Analysis and Design of Large-scale Online Classes

Tao C. Lee
LCAV, EPFL

June 26, 2014

EDIC research project

Supervised by

Dr. Paolo Prandoni
Professor Martin Vetterli

Lausanne, academic year 2013 – 2014

EDUCATING A MODEL STUDENT WITH MONOTONICALLY DECREASING RECAP

Tao C. Lee, Paolo Prandoni, and Martin Vetterli

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
Email: {tao.lee, paolo.prandoni, martin.vetterli}@epfl.ch

ABSTRACT

Massive Open Online Courses (MOOCs) have become an emergent paradigm of large-scale knowledge distribution. One of the dilemma faced by MOOC teachers is the trade-off between introducing new material and reviewing taught material. We model the recap activities as units of monotonically decreasing lengths and study the their schedule. We validate our model with the data of a Digital Signal Processing (DSP) MOOC. Our results include algorithms to construct schedules adhering to some spacing constraints, bounds on the rate at which new material can be introduced, and algorithms for MOOC data processing. We give insights to the design of the DSP MOOC using the Monotonically Decreasing Recap (MDR) schedule.

Index Terms— Scheduling, asymptotic analysis, signal processing, MOOCs, learning analytics.

1. INTRODUCTION

Massive Open Online Courses (MOOCs), as an emergent paradigm of massive knowledge distribution, took off in 2012 and attracted wide attention in the higher education community. Partially accelerated by today's computing technology, high-speed Internet access, wide availability of computers and mobile devices, MOOCs could reach more students than ever. By the early of 2013, with more than 2.8 million registered users on Coursera, star classes usually have more than tens of thousands learners each term. The wide popularity of MOOCs not only attracts the eyes of learners, but also attracts many researchers to study the MOOC phenomena [1, 2, 3, 4, 5].

A MOOC contains several components: Video lectures, exercises, exams and online forums. Large number of students can interact with teachers on the online forums, and this interaction deserves investigation [6]. One interesting aspect of online forums is that they manifest the interaction of knowledge networks and social networks. Data mining on this dataset can answer questions such as what are the learning difficulties, what are the interested topics, and how to present the material. The aggregated statistics can shed light on how to contain the instructional complexity of education [7].

An important challenge for the analysis and design of MOOCs is how to model the learning process of students by incorporating the results of empirical research. An example is how to present the material considering students' learning difficulties. Teachers must trade-off between the needs of the students and the standards of the class. A fundamental class of students' needs is the schedules of new material and review, which determines the pace of the class and classify fast and slow students. A mathematical model was proposed by researchers to capture these issues [8] .

Novikoff and Kleinberg [8] proposed to model the education process as a sequence of abstract units: New material and review are units spaced over time, with the students' needs modeled as constraints on the schedules. Their main purpose is to help the design of educational software for personalized education. Applications include building vocabularies, learning mathematics, and teaching piano lessons. The ultimate goal is a software system where the designer can control the schedule for the introduction of new material and the review of taught material. In their model, new material and review are modeled as units of the same length.

The study of human brain shows that human memory has two stages: the short-term memory and the long-term memory [9, 10]. The short-term memory is the place where new learned information is stored, which needs to be continuously refreshed. The acquired information then gradually moves to the long-term memory, and might not be accessible over time. The refreshing process in the short-term memory is faster than the initial learning process, as information is already kept in the short-term memory or even in the long-term memory.

In this paper we model the recap modules as units of monotonically decreasing lengths and study their schedule. This model is motivated by the teaching of a DSP MOOC on Coursera [11], where recap units are scheduled at the beginning of new modules. The educational data generated by the large number of participants in the DSP MOOC provides a virtual laboratory to validate our model. The length of the recap decreases in each reappearance, as students are expected to make progress through time. This kind of review is not

This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications - SPARSAM - no. 247006.

only helpful for students but also useful for teachers, as it can bridge the taught material and the material to be introduced.

The recap schedule is designed for the DSP MOOC [11]. One of the data of particular interest is the data of online forums as forum posts give direct feedback to the course. Compared with previous studies using statistical analysis of grades and other numerical data [1, 2], we develop computational tools to process the data of online forums: We model the forum activities as a graph, with students and teachers as nodes, and exchanged messages as edges. We use keyphrase extraction to label the concepts in the forum posts. Concepts that generate the more interest have higher priority for review.

Our main contribution is to develop an approach to integrate the MDR schedule in MOOCs. We begin by introducing the Fixed Length Recap (FLR) schedule, and then consider the design of the MDR schedule that can be integrated in the DSP MOOC [11].

2. MODELS

Novikoff et al. [8] proposed to model the education process as a sequence of abstract educational units. For example, the sequence u_1, u_2, u_3, u_1, u_4,..., indicates that educational unit u_1 was introduced at the first time step and reviewed at the fourth time step. The students' needs are modeled using two sequences, $\{a_k\}$ and $\{b_k\}$: After an educational unit has been introduced, the ideal time for the student to see it for the $(k+1)^{st}$ time is between a_k and b_k time steps after seeing it for the kth time. The numbers b_k model how long the student can retain learned material. The numbers a_k model how long a student should wait before review is beneficial, capturing the notion that there is an ideal time to review.

Their key definition is that a schedule satisfies a set of spacing constraints if, for each u_i in the schedule, the $(k+1)^{st}$ occurrence of u_i in the schedule comes between a_k and b_k positions (inclusive) after the kth occurrence. The fact that a student can go longer between occurrences as they gain familiarity with the educational unit is represented by the assumption that the numbers a_k and b_k are weakly increasing, and $a_k \leq a_{k+1}$ and $b_k \leq b_{k+1}$ for all k.

One key insight of their analysis is the asymptotic behavior of t_n, the position in the schedule of the first occurrence of the nth distinct educational unit: the slower the growth of t_n, the faster new educational units are being introduced. When each one educational unit takes one time step, it is clear that $t_n \geq$ n for any schedule. They gave asymptotic analysis of t_n for different educational goals and different types of students. In their analysis, t_n takes only integer values, with each educational unit having the same length.

We extend the analysis to t_n taking not only integer values but all real values, with each educational unit can have fractional length. We show that the asymptotic learning behavior of a model student changes considerably under this condition. Before going further, we review the definition of infinite perfect learning and the FLR schedule. We also review some theoretical results.

Infinite Perfect Learning. A schedule exhibits infinite perfect learning with respect to some spacing constraints if (i) it satisfies the spacing constraints, and (ii) it contains infinitely many educational units, each of which occurs infinitely often. Thus if the constraints represented the needs of a student, then with such a schedule the student would, over the course of the infinite sequence, learn an infinite set of educational units without ever forgetting anything.

The FLR schedule. Consider the spacing constraints $a_k = 2^k$ and $b_k = 2^{k-1}(k+1)$. A schedule that allows for infinite perfect learning with respect to these spacing constraints can be described as follows: To find the first $2^k(k+1)$ entries of the schedule, consider a depth-first postorder traversal of a full binary tree of height k with $2k$ leaves labeled u_0 , u_1 ,..., u_{2k-1} from left to right. Begin with an empty sequence. Every time a leaf is visited, append the sequence with the corresponding educational unit. Every time a nonleaf node is visited (after both children have been visited), append the sequence with the units corresponding to all of the descendant leaves, in left-to-right order.

Thus, using $k = 2$, we have that the first 12 entries of the schedule are

$u_0, u_1, u_0, u_1, u_2, u_3, u_2, u_3, u_0, u_1, u_2, u_3$, in Figure 1.

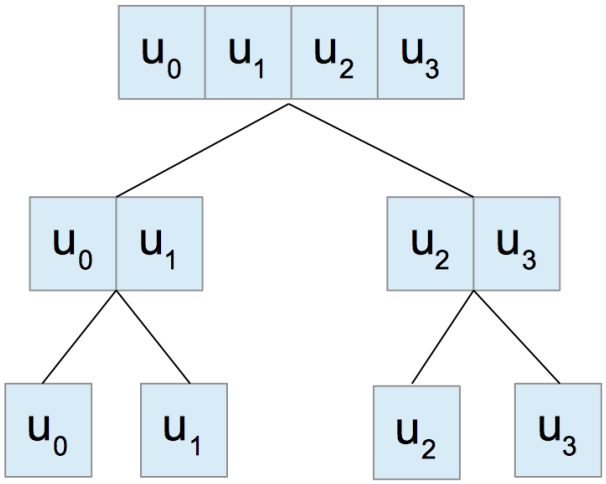


Fig. 1. A FLR schedule

It should be noted that, by the properties of depth-first post-order traversal, this description defines a unique sequence, because the first $(k+1)2^k$ elements of the sequence are the same regardless of whether one considers a tree of height k or one of height greater than k.

Theorem 1. (Asymptotics of the Introduction Time Function.) We define the notation, $T_n(k)$, as the kth occurrence

of the nth educational unit. In the FLR schedule, $t_n = T_n(1)$ grows as $\Theta(n\ log_2\ n)$.
Proof: See [8].
Theorem 2. (Bounds on the Introduction Time Function.) In the FLR schedule,

$$1/2 \cdot n \cdot (\lfloor log_2 n \rfloor + 1) \leq T_n(1) \leq n \cdot (\lfloor log_2 n \rfloor + 1)$$

for all n.
Proof: See [8].

