%) were attributable to a registered user. Thus, since reviews voted "helpful" are likely of higher quality, it can be concluded that getting registered users to review regularly would increase the overall value of CourseTalk's review content.

2.) User Profiling: A focus should be placed on profile completion to allow greater insight on core user groups. The research showed a small percentage of active registered users had completed profiles. Because the majority of CourseTalk's registered users have limited or no profile completion, there is limited insight as to the types of users submitting reviews and what insights we can draw from their reported experiences. Adding/refining required fields in the user profile will be key, but it requires a balance between collecting more information and maintaining an optimal user experience. Prioritization of top user profile criteria will be critical to this balance, as will leveraging social media login data. As user

profile data becomes more robust, it will be possible to further segment specific user groups (by occupation, education level, age, etc.) and compare it with key activity data (such as time spent on CourseTalk, number of reviews, types of courses reviewed, etc.). By combining this data, usage patterns may become apparent and allow for broader application of findings.

3.) Fostering a Reviewer Community: CourseTalk should prioritize reviewer engagement to continue to foster a sense of community. This will likely lead to improved review content by increasing the number and quality of review submissions. In addition, an active reviewer community could promote interaction with other reviews through features like the existing "helpful" vote button, further improving review quality through user feedback. A sense of community will also enable the collection of additional data on registered users, including user profile information and key site engagement data, both of which will foster greater review insights and better-informed UX/UI changes. Furthermore, drawing upon social computing dynamics, CourseTalk could devise strategies to bring collective meaning to the notion of participation in the CourseTalk community. Promoting the CourseTalk leaderboard and adding rewards for highly active/rated reviewers could further build the CourseTalk community and incentivize review activity

4.) Search Data Integration: Finally, CourseTalk should enhance its data analytics capabilities to assess user engagement data across core activities by key user groups. For example, in this analysis, the log of search terms is isolated from the rest of the findings. With no identifying information, little more than a general descriptive summary was possible. If, however, there was some way to link searches to registered users and/or website sessions, this analysis could potentially be refined to see if any differences exist between registered users and non-registered visitors. Furthermore, if the originating IP address was recorded more consistently for key activities like the course "enroll" option, the geographical diversity of CourseTalk's user base could be better quantified and mapped, as well as compared with course engagement.

END