Modélisation et analyse de réseaux d'apprentissage à partir du dataset OULAD

Data Mining

Katia Lounas, Linda Chouati

Université Claude Bernard Lyon 1 novembre 2025

Table des matières

1 Introduction

Dans le cadre de ce projet de data mining, nous nous intéressons à l'exploitation de données afin d'extraire des connaissances à partir des traces d'apprentissage d'étudiants.

Pour cela, nous utilisons le dataset **OULAD (Open University Learning Analytics Dataset)** qui rassemble les traces d'interaction de plusieurs milliers d'étudiants inscrits à des cours en ligne (clickstream, évaluations, informations démographiques, métadonnées de ressources). Ce jeu de données se prête particulièrement bien à une analyse exploratoire visant à comprendre comment les comportements d'interaction influencent la réussite académique, par exemple.

Notre objectif à nous est d'appliquer des techniques de réseaux et de clustering sur ceux-ci afin d'identifier des profils d'apprenants et des structures de relations entre étudiants et ressources pédagogiques.

Cette démarche s'inscrit dans la logique du data mining éducatif : découvrir, à partir de grandes quantités de données, des patrons cachés et des comportements significatifs susceptibles d'améliorer la compréhension des processus d'apprentissage et de soutenir la prise de décision pédagogique.

2 Présentation et préparation des données

2.1 Description des données

Le Open University Learning Analytics Dataset (OULAD) [1] [2] regroupe des données anonymisées issues de The Open University (Royaume-Uni), collectées au cours des années académiques 2013–2014. Il contient des informations détaillées sur plus de 32 000 étudiants, 22 modules d'enseignement, et plus de 10 millions d'interactions au sein de l'environnement d'apprentissage virtuel (VLE).

Ce jeu de données combine des informations démographiques, académiques et comportementales, permettant d'étudier les trajectoires d'apprentissage, la réussite étudiante et la personnalisation des parcours pédagogiques. Le tableau 1 resume les attributs de chaque table du dataset. Il constitue une référence majeure dans la recherche en learning analytics et en data mining éducatif. (Voir 1 pour le schéma de la BD).

Table 1 – Description des fichiers et attributs du dataset OULAD

Dataset	Colonne	Type	Description	N manq.	%
courses	code_module	object	Code du module (identifiant)	0	0.00
	code_presentation	object	Code de la session (ex : 2013B, 2013J)	0	0.00

Suite à la page suivante

Table 1 – suite de la page précédente

Dataset	Colonne	Type	Description	N manq.	%
	module-presentation-length	int64	Durée de la session (en jours)	0	0.00
assessments	code_module	object	Code du module	0	0.00
	code_presentation	object	Code de la session	0	0.00
	id_assessment	int64	Identifiant unique de l'évaluation	0	0.00
	assessment_type	object	Type d'évaluation (TMA, CMA, examen final)	0	0.00
	date	float64	Date de soumission (jours depuis début)	11	5.34
	weight	float64	Poids de l'évaluation (%)	0	0.00
studentAssessment	id_assessment	int64	Identifiant de l'évaluation	0	0.00
	id_student	int64	Identifiant unique de l'étudiant	0	0.00
	date_submitted	int64	Date de soumission par l'étudiant	0	0.00
	is_banked	int64	Résultat transféré d'une session précédente	0	0.00
	score	float64	Note de l'étudiant (0-100)	173	0.10
studentInfo	code_module	object	Code du module	0	0.00
	code_presentation	object	Code de la session	0	0.00
	id_student	int64	Identifiant de l'étudiant	0	0.00
	gender	object	Sexe de l'étudiant	0	0.00
	region	object	Région de résidence	0	0.00
	highest_education	object	Niveau d'éducation à l'entrée	0	0.00
	imd_band	object	Indice de privation socio-économique	1111	3.41
	age_band	object	Tranche d'âge	0	0.00
	num_of_prev-attempts	int64	Nombre de tentatives antérieures	0	0.00
	studied_credits	int64	Nombre total de crédits étudiés	0	0.00
studentRegistration	code_module	object	Code du module	0	0.00
	code_presentation	object	Code de la session	0	0.00
	id_student	int64	Identifiant de l'étudiant	0	0.00
	date_registration	float64	Date d'inscription (relative au début du module)	45	0.14
	date-unregistration	float64	Date de désinscription (relative au début du module)	22521	69.10
studentVle	code_module	object	Code du module	0	0.00
	code_presentation	object	Code de la session	0	0.00
	id_student	int64	Identifiant de l'étudiant	0	0.00
	id_site	int64	Identifiant du matériel VLE	0	0.00
	date	int64	Date d'interaction (jours depuis début)	0	0.00
	sum_click	int64	Nombre total de clics	0	0.00
vle	id_site	int64	Identifiant unique du matériel	0	0.00

Suite à la page suivante

Table 1 – suite de la page précédente

Dataset	Colonne	Type	Description	N manq.	%
	code_module	object	Code du module	0	0.00
	code_presentation	object	Code de la session	0	0.00
	activity_type	object	Type d'activité du matériel pédagogique	0	0.00

2.2 Traitement des données

2.2.1 Valeur manquantes

Dans le tableau 1 et la figure 1, on présente le nombre et le pourcentage de valeurs manquantes pour chaque dataset. On remarque que la majorité des attributs sont complets, à l'exception de quelques variables nécessitant un traitement particulier : notamment l'attribut *imd band* dans le fichier *studentInfo.csv*, *date unregistration* dans *studentRegistration.csv*, et *score* dans *studentAssessment.csv*.

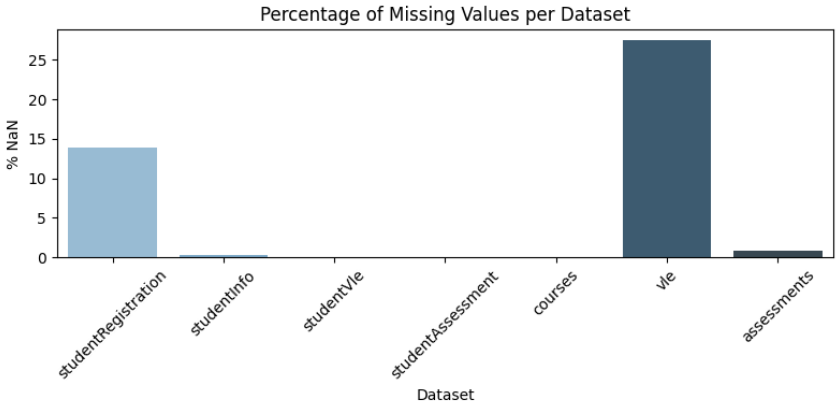

FIGURE 1 – Percentage des valeurs manquantes par dataset

Afin d'obtenir un traitement plus précis des données, on a appliqué un prétraitement spécifique à chaque variable contenant des valeurs manquantes :

- `imd_band` **(studentInfo)** : Cet attribut indique le *niveau de déprivation multiple* (*Index of Multiple Deprivation*) correspondant à la région où l'étudiant résidait pendant la présentation du module. Les valeurs manquantes ont été imputées en utilisant la modalité la plus fréquente (*mode*) observée pour la même région, afin de maintenir une cohérence géographique.
- `date_unregistration` **(studentRegistration)** : Cette variable décrit la date à laquelle un étudiant s'est désinscrit d'un module. Les statuts *Pass*, *Fail* et *Distinction* indiquent que l'étudiant a terminé le module, tandis que le statut *Withdrawn* signifie qu'il l'a abandonné (voir figure 2). Étant donné que les valeurs manquantes ne peuvent pas être imputées de manière fiable, et que la distribution des valeurs non nulles (voir figure 3) montre une forte concentration chez les étudiants ayant terminé le module, on a choisi de supprimer cette variable, jugée peu représentative.