Superlinearity of the Introduction Time Function. For any FLR schedule that exhibits infinite perfect learning with respect to any spacing constraints $\{(a_k,b_k)\}$, there cannot be a constant c such that $t_n \leq c \cdot n$ for all n.
Proof: See [8].

The MDR schedule. Consider the spacing constraints $a_k = 2^k$ and $b_k = 2^{k-1}(k+1)$, and the FLR schedule [8]. With the length of recap decreases through time, the number of units that can be included into the schedule under the same spacing constraint increases. We define a decreasing factor α such that the length of a unit is α^n, $0 < \alpha \leq 1$, and n is the number of time a unit is recapped. Given a full binary tree of height k with $2k$ leaves, we insert a new unit at a nonleaf node when the total length of units of this node is one unit less than the FLR schedule. The spacing constraints for the units introduced in the nonleaf nodes are relaxed. These units, in practice, can be understood as optional material that serve as deserts for students.

Thus, using k = 2, $\alpha = 0.5$, we have that the first 16 entries of the schedule are

$u_0, u_1, u_0, u_1, u_2, u_3, u_4, u_3, u_4, u_5, u_0, u_1, u_2, u_3, u_4, u_5,$
u_6, u_7, in Figure 2.

Units u_2, u_5, u_6 and u_7 are introduced in the nonleaf nodes (shown in blue) to maintain the total length of of each nonleaf node the same as that of the FLR schedule. The spacing constraints for u_2, u_5, u_6 and u_7 are relaxed. In this schedule, units introduced in the leaf nodes satisfy the spacing constraints under the condition that the units introduced in the nonleaf nodes can be arbitrarily distributed in the nodes they are present.
Theorem 3. (Asymptotics of the Introduction Time Function.) In the MDR schedule with a decreasing factor $\alpha \leq 0.5$,

$$T_n(1) \text{ grows as } O(n).$$

Proof: We first prove the case of $\alpha = 0.5$. Let us recall the FLR schedule. By the recap lemma and properties of depth-first postorder traversal [8], at time step $T_{2^k}(1)$ units u_0, u_1,..., u_{2k-1} have each occurred exactly $k+1$ times, and nothing else has occurred at all. Therefore,

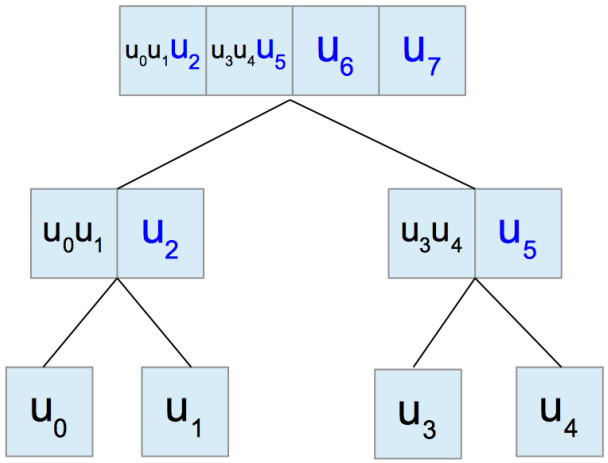


Fig. 2. A MDR schedule with $\alpha = 0.5$

$$T_{2^k}(1) = 2^k \cdot (k+1)$$

with decreasing factor $\alpha = 0.5$, additional units $2^{k-1} \cdot k$ are inserted in the binary tree. Therefore,

$$T_{2^k+2^{k-1}\cdot k}(1) = 2^k \cdot (k+1)$$

and so

$$T_{2^{k-1}(k+2)}(1) = 2^k \cdot (k+1)$$

and

$$T_n(1) = 2 \cdot n - 2^k < 2 \cdot n$$

for n of the form $n = 2^{k-1}(k+2)$, which establishes that $T_n(1)$ grows as $O(n)$ when considered as a function of integers of the form $n = 2^{k-1}(k+2)$. Because $T_n(1)$ increases monotonically in n, it follows that that $T_n(1)$ grows as $O(n)$ when considered as a function of all positive integers.
For the case $0 < \alpha < 0.5$, we have $T_n(1)_{\alpha<0.5} < T_n(1)_{\alpha=0.5}$, so it follows that the theorem also holds.
Theorem 4. (Bounds on the Introduction Time Function.) In the MDR schedule with a decreasing factor $\alpha \leq 0.5$,

$$n \leq T_n(1) < 2 \cdot n$$

for all n. In other words, $T_n(1)$ grows as $\Theta(n)$.
Proof: By Theorem 3,

$$T_n(1) < 2 \cdot n.$$

Furthermore, by construction, all units appearing for the first time have unit length. Therefore,

$$n \leq T_n(1).$$

Theorem 5. (Adherence to spacing constraints.) In the MDR schedule with a decreasing factor $\alpha = 0.5$, the units introduced in the leaf nodes adheres to the spacing constraints

$$a_k = 2^k$$
$$b_k = 2^{k-1}(k+1).$$

under the condition that the units introduced in the nonleaf nodes can be arbitrarily distributed in the nodes they are present.

Proof: Our goal is to show that

$$a_k \leq T_i(k+1) - T_i(k) \leq b_k$$

for i introduced in the leaf nodes, and all k, under the condition that the units introduced in the nonleaf nodes can be arbitrarily distributed.

We observe that the length of each node in the MDR schedule is the same as that of the FLR schedule. Therefore, $T_i(k+1) - T_i(k)$ is the same for the inter-node spacing, it remains to show that the intra-node spacing satisfy the constraints. The monotonically decreasing lengths of units give space for units introduced in the nonleaf nodes, and the condition that these units can be arbitrarily distributed in the nodes they are present guarantees that the leaf units can be scheduled in precisely the same place as that of the FLR schedule. Note that, for some k, distribution of nonleaf units is not needed.

Superlinearity of the Introduction Time Function For the MDR schedule with a decreasing factor $\alpha \leq 0.5$, the introduction time function grows linearly. We now show that this linearity is not universal. For some MDR schedules that exhibits infinite perfect learning with respect to certain spacing constraints $\{(a_k,b_k)\}$, there cannot exist a constant c such that $t_n \leq c \cdot n$ for all n.

Proof: To prove this statement, we consider an arbitrary set of spacing constraints (a_k, b_k) and an arbitrary schedule that exhibits infinite perfect learning with respect to these constraints, and assume for the sake of contradiction that there is a constant c such that $t_n \leq c_n$ for all n. Let $b^k = \sum_{j=1}^{k} b_j$ and let n_0 be any integer such that $n_0 > b^{c+1}$. By our assumption, at least n_0 educational units have been introduced by the time step $c \cdot n_0$. Schedule that exhibits infinite perfect learning, any unit that has been introduced by time step t will have occurred k times by time step $t + b^k$, by the definition of b^k. Thus at least n_0 units will have occurred $c+1$ times by time step $c \cdot n_0 + b^{c+1}$. So $c \cdot n_0 + b^{c+1} \geq c \cdot n_0 + \sum_{j=1}^{c+1} a_j \cdot n_0 \geq (c+1) \cdot n_0$. Because $\{a_k\}$ represent the lower bounds of the time steps between the occurrence of the same educational unit, and $a_k \geq 1, \forall k \in N$. Subtracting $c \cdot n_0$ from both sides, we have that $b^{c+1} \geq n_0$, which contradicts our choice of n_0. Thus, there cannot be a constant c such that $t_n \leq c \cdot n$ for all n.

One key observation is that even with the MDR schedule, the lower bounds $\{a_k\}$ still preserve the superlinearity of the introduction function.

When organizing a MOOC, we created an outline of the course content, and then refined the outline in more detail. The draft plan was linear at first, and then followed by insertion of some example modules. We showed that with the MDR schedule, we could achieve linear schedule with a constant factor of 2. The issue that the MDR schedule would reduce the units into meaningless short lengths in infinite perfect learning is only a theoretical pitfall, since in practice courses always have finite lengths. In the following sections, we consider the design of the MDR schedule that can be integrated in the DSP MOOC.

3. DATA PROCESSING

With large number of participants, MOOCs generate unprecedented educational data. In the DSP MOOC of the 2013 spring, around 40,000 students enrolled and 2,000 students stayed active in the final module. Some of the previous studies perform statistical analysis of grades or other numerical data, but cannot avoid the statistical biases stemmed from the open and online nature of MOOCs. On the other hand, online forums of MOOCs are the places where teaching staffs get feedback from massive number of students [11, 6]. Forum posts can be referenced when adjusting the pace of classes, fixing defects of teaching material, and scheduling of review. In the DSP MOOC of the 2013 spring, around 1,000 threads and 4,000 posts were created on the forum. We develop computational tools to process the forum data of the DSP MOOC, and consider the design of the MDR schedule in the sequel.

Definition 1. (The structure of forum data.) Before developing computational tools to process forum data, we first introduce the structure of forum data,

$$forum\ data = \{thread_i\}$$
$$thread_i = \{post_j\}$$
$$post_j = post_j | \{comment_k\}.$$

To understand the spacing constraints of massive students in a DSP MOOC, we performed text data mining using machine learning tools [12]. We used automatic keyphrase extraction to extract the keyphrases of forum posts. We identified the concepts of interest by the keyphrases and used these concepts as the guidance of scheduling recap. Due to short text in forum posts, we screened the extracted keyphrases and ranked them by frequency. The top-ten concepts of interest are shown in Table 1.

