

Titre courant : Analyse statistique des profils et de l'activité des participants d'un MOOC

Analyse statistique des profils et de l'activité des participants d'un MOOC

Avner Bar-Hen, Université Paris Descartes, Laboratoire MAP5 (CNRS, UMR8145), 45 rue des
Saints-Pères, 75006 Paris – France, avner.bar-hen@mi.parisdescartes.fr, 30

Hubert Javaux, Université Sorbonne Paris Cité, Service SAPIENS, 10 rue Charles V,
75004 Paris – France, hubert.javaux@parisdescartes.fr

Nathalie Villa-Vialaneix, INRA, UR875 MIA-T, BP 52627, 31326 Castanet-Tolosan
cedex – France, nathalie.villa@toulouse.inra.fr

Titre courant : Analyse statistique des profils et de l'activité des participants d'un MOOC

Analyse statistique des profils et de l'activité des participants d'un MOOC

Résumé

L'acronyme MOOC (Massive Online Open Course) est utilisé pour désigner les plateformes d'enseignement en ligne qui proposent des cours ouverts et qui peuvent s'adresser à des centaines, des milliers, voire des dizaines de milliers d'étudiants simultanément. Le but du présent article n'est pas de proposer une analyse du phénomène MOOC dans sa généralité, mais d'offrir une vision originale d'un MOOC, au travers de ses étudiants, de leur profil et de leur activité lors du cours. Pour ce faire, nous analysons les données collectées lors d'un des premiers cours lancés sur la plateforme française FUN, le cours « Fondamentaux en statistique ». Nous y décrivons la communauté des étudiants, leur profil socioéconomique, leurs motivations, leur activité sur le forum du cours. Nous étudierons comment les échanges sur le forum se structurent lors du cours.

Mots-clés

MOOC, statistique, analyse de données, typologie d'étudiants, réseaux sociaux

Abstract

The acronym MOOC (Massive Open Online Course) refers to platforms that offer on-line education and open courses to hundreds, thousands or even tens of thousands of students simultaneously. The purpose of this article is not to provide an analysis of the phenomenon MOOC in its generality but to offer an original vision of a given MOOC, through its students, their profile and their activity during the course. To do so, we analyze the data collected during one the first course of the French platform FUN, "Fundamentals in Statistics". We describe the community of students, their socioeconomic profile, their motivations, their activities on the course forum. We also study how the forum exchanges are structured in the course.

Keywords

MOOC, statistics, data mining, typology of students, social network

Analyse statistique des profils et de l'activité des participants d'un MOOC

Introduction

L'acronyme MOOC (Massive Online Open Course) est utilisé pour désigner les plateformes d'enseignement en ligne qui proposent des cours ouverts et qui peuvent s'adresser à des centaines, des milliers, voire des dizaines de milliers d'étudiants simultanément. Le sens du terme a peu à peu glissé pour désigner à présent également les cours eux-mêmes et c'est dans ce sens que nous l'utiliserons dans la suite de cet article.

L'engouement massif que ces plateformes ont suscité a fait des MOOC un objet d'attention : les institutions publiques nationales et internationales ainsi que de nombreuses écoles, universités et grandes entreprises investissent dans la création de ce type de cours avec des objectifs divers : aider au développement de la formation continue, de la formation des pays en voie de développement, accroître leur visibilité... le nombre croissant d'écrits sur les MOOC témoigne de la grande actualité du sujet, que ces publications soient à destination du grand public ou d'un public plus spécialisé (voir par exemple Rolfe, 2015 pour une revue récente de la littérature).

Le but de cet article n'est pas de proposer une analyse du phénomène MOOC, de ses enjeux, de ses réussites, de ses échecs ou de son avenir, mais d'offrir une vision originale d'un MOOC, au travers de ses étudiants, de leur profil et de leur activité lors du cours. Les deux premiers auteurs du présent article sont à l'origine d'un des premiers cours lancés sur la plateforme FUN, le cours « Fondamentaux en statistique ». Ce cours de statistique élémentaire, annoncé sans prérequis nécessaire, s'adressait à un public large. Le but du présent article est d'analyser la communauté des étudiants, leur profil socioéconomique, leurs motivations, leur activité dans le forum du cours. Nous discuterons celle-ci pour montrer de qui la communauté

des étudiants d'un MOOC donné est constituée et comment celle-ci, qui fait souvent preuve d'une grande activité, se structure lors du cours. En particulier, nous nous focaliserons sur l'activité du forum de discussion : en effet, les forums sont souvent décrits comme un ingrédient essentiel d'un cours en ligne efficace, car cette communication asynchrone représente la principale interaction éducative sur ce type de cours (pour des discussions sur ce point, voir par exemple Anderson et Kanuka, 1997; Berge, 1995; Mak, Williams et Mackness, 2010; Salmon, 2004). Un nombre croissant d'études posent ainsi la question de l'utilisation du forum, de son rôle (forum de discussion ou bien site de questions/réponses) ou du rôle des modérateurs et des divers participants (par exemple, Anderson, Huttenlocher, Kleinberg et Leskovec, 2014; Baxter et Haycock, 2014; Onah, Sinclair et Boyatt, 2014). Les résultats de ces divers travaux sont parfois contradictoires et le but de cet article est donc d'apporter une contribution à ce débat afin de fournir des éléments pour une meilleure utilisation des forums ou pour réfléchir à des solutions de rechange. Notre travail vise à analyser, à partir de l'étude du forum d'un MOOC, les comportements des divers participants pour fournir un diagnostic permettant à terme une utilisation plus efficace de cet outil de communication entre étudiants et entre étudiants et équipe enseignante. Notons que cette analyse est, à notre connaissance, une des premières analyses du comportement des étudiants sur un MOOC français : cette étude a donc aussi valeur de comparaison entre les études précédentes, principalement effectuées sur les participants des grandes plateformes américaines (edX, voir Breslow *et al.*, 2013; Brinton *et al.*, 2013) et les participants de la plateforme française FUN.