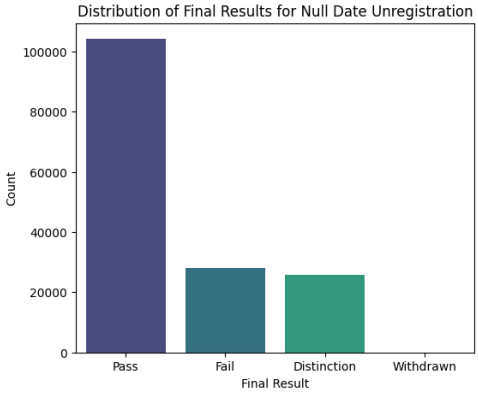

FIGURE 2 – Distribution des résultats finaux pour les dates de désinscription nulles

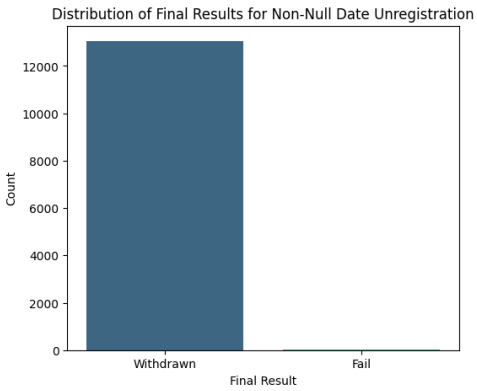

FIGURE 3 – Distribution des résultats finaux pour les dates de désinscription non nulles

- `score` **(studentAssessment)** : Cette variable correspond à la note obtenue à une évaluation. Lorsqu'aucune valeur n'est renseignée, on considère que l'étudiant n'a pas soumis l'évaluation. Dans ce cas, la valeur manquante a été remplacée par **0**, afin de refléter une absence de participation équivalente à une note nulle.
- `week_from et week_to` **(vle)** : Ces deux variables ont été jugées peu pertinentes pour notre étude. Par conséquent, on a décidé de supprimer ces colonnes.

2.2.2 Valeur aberrantes

Ci-dessous (figure 4), on présente les boîtes à moustaches des variables numériques pour chaque jeu de données.

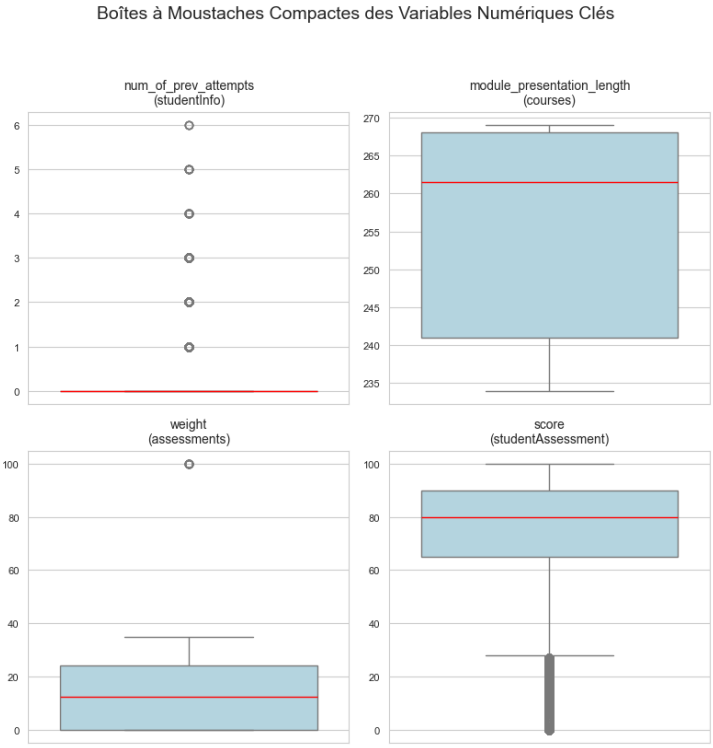

FIGURE 4 – Boîtes à moustaches des variables numériques

Pour le traitement des valeurs aberrantes, on a appliqué une imputation par la médiane pour la

variable `module_presentation_length`. Concernant les autres variables, les valeurs extrêmes ont été conservées, car elles restent cohérentes avec la nature des données. Par exemple, la variable `score` varie naturellement entre 0 et 100, ce qui rend légitime la présence de valeurs faibles (comme entre 0 et 20).

2.3 Analyse statistique préliminaire

L'analyse des histogrammes (Figure 5 met en évidence une forte hétérogénéité entre les variables numériques, révélant la présence de sous-populations distinctes et un déséquilibre notable dans les données. La variable `num_of_prev_attempts` montre une forte asymétrie : la majorité des étudiants s'inscrivent pour la première fois, tandis qu'une minorité ayant plusieurs tentatives tire la moyenne vers le haut. De même, la variable `weight` présente une distribution très déséquilibrée, la plupart des évaluations ayant un faible poids alors que quelques-unes atteignent 100 La `module_presentation_length` présente une distribution multimodale, centrée sur quelques durées standardisées, suggérant des différences structurelles entre les modules. Enfin, la variable `score` adopte une forme bimodale : un groupe d'étudiants obtient de très bons résultats, tandis qu'un autre présente des scores faibles, reflétant une distinction claire entre réussite et échec.