Forum data can be seen as a sequence of posts in time, associated with participants and keyphrases.

Table 1. Keyphrases of forum posts

Concept	Keyphrases
Signals	Gaussian, sinc, ...
Basic mathematics	summation, polynomial, ...
Filters	filter, convolution, FIR, ...
Fourier analysis	Fourier, DFS, DFT, ...
Periodicity	periodic, period
Vector space	Hilbert space, basis, subspace, ...
Sampling	sampling frequency/theorem
Orthogonality	orthonormal, orthogonal
Interpolation	Lagrange/linear interpolation, ...
Wavelets	wavelet, wavelets

$$forum\ data = \{p_1, p_2, ..., p_n\},$$
where $p_k = p_k(timestamp, participant, keyphrases)$.

On the other hand, an online forum can also be seen as a social network, with participants (students, teachers, assistants) as nodes, exchanging messages with each other. For example, the following posts in the same thread,

$$p_1(t_1, S_1, key), p_2(t_2, S_2, key), p_3(t_3, T_1, key),$$
where $t_1 < t_2 < t_3$, S_1, S_2 are two students
T_1 is a teacher, key = sampling

can be graphically illustrated as Figure 3,

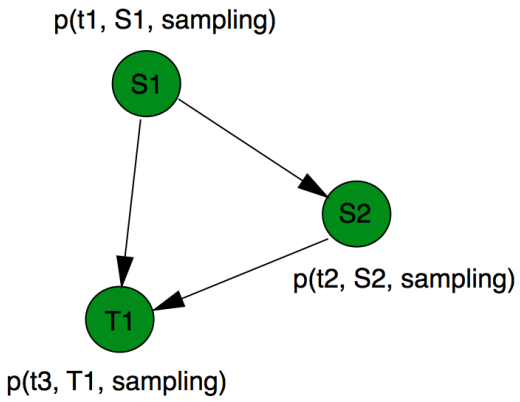


Fig. 3. A graph of forum activities

where the directions of edges follow a chronological order from earlier posts to later posts. In this example, we assume all forum posts are related to the concept, *sampling*, and later posts are more authoritative than earlier posts, so that edges from students S_1 and S_2 point to the last post (by teacher T_1) as the authoritative source. This simple example only shows the graph of a single thread, multiple threads contributed by the same participants could render the edges bi-directional. The following is an algorithm to construct a graph of forum data.

Algorithm 1. (Constructing a graph of forum data.)
Graph (F, k)
F: forum data
k: number of forum threads, posts and comments
Let $S_\sigma \leftarrow \emptyset$
For i = 1, 2, ..., k
Collect participants of forum posts and comments.
For every participant
Add one node to S_σ.
For i = 1, 2, ..., k
Connect two participants in S_σ in a chronological order if they are in the same forum thread.
Return S_σ.

Experiments showed that the raw sequence of forum posts following a chronological order was not very informative. To shed more light on the spacing constraints of students, we need to develop computational tools to rank the posts to find the informative ones.

Definition 2. (Filtering the sequence of forum posts.) A filtering operation F on the sequence of forum posts $\{p_k\}$, ranks and keeps the n_{top} top-ranked p_k as $\{p'_k\}$

$$F(\{p_k\}, n_{top}) = \{p'_k\}.$$

As an example, we can filter the posts by the timestamps.

Definition 3. (Timestamp filtering.) A timestamp filtering operation TF on the sequence of forum posts $\{p_k\}$, keeps the posts with the timestamps between t_{from} and t_{to} as $\{p'_k\}$,

$$TF(\{p_k\}, t_{from}, t_{to}) = \{p'_k\}.$$

It can be easily verified that timestamp filtering is a linear operation.

Filtering can also be carried out on the graph of forum data. For example, we can filter the forum posts to contain only those by the important participants. Kleinberg [13] developed a computational tool to find authoritative sources in a network S_σ. With each participant p, he associated a non-negative authority weight x_p and a non-negative hub weight y_p. He maintained the invariant that the weights of each type are normalized so their squares sum to 1: $\sum_{p \in S_\sigma} x_p^2 = 1$, and $\sum_{p \in S_\sigma} y_p^2 = 1$. He viewed the participants with larger x- and y-values as being "better" authorities and hubs respectively.

Numerically, it is natural to express the mutually reinforcing relationship between hubs and authorities as follows: If p points to many pages with large x-values, then it should receive a large y-value; and if p is pointed to by many pages with large y-values, then it should receive a large x-value. This motivates the definition of two operations on the weights, which he denotes by I and O. Given weights $\{x_p\}$, $\{y_p\}$, the I operation updates the x-weights as follows.

$$x_p \leftarrow \sum_{q:(q,p) \in E} y_q.$$

The O operation updates the y-weights as follows.

$$y_p \leftarrow \sum_{q:(p,q)\in E} x_q.$$

Thus I and O are the basic means by which hubs and authorities reinforce one another. He showed that as one applies Iterate with arbitrarily large values of k, the sequences of vectors $\{x_k\}$ and $\{y_k\}$ converge to fixed points x^* and y^*.

Algorithm 2. (Computing x_k and y_k.)
Iterate (G, k)
G: a collection of n linked nodes
k: a natural number
Let z denote the vector $(1, 1, 1, ..., 1) \in R_n$. Set $x_0 \leftarrow z$.
Set $y_0 \leftarrow z$.
For i = 1, 2, ..., k
Apply the I operation to (x_{i-1}, y_{i-1}), obtaining new x-weights $x_{i,new}$.
Apply the O operation to ($x_{i,new}$, y_{i-1}), obtaining new y-weights $y_{i,new}$.
Normalize $x_{i,new}$, obtaining x_i.
Normalize $y_{i,new}$, obtaining y_i.
End
Return (x_k, y_k).

Theorem 6. (Convergence of $\{x_k\}$ and $\{y_k\}$)The sequences x_1, x_2, x_3, ... and y_1, y_2, y_3, . . . converge (to limits x^* and y^* respectively).
Proof: See [13].

We now introduce some notations. Let $G = (V, E)$, with $V = \{p_1, p_2, ..., p_n\}$, and let A denote the adjacency matrix of the graph G; the (i, j) th entry of A is equal to 1 if (p_i, p_j) is an edge of G, and is equal to 0 otherwise. One easily verifies that the I and O operations can be written $x \leftarrow A^T y$ and $y \leftarrow Ax$ respectively. one can denote these eigenvalues by $\lambda_1(M)$, $\lambda_2(M)$, ..., $\lambda_n(M)$, indexed in order of decreasing absolute value, and with each eigenvalue listed a number of times equal to its multiplicity. For each distinct eigenvalue, one can choose an orthonormal basis of its eigenspace; considering the vectors in all these bases, one can obtain a set of eigenvectors $\omega_1(M)$, $\omega_2(M)$, ..., $\omega_n(M)$ that one can index in such a way that $\omega_i(M)$ belongs to the eigenspace of $\lambda_i(M)$. For the sake of simplicity, one can make the following technical assumption about all the matrices we deal with:

$$|\lambda_1(M)| > |\lambda_2(M)|.$$

When this assumption holds, one refers to $\omega_1(M)$ as the principal eigenvector, and all other $\omega_i(M)$ as non-principal eigenvectors.
Theorem 7. (Interpretation of x^* and y^*.) x^* is the principal eigenvector of $A^T A$, and y^* is the principal eigenvector of AA^T.
Proof: See [13].

Although originally developed as a tool to find authoritative sources in a hyperlinked environment [13], Algorithm 2 can be used to filter the forum posts to contain only those by the important participants. We define the authority filtering operation as follows.
Definition 4. (Authority filtering.) An authority filtering operation AF on the sequence of forum posts $\{p_k\}$, ranks and keeps the posts of the n_{top} top-ranked nodes as $\{p'_k\}$ by authority weights using Algorithm 2,

$$AF(\{p_k\}, n_{top}) = \{p'_k\}.$$

As an extension of Algorithm 2, we want to emphasize the nodes with posts containing keyphrase K. Given weights $\{x'_p\}$, $\{y'_p\}$, and an amplification factor a ($a \geq 1$), the I' operation updates the x-weights as follows.

$$x'_p \leftarrow \sum_{q:(q,p)\in E} \theta(K) \cdot y'_q.$$

The O' operation updates the y-weights as follows.

$$y'_p \leftarrow \sum_{q:(p,q)\in E} \theta(K) \cdot x'_q,$$

where

$$\theta(K) = \begin{cases} a & : p \text{ or } q \text{ has posts containing } K \\ 1 & : \text{otherwise} \end{cases}$$

Thus I' and O' are the amplified means by which hubs and authorities reinforce one another.

Algorithm 3. (Computing x'_k and y'_k.)
AmplifyAndIterate (G, k, K)
G: a collection of n linked nodes
k: a natural number
K: a keyphrase
Let z denote the vector $(1, 1, 1, ..., 1) \in R_n$. Set $x_0 \leftarrow z$.
Set $y_0 \leftarrow z$.
For i = 1, 2, ..., k
Apply the I' operation to (x_{i-1}, y_{i-1}), obtaining new x-weights $x'_{i,new}$.
Apply the O' operation to ($x'_{i,new}$, y_{i-1}), obtaining new y-weights $y'_{i,new}$.
Normalize $x'_{i,new}$, obtaining x'_i.
Normalize $y'_{i,new}$, obtaining y'_i.
End
Return (x'_k, y'_k).

