

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

Avner Bar-Hen¹, Hubert Javaux², Nathalie Villa-Vialaneix³

¹ Université Paris Descartes, MAP5 UFR Math-Info, 45, rue des Saints-Pères, 75270 Paris, France

² Université Sorbonne Paris Cité, Service SAPIENS, 10 rue Charles V, 75004 Paris

³ INRA, UR 875 MIA-T, Castanet-Tolosan cedex - France

Mooc "Fondamentaux en statistique"

Entre le 16 janvier 2014 et le 16 mars 2014

Programme général :

- ▶ 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.

Programme de la semaine :

- ▶ Vidéos ;
- ▶ quizz ;
- ▶ exercice (sur Moodle Sorbonne Paris Cité).

+ Forum + Wiki + Synthèse de la semaine.

Contexte

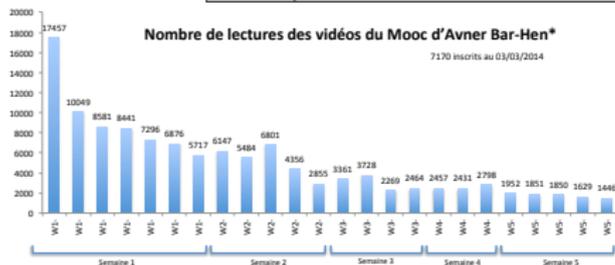
- ▶ 7997 inscrits sur FUN dont "seulement" 4600 le 16 janvier ;
- ▶ 85% des inscrits du cours (6 918 étudiants dont 32% de femmes) ont complétés le questionnaire ;
- ▶ 1832 inscrits sur la plate-forme Moodle (22.91% des inscrits) dont 74.34% en France ;
- ▶ 251 attestations finales.

Qui sont les étudiants ?

- ▶ Localisation (API Google Map Open Street Map) : 3 730 étudiants (54% des questionnaires) dont 2 826 (76%) provenaient de France ;
- ▶ âge ;
- ▶ niveau d'études ;
- ▶ motivation

Vidéos

	W1	W2	W3	W4	W5
S1	18445	6794	3763	2908	2457
S2	10608	5952	4149	2885	2337
S3	9000	7387	2715	3272	2305
S4	8969	4747	2836		2036
S5	7708	3255			1845
S6	7312				
S7	6096				

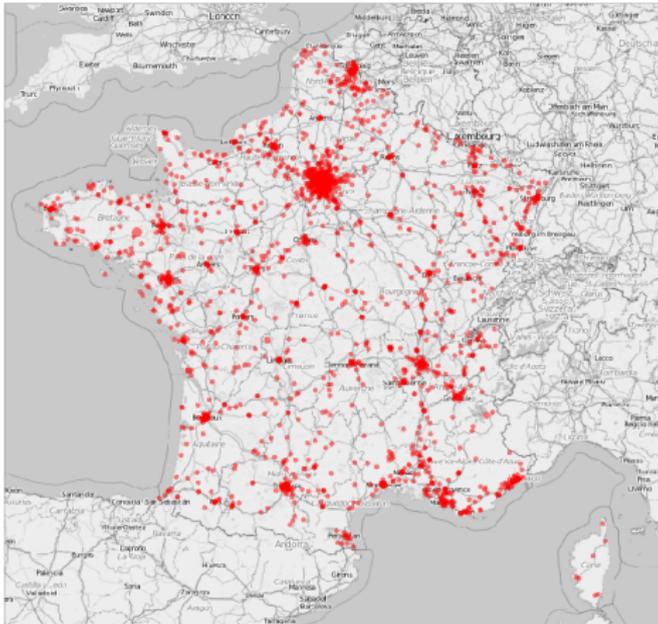


N° de semaine (week)
 ↓
 W7-47
 ↑
 s : séquence vidéo



* Source : nombre de lectures indiqués pour chaque vidéo sur le compte DMCloud.
 Pour qu'une lecture soit comptée, il suffit de la visionner même une seconde (pas d'obligation de la voir en entier)

Localisation : France



ville	Freq
Bordeaux	26
Lyon	49
Marseille	46
Montpellier	33
Nantes	23
Nice	20
Paris	388
Rennes	20
Strasbourg	31
Toulouse	41
Versailles	21

Localisation : Monde



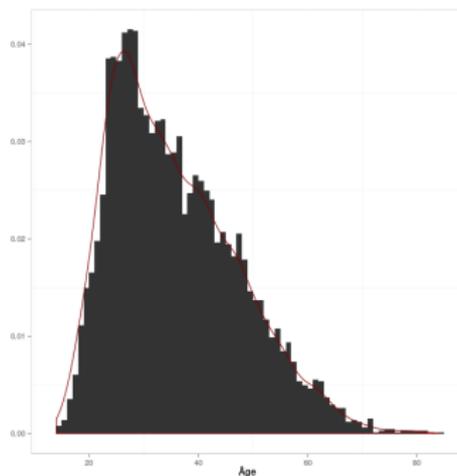
Pays	Freq
Algérie	60
Belgique	45
Bénin	20
Burkina Faso	20
Cameroun	36
Canada	38
Côte d'Ivoire	28
États-Unis	78
Mali	21
Maroc	97
Sénégal	21
Suisse	29
Tunisie	33

Diplômes

Aucun	27
Brevet	65
Baccalauréat	497
DUT/BTS	546
Licence professionnelle	178
Licence	967
Master	3208
Doctorat	937
Autre	188
Total	6613

Âge

6 479 réponses entre 13 et 100 ans

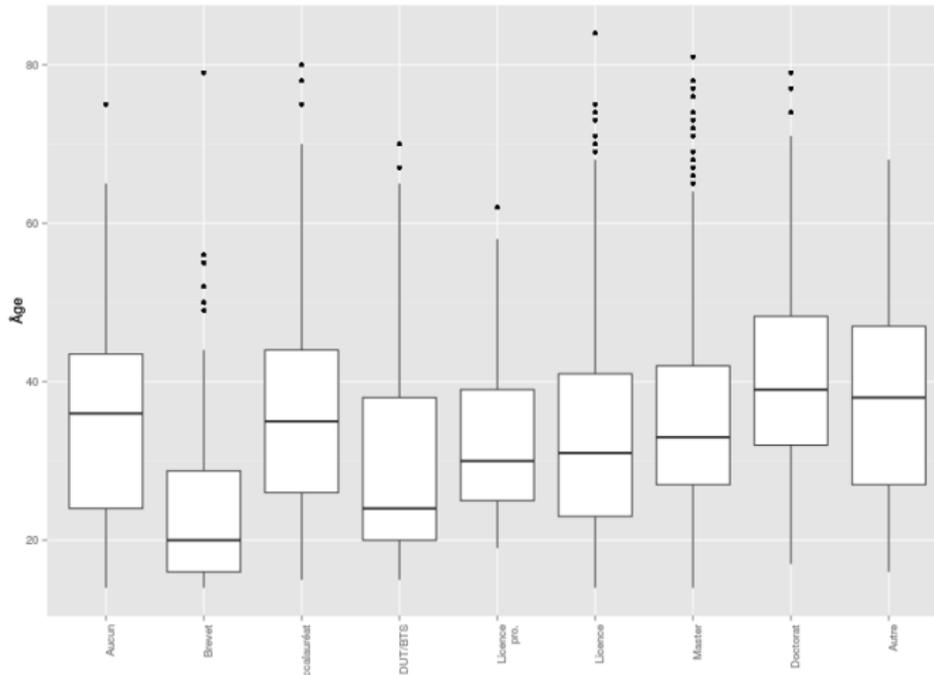


- ▶ 1/4 des étudiants ont moins de 26 ans ;
- ▶ la moitié des étudiants a moins de 34 ans ;
- ▶ l'âge moyen est d'environ 36 ans ;
- ▶ fort étalement à droite de la distribution avec un âge maximal de 84 ans.

⇒ Minorité d'étudiants

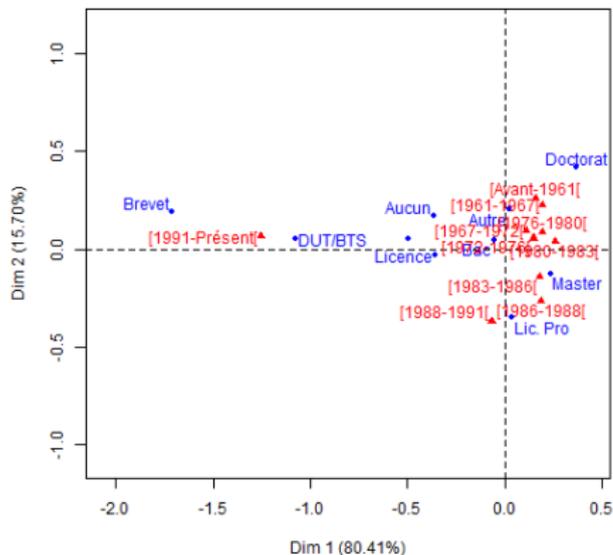
Lien entre variables

- ▶ Pas vraiment de lien entre le sexe et diplôme ;
- ▶ pas vraiment de lien entre le sexe et l'âge ;
- ▶ âge médian généralement croissant selon le niveau d'études.

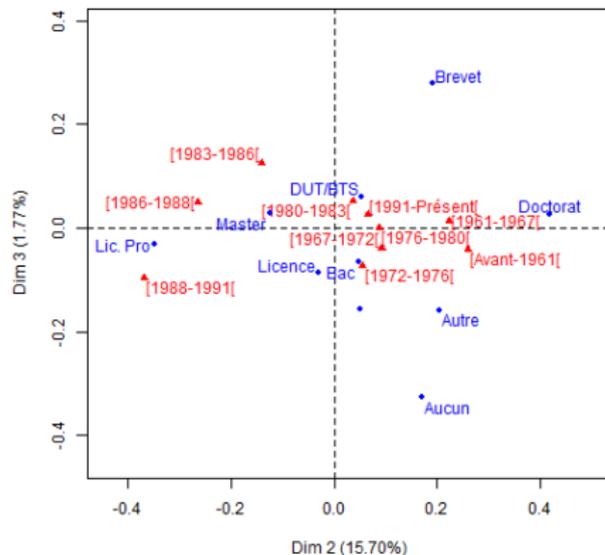


Projection : AFC

CA factor map



CA factor map



Projection : AFC

4 groupes :

- ▶ Après 1991 peu diplômés ;
- ▶ avant 1967 (avec une sur-représentation d'étudiants en doctorat) ;
- ▶ entre 1968 et 1983 (les Bac et DUT/BTS) ;
- ▶ entre 1984 et 1991 (avec une sur-représentation d'étudiants en licence pro ou master/ingénieur).

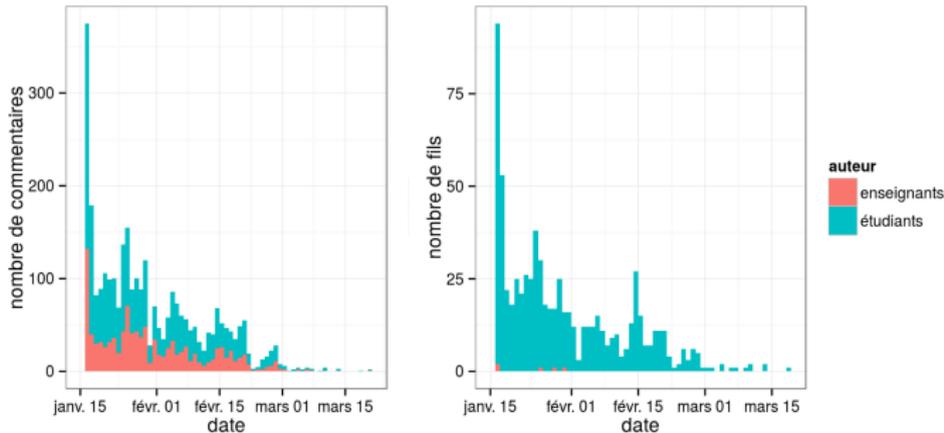
Forum

- ▶ 716 **files** de discussions (dont 5 de l'équipe pédagogique) ;
- ▶ 3024 **posts** dont 1 080 par l'équipe pédagogique ($\approx 36\%$ des commentaires) ;
- ▶ 502 **étudiants** différents (+ 3 membres équipes pédagogiques +23 commentaires anonymes) : 6.6% d'actifs.

- ▶ Les participants au forum sont **plus vieux** (l'âge moyen de 40 ans contre 35 ans pour l'ensemble des étudiants).

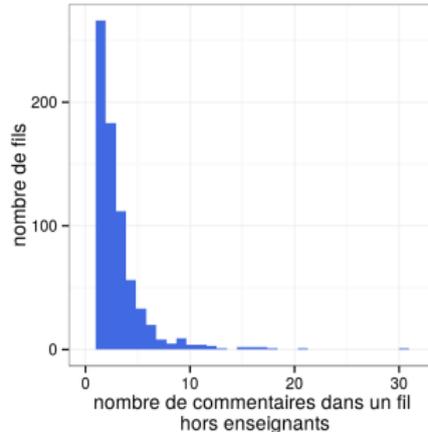
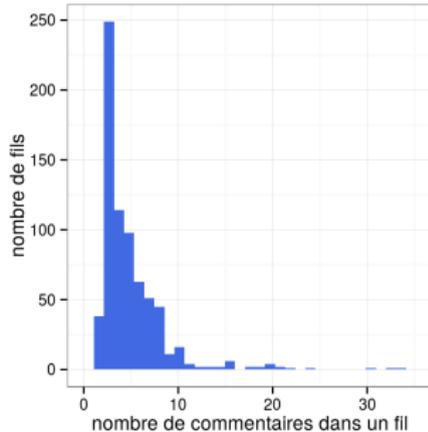
- ▶ Les étudiants **diplômés** d'une licence ou plus sont plus présents sur le forum (19% des participants du forum sont des docteurs versus $\approx 14\%$ des inscrits

Participation au forum



- ▶ Tendance globale à la **diminution de l'activité** : décrochage et abandon ;
- ▶ 4 fois moins de commentaires après 5ème semaine qu'après 1ère ;
- ▶ les nouvelles vidéos sont suivies d'un rebond d'activité ;
- ▶ un décrochage net s'opère après la deuxième série de vidéos ;
- ▶ l'évolution du nombre de commentaires de **l'équipe enseignante** suit assez fidèlement celle des commentaires des **étudiants**.

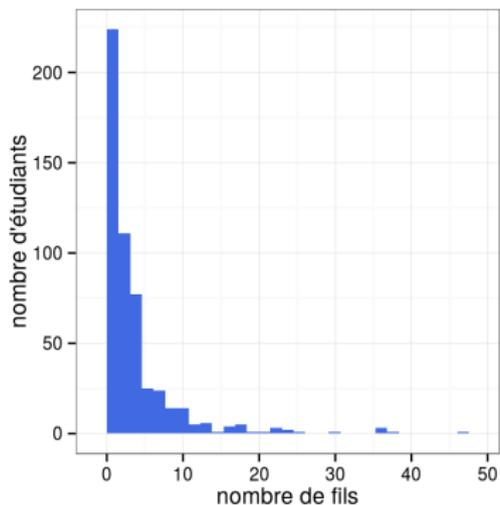
Longueur des fils



- ▶ Le nombre de commentaires par fil est plutôt faible en général ;
- ▶ le fil le plus long : 30 commentaires ;
- ▶ distribution très asymétrique.

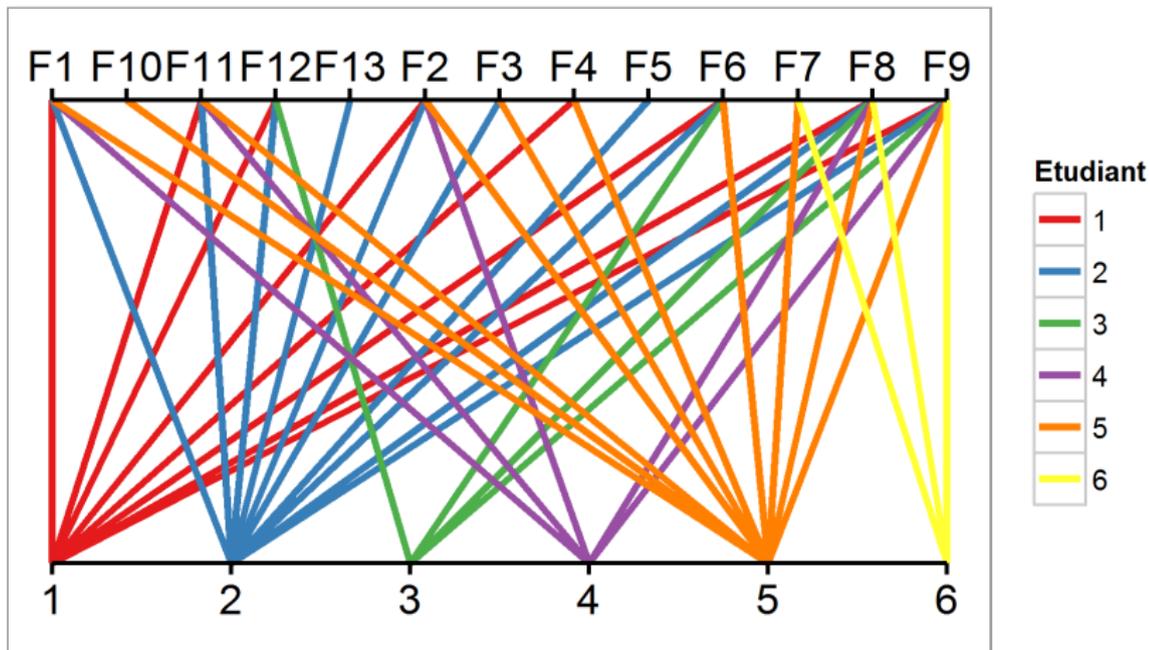
La comparaison des deux distributions montre que beaucoup de fils à deux commentaires semblent correspondre à une réponse satisfaisante d'un étudiant, sans intervention de l'équipe enseignante.

Activité des étudiants

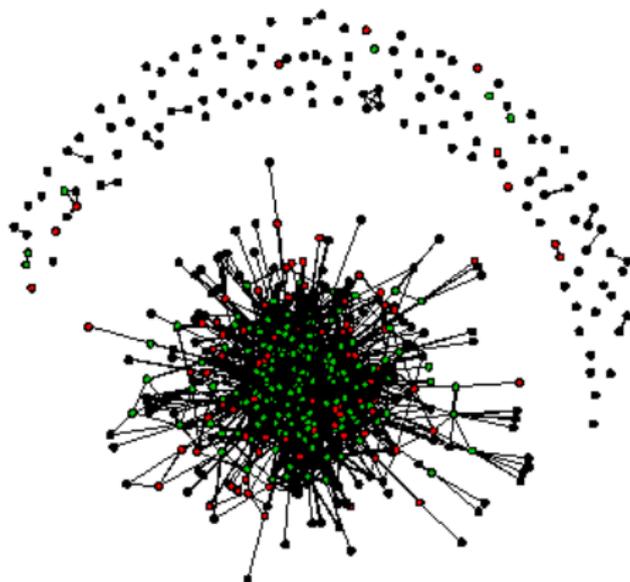


- ▶ La plupart des étudiants n'interviennent que très peu : 224 étudiants (soit plus de 40%) n'ont posté qu'un commentaire et 111 étudiants n'en ont posté que deux.
- ▶ Certains étudiants interviennent très fréquemment sur le forum (avec un maximum de 48 commentaires).

Activité des étudiants : première représentation

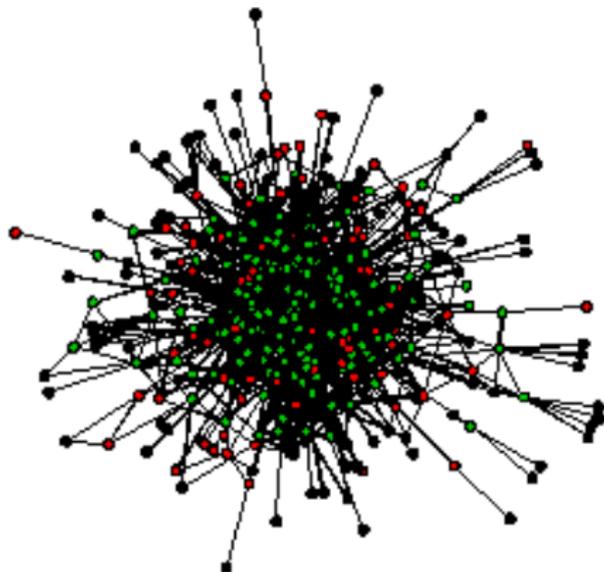


Activité des étudiants : graphe



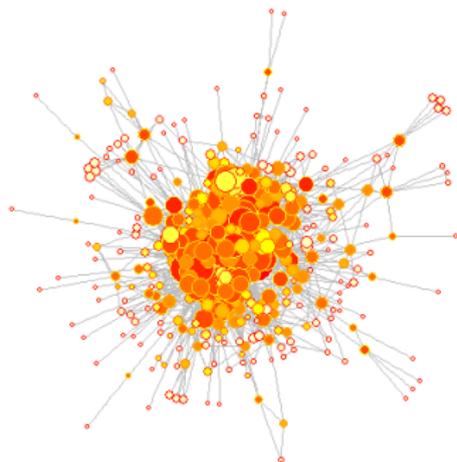
- ▶ 524 sommets et 2 165 arêtes ;
- ▶ 104 composantes connexes ;
- ▶ une composante géante qui contient 401 étudiants.

Principale composante connexe



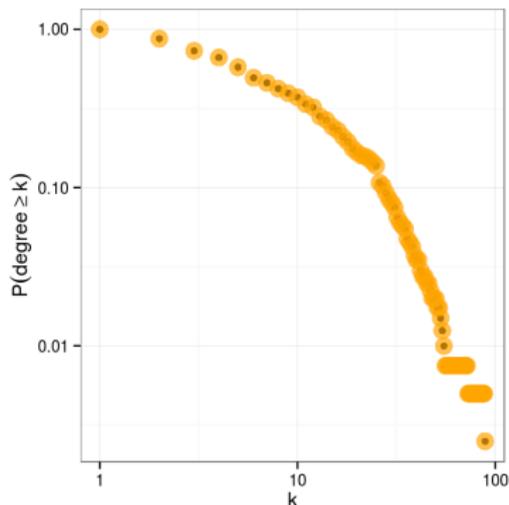
- ▶ Une composante avec 401 étudiants ;
- ▶ densité : 2.7% ;
- ▶ transitivité : 38.4% : attachement préférentiel.

Principale composante connexe



- ▶ Les sommets les plus centraux sont de couleur foncée ;
- ▶ les sommets de plus fort degré sont de taille plus grande ;
- ▶ l'indice de Moran reliant l'âge aux arêtes du graphe a été trouvé significativement élevé (test de permutation).

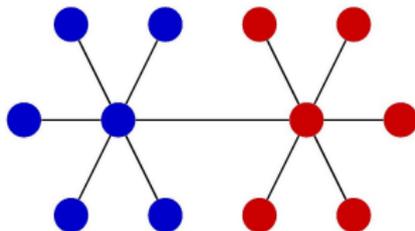
Distribution des degrés



- ▶ Pas une loi puissance mais distribution très asymétrique ;
- ▶ les trois sommets de degré supérieur à 60 sont trois actifs de sexe masculin, nés en 1950, 1955 et 1979 ;
- ▶ les seuls à avoir un indice de centralité supérieur à 6 000.

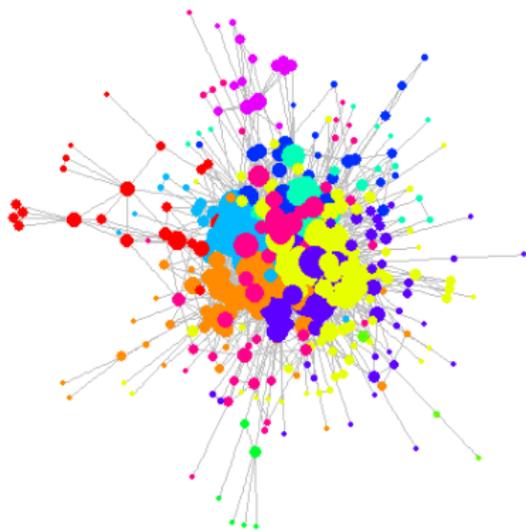
Communautés : sous-groupes sociaux cohésifs

Communautés : partition des sommets du graphe, telle que les groupes de sommets sont denses (ie avec un grand nombre d'arêtes à l'intérieur du groupe) et connectés entre eux par un nombre faible (comparativement) d'arêtes.



- ▶ La **modularité** : compare le nombre d'arcs dans les sous-graphes avec le nombre de liens d'un graphe aléatoire de même taille, avec même répartition des degrés ;
- ▶ la modularité prend ses valeurs entre -1 et 1 inclus (différence de proportion) ;
- ▶ bon partitionnement d'un graphe implique un nombre d'arêtes intra-communautaires important et un nombre d'arêtes inter-communautaires faible.

Communautés



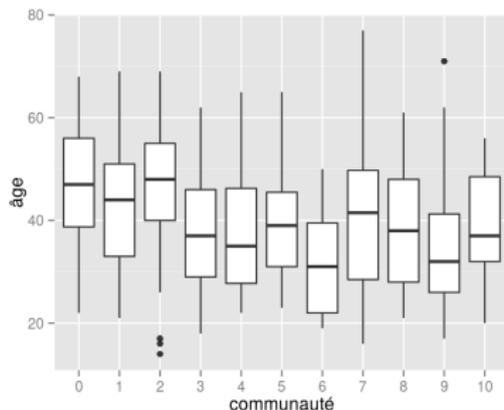
0	1	2	3	4	5	6	7	8	9	10
39	46	35	21	30	52	15	49	44	38	32

Modularité : 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.

Communautés

Pas de lien significatif entre communautés et sexe, localisation géographique (France ou étranger) ou niveau d'études

Lien significatif entre communautés et âge :



- ▶ Plus petite communauté (la 6, 15 étudiants) a un âge médian ≈ 30 ans.
- ▶ La communauté 2 (35 étudiants) a un âge médian plus élevé, ≈ 50 ans.

Analyse des messages

Les 3024 messages ont été **nettoyés** pour en retirer les signes de ponctuation, les accents, les numéros ainsi que les mots d'arrêt tels que définis dans le projet « Snowball stemmer »

8 256 mots différents ont ainsi été conservés, dont 3 792 ne sont utilisés qu'une fois dans le forum et 1 237 sont utilisés deux fois.

À l'inverse, les mots les plus utilisés sont

- ▶ « **plus** » (692 occurrences dans 492 messages) ;
- ▶ « **bien** » (786 occurrences dans 664 messages) ;
- ▶ « **bonjour** » (802 occurrences dans 800 messages) ;
- ▶ « **cours** » (811 occurrences dans 607 messages) ;
- ▶ « **merci** » (840 occurrences dans 776 messages) ;

Analyse des messages

	moyenne	médiane	mode	Max	Min	$Q_{75\%}$	Écart-type
Occurrences/mot	8.33	1	2	840	1	5	31.38
Mot/message	22.74	15	7	303	0	28	25.84
Mot/début de fil	30.19	20	9	285	1	36.25	32.52
Mot/réponse sur un fil	20.43	13.5	7	303	0	25	22.9
Mot/étudiants	93.65	32	8	2532	0	81	194.7
Mots différents/étudiants	66.09	28	8	1205	0	69	109.2
Message/étudiant	3.88	2	1	47	1	4	5.45
Mot/étudiants/message	21.88	17	6	27	0	27	19.25

Quelques résultats

- ▶ Les étudiants les plus actifs utilisent plus de mots différents, mais il n'y a pas de lien entre le nombre de messages, la longueur des messages et le nombre moyen de mots par message. En résumé [la lexicographie ne caractérise pas l'activité des étudiants sur le forum.](#)
- ▶ Les facteurs sexe, âge, et niveau d'études n'ont pas d'influence sur le nombre de messages.
- ▶ Un nombre moyen de mots plus important pour les étudiants ayant déclaré un faible niveau d'études ou ayant posté de façon anonyme.

De la lexicographie à la linguistique ?

Apprentissage supervisé (ou non) ?

Merci de votre attention (et de vos questions)