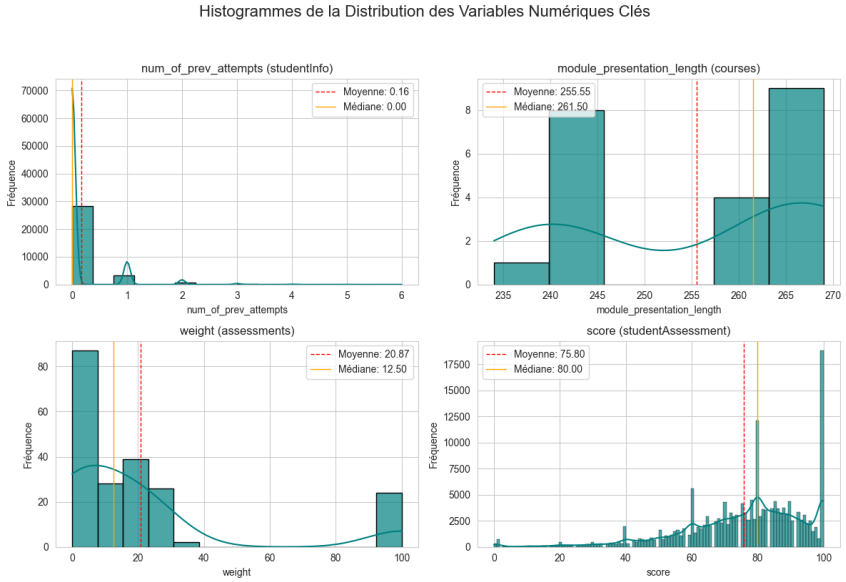

FIGURE 5 – Histogrammes de la Distribution des Variables Numériques Clés

3 Analyse par graph

Après l'analyse exploratoire des données, plusieurs questions de recherche se sont imposées :

- Quels types de ressources sont les plus étroitement associés à la réussite académique ?
- Existe-t-il des profils d'apprentissage distincts parmi les étudiants (par exemple, « orientés forum » vs « orientés quiz ») ?
- Peut-on identifier les étudiants à risque à partir de leur connectivité dans le réseau d'interactions ?
- Quelles sont les ressources critiques, incontournables pour la réussite ?

Ces interrogations nous ont donc naturellement conduit à adopter des approches basées sur les graphes.

3.1 Approche 01 : Détection de communautés d'étudiant sur graphe

L'objectif de cette partie est de déterminer s'il est possible de faire émerger des **profils d'usage d'apprenants** à partir de leurs interactions avec des ressources. Notre hypothèse est qu'un graphe étudiant–étudiant basé sur la similarité cosinus, construit depuis des vecteurs d'activité pondérés en tf–idf, révèle des communautés interprétables et informatives pour l'analyse pédagogique.

L'idée est donc de représenter la cohorte d'étudiants sous la forme d'un **graphe de similarité apprenant–apprenant**. Chaque étudiant est un *nœud* du graphe, et une *arête* relie deux étudiants lorsqu'ils ont un comportement d'usage proche. Ainsi, un graphe bien construit permet de visualiser les relations de proximité dans les comportements d'étude : des étudiants fortement connectés correspondent à des pratiques similaires.

3.1.1 Protocole

Chaque étudiant est représenté par un **vecteur d'activité** qui regroupe le nombre total de clics sur chaque ressource (`id_site`) de la plateforme. Ces vecteurs traduisent la manière dont chaque apprenant interagit avec les contenus du cours. Nous calculons ensuite la **similarité cosinus** entre tous les couples d'étudiants afin d'estimer à quel point leurs comportements d'usage sont proches.

Avant ce calcul, trois stratégies de **normalisation** ont été testées : (i) la version *raw*, où les données brutes sont utilisées telles quelles ; (ii) la version *rowmax*, qui ramène chaque profil dans l'intervalle $[0, 1]$ en divisant par la valeur maximale de la ligne ; et (iii) la version *tf–idf*, qui atténue l'influence des ressources trop consultées et valorise les interactions plus spécifiques. Cette dernière permet de distinguer des comportements d'apprentissage différents, même lorsque les étudiants consultent globalement les mêmes ressources.

À partir de la matrice de similarité obtenue, nous construisons un **graphe des k plus proches voisins** (k–NN). Chaque nœud correspond à un étudiant, et une arête est ajoutée lorsqu'un autre étudiant figure parmi ses k voisins les plus similaires. Un **seuil de similarité** τ est également appliqué pour filtrer les connexions trop faibles et ne garder que les liens significatifs. Nous faisons varier ces deux paramètres selon $k \in \{5, 10, 15, 20, 30\}$ et le **seuil** $\tau \in \{0.0, 0.1, 0.2, 0.3\}$. Cela permet d'évaluer l'effet de la densité du graphe (k) et du degré de tolérance à la similarité (τ) sur la structure communautaire obtenue.

Les communautés sont ensuite détectées à l'aide de l'algorithme **Louvain**, reconnu pour son efficacité sur des graphes de grande taille. Chaque partition est évaluée selon quatre indicateurs : la **modularité** Q, qui mesure la cohérence interne des groupes ; la **pureté**, qui indique leur correspondance avec les résultats académiques (`final_result`) ; ainsi que la **NMI** et l'**ARI**, qui évaluent la stabilité et la qualité globale des partitions.

La configuration optimale est définie comme celle qui maximise la modularité tout en maintenant un bon équilibre entre **séparation des groupes et lisibilité des résultats**. Autrement dit, nous cherchons un graphe suffisamment structuré pour faire émerger des profils distincts, sans créer de micro-communautés difficilement interprétables.

3.1.2 Résultats et analyse des paramètres

Le tableau 2 présente un extrait des résultats obtenus. Nous observons que le prétraitement **tf-idf** offre systématiquement les meilleures performances, avec une modularité proche de 0.66 pour $k = 5$ et $\tau = 0.2$. Les méthodes *raw* et *rowmax*, en revanche, atteignent difficilement $Q = 0.60$, traduisant des graphes moins structurés.

Norm.	k	τ	n_nodes	n_edges	Q	n_comm.	Pureté	NMI	ARI
tf-idf	5	0.2	378	1498	**0.659**	10	0.817	0.113	0.058
tf-idf	5	0.1	378	1502	0.654	9	0.817	0.114	0.059
tf-idf	5	0.3	376	1483	0.649	9	0.824	0.116	0.055
rowmax	5	0.2	378	1531	0.603	9	0.807	0.116	0.036
raw	5	0.1	378	1531	0.603	9	0.807	0.116	0.036

TABLE 2 – Indicateurs calculés pour différentes configurations du graphe (Louvain).

La Figure 6 montre la variation de la modularité moyenne selon le nombre de voisins k, moyennée sur les valeurs de τ. On constate une **diminution nette de** Q lorsque k augmente : des graphes trop denses font disparaître les frontières entre groupes. La normalisation **tf-idf** se distingue nettement, confirmant qu'elle permet de mieux capturer la diversité des comportements d'apprentissage.