Theorem 8. (Convergence of $\{x'_k\}$ and $\{y'_k\}$) The sequences x'_1, x'_2, x'_3, ... and y'_1, y'_2, y'_3, . . . converge (to limits x'^* and y'^* respectively).
Proof: Let $G = (V, E)$, with $V = \{p_1, p_2, ..., p_n\}$, and let B denote the weighted adjacency matrix of the graph G; the (i, j) th entry of B is equal to $\theta(K)$ if (p_i, p_j) is an edge of G, and is equal to 0 otherwise. One easily verifies that the I' and O' operations can be written $x \leftarrow B^T y$ and $y \leftarrow Bx$ respectively. Thus x_k is the unit vector in the direction

of $(B^TB)^{k-1}B^Tz$, and y_k is the unit vector in the direction of $(BB^T)^kz$. Now, a standard result of linear algebra [14] states that if M is a symmetric $n \times n$ matrix, and v is a vector not orthogonal to the principal eigenvector $\omega_1(M)$, then the unit vector in the direction of M^kv converges to $\omega_1(M)$ as k increases without bound. Also (as a corollary), if M has only non-negative entries, then the principal eigenvector of M has only non-negative entries. Consequently, z is not orthogonal to $\omega_1(BB^T)$, and hence the sequence $\{y'_k\}$ converges to a limit y'^*. Similarly, one can show that if $\lambda_1(B^TB) \neq 0$, then B^Tz is not orthogonal to $\omega_1(B^TB)$. It follows that the sequence $\{x'_k\}$ converges to a limit x'^*.

Theorem 9. (Interpretation of x'^* and y'^*.) x'^* is the principal eigenvector of B^TB, and y'^* is the principal eigenvector of BB^T.

Proof: Similar to Theorem 7.

Definition 5. (Concept filtering.) A concept filtering operation CF on the sequence of forum posts $\{p_k\}$, ranks and keeps the posts of the n_{top} top-ranked nodes as $\{p'_k\}$ by authority weights using keyphrase K and Algorithm 3,

$$CF(\{p_k\}, n_{top}, K) = \{p'_k\}.$$

Theorem 10. (Nonlinear filtering.) Authority filtering and concept filtering are nonlinear operations.

Proof: Let S_1 and S_2 be two graphs. The filtering operations find the eigenvectors x_1^* and x_2^* of the matrices $C_{S_1}^TC_{S_1}$ and $C_{S_2}^TC_{S_2}$ of S_1 and S_2. The eigenvector x_3^* of the matrix $C_{S_1+S_2}^TC_{S_1+S_2}$ of the graph $S_1 + S_2$ is, in general, not equal to $x_1^* + x_2^*$. Also ranking and keeping the n_{top} entries of an eigenvector are nonlinear operations.

Filtering operations can be cascaded to form useful filters. For instance, we can filter the forum posts by timestamps and then by concepts.

$$CF(TF(\{p_k\}, t_{from}, t_{to}), n_{top}, K).$$

It is easy to see that the cascaded filter is also nonlinear.

4. MODEL VALIDATION

Scheduling Recap for a DSP MOOC. In the spring 2013, the schedule of education units of the DSP MOOC is shown in Table 2. 17 units are spaced in a 9-week time frame. We rule out the option to design a global MDR schedule because that would require 17 units to be taught in 4.5 weeks, which is a much heavier workload. Instead, we observe the topics of interest in the online forum and design the MDR schedules for three groups of educational units, respectively. Concept filtering is used to process the forum data and highlight the interesting forum posts that could be addressed in the MDR schedules.

$u_0, u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}$.

Table 2. Educational units of the DSP MOOC

Unit	Description
u_0	Introduction to DSP, basic mathematics
u_1	Hilbert space and approximation
u_2	Introduction to DFT
u_3	DFT examples
u_4	DFT, DFS, DTFT
u_5	Relationship between transforms
u_6	Linear filters
u_7	Frequency response
u_8	Realizable filters
u_9	Filter design
u_{10}	Interpolation and bandlimited signals
u_{11}	Sampling and aliasing
u_{12}	Stochastic signal processing
u_{13}	Image processing
u_{14}	Image filtering
u_{15}	Digital communication systems
u_{16}	Modulation and demodulation

Recap for $u_0, u_1, u_2, u_3, u_4, u_5$. The first MDR schedule is designed for units $u_0, u_1, u_2, u_3, u_4, u_5$ as shown in Figure 4. Note that u_0, u_1, u_2, u_4 are introduced in the leaf nodes, as they are considered as important units that need more attention. Units u_3, u_5 are introduced in the nonleaf nodes as they are example (DFT examples) and recap (relationship between transforms) modules, which would slightly change the 2013 spring schedule. Units X_0, X_1 are hinted by concept filtering.

$u_0, u_1, u_0, u_1, X_0, u_2, u_4, u_2, u_4, u_3, u_0, u_1, X_0, u_2, u_4, u_3, u_5, X_1$.

X_0 recap. The educational units and the respective concept filtering of X_0 recap are shown in Table 3. Some of the filtered forum posts are listed as Q & A pairs below.

Table 3. X_0 recap

Unit	Description	Concept filtering
u_0	Introduction to DSP	Signals, basic mathematics
u_1	Hilbert space	Vector space, orthogonality

- Q: Floor (w/ discontinuity), constant functions periodic?
- A: Yes ...

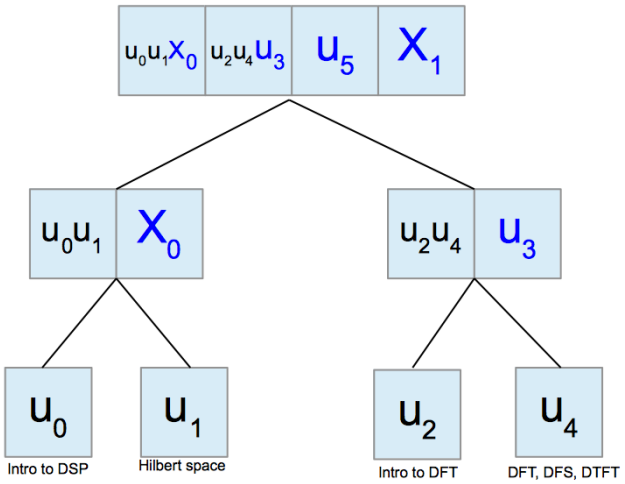


Fig. 4. A MDR schedule for u_1, u_2, u_3, u_4, u_5

- Q: Properties of sinusoids ...
- A: $W_N^{N/2} = \exp(-j2\pi\frac{N}{2N}) = \exp(-j\pi) = -1$ ...
- Q: Properties of the inner product ...
- A: Check the axioms of the inner product ...

The questions posted by the students include the periodic functions, some basic mathematics, the properties of the inner product. Our experience suggests that these questions are common among students and they should be included in the X_0 recap.

X_1 **recap**. The educational units and the respective concept filtering of X_1 recap are shown in Table 4. Some of the filtered forum posts are listed as Q & A pairs below.

Table 4. X_1 recap

Unit	Description	Concept filtering
u_2	Introduction to DFT	Fourier analysis, periodicity
u_4	DFT, DFS, DTFT	Fourier analysis, periodicity

- Q: Calculation of DTFT ...
- A: The modulation/convolution/linearity properties of DTFT pairs
- A: DTFT pairs of step/sine/sawtooth functions
- A: Properties of a Hermitian operator
- Q: Discrete Fourier Series vs. Discrete Fourier Transform
- A: Fourier analysis for periodic functions and discrete sequences ...
- Q: A factor $\frac{1}{\sqrt{N}}$ in front of both DFT and IDFT?
- A: Factor $\frac{1}{N}$ considered in the IDFT ...

The questions posted by the students include the calculation of DFT pairs, DFS and DFT, the mathematics of DFT and IDFT. Our experience suggests that these questions are common among students and they should be included in the X_1 recap.

Recap for u_6, u_7, u_8, u_9**.** The second MDR schedule is designed for units u_6, u_7, u_8, u_9 as shown in Figure 5. Units X_2, X_3, X_4, X_5 are hinted by concept filtering.

$u_6, u_7, u_6, u_7, X_2, u_8, u_9, u_8, u_9, X_3, u_6, u_7, X_2, u_8, u_9, X_3, X_4, X_5$.

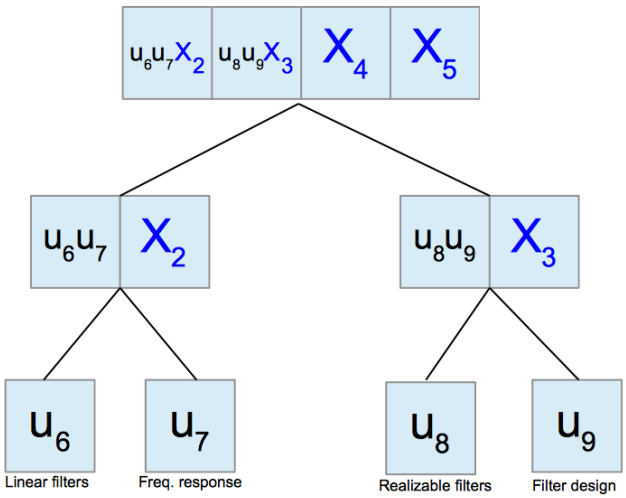


Fig. 5. A MDR schedule for u_6, u_7, u_8, u_9

X_2, X_3, X_4, X_5 **recap**. The educational units and the respective concept filtering of X_2, X_3, X_4, X_5 recap are shown in Table 5. Since X_2, X_3, X_4, X_5 are all related to digital filters, the concept filtering is targeted on keyphrases related to filters in Table 1. Some of the filtered forum posts are listed as Q & A pairs below.

Table 5. X_2, X_3, X_4, X_5 recap

Unit	Description	Concept filtering
u_6	Linear filters	Filters
u_7	Freq. response	Filters
u_8	Realizable filters	Filters
u_9	Filter design	Filters

- Q: Properties of a filter ...
- A: Linearity, stability, causality, linear phase ...
- Q: The transfer function of a filter ...
- A: Analyze the filter regardless of the input ...
- Q: Poles and zeros and filter properties ...
- A: Use ROC inside/outside the unit circle to prove ...
- Q: Moving the poles of a leaky integrator ...
- A: Resulting in a narrow-pass filter ...

The questions posted by the students include the properties of a filter, the transfer function of a filter, poles, zeros and filter design. Our experience suggests that these questions are common among students and they should be included in the $X_2, X_3, X_4, X5$ recap.

Recap for u_{10}, u_{11}. The third MDR schedule is designed for units u_{10}, u_{11} as shown in Figure 6. Unit X_6 is hinted by concept filtering.

$$u_{10}, u_{11}, u_{10}, u_{11}, X_6.$$

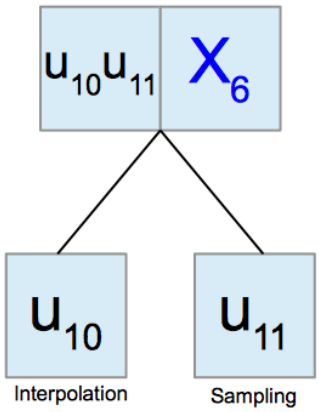


Fig. 6. A MDR schedule for u_{10}, u_{11}

X_6 recap. The educational units and the respective concept filtering of X_6 recap are shown in Table 6. Some of the filtered forum posts are listed as Q & A pairs below.

Table 6. X_6 recap

Unit	Description	Concept filtering
u_{10}	Interpolation	Interpolation
u_{11}	Sampling and aliasing	Sampling

- Q: Linear and Lagrange interpolation
- A: Use the samples as bases ...
- Q: Sampling at exactly the Nyquist rate ...
- A: Pointwise effect ...
- Q: How to determine the sampling freq?
- A: BW is known at the receiver side ...
- Q: What is the time duration between two samples ...
- A: $\frac{1}{f}$ ...

The questions posted by the students include the bases of interpolation, the determination of the sampling frequency, the sampling at the Nyquist rate, and the time duration between samples. Our experience suggests that these questions are common among students and they should be included in the X_6 recap.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated the potential of combining the scheduling problem in education and the MOOC data processing to analyze and design MOOCs. Our results include algorithms to construct the MDR schedules with some proven properties, and the design of the MDR schedules for the DSP MOOC. A number of interesting observations have been made and future directions opened.

First, the scheduling problem in education provides rich grounds for modellization. In this paper, two models, the FLR schedule and the MDR schedule are discussed in depth, and the benefits of MDR are illustrated. A number of extensions of MDR seems plausible. 1) Education units have dependency and it would be interesting to see how to model such dependency in the scheduling problem. 2) Student-dependent models, such as flexible & slow students, are also worth exploring. 3) In addition to infinite perfect learning, other types of learning, such as cramming, are also worth investigating.

Second, MOOC data processing provides rich grounds for model validation. Filtering forum data plays a key role in extracting the hints of recap, and a central role in the design of the MDR schedule for the DSP MOOC. More filtering algorithms for MOOC data processing can be further developed for different scheduling models.

To sum up, the scheduling problem in education and the MOOC data processing, as shown in this paper, have complementary strengths in the research of educating a model student in MOOCs.

Acknowledgement

We thank LCAV members for several helpful suggestions, and anonymous contributors to the DSP MOOC on Coursera.

6. REFERENCES

[1] L. Breslow, D. Pritchard, J. DeBoer, G. S. Stunmp, A. Ho, and D. Seaton, "Studying learning in the worldwide classroom research into edx's first mooc," *Journal of research and practice assessment*, 2013.

[2] J. Ahn, B. Butler, A. Alam, and S. Webster, "Learner participation and engagement in open online courses: Insights from the peer 2 peer university," *Journal of online learning and teaching (JOLT)*, vol. 9, 2013.

[3] D. Bruff, D. Fisher, K. McEwen, and B. Smith, "Wrapping a mooc: Student perceptions of an experiment in blended learning," *Journal of online learning and teaching (JOLT)*, vol. 9, 2013.

[4] V. Irvine, J. Code, and L. Richards, "Realigning higher education for the 21st-century learner through multi-access learning," *Journal of online learning and teaching (JOLT)*, vol. 9, 2013.

[5] Ezekiel J. Emanuel, "Online education: Moocs taken by educated few," *Nature*, 2013.

[6] Arpita Ghosh and Jon Kleinberg, "Incentivizing participation in online forums for education," in *Proceedings of the fourteenth ACM conference on Electronic commerce*, New York, NY, USA, 2013, EC '13, pp. 525–542, ACM.

[7] Booth J. L. Koedinger K. R. and Klahr D., "Education research. instructional complexity and the science to constrain it," *Science*, 2013.

[8] Timothy Novikoff, Jon Kleinberg, and Steven Strogatz, "Education of a model student," in *Proceedings of National Academy of Sciences*, New York, NY, USA, 2012, PNAS '12, pp. 525–542, PNAS.

[9] R. Caine and G. Caine, *Making connections: Teaching and the human brain*, Association for supervision and curriculum development, Alexandria, Va., 1991.

[10] F. Bernardon, "When neuroscience guides education," *Science*, 2013.

[11] Paolo Prandoni and Martin Vetterli, "Digital signal processing," `https://www.coursera.org/course/dsp`.

[12] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning, "Kea: Practical automatic keyphrase extraction," in *IN PROCEEDINGS OF THE 4TH ACM CONFERENCE ON DIGITAL LIBRARIES*, 1998, pp. 254–255.

[13] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sept. 1999.

[14] G. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.

CCDFs in MOOCs

Tao C. Lee
EDIC, LCAV, EPFL

June 6, 2014

Abstract

Massive Open Online Courses (MOOCs) have generated wide interest in the higher education community and are anticipated by many to bring impact to the future of higher education. In this research, we study the Complementary Cumulative Distribution Functions (CCDFs) of grades distribution of the Digital Signal Processing (DSP) MOOC on Coursera. We attempt to model and explain the CCDFs and use CCDFs as tools to gain insight into the MOOC phenomena.

Keyword: MOOCs, learner dynamics, CCDFs

1 Introduction

Since 2012, MOOCs have generated a learning wave across the world, and have recorded the learning trajectories of many learners on a diverse set of subjects, primarily on the level of university courses. While the future of this wave of free education beyond campus boundaries is still unknown, the learning trajectories left by learners have attracted the attention of several research communities. The central question of interest is: What can we learn from the learners?

The MOOC research, as an emergent field, is shaping into a new interdisciplinary research field blended with flavors of computer science, education, and psychology. With computer science (machine learning) researchers taking the data-driven approach, while education and psychology researchers taking the qualitative approach of interpreting the results of data analysis. Among the many emergent problems in the MOOC research, several problems have received wide attention in recent years: learner subpopulation analysis, dropout prediction and learning progress modeling, the incentives of forum participation, and the schedule of review.

The research of learner subpopulation analysis [Kizilcec13] intends to demystify the large learner population of MOOCs. The central claim is to show that the monolithic view to classifying learners between those who pass and those who do not is not sufficient to understand a large part of the MOOC learners.

The research of dropout prediction [Halawa14] stems from one of the major criticisms of MOOCs: low completion rate. A natural question arises as how to predict potential dropouts based on the learning trajectories of the learners, and how to engage the potential dropouts as early as possible. The question itself can be seen as a special case of learner subpopulation analysis, with the learners to be analyzed focused on the dropouts.

The research of learning process modeling has a long history in learning sciences and other scientific disciplines [Piech12]. The renewal interest in this direction is the expectation that the large datasets of learning trajectories can shed new light on the finesses and the accuracy of modeling.

The research of the incentives of forum participation has received attention partly because online forums serves as the main media of student feedback [Ghosh13, Rene14]. Without physical classrooms, online forums play the role of interaction between students and teaching staffs, as well as among students. Research in this direction focuses on how to increase the incentives of forum participation.

In this research, we study the CCDFs of final grades distribution in the DSP MOOC, and use CCDFs as tools to understand the behavior of the learners. While CCDFs have been widely used in network sciences [Wu10, Dodds03, Barabasi05], CCDFs have not been extensively discussed in the context of MOOCs. Comparing to the power-law like CCDFs in network sciences, we have found CCDFs in MOOCs usually exhibit multi-modal behaviors that are particularly interesting. This observation is not addressed in any of the previous MOOC research. We attempt to model and explain the CCDFs and use CCDFs as tools to gain insights to the MOOC phenomena.

2 Datasets

The MOOC dataset under investigation is the learning trajectories of the Digital Signal Processing (DSP) class, spring and fall 2013 semesters, on the Coursera platform. The dataset consists several parts as shown in Table 1.

Table 1: Description of the MOOC dataset

Source data	Description
Database dumps	Quiz submissions and grades, forum posts and comments
Clickstreams	Time-stamped video, quiz, wiki, forum click events
Other data	Survey, quiz responses

Database dumps provide data tables of quiz submissions, grades and final grades of the learners. These data tables can be used to measure the performance of the learners. Also forum posts and comments can be used to measure the engagement levels and feedback levels of learners.

Clickstreams provide serial data of time-stamped video, quiz, wiki, forum events. These events can be used to measure the engagement levels, learning methods, learning habits, learning purposes, learning interests of learners.

Other data sources include survey and quiz responses. The quiz and survey responses can be used to measure learners' performance, engagement levels and feedback levels.

Table 2: Description of the survey data

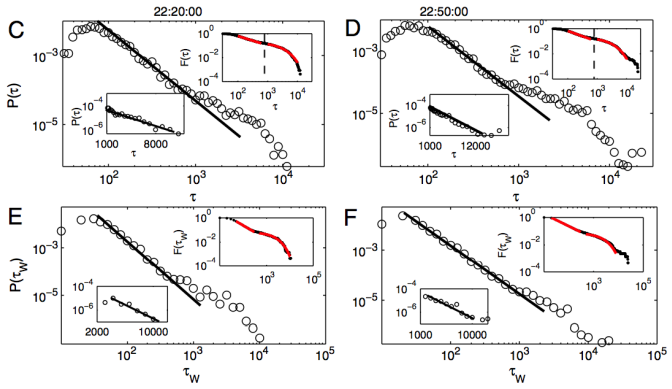
Survey data	Description (participated / total learners)
Spring'13	9181 / 48401 (19%)
Fall'13	6319 / 35318 (18%)

In this research, we use the survey data to acquire the background information of learners, including age, occupation, gender, geographical location, motivation, etc. The collected information is used to produce CCDFs for different groups of learners. As well as to gain insight to the particular mode of distribution in the CCDFs. Final grades are used a the indicative metrics to measure the learning performance. Since in the DSP MOOC, final grades is calculated from 17 homework grades, they also bear the meaning of participation levels in the class.

Learner information is obtained from an initial survey filled by the learners on a voluntary basis. As shown in Table 3, approximately 20% of learners participated in the survey (19% for spring'13 and 18% for fall'13). Although only a fraction of the total learners, the number of learners that provided the information was still huge compared to on-campus courses (Considering a typical on-campus course with 50 students, the number of learners submitted the survey is 120x to 180x more). We will analyze the information contained in the survey data in the subsequent sections.

3 Complementary Cumulative Distribution Functions (CCDFs)

CCDFs have been widely studied in network sciences [Wu10, Dodds03] and the study of human dynamics [Barabasi05]. The distributions have been classified by their tails in several types: light-tail distribution, heavy-tail distribution and the mix of the two. Many models have been proposed to explain and simulate the types of distribution of interest. In particular, a model referred to as the preferential attachment is widely used to explain the emergence of heavy-tail distribution, and similar models also appear in social sciences and economical sciences.



(a)

Figure 1: Interval distribution of human communications in log-log scale, [Wu10]

A CCDF of x is defined as

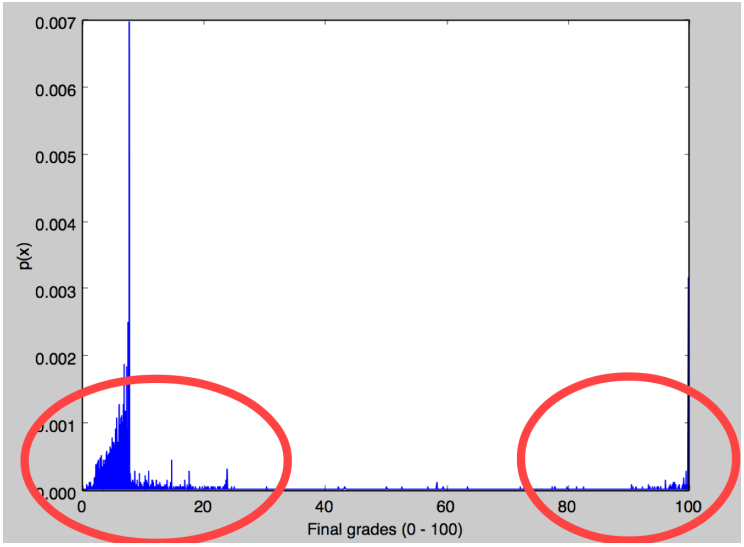
$$CCDF(x) = 1 - \sum_{j=1}^{x-1} p(j) = \sum_{j=x}^{\infty} p(j) \tag{1}$$

From its definition, we see the derivative of a CCDF is a histogram.

$$-\frac{d(CCDF(x))}{dx} = p(x) \tag{2}$$

Since integration smooths out the noise, CCDFs are more suitable for modeling than histograms. In network sciences, CCDFs are usually shown in log-log scale, partly because the numbers are large enough. In the DSP MOOC, the final grades are in the range of [0, 100], so we show the CCDFs in linear-linear scale.

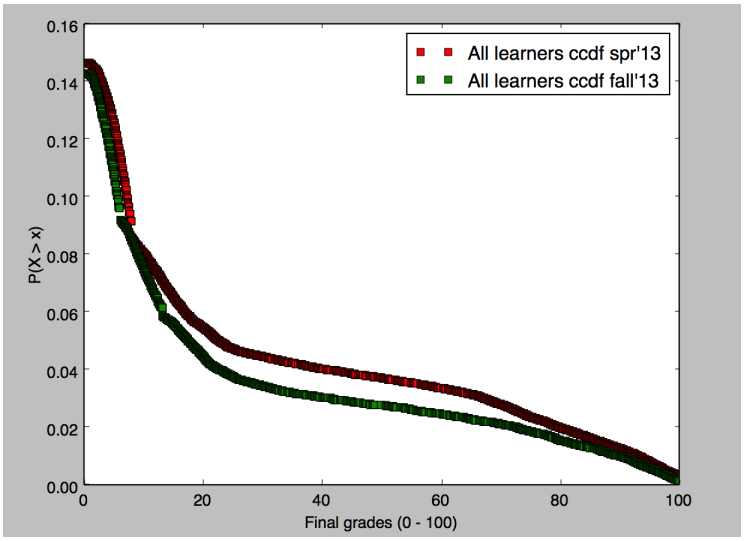
4 CCDFs of Final Grades



(a)

Figure 2: Final grades histogram of all learners, Spring'13

A final grades histogam, as shown in Figure 2, is very noisy and difficult to model. A rough quantitative observation is that many learners concentrate on the two extremes of the grades distribution, a phenomenon we refer to as bimodal distribution of learning performance in MOOCs. For the rest of this paper, we study this bimodal distribution using CCDFs.



(a)

Figure 3: CCDFs of all learners

The CCDFs of all learners with final grades greater than 0 (learners without any submissions of homeworks are excluded) are shown in Figure 4. Interestingly, the CCDFs exhibit piecewise patterns that can be further analyzed. As a simple model, we model the CCDFs by piecewise linear functions.

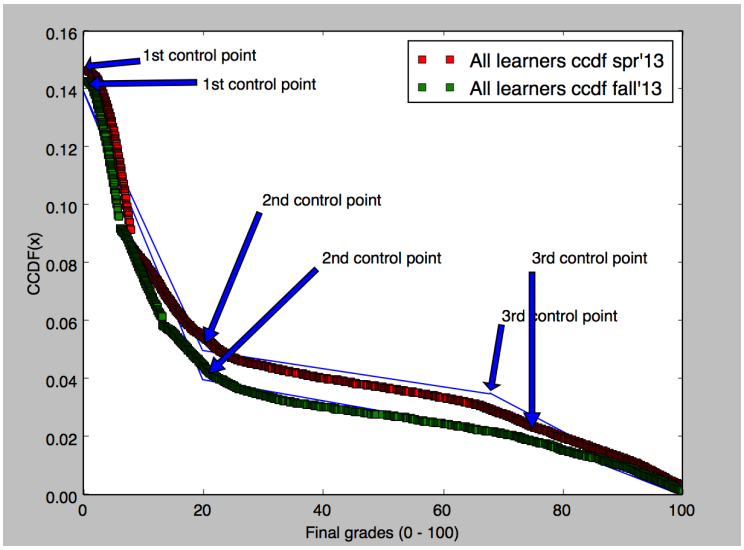
$$CCDF(x) = \cup_i f_i(x), x \in [0, 100]$$

where

$$f_i(x) = a_i x + b_i, x \in [l_i, u_i]$$
$$\cup_i [l_i, u_i] = [0, 100]$$

As a first attemp, we model the CCDFs of all learners by linear functions in three regions: low-score, middle-score, and high-score regions.

$$CCDF(x) = \begin{cases} a_1 x + b_1 & : x < \delta_1 \\ a_2 x + b_2 & : \delta_1 \leq x < \delta_2 \\ a_3 x + b_3 & : \delta_2 \leq x \leq 100 \end{cases} \tag{3}$$



(a)

Figure 4: CCDFs of all learners

Where $a_1, b_1, a_2, b_2, a_3, b_3$ are determined by three control points $\alpha_1, \alpha_2, \alpha_3$.

Table 3: Description of the survey data

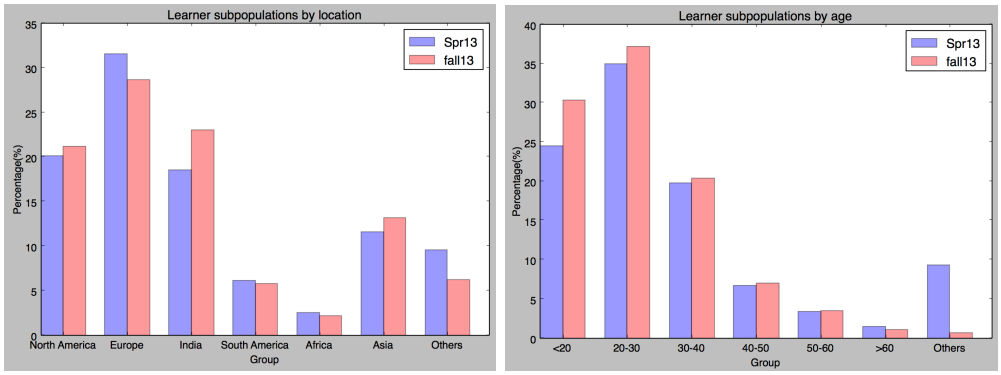
CCDFs of final grades	Control points $\alpha_i(x, CCDF(x))$
Spring'13	$\alpha_1 = (0, 0.14)$, $\alpha_2 = (20, 0.05)$, $\alpha_3 = (68, 0.35)$
Fall'13	$\alpha_1 = (0, 0.14)$, $\alpha_2 = (20, 0.04)$, $\alpha_3 = (75, 0.02)$

Comparing the three control points of the two CCDFs, we have several interesting observations. 1) α_1s of spring and fall semesters almost overlap with each other. 2) α_2 of the Fall'13 class, drops lower than that of Spring'13 class. The steeper slop indicates more learners in the Fall'13 class falls in the first segment of the CCDF, between (0, 20]. 3) α_3 of the Fall'13 class shifts towards 100 and produces gradual slope in the third segment, between (75, 100]. The gradual slope indicates less learners made to this segment.

One interesting question is what caused the control points to shift locations and produced different CCDFs? Is the difference caused by differet student subpopulations in different semesters? We investigate this question in the following sections.

5 Learner subpopulations

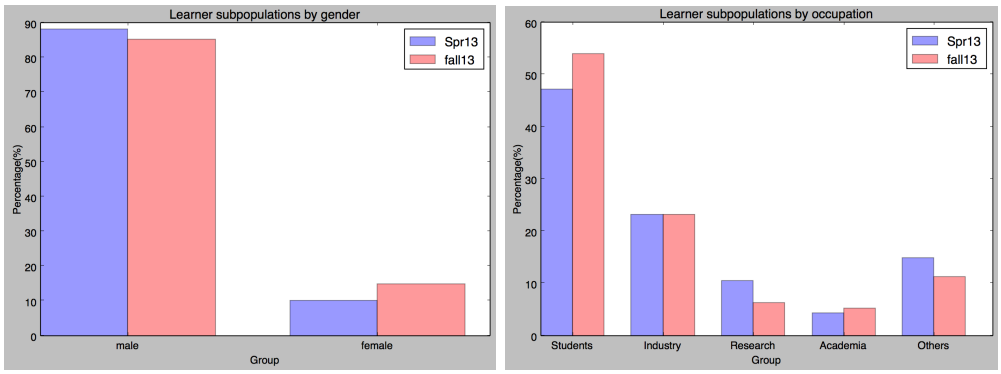
Before investigating the CCDFs by learner subpopulations, let us first discuss the distribution of the learner subpopulations in more detail.



(a) Region (b) Age

Figure 5: Subpopulation by region and age

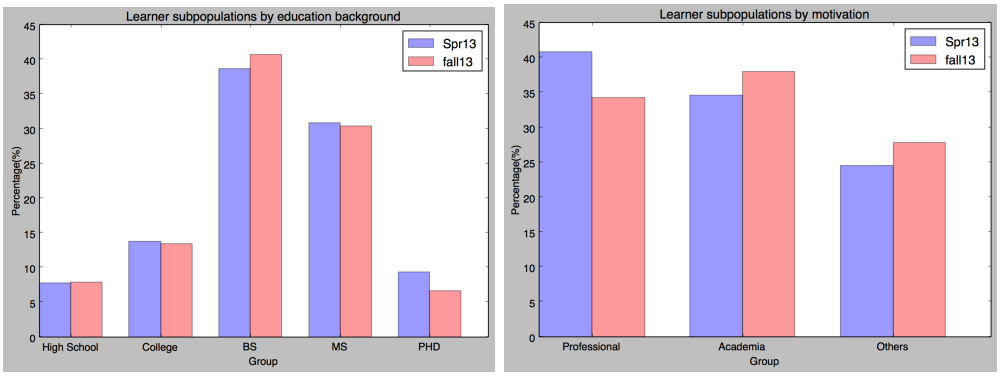
The distributions of the learner subpopulations by region and age are shown in Figure 5. Comparing learners by region, we see more Asian learners joined the class in the fall semester, and the ratio of European and North American learners dropped. Comparing learners by age, we see more young learners joined the class in the fall semester, in particular in the age below 20, and between 20 and 30.



(a) Gender (b) Occupation

Figure 6: Subpopulation by gender and occupation

The distributions of the learner subpopulations by gender and occupation are shown in Figure 6. Comparing learners by gender, we see more female learners joined the class in the fall semester, and the ratio of male learners dropped. Comparing learners by occupation, we see more student learners joined the class in the fall semester.



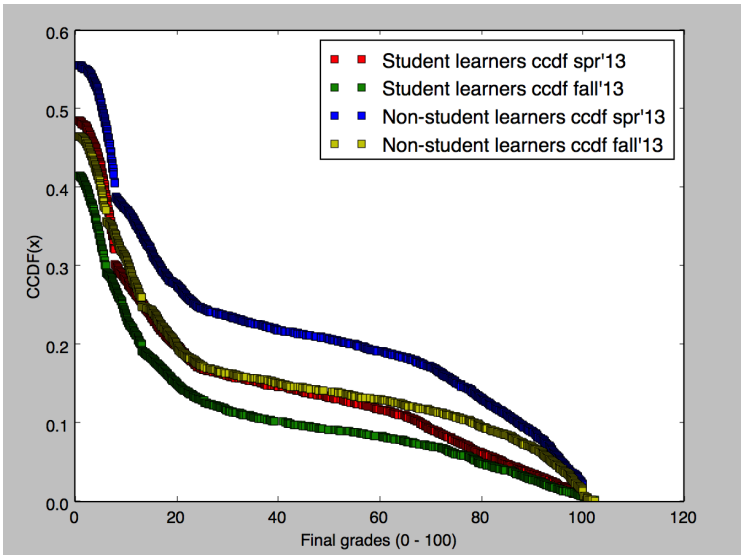
(a) Education background (b) Motivation

Figure 7: Subpopulation by education background and motivation

The distributions of the learner subpopulations by education background and motivation are shown in Figure 6. Comparing learners by education background, we see more learners with BS degree joined the class in the fall semester, while the ratio of the learners with other degrees dropped. Comparing learners by motivation, we see more learners from academia joined the class in the fall semester.

6 Student vs. Non-student Learners

The first question pops out is why student or non-student learners have very different CCDFs? Student learners are learners that have other learning tasks at school, whereas the non-student learners are professionals from the industry or retired individuals. We compute the CCDFs of the student and non-student learners and show the results in Figure 8.



(a)

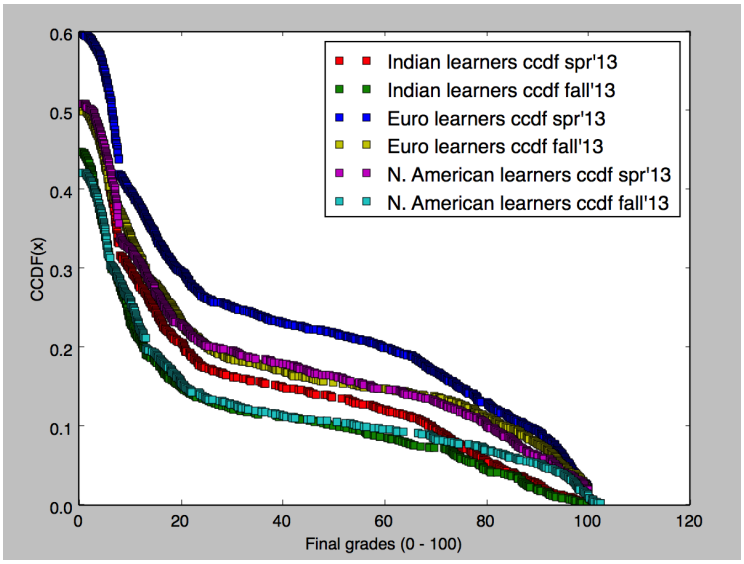
Figure 8:

As we can see, the non-student learners consistently outperform the student learners in both the spring and the fall semesters. Using the three region model developed previsouly, we can observe more non-student learners in the high-score region, less learners in the low-score region. Spring semester is especially different, in that non-student learners perform the best, and there is a large gap between learning performance of the spring and the fall semesters.

Learning performance degrades from the spring to the fall semester. This could be explained by the observation that spring 2013 was the first offering of the class, and there was more attraction and more media coverage of the MOOC tidal wave. so we can think of the spring semester as the excited state, and the drop in the fall semester is a normal process of converging to a stable state.

7 European, North American and Indian Learners

Since the top three regions where the learners came from are Europe, North America and India, we compute the CCDFs for these three regions in Figure 9.



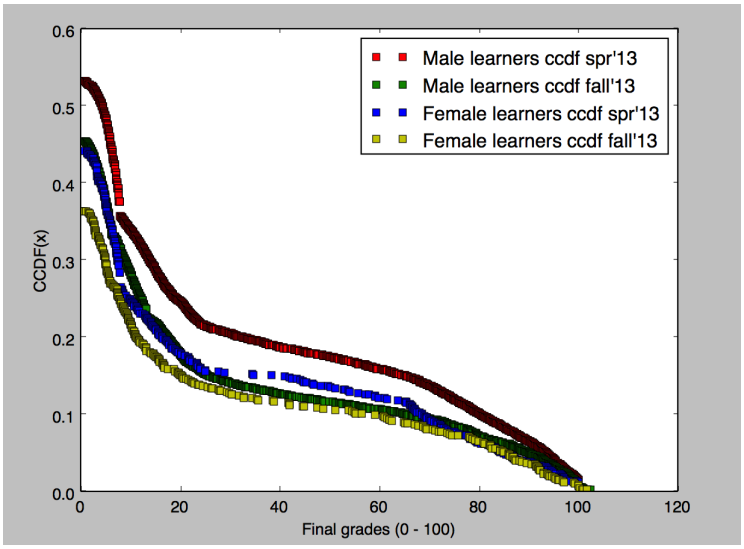
(a)

Figure 9:

As we can see, the European learners consistently outperform the North American and Indian learners in both the spring and the fall semesters. North American learners are on par with Indian learners. Using the three-region model, we observe more European learners in the high-score region and less in the low-score region. Excited states are still observed in the spring semester, and drop are observed in the fall semester, for European, North American and African learners.

8 Male vs. Female Learners

Another interesting question is whether there is a gender gap in learning performance in the DSP MOOC? We compute the CCDFs for male and female learners in Figure 10.



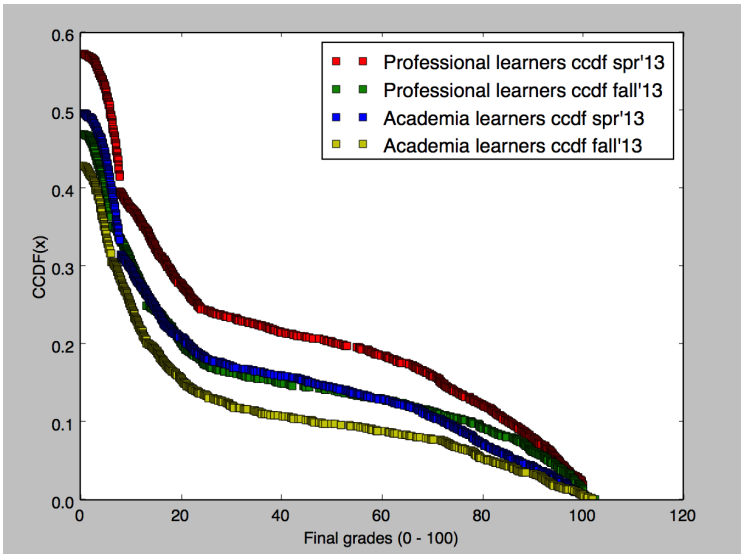
(a)

Figure 10:

As we can see, except for the spring class where male learners stand out, both genders perform equally well in the fall semester. The hype in the spring semester is remarkable in that male learners outperformed female learners by a large gap, but the difference decreased in the fall semester, especially in the middle-score and high-score regions, where the CCDFs almost overlap.

9 Professional vs. Academia learners

Another interesting question is whether motivation produces a difference in the learning performance? We compute the CCDFs for professional learners and academia learners in Figure 11.



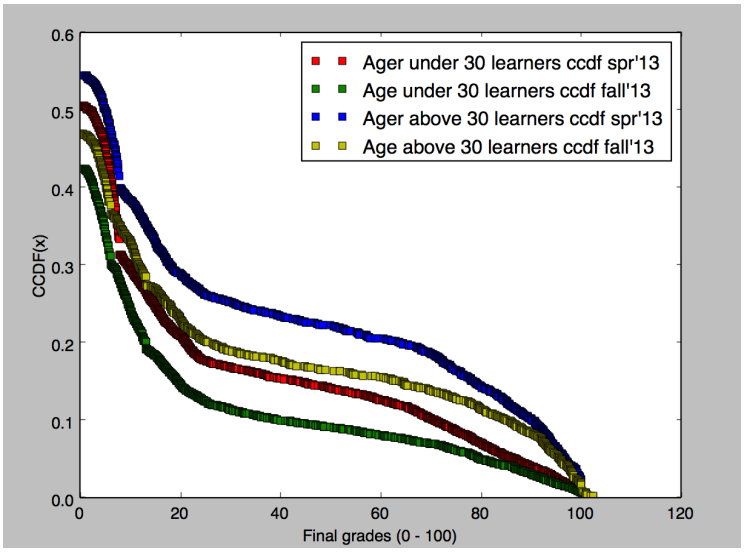
(a)

Figure 11:

As we can see, professional learners consistently outperform academia learners in both the spring and the fall semesters. The result is no surprise, since many of the professional learners are non-students, and we have discovered in the previous section that non-studnet learners outperform student learners. The hype in the spring semester can be observed and explained that more professional learners joined the class and were active throughout.

10 Age below 30 vs. above 30 learners

Another interesting question is whether age makes a difference in the learning performance. We compute the CCDFs for learners with age below 30 and age above 30 in Figure 12.



(a)

Figure 12:

As we can see, aged learners consistently outperform the young learners in both the spring and the fall semesters. The result is also no surprise, since non-student or professional learners are usually older and oftentimes with age above 30. The hype in the spring semester can also be explained by the observation that more aged learners joined the class and were active throughout. The hype decreased in the fall semester and could be understood as a normal process to converge to a stable state.

11 Conclusions

Based on the results of CCDFs by subpopulations, we see, in the fall semester, the DSP MOOC reached out to more young student learners and more learners outside of Europe, which is certainly a good thing in terms of the diversity coverage of the class. However, because non-student learners have better learning performance, the increase of the student learners actually contributed to the drop of the final grades CCDFs in the fall semester. This signal might have deep implications in online education, as we should decide to cater the class to the student learners or non-student learners. These two groups of learners have very different learning motivation and education background,

Another thing worth noticing is that geographical locations do make differences to the performance of the students. While MOOCs promise to bring down the walls imposed by geographical locations, there is still much to do in this regard. In particular, EPFL is a European school, and there are more things to do to reach out to the learners outside Europe.

12 Future work

In this research, we investigate the use of CCDFs in MOOCs, in particular in the modeling of the learning performance of the learners. We study the CCDFs of the spring and fall semesters by learner subpopulations. The subpopulations are characterized by age, gender, education background, occupation and geographical location. In the future, more ways to classify learners into different subpopulations can be investigated and coupled with CCDFs to gain more insight into the MOOC phenomena.

References

[Piech12] C. Piech, Mehran Sahami, Daphne Koller, Stephen Cooper, and Paulo Blikstein. *Modeling how students learn to program.* Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, 2012.

[Novikoff12] T. Novikoff, J. Kleinberg, and S. Strogatz. *Education of a model student.* Proceedings of National Academy of Sciences, 2012.

[Kizilcec13] R. Kizilcec, C. Piech and E. Schneider. *Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses* Proceedings of the Third International Conference on Learning Analytics and Knowledge, 2013.

[Halawa14] Sherif Halawa, Daniel Greene and John Mitchell. *Dropout Prediction in MOOCs using Learner Activity Features*, eMOOCs, 2014.

[Ghosh13] Arpita Ghosh and Jon Kleinberg. *Incentivizing participation in online forums for education*, EC, 2013.

[Rene14] Rene Kizilcec, Emily Schneider, Geoffrey Cohen and Daniel McFarland. *Encouraging forum participation in online courses with collectivist, individualist and neutral motivational framings*, eMOOCs, 2014.

[Wu10] Ye Wu, , Changsong Zhoud, Jinghua Xiaob, Jürgen Kurthsa, and Hans Joachim Schellnhuber *Evidence for a bimodal distribution in human communication.* Proceedings of National Academy of Sciences, 2010.

[Dodds03] Peter Sheridan Dodds, Duncan J. Watts, and Charles F. Sabel *Information exchange and the robustness of organizational networks.* Proceedings of National Academy of Sciences, 2003.

[Barabasi05] Albert-Laszlo Barabasi *The origin of bursts and heavy tails in human dynamics* . Nature, 2005.

[Raissa07] Raissa M. D'Souza, Christian Borgs, Jennifer T. Chayes, Noam Berger, and Robert D. Kleinberg *Emergence of tempered preferential attachment from optimization.* Proceedings of National Academy of Sciences, 2007.