La suite du présent article est organisée comme suit : dans la section 2, nous présenterons brièvement le cours « Fondamentaux en statistique » pour situer le contexte de notre étude et donnons un aperçu descriptif du profil des étudiants qui s'y sont inscrits. Dans la section 3, nous

nous concentrerons sur l'analyse statistique des données collectées sur le forum de discussion en nous focalisant, en particulier, sur la structure communautaire des échanges.

Enfin, nous signalons en préambule que l'anonymat des étudiants a été scrupuleusement respecté : les données présentées ici ne concernent que les profils des étudiants et leur activité, sans possibilité de les identifier.

Présentation générale du MOOC « Fondamentaux en statistique »

FUN est la plateforme nationale française visant à promouvoir l'utilisation des cours en ligne ouverts et massifs. Le cours « Fondamentaux en statistique » a fait partie de la première série de huit cours proposés par FUN : il a démarré le 16 janvier 2014 pour une durée de 5 semaines. Ce cours se proposait de fournir une introduction aux concepts de base en statistique. L'angle de présentation du cours était de s'appuyer sur des problématiques d'intérêt général (par exemple, l'accent a été mis sur l'application pratique aux questions de santé publique) plutôt que sur le développement théorique et mathématique.

Le contenu du cours était structuré en 5 séances principales, correspondant aux 5 semaines de cours, et dévoilées progressivement. Les concepts abordés étaient les suivants :

- semaine 1 : statistique descriptive univariée;
- semaine 2 : statistique descriptive bivariée et ACP;
- semaine 3 : classification non supervisée;
- semaine 4 : tests non paramétriques;
- semaine 5 : tests paramétriques.

Chacune de ces semaines était elle-même structurée en sous-séquences courtes comportant chacune une vidéo d'une vingtaine de minutes et un quiz d'auto-évaluation des notions abordées dans la vidéo.

En outre, chaque ensemble de séquences était accompagné d'une introduction qui précisait les objectifs de la semaine, les prérequis nécessaires et le temps nécessaire pour, d'une part, écouter les vidéos, et d'autre part, réaliser le travail personnel. Une partie synthèse concluait chaque semaine et reprenait les notions abordées en les approfondissant sur certains aspects pour les personnes souhaitant aller plus loin. Dans ces synthèses, les données de profil des étudiants ont parfois été utilisées pour illustrer le cours. Enfin, un exercice de synthèse demandait chaque semaine aux étudiants d'utiliser les notions abordées pour analyser un jeu de données réel et de rédiger les conclusions de leurs analyses sous forme d'une courte dissertation. Pour des questions techniques, les exercices ont été gérés à l'extérieur de FUN, sur la plateforme Moodle de l'Université Sorbonne Paris Cité : les étudiants y récupéraient leurs énoncés, y déposaient leurs solutions et, dans une seconde phase, évaluaient le travail de leurs pairs¹. Au total, le cours nécessitait une charge de travail hebdomadaire moyenne estimée à environ 6 à 8 heures par semaine, c'est-à-dire 30 à 40 heures de disponibilité sur les 5 semaines.

Enfin, en parallèle du matériel de cours mis à disposition des étudiants, le cours contenait des fonctionnalités pour permettre l'échange entre étudiants et avec l'enseignant : un wiki, qui permettait à l'enseignant et aux étudiants d'amener des informations supplémentaires sur les notions du cours (liens extérieurs, documents additionnels, outils utiles...) et un forum permettant les discussions.

Du point de vue des choix pédagogiques, l'objectif du cours était de présenter à un public totalement novice des notions de statistiques permettant de mener une analyse de données élémentaires. Le cours a été orienté de manière à présenter les aspects mathématiques des méthodes et à les illustrer sur des cas d'études simplifiés et génériques. La mise en œuvre pratique, qui aurait nécessité la présentation d'un logiciel statistique, n'a pas été abordée.

Certains des étudiants qui ont rendu, durant les premières semaines, les réponses aux exercices ont donc traité ceux-ci « à la main », c'est-à-dire avec une calculatrice et des graphiques réalisés sur une feuille de papier au crayon.

Le MOOC « Fondamentaux en statistique » a obtenu 7 997 inscrits dont « seulement » 4 600 étaient déjà inscrits au premier jour du cours. En comparaison, seuls 1 832 étudiants (22,61 %) ont ouvert un compte Moodle pour pouvoir déposer les exercices. Ceci signifie qu'au moins un quart des étudiants a eu une démarche active par rapport au cours. Enfin, 251 étudiants ont achevé l'ensemble des cinq exercices. Un questionnaire autodéclaratif conçu par FUN et proposé aux étudiants lors de leur inscription a permis d'obtenir des informations sur le profil des étudiants : 85 % des inscrits du cours (6 918 étudiants dont 32 % de femmes) ont terminé celui-ci, ce qui en fait une source de données que l'on peut considérer comme fiable. L'analyse descriptive détaillée de celui-ci peut être consultée à <http://tuxette.nathalievilla.org/?p=1614>. Parmi les faits marquants, on peut noter qu'environ trois quarts des étudiants dont il a été possible de trouver la localisation géographique étaient domiciliés en France, les autres étant majoritairement des étudiants étrangers provenant de l'Afrique francophone. Les MOOC, ouverts et gratuits et qui ne requièrent que d'avoir accès à une connexion Internet, revêtent en effet un intérêt particulier pour la formation dans les pays du tiers monde (voir par exemple Cisel, 2014).

Les étudiants inscrits ont en moyenne environ 36 ans, ce qui indique qu'une proportion assez faible d'étudiants sont des étudiants qui souhaitent approfondir ou compléter un cours qu'ils sont eux-mêmes en train de suivre de manière classique (c'est-à-dire en formation initiale à l'université, voir par exemple Alario-Hoyos *et al.*, 2013). Ce fait est corroboré par une très forte présence d'inscrits ayant déjà un niveau master (près de 50 % des réponses), qui est

supérieur au niveau d'études classique auquel se situe habituellement le cours « Fondamentaux en statistique » (premier cycle universitaire en France).

Analyse du forum

Un des outils mis à la disposition des étudiants sur la plupart des MOOC est le forum de discussion : celui-ci est le lieu d'échanges entre étudiants, mais aussi entre les étudiants et l'équipe enseignante et permet donc d'avoir une image de l'interactivité lors du déroulement du MOOC. Sur cet outil, n'importe quelle personne, étudiant ou enseignant, peut ouvrir un fil de discussion qui se caractérise par un nouveau message avec un nouveau titre sur le forum. N'importe quelle autre personne peut répondre à ce premier message dans le même fil de discussion par un deuxième message et le nombre de messages de réponse pour un fil de discussion donné n'est pas borné. Le fil de discussion a donc, en règle générale, une homogénéité thématique et, comme tous les fils sont accessibles à tous, une thématique donnée est normalement abordée dans un seul fil (cette règle n'étant que théorique, car en pratique, il est courant qu'une thématique soit abordée deux ou plusieurs fois dans des fils différents, lorsque la personne qui ouvre un nouveau fil n'a pas pris la peine de vérifier, en lisant l'intégralité des autres fils, si la réponse à sa question ou à son commentaire était déjà présente dans un autre fil).

Dans cette section, nous nous proposons d'analyser ces fils de discussion, en décrivant tout d'abord le forum d'un point de vue quantitatif, puis en nous intéressant à la structuration des groupes d'étudiants et enfin, en analysant les textes des commentaires eux-mêmes.

Analyse descriptive du forum

Le forum a été actif du 16 janvier 2014 au 20 mars 2014. Il a été utilisé par 502 étudiants identifiés, trois membres de l'équipe enseignante et 23 personnes anonymes, qui ont ouvert

716 fils de discussion correspondant à 3 024 messages distincts. Le nombre de participants du forum représente donc un pourcentage assez faible du nombre d'inscrits (environ 6,6 %), qui est même faible devant le nombre d'étudiants ayant ouvert un compte Moodle pour pouvoir déposer des exercices (pour rappel, 1 832 étudiants). Toutefois, si on le compare au nombre d'étudiants ayant proposé une solution pour les cinq exercices à rendre sur la plateforme Moodle (251 étudiants), on constate que les étudiants actifs, qui cherchent à s'investir dans le cours en participant par l'intermédiaire du forum, sont deux fois plus nombreux que les étudiants qui ont complété intégralement le cours. Ces résultats sont consistants et plutôt même le signe d'une activité assez importante dans le MOOC « Fondamentaux en statistique » si on les compare aux études de Breslow *et al.* (2013) et Manning et Sanders (2013) qui montrent qu'en général moins de 10 % des inscrits (et le plus souvent moins de 5 % de ceux-ci) participent aux forums.

Des études précédentes montrent que le rôle de l'instructeur ou du modérateur est important pour le succès des forums en ligne (Anderson et Kanuka, 1997; Berge, 1995). De même, Salmon (2004) fonde l'apprentissage en ligne et l'interaction des forums de discussion sur le fait que l'instructeur ou le modérateur encourage les activités et discussions (Harasim *et al.*, 1995; Mak *et al.*, 2010). Notons que d'autres auteurs sont moins enthousiastes et font remarquer que peu d'enseignants sont impliqués dans les forums, voire que leur rôle peut être négatif en matière d'engagement étudiant (voir Baxter et Haycock, 2014; Onah *et al.*, 2014). Sur le MOOC « Fondamentaux en statistique », le parti pris de l'équipe enseignante a été de ne pas ou peu ouvrir de fils de discussion (seulement 5) pour laisser la liberté aux étudiants d'aborder les thèmes de leurs choix : l'analyse qui suit est donc celle d'un forum qui n'a pas été structuré a priori. Toutefois, l'équipe pédagogique a été très active avec 1 080 commentaires (environ 36 % des commentaires) et a concentré ses interventions sur les fils un peu plus longs (les fils très

courts, de 2 ou 3 messages, correspondent souvent à une question posée par un étudiant à laquelle un autre étudiant répond de manière appropriée). En complément, la figure 1 montre la distribution du nombre de commentaires par étudiant. Si certains étudiants interviennent très fréquemment sur le forum (avec un maximum de 48 commentaires pour un étudiant), la plupart des étudiants n'interviennent que très peu : 224 étudiants (soit plus de 40 % des participants du forum) n'ont posté qu'un commentaire et 111 étudiants n'en ont posté que deux.

Note à la revue : insérer ici la figure 1

On voit donc la communauté se structurer autour de quelques étudiants très actifs qui sont engagés dans les forums et qui, tout en ayant un rôle moindre par rapport à l'équipe enseignante, dynamisent les échanges. Il faut aussi noter que le profil des étudiants participant aux forums diffère légèrement du profil de l'ensemble des étudiants : les participants au forum sont plus vieux (l'âge moyen et l'âge médian sont de l'ordre de 40 ans contre 35 ans pour l'ensemble des étudiants) et les étudiants diplômés d'une licence ou plus sont plus présents sur le forum (en particulier, 19 % des participants du forum sont des docteurs alors que ceux-ci ne représentent que 14 % des inscrits environ). Ces résultats sont en partie similaires à ceux trouvés par Gillani et Eynon (2014) qui montrent que les participants des forums sont plus diplômés. Toutefois, dans cette même étude, ceux-ci apparaissent comme plus jeunes que l'ensemble des étudiants, ce qui est une différence notable par rapport au MOOC « Fondamentaux en statistique ». Une hypothèse permettant d'expliquer ce phénomène pourrait être que, l'équipe éducative ayant laissé l'initiative des discussions, des étudiants plus expérimentés ont pris une place plus importante, jouant le rôle d'animateur au sein du forum.

Le nombre relativement élevé d'étudiants actifs sur les forums en comparaison des étudiants ayant terminé le cours est en partie expliqué par la figure 2 qui montre l'évolution du

nombre de commentaires effectués au cours du temps. Y sont clairement identifiés plusieurs phénomènes : le premier visible est une tendance globale à la diminution de l'activité sur le forum, déjà mis en valeur dans Brinton *et al.* (2013) et facilement imputable aux abandons et décrochages des étudiants, au fur et à mesure que le cours devient plus difficile.

Note à la revue : insérer ici la figure 2

Également, il est clairement visible que l'apparition de nouvelles vidéos sur la plateforme est systématiquement suivie d'un rebond d'activité sur le forum : les cinq séries de vidéos sont identifiables par une augmentation du nombre de fils et de commentaires sur le forum. On remarque donc que, même si une partie non négligeable des commentaires de la première semaine concerne des aspects purement techniques d'utilisation de la plateforme, le nombre de commentaires est moins de quatre fois moindre après la cinquième semaine de cours (fin février) qu'après la première. En particulier, un décrochage net s'opère après la deuxième série de vidéos (qui aborde la statistique bivariée et l'ACP) et, dans une moindre mesure, après la quatrième. Enfin, l'évolution du nombre de commentaires de l'équipe enseignante suit assez fidèlement celle des commentaires des étudiants. Ces remarques montrent que l'analyse de l'activité du forum, effectuée en temps réel, est un bon indicateur de l'activité des étudiants en général, et ce, même si une faible proportion de ceux-ci y participent. En particulier, ils permettent à l'enseignant d'identifier facilement quelles parties de son cours ont été les moins bien reçues et pourraient donc être utilisées pour proposer des compléments ou clarifications permettant aux étudiants de franchir le cap de cette difficulté.

Analyse du graphe de discussion

La question de savoir si les forums structurent la communauté des étudiants n'est pas claire : Anderson *et al.* (2014) montrent que les discussions sur les forums de cours se déroulent d'une manière plutôt linéaire : chaque fil se développe avec l'arrivée de nouveaux contributeurs plutôt que par le dialogue entre les mêmes contributeurs. Pour McGuire (2013), le forum n'offrant aucun outil de classement du fourmillement des fils, il étouffe plus la formation de communautés d'étudiants qu'il ne la favorise. Gillani et Eynon (2014) utilisent une technique multidimensionnelle (Bayesian non-negative matrix factorization) pour extraire des communautés d'étudiants à partir des messages du forum. Le but est de proposer une typologie de l'engagement des étudiants dans les MOOC : en particulier, les auteurs notent une structuration géographique dans la communauté des étudiants. Notons enfin que le projet final était discuté dans le forum et donc la notion d'engagement dans le forum plus fortement reliée à la validation du MOOC que dans notre cas, ce qui peut avoir biaisé un peu l'analyse. L'idée de l'étude que nous proposons ici est assez proche en cherchant à interpréter la pertinence des communautés à partir des caractéristiques des étudiants (telles que déclarées sur les questionnaires).

Pour ce faire, nous avons construit le graphe (réseau) suivant :

- les sommets du graphe sont les étudiants ayant participé au forum et qui se sont correctement identifiés (c'est-à-dire qu'à la fois l'équipe enseignante et les étudiants non identifiés ont été exclus de l'analyse);
- les arêtes du graphe correspondent à l'interaction simultanément de deux étudiants sur un même fil de discussion. Les arêtes sont pondérées par le nombre de fils communs entre les deux étudiants de manière à ce que la fréquence de

l'interaction entre les deux personnes, et non seulement son existence, soit prise en compte dans l'analyse.

Ce type de graphe est une manière classique de représenter les interactions entre individus et correspond à la projection sur les étudiants du graphe biparti étudiants/fils qui est le graphe dans lequel les sommets représentent les étudiants ou les fils de discussion et les arêtes relient les étudiants aux fils dans lesquels ils ont été actifs (pour des précisions et des exemples sur ce type de graphes, voir Barber, 2007; Rossi, Villa-Vialaneix et Hautefeuille, 2013). Il fournit une vision objective des interactions directes dans la communauté des étudiants au travers du forum. Nous avons choisi d'analyser celui-ci à partir de méthodes statistiques de fouille de réseaux afin de tester autant que possible la significativité des phénomènes que nous mettons en valeur. Nous renvoyons le lecteur à la littérature citée sur le sujet pour de plus amples détails techniques.

Le graphe projeté contient donc 524 sommets et 2 165 arêtes. Il est composé de 104 composantes connexes donc une composante géante qui contient 401 étudiants (plus des 3/4 des étudiants ayant participé aux forums). Les autres étudiants sont majoritairement isolés (ce sont majoritairement ceux qui n'ont posé qu'une seule question sur le forum, question à laquelle un autre étudiant ou un enseignant a répondu). La plus grande composante connexe du graphe des étudiants, c'est-à-dire la composante maximale telle que toutes les paires de sommets soient connectées par un chemin le long des arêtes, est donnée dans la figure 3.

Note à la revue : insérer ici la figure 3

Cette composante a une densité (ratio du nombre d'arêtes sur le nombre de paires de sommets) de 2,7 % environ, ce qui est classique pour un réseau social (le nombre d'arêtes étant généralement faible devant le nombre de paires de sommets : voir Dorogovtsev et Mendes, 2003,

pour des exemples réels). La transitivité, qui est la fréquence, parmi les paires de sommets ayant une connexion commune, de ceux qui sont eux-mêmes connectés (plus précisément, c'est le nombre de triangles divisé par le nombre de triplets de sommets contenant au moins deux arêtes) est égale à environ 38,4 % : cela signifie que deux sommets ont approximativement 10 fois plus de chances d'être connectés s'ils ont un voisin en commun et indique un phénomène d'attachement préférentiel dans le réseau (Barabasi et Albert, 1999).

Le degré d'un sommet est le nombre d'arêtes afférentes à ce sommet. Sur le forum, la distribution des degrés est très asymétrique (voir la figure 4) : un faible nombre d'étudiants a un degré très important comparativement à la moyenne des degrés. Les trois sommets de plus fort degré (les seuls à avoir un degré supérieur à 60) sont trois actifs de sexe masculin, nés en 1950, 1955 et 1979 respectivement. Ces trois étudiants sont aussi ceux qui ont la plus grande centralité², les seuls à avoir un indice de centralité supérieur à 6 000. Notons que ceux-ci font partie des étudiants plutôt âgés, ce qui corrobore les conclusions déjà données sur la participation plus forte des plus expérimentés au forum, mais aussi sur leur rôle central en matière d'animation.

Note à la revue : insérer ici la figure 4

Comme dans Anderson *et al.* (2014), nous remarquons que ces étudiants centraux et très actifs ont moins tendance que les autres à ouvrir des fils de discussion : s'ils fournissent à eux seuls 36,6 % des réponses aux fils de discussion (hors équipe pédagogique), ils n'ouvrent que 23,7 % des discussions (un test du chi-deux montre que cette tendance est significative au seuil de 1 %).

Enfin, des tests de permutations ont été pratiqués pour vérifier si les interactions avaient lieu préférentiellement entre personnes de même âge, de même sexe ou de même niveau d'études

(la méthodologie décrivant ce type de tests est expliquée dans Laurent et Villa-Vialaneix, 2011). Seul l'indice de Moran reliant l'âge aux arêtes du graphe a été trouvé significativement élevé, ce qui indique une tendance significative des étudiants à interagir avec d'autres étudiants de la même classe d'âge. Les étudiants n'ayant, a priori, aucune indication de l'âge des autres étudiants, il est intéressant de se demander pourquoi deux étudiants de même âge ont une tendance plus forte à communiquer via le forum que deux étudiants d'âges distincts : les hypothèses pourraient être liées à des différences de vocabulaire ou des différences de maturité de la réflexion. On peut imaginer que le rôle de l'équipe enseignante pourrait être, en particulier, de veiller à l'établissement d'un socle commun de communication pour améliorer la discussion entre classes d'âge ou bien de tirer parti de la structuration naturelle des étudiants en classes d'âge pour avoir un discours personnalisé et adapté pour celles-ci.

Les communautés

La notion de communautés est centrale en analyse de graphes, notamment de réseaux sociaux. En effet, il est généralement admis de manière assez naturelle (Freeman, 2004) que les groupes humains sont structurés en sous-groupes sociaux cohésifs. Du point de vue de l'objet mathématique « graphe », la définition de ces communautés n'est pas complètement uniforme et peut varier selon le domaine d'application. Cependant, de manière assez consensuelle, la notion de communautés fait référence à une partition des sommets du graphe, telle que les groupes de sommets sont denses (c'est-à-dire avec un grand nombre d'arêtes à l'intérieur du groupe) et connectés entre eux par un nombre faible (comparativement) d'arêtes.

Nous avons recherché des communautés dans le graphe des étudiants en maximisant la modularité. C'est une des approches les plus classiques (Newman et Girvan, 2004) et nous avons utilisé la méthode d'optimisation approchée décrite dans Rossi et Villa-Vialaneix (2011)³. Nous

avons, de plus, testé la significativité de cette partition en la comparant à la distribution de la modularité maximale d'un ensemble de graphes aléatoires de même distribution de degrés que notre graphe d'étudiants. La partition optimale que nous avons obtenue contenait 11 classes dont les effectifs (nombre de sommets) sont donnés dans le tableau 1 et qui avaient une modularité égale à 0,507, bien supérieure à la modularité maximale trouvée pour 100 graphes aléatoires (0,364) : la structure en communautés du graphe est donc assez marquée, contrairement à ce que conclut McGuire (2013).

Note à la revue : insérer ici le tableau 1

Il semble donc y avoir, comme le soulignent également Gillani et Eynon (2014), une organisation des étudiants en groupes de personnes qui prennent l'habitude de discuter plus fréquemment entre eux. Gillani et Eynon (2014) montrent qu'il existe une structuration géographique au sein de ces groupes. De manière similaire, nous avons étudié les classes pour trouver des différences de profils entre étudiants du point de vue de leur âge, leur sexe, leur localisation géographique (France ou étranger) et leur niveau d'études. Seul l'âge est significativement différent selon la communauté, confirmant les résultats des tests effectués sur la corrélation entre structure du graphe et description des individus décrits plus haut. La distribution de l'âge par communauté est donnée dans la figure 5. On y constate, en particulier, que la plus petite communauté (la 6, contenant 15 étudiants) est celle dont l'âge est le plus faible (l'âge médian y est environ égal à 30 ans) alors que la communauté 2 (35 étudiants) a un âge médian plus élevé, proche de 50 ans, malgré trois valeurs atypiques dont l'âge est inférieur à 20 ans. Il semble à nouveau que la stratégie visant, pour l'équipe pédagogique, à permettre une meilleure communication entre les diverses classes d'âges (en veillant, par exemple, à limiter les effets de différence d'expression ou de référentiel culturel) pourrait être une clé pour alimenter la

vie du forum. Une autre solution pourrait être, au contraire, de repérer et comprendre les besoins spécifiques de ces différentes communautés pour produire un discours personnalisé en direction de chacune d'elle.

Note à la revue : insérer ici la figure 5

Conclusion

Cet article a présenté une étude quantitative des participants du MOOC « Fondamentaux en statistique » et en particulier des participants du forum. L'analyse a montré que les étudiants inscrits au cours ont un profil et des comportements qui ressemblent à ceux relevés dans les études précédentes effectuées sur les cours des grandes plateformes américaines. En particulier, leur profil est assez différent de l'étudiant d'université auquel ce cours pourrait s'adresser : ils sont plus âgés, déjà diplômés et un nombre non négligeable d'entre eux suivaient le cours depuis des pays étrangers francophones. L'activité déployée autour du cours a été essentiellement produite par un « noyau dur » d'étudiants qui eux-mêmes avaient une activité très contrastée (certains extrêmement actifs et d'autres actifs). Les étudiants les plus actifs se sont avérés être aussi les plus âgés, contrairement à ce que d'autres travaux avaient montré, et semblent jouer un rôle proche de celui de l'équipe pédagogique en fournissant une part importante des réponses plutôt qu'en ouvrant eux-mêmes des fils.

En revanche, nous avons montré que, contrairement à des études antérieures, les participants se structurent en groupes d'échanges indépendants de la plupart des caractéristiques de leurs profils, hormis l'âge. C'est un point qui devrait être pris en compte par l'équipe enseignante pour gérer le forum et favoriser l'engagement des étudiants, soit en jouant un rôle de liant, soit en produisant une réponse personnalisée en direction de chaque groupe.

L'analyse de l'activité fait également ressortir un phénomène de décrochage progressif déjà observé par ailleurs. Celui-ci est un bon vecteur pour évaluer quelles parties du cours ont été jugées trop difficiles. Il aurait été intéressant de corréler ces mesures d'activité à la participation aux exercices et à la réussite au devoir, mais les données disponibles ne permettaient pas de faire un lien clair entre étudiants inscrits sur FUN et les exercices déposés sur une plateforme externe, comme le soulignent aussi Anderson *et al.* (2014) dans leur étude.

Références

- Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada-Gélvez, H. A., Muñoz-Organero, M. et Rodríguez-de-las-Heras, A. (2013). Analysing the impact of built-in and external social tools in a MOOC on educational technologies. Dans D. Hernández-Leo, T. Ley, R. Klamma et A. Harrer (dir.), *Scaling up Learning for Sustained Impact : Proceedings of the 8th European Conference on Technology Enhanced Learning* (p. 5-18). Berlin, Allemagne : Springer.
- Anderson, A., Huttenlocher, D., Kleinberg, J. et Leskovec, J. (2014). Engaging with massive online courses. Dans C.-W. Chung (dir.), *Proceedings of the 23rd international conference on World Wide Web* (p. 687-698). doi:10.1145/2566486.2568042
- Anderson, T. et Kanuka, H. (1997). On-line forums: new platforms for professional development and group collaboration. *Journal of Computer-Mediated Communication*, 3(3), 0. [doi:10.1111/j.1083-6101.1997.tb00078.x](https://doi.org/10.1111/j.1083-6101.1997.tb00078.x)
- Barabasi, A. et Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286 (5439), 509-512. [doi:10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509)
- Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6). doi:10.1103/PhysRevE.76.066102 [Récupéré](#) de Cornell University Library : <http://arxiv.org>
- Baxter, J. A. et Haycock, J. (2014). Roles and student identities in online large course forums: implications for practice. *The International Review of Research in Open and Distance Learning*, 15(1), 20-40. [Récupéré](#) de <http://www.irrodl.org>

- Berge, Z. L. (1995). Facilitating computer conferencing: recommendations from the field. *Educational Technology*, 35(1), 22-30.
- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D. et Seaton, D. T. (2013). Studying learning in the worldwide classroom: research into edX's first MOOC. *Research & Practice in Assessment*, 8(1), 13-25. [Récupéré](#) du site de MOOCs @ Peking University : <http://mooc.pku.edu.cn>
- Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z. et Wong, F. M. F. (2013). Learning about social learning in MOOCs: from statistical analysis to generative model. *IEEE Transactions on Learning Technologies*, 7(4), 346-359. doi:10.1109/TLT.2014.2337900 [Récupéré](#) de Cornell University Library : <http://arxiv.org>
- Cisel, M. (2014). Analyzing completion rates of first French xMOOC. Dans U. Cress et C. Delgado Kloos (dir.), *Proceedings of the European MOOC Stakeholder Summit 2014* (p. 26-32). Barcelone, Espagne : P. A. U. Education.
- Dorogovtsev, S. N. et Mendes, J. F. F. (2003). *Evolution of networks. From biological nets to the Internet and WWW*. Oxford, R.-U. : Oxford University Press.
- Freeman, L. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, Canada : Booksurge.
- Fruchterman, T. M. J. et Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software, Practice and Experience*, 21(11), 1129-1164. doi:10.1002/spe.4380211102 [Récupéré](#) du site personnel d'Edward M. Reingold : <http://emr.cs.iit.edu/~reingold/>

- Gillani, N. et Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education*, 23, 18-26. doi:10.1016/j.iheduc.2014.05.004 [Récupéré](#) du site ResearchGate : <http://www.researchgate.net>
- Harasim, L., Hiltz, S.R., Teles, L. et Turoff, L. (1995). *Learning networks*. Cambridge, MA: The MIT Press.
- Laurent, T. et Villa-Vialaneix, N. (2011). Using spatial indexes for labeled network analysis. *Information, Interaction, Intelligence (I3)*, 11(1). [Récupéré](#) de l'archive ouverte HAL : <http://hal.archives-ouvertes.fr>
- Mak, S. F. J., Williams, R. et Mackness, J. (2010). Blogs and forums as communication and learning tools in a MOOC. Dans L. Dirckinck-Holmfeld, V. Hodgson, C. Jones, M. De Laat, D. McConnell et T. Ryberg (dir.), *Proceedings of Networked Learning Conference* (p. 275-284). [Récupéré](#) du dépôt Parade@Portsmouth : <http://eprints.port.ac.uk/>
- Manning, J. et Sanders, M. (2013). *How widely used are MOOC forums? A first look* [billet de blogue]. [Récupéré](#) le 22 avril 2015 du blogue Signals – Thoughts on Online Learning : <http://signalblog.stanford.edu>
- McGuire, R. (2013). Building a sense of community in MOOCs. *Campus Technology*, 26(12), 31-33. [Récupéré](#) du site de la revue : <http://online.qmags.com/CPT0813>
- Newman, M. E. J. et Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2). doi:10.1103/PhysRevE.69.026113 [Récupéré](#) du site de Computational Epidemiology Research : <http://compepi.cs.uiowa.edu>

- Onah, D. F. O., Sinclair, J. E. et Boyatt, R. (2014). Exploring the use of MOOC discussion forums. Dans *Proceedings of London International Conference on Education* (p. 1-4).
[Récupéré de Warwick Research Archive Portal : http://wrap.warwick.ac.uk](http://wrap.warwick.ac.uk)
- Rolfe, V. (2015). A systematic review of the socio-ethical aspects of massive online open courses. *European Journal of Open, Distance and E-Learning*, 18(1), 53-72. [Récupéré de http://www.eurodl.org](http://www.eurodl.org)
- Rossi, F. et Villa-Vialaneix, N. (2011). Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets. *Journal de la société française de statistique*, 152(3), 34-65. [Récupéré de http://journal-sfds.fr](http://journal-sfds.fr)
- Rossi, F., Villa-Vialaneix, N. et Hautefeuille, F. (2013). Exploration of a large database of French notarial acts with social network methods. *Digital Medievalist*, 9. [Récupéré de l'archive ouverte HAL : http://hal.archives-ouvertes.fr](http://hal.archives-ouvertes.fr)
- Salmon, G. (2004). *e-moderating. The key to teaching and learning online* (2^e éd.). Londres, R.-U. : RoutledgeFalmer.

Notes

¹Les étudiants qui souhaitaient réaliser les exercices devaient s'inscrire, parallèlement à FUN, sur la plateforme Moodle de l'Université Sorbonne Paris Cité : certains ne l'ont pas fait, certains ont utilisé un e-mail d'identification différent... Tout ceci explique qu'il est difficile de croiser l'activité des étudiants sur FUN, ainsi que leurs profils socioprofessionnels, avec la participation aux exercices.

²La centralité d'un sommet est le nombre de plus courts chemins entre deux sommets du graphe passant par ce sommet. Les sommets à forte centralité sont des vecteurs de communication importants dans le graphe et sont susceptibles de déconnecter le graphe s'ils sont supprimés.

³Le programme ayant permis de générer la partition des sommets est disponible à <http://apiacoa.org/research/software/graph/index.en.html>.

Tableau 1

Nombre de sommets dans chacune des classes (communautés) trouvés par optimisation de la modularité.

Communauté	0	1	2	3	4	5	6	7	8	9	10
Nombre de sommets	39	46	35	21	30	52	15	49	44	38	32

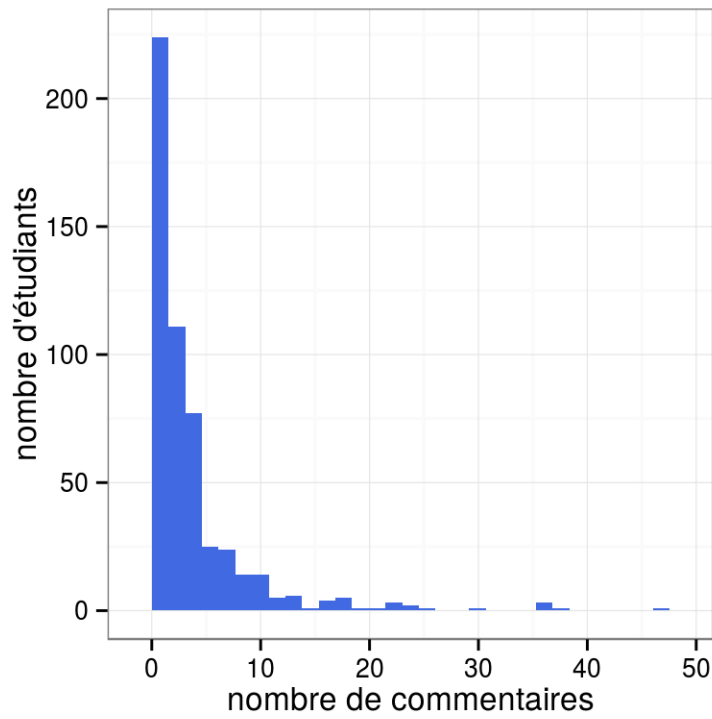


Figure 1. Distribution du nombre de commentaires postés par étudiant (l'équipe enseignante exclue).

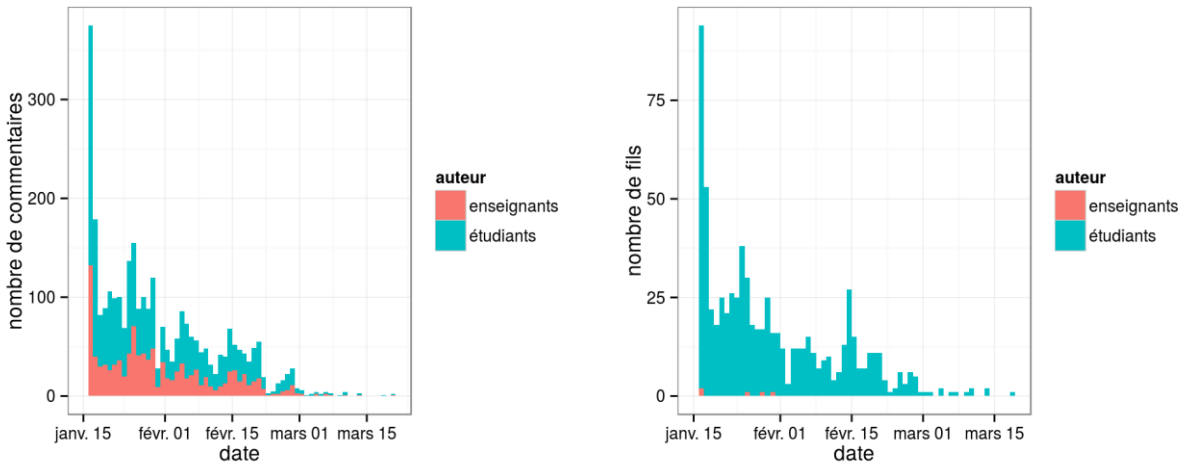


Figure 2. Évolution du nombre de commentaires (à gauche) et de fils de discussion (à droite) au fur et à mesure de l'avancement du MOOC.

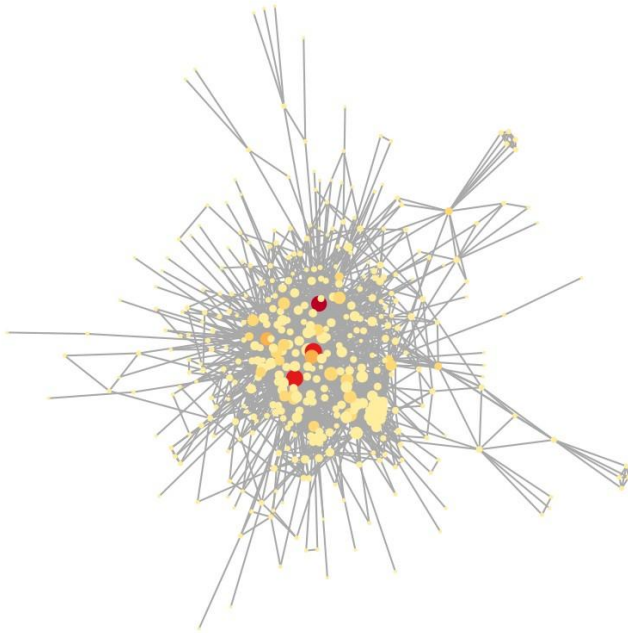


Figure 3. Représentation par un algorithme de forces (Fruchterman et Reingold, 1991) de la plus grande composante connexe du graphe d'interaction entre étudiants. Les couleurs indiquent la centralité des sommets (les sommets les plus centraux sont de couleur foncée) et la taille le degré (les sommets de plus fort degré sont de taille plus grande).

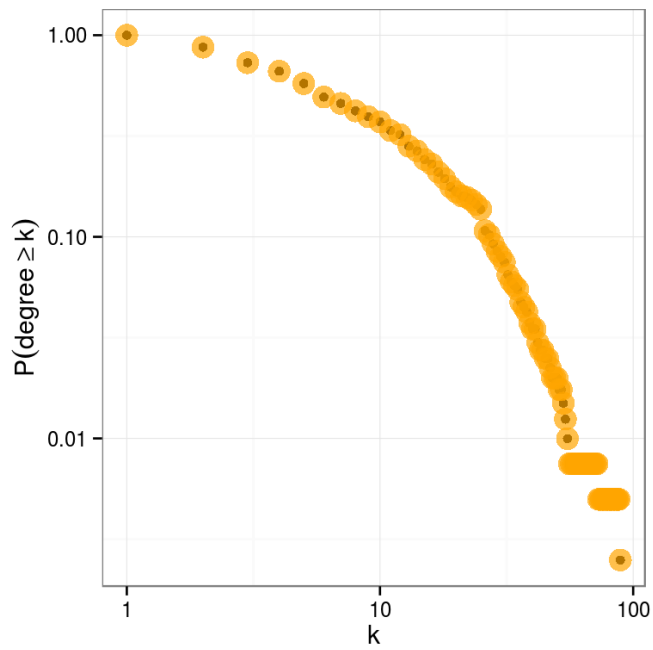


Figure 4. Distribution des degrés du graphe : fréquence de sommets de degrés supérieurs ou égale à k en fonction de k . Les échelles des deux axes sont logarithmiques.

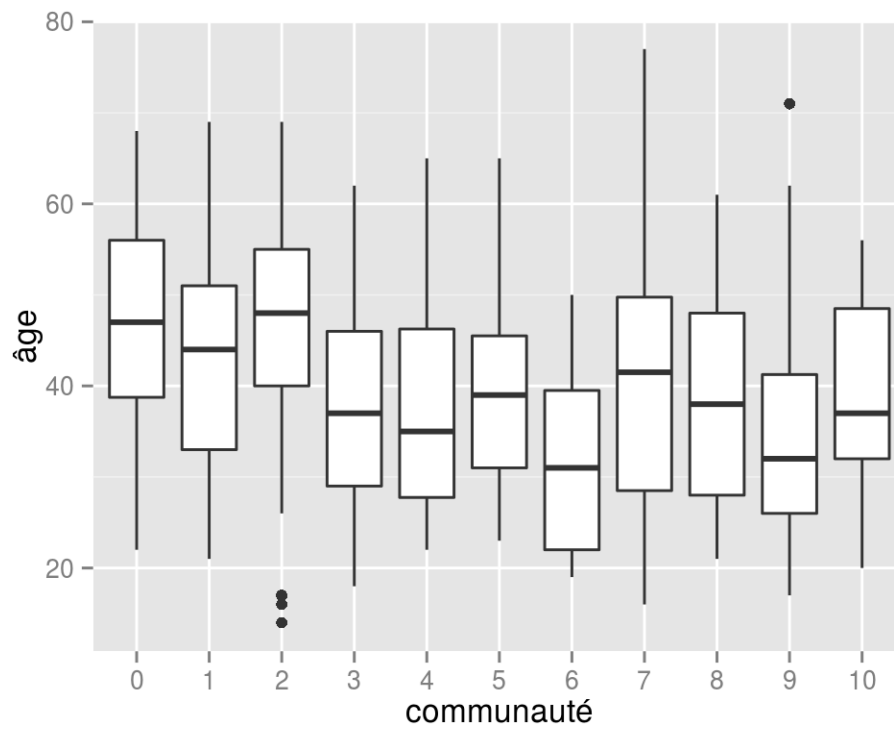


Figure 5. Distribution de l'âge selon la communauté.