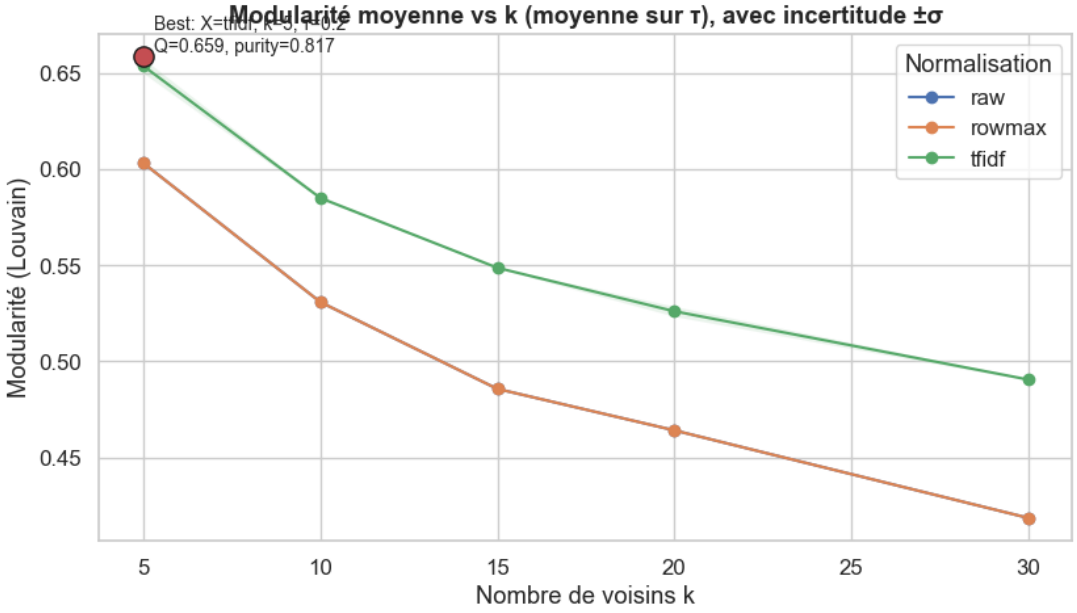

FIGURE 6 – **Modularité moyenne vs** k (moyenne sur τ).

Bien plus, la Figure 7 illustre la modularité Q pour chaque combinaison de paramètres (k, τ). L'effet du seuil τ reste limité comparé à celui de k, ce qui montre que le nombre de voisins est le paramètre le plus déterminant. Une nouvelle fois, la normalisation *tf-idf* produit les valeurs les plus élevées de modularité.

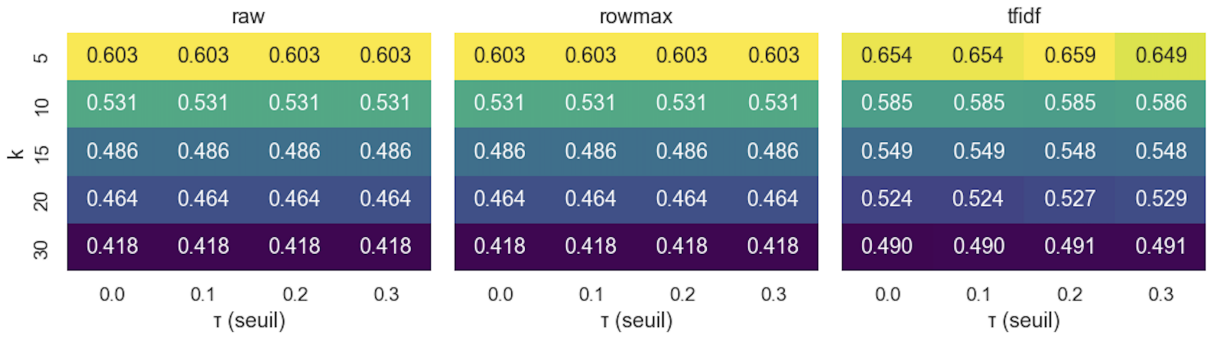

FIGURE 7 – **Cartes** $k \times \tau$ de la modularité (Louvain)

La Figure 8 illustre la partition Louvain sur le graphe optimal. Chaque couleur correspond à une communauté ; et nous observons des amas denses bien séparés, indiquant une partition nette et exploitable

FIGURE 8 – **Partition Louvain** (tf-idf, k=5, τ=0.2).

Configuration retenue.

$$\boxed{\text{tf-idf}, \quad k = 5, \quad \tau = 0.2}$$

Cette configuration présente la meilleure structure communautaire avec une modularité de 0.659, une pureté de 0.817. Nous obtenenons aussi pour ce grpahe de **378** nœuds et **1498** arêtes une densité 0.021 et un degré moyen d'environ 7.9.

3.1.3 Comparaison des méthodes de détection de communautés

Une fois les paramètres du graphe optimisés (`tf-idf`, $k = 5$, $\tau = 0.2$), l'étape suivante consiste à vérifier si la structure communautaire obtenue dépend du choix de l'algorithme de détection. Autrement dit, nous cherchons à savoir si les communautés mises en évidence par Louvain reflètent réellement une organisation stable du graphe, ou si elles résultent d'un biais propre à cette méthode.

Pour cela, nous avons appliqué plusieurs algorithmes de référence sur le même graphe optimal :

- **Louvain** : approche hiérarchique d'optimisation de la modularité, utilisée comme référence principale ;
- **Greedy (Clauset–Newman–Moore)** : stratégie gloutonne qui fusionne itérativement les groupes pour maximiser Q ;

- **Infomap** : méthode informationnelle, fondée sur la probabilité qu'un *random walk* reste à l'intérieur d'une même communauté ;
- **Girvan–Newman** : approche plus classique, reposant sur la suppression progressive des arêtes à forte centralité d'intermédiarité.

Le tableau 3 et la figure 9 présentent les scores obtenus selon différents critères (modularité, pureté, NMI et ARI).

Méthode	Q	Communautés	Pureté	NMI	ARI
Louvain	**0.659**	10	0.817	0.113	0.058
Greedy	0.641	8	0.825	0.132	0.080
Infomap	0.627	26	**0.844**	0.119	0.040
Girvan–Newman	0.533	5	0.833	**0.152**	**0.144**

TABLE 3 – Comparaison des principales méthodes de détection de communautés sur le graphe optimal (`tf-idf`, k=5, τ=0.2).

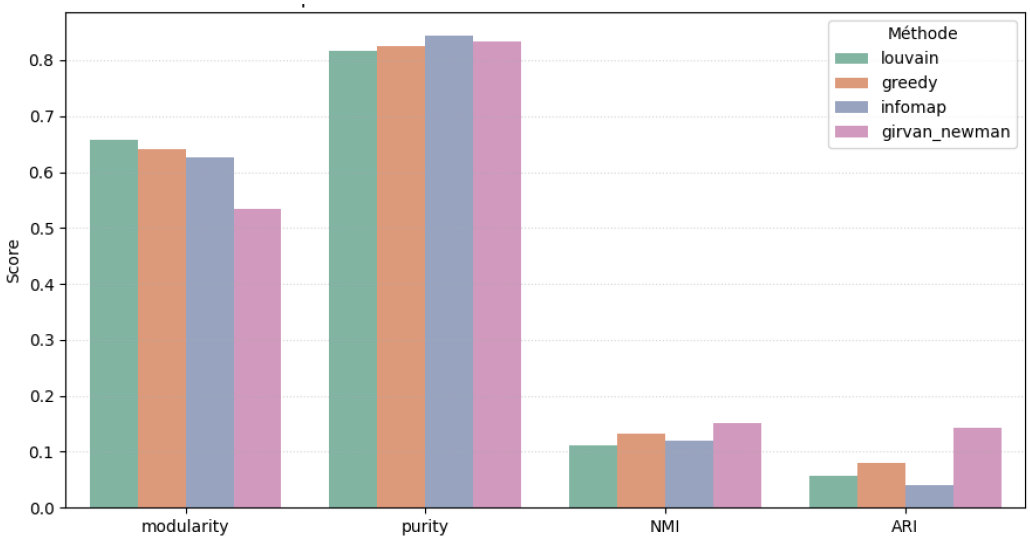

FIGURE 9 – Comparaison visuelle des méthodes selon différentes métriques (modularité, pureté, NMI, ARI).

On note que les performances varient selon les algos. En effet, l'algorithme **Louvain** présente la **meilleure modularité** ($Q = 0.659$) tout en conservant un nombre raisonnable de communautés (10), garantissant une bonne lisibilité. **Greedy** obtient une modularité légèrement inférieure mais une pureté plus forte, ce qui suggère des groupes un peu plus homogènes vis-à-vis de la réussite académique. **Infomap** identifie 26 micro-communautés, vela permet d'obtenir une bonne pureté mais nuit à l'interprétation pédagogique. Enfin, **Girvan–Newman** offre les NMI et ARI les plus élevés, traduisant une meilleure correspondance locale aux résultats étudiants, mais au prix d'une modularité nettement plus faible.

Ces observations montrent que les algorithmes ne construisent pas la même vision du graphe : Louvain et Greedy privilégient une **cohérence structurelle globale**, tandis qu'Infomap et Girvan–Newman accentuent la **précision locale**.

Dans notre contexte, où l'objectif est de mettre en évidence des **profils d'usage cohérents et interprétables**, Louvain apparaît comme le meilleur compromis. Il conserve une structure claire (10 communautés bien séparées) tout en maintenant une correspondance satisfaisante avec la réussite académique.

3.1.4 Interprétation des communautés (pattern mining)

Après avoir identifié la structure du graphe, nous avons cherché à **caractériser le comportement interne** de chaque communauté. Pour cela, une analyse de **règles d'association** (algorithme *Apriori*) a été appliquée sur les traces d'usage des étudiants regroupés dans chaque cluster. L'objectif est de révéler des **co-occurrences d'activités significatives**, c'est-à-dire des ressources souvent consultées ensemble au sein d'une même communauté.

Pour cela, pour chaque communauté, nous extrayons les associations du type $A \rightarrow B$ avec un **support minimal de 0.25** et un **lift supérieur à 1.1**. Bien plus, les indicateurs utilisés sont :

- le **support**, qui indique la fréquence du motif dans la communauté ;
- la **confiance**, qui mesure la probabilité que B soit utilisé si A l'est ;
- le **lift**, qui évalue la force réelle de la corrélation (lift > 1 = usage conjoint plus fréquent que le hasard).

Ces métriques visent à éliminer les associations faibles et à ne conserver que les liens récurrents et pertinents.

3.1.5 Interprétation des communautés (pattern mining)

Après la détection des communautés sur le graphe d'apprenants, il est essentiel de comprendre **pourquoi** ces regroupements se sont formés et **quelles pratiques d'usage** ils reflètent réellement. Pour cela, nous avons appliqué une **fouille de règles d'association** (*Apriori*) à l'intérieur de chaque communauté détectée par l'algorithme Louvain. Cette étape vise à identifier les **combinaisons d'activités fréquemment co-utilisées** par les étudiants d'un même groupe. En pratique, l'analyse a été conduite pour l'ensemble des communautés, mais seul l'exemple de la **communauté 4** est détaillé ici, car il illustre clairement le type de profil comportemental que cette méthode permet de révéler.

Donc pour chaque communauté, nous extrayons les règles d'association de la forme $A \rightarrow B$. Bien plus, les indicateurs utilisés sont :

- le **support** : fréquence du motif au sein de la communauté ;
- la **confiance** : probabilité que B soit utilisé lorsque A l'est ;
- le **lift** : intensité réelle de la corrélation (valeurs supérieures à 1 = usage conjoint non aléatoire).

Exemple : Communauté 4. D'après les résultats obtenues, la communauté 4 regroupe un sous-ensemble d'étudiants caractérisés par une activité particulièrement marquée autour des ressources collaboratives et synchrones. En effet, les principales règles extraites sont :

$$\text{classe virtuelle} \rightarrow \text{activité collab.} \quad (\text{lift} = 1.35, \text{conf} = 0.90)$$

$$\text{activité collab.} \rightarrow \text{classe virtuelle} \quad (\text{lift} = 1.35, \text{conf} = 0.64)$$

Ces relations bidirectionnelles traduisent donc une forte interdépendance entre la participation à des **classes virtuelles** et l'engagement dans des **activités collaboratives**. Autrement dit, les étudiants de cette communauté tendent à combiner les deux modes d'interaction, synchrone et collectif, plutôt qu'à les utiliser séparément.

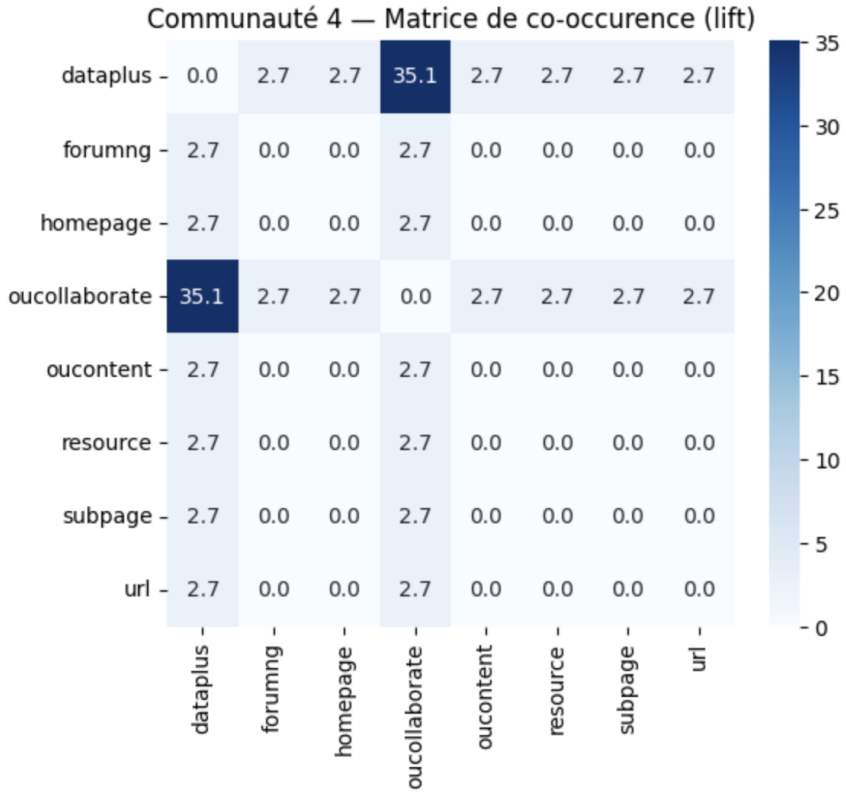

FIGURE 10 – **Communauté 4 — Matrice de co-occurrence (lift).**

Bien plus, la matrice de co-occurrence atteste ces résultats (Figure 10). Effectivement, cela révèle bien que les modules `oucollaborate` et `dataplus` présentent un lift particulièrement élevé et que donc ces outils sont tous deux liés à des usages collectifs (forums de travail, dépôts partagés, activités collaboratives).

En résumé, cette communauté regroupe des apprenants **coopératifs et synchrones**, qui privilégient les échanges et la co-construction plutôt que la simple consommation de ressources.

Le pattern mining nous a donc aidé à comprendre comment la **formation du cluster** s'est opérée dans le graphe initial. Autrement dit, ce cluster ne résulte pas d'un artefact de l'algorithme, mais bien d'une proximité comportementale réelle.

Ce type d'analyse démontre que la fouille de motifs ne sert pas seulement à décrire les comportements, mais à **interpréter la structure du graphe** en reliant les connexions détectées à des dynamiques d'apprentissage observables.

3.2 Approche 02 : Réseau bipartite étudiant–ressource

L'approche précédente s'intéressait aux relations entre étudiants, en identifiant des communautés de comportements similaires. Pour compléter cette analyse, nous proposons ici de relier directement les étudiants aux **ressources pédagogiques** qu'ils consultent, afin d'examiner les interactions au cœur même de leur activité d'apprentissage.

L'objectif n'est plus seulement d'observer des proximités entre apprenants, mais de comprendre **quelles ressources structurent leurs pratiques** et comment ces usages varient selon les profils de réussite académique. Pour cela, nous construisons un **réseau bipartite étudiant–ressource**. Ce modèle permet de représenter les interactions entre les étudiants et les ressources pédagogiques, afin d'identifier les patterns d'usage caractéristique associés aux différents niveaux de performance académique.

3.2.1 Construction du réseau

Le réseau bipartite construit relie deux ensembles de nœuds distincts :

- **Étudiants** (S) : identifiés à partir de la table `studentInfo`, contenant leurs caractéristiques démographiques et le résultat final (*Pass*, *Fail*, *Distinction*, *Withdrawn*) ;
- **Ressources pédagogiques** (A) : extraites de la table `vle`, décrivant les types d'activités disponibles sur la plateforme VLE.

Les **arêtes** représentent les interactions entre les étudiants et les ressources, pondérées par le **nombre de clics** enregistré dans la table `studentVle`.

La représentation formelle du graphe est donnée par :

$$G = (V, E, W), \quad V = S \cup A, \quad E \subseteq S \times A, \quad W : E \to \mathbb{R}^+,$$

où W correspond au poids de l'interaction (nombre de clics).

Les principaux attributs des nœuds qu'on garde sont alors pour :

- **Étudiant :** `final_result`, `gender`, `age_band`, `bipartite = 0`
- **Activité :** `activity_type`, `bipartite = 1`

3.2.2 Résultats et analyses

Visualisation du graphe

La figure 11 illustre le graphe bipartite représentant les interactions entre les étudiants et les types de ressources du module `"BBB"` durant l'année 2013. En raison de la complexité du réseau global, seule la visualisation d'un module a été retenue pour une meilleure lisibilité, comme pour la premiere approche. On observe une forte convergence vers les forums (`forumng`), indiquant leur rôle central dans les interactions étudiantes. Les autres types d'activités, tels que `glossary`, `ouwiki` ou `questionnaire`, présentent des connexions plus fines et dispersées, traduisant une utilisation plus ponctuelle. Enfin, on note qu'aucun étudiant n'est isolé, ce qui témoigne d'une participation minimale de tous les apprenants aux activités du module.

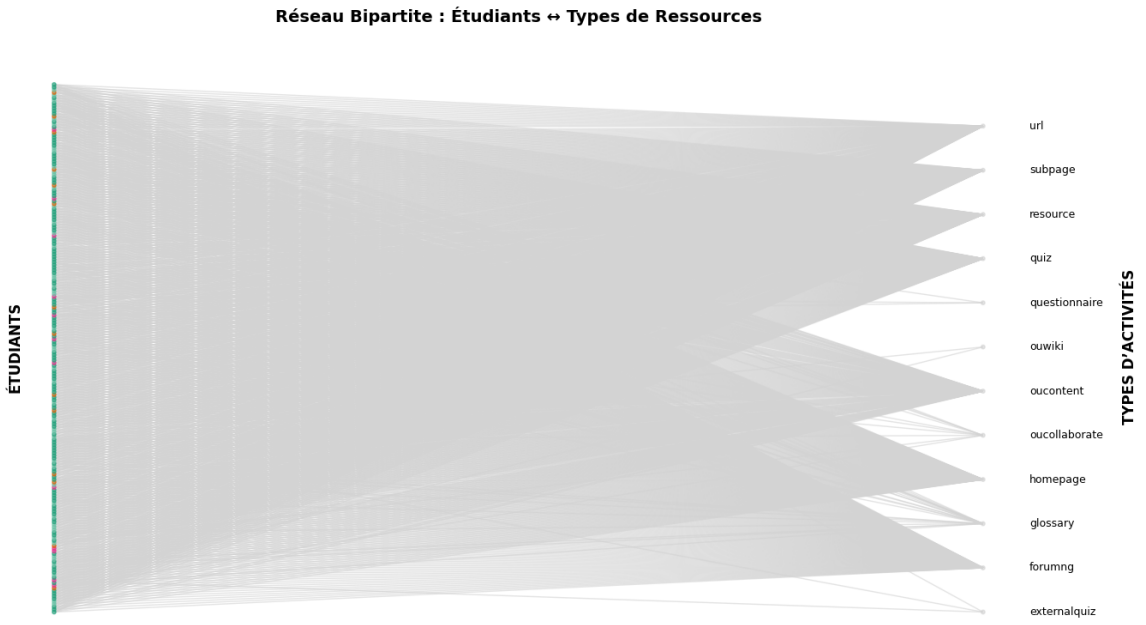

FIGURE 11 – Graph bipartite Étudiants - Type de ressource

Analyse structurelle du réseau

Le réseau biparti construit relie les étudiants aux différents types d'activités pédagogiques de la plateforme. Après filtrage, nous avons obtenu un graphe composé de **212 nœuds** (dont 200 étudiants et 12 types d'activités) et de **1413 arêtes**. Le graphe présente une **densité de 0.5887**, indiquant un niveau de connectivité élevé entre étudiants et ressources. Il est entièrement **connexe**, ce qui signifie qu'il existe un chemin entre chaque paire de nœuds, traduisant ainsi une forte cohésion du réseau.

Le **degré moyen global** du réseau est de **13.33**. Plus précisément, chaque étudiant est connecté en moyenne à **7.07 types d'activités**, tandis que chaque activité est reliée à environ **117.75 étudiants**. Cette asymétrie reflète la nature bipartite du réseau : les ressources sont massivement sollicitées par de nombreux étudiants, tandis que chaque étudiant interagit de manière sélective avec un sous-ensemble restreint d'activités. Le **diamètre du réseau** est de **4**, ce qui indique que deux nœuds quelconques peuvent être reliés en quatre étapes au maximum. Le **chemin moyen**, évalué à **1.98**, témoigne d'une structure compacte où les interactions entre étudiants et activités restent très rapprochées.

Visualisation avec Gephi

Afin d'obtenir une visualisation plus claire et interactive du réseau, nous avons utilisé **Gephi**. Sur le graphe présenté en figure 12, nous avons appliqué la mesure de *Network Diameter* pour analyser la structure globale du réseau, ainsi que la *Betweenness Centrality* afin de dimensionner la taille des nœuds selon leur importance. Les nœuds ont été colorés en fonction du *résultat final* des étudiants : chaque couleur représente une catégorie (par exemple, violet pour `Pass`, rouge pour `Fail`, etc.), tandis que les grands nœuds verts correspondent aux différents *types de ressources*. Les arêtes sont colorées et pondérées selon leur intensité d'interaction (nombre de clics).

FIGURE 12 – Gephi : Graph bipartite Étudiants - Type de ressource

Cette visualisation met clairement en évidence que la majorité des étudiants ayant réussi sont ceux qui interagissent le plus fréquemment avec les ressources pédagogiques, en particulier avec le `forumng`, qui apparaît comme une ressource centrale associée à la réussite académique.

3.2.3 Synthèse de l'analyse des interactions étudiant–ressource

L'analyse des quatre dimensions (popularité, engagement, intensité et score composite) (voir figure 13) met en évidence une hiérarchie claire dans l'usage des ressources du VLE.

- **Popularité :** Les ressources `forumng`, `homepage`, `subpage`, `quiz`, `resource`, `oucontent` et `url` sont utilisées par la quasi-totalité des étudiants (environ 200). À l'inverse, les activités telles que `externalquiz`, `questionnaire`, `oucollaborate` et `ouwiki` restent très marginales.
- **Engagement :** Le volume total d'interactions confirme la domination du `forumng` (450 000 clics) et de la `homepage` (200 000 clics), soulignant leur rôle central dans la navigation et les échanges.
- **Intensité :** En moyenne, un étudiant effectue plus de 2000 clics sur le `forumng` et environ 500 sur la `homepage`, illustrant une implication importante dans les activités collaboratives.
- **Score composite :** Le croisement entre la popularité et l'intensité confirme la position dominante de `forumng`, suivie de `homepage` et `subpage`, comme ressources clés du dispositif d'apprentissage.

Le `forumng` émerge comme la ressource la plus stratégique, concentrant à la fois la participation, l'engagement et la durée d'interaction des étudiants. L'apprentissage en ligne semble ainsi fortement médié par la collaboration et l'échange d'informations.

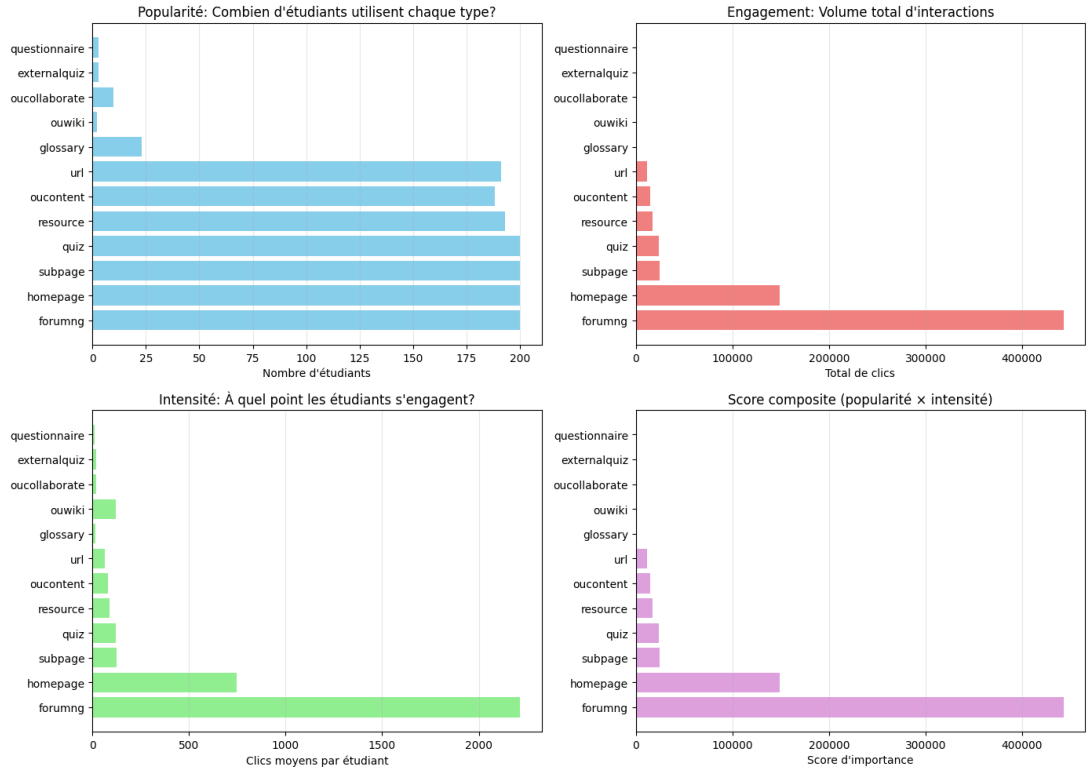

FIGURE 13 – Aanalyse des quatre dimensions

3.2.4 Détection de communautés et profils d'apprentissage

Méthodes et approche

Le réseau bipartite *Étudiants–Types de ressources* présente une densité élevée ($\rho = 0.59$) et une structure quasi-complète, ce qui limite l'efficacité des algorithmes classiques de détection de communautés (tels que *Louvain* ou *Label Propagation*). Pour pallier cette contrainte structurelle, nous avons adopté une **projection unipartite**, transformant le réseau en graphe *Étudiant–Étudiant*, où deux étudiants sont reliés lorsqu'ils présentent des comportements d'usage similaires.

La similarité entre deux étudiants i et j est mesurée à l'aide du **cosinus de similarité** :

$$\text{sim}(i,j) = \frac{M_i \cdot M_j}{\|M_i\|\|M_j\|}$$

où M représente la matrice normalisée (200×12) décrivant le nombre de clics par type d'activité. Nous avons testé plusieurs **seuils de similarité** afin d'identifier le compromis optimal entre densité du graphe et structure informative :

Seuil	Nombre d'arêtes	Densité	Interprétation
0.5	∼35 000	0.98	Trop permissif, graphe surconnecté
0.7	18 546	0.93	Équilibre entre connectivité et différenciation
0.9	∼5 000	0.25	Trop strict, graphe fragmenté

Le seuil **0.7** a finalement été retenu, garantissant un réseau suffisamment connecté tout en préservant des différences comportementales exploitables.

Deux approches complémentaires de détection ont ensuite été appliquées :

- **Label Propagation** : méthode non supervisée et rapide, adaptée aux grands graphes mais sensible à l'homogénéité.
- **K-Means (k = 4)** : approche supervisée permettant de forcer une partition en groupes interprétables, même dans un graphe dense.

Le choix du nombre de clusters k dans K-Means a été guidé par une double logique :

- **Intuition pédagogique** : quatre profils d'apprentissage étaient attendus a priori (étudiants très engagés, réguliers, collaboratifs, et désengagés).
- **Validation empirique** : les tests réalisés ont montré que $k = 2$ produisait une classification trop grossière, tandis que $k = 6$ fragmentait excessivement les groupes, générant des clusters très petits et peu significatifs. Le choix de $k = 4$ constitue ainsi un compromis cohérent entre granularité et interprétabilité.

Résultats et analyse

L'application de **Label Propagation** a conduit à la détection de seulement **deux communautés**, (Voir figure 14) dont une dominante (198 étudiants, 99%) et une marginale (2 étudiants). La **modularité nulle** ($Q = 0.000$) confirme l'absence de structure communautaire distincte, signe d'une forte homogénéité des comportements d'apprentissage parmi les étudiants les plus actifs. Les deux étudiants isolés correspondent à des profils atypiques, tous deux en situation d'échec ou de retrait.

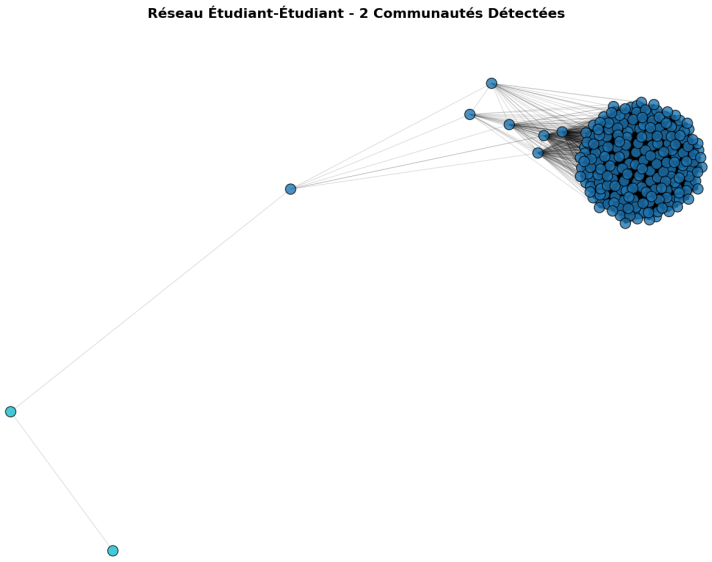

FIGURE 14 – Visualisation des 2 communautés - Label Propagation

En revanche, le **clustering K-Means** ($k = 4$) a permis de révéler des profils d'apprentissage plus fins malgré la densité élevée du graphe. (Voir figure 15) Les quatre clusters identifiés sont les suivants :

- **Cluster 0 "Super-engagés" (77%)** : usage intensif du `glossary`, `ouwiki` et `externalquiz`. Taux de réussite : 89.6%.
- **Cluster 3 "Navigateurs classiques" (20%)** : dépendance à `homepage`, `subpage` et `quiz`. Taux de réussite : 85.0%.
- **Cluster 2 "Explorateurs collaboratifs" (1.5%)** : interaction forte via `ouwiki` et `oucollaborate`. Taux de réussite : 66.7%.
- **Cluster 1 "Désengagés" (1.5%)** : activité très faible, centrée sur `questionnaire` et `oucontent`. Taux de réussite : 0%.

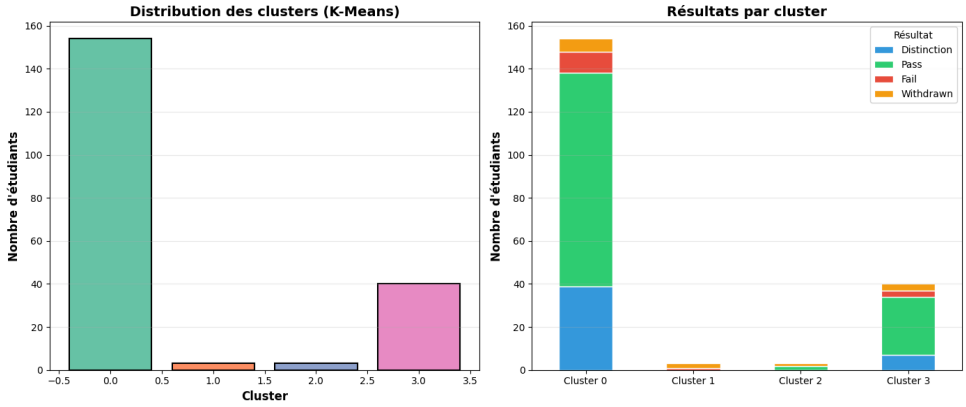

FIGURE 15 – distribustion et résultats par cluster

Ces résultats montrent que, même en l'absence de modularité globale, il est possible d'identifier des **profils comportementaux** associés à des niveaux de réussite distincts. L'usage d'outils collaboratifs comme le `glossary` et le `ouwiki` est fortement corrélé à la performance académique, tandis qu'une navigation centrée sur la `homepage` ou un engagement sporadique constitue un signal d'alerte.

Résumé

La détection classique de communautés s'avère inefficace sur un réseau bipartite dense ($Q = 0$). Cependant, la combinaison **projection unipartite + clustering K-Means** a permis de mettre en évidence des **profils d'apprentissage différenciés** :

- Les étudiants performants exploitent activement les outils collaboratifs et de référence.
- Les échecs sont associés à un engagement faible ou non collaboratif.

Les expérimentations sur différents seuils de similarité ont confirmé la robustesse du choix sim $= 0.7$, garantissant un équilibre optimal entre lisibilité structurelle et connectivité. Ces observations ouvrent la voie à une détection précoce du désengagement et à la mise en place d'actions pédagogiques ciblées. Une extension future consisterait à inclure l'ensemble des 2 237 étudiants et à intégrer une dimension temporelle pour suivre l'évolution des comportements au fil du semestre.

4 Discussion et limites

Les deux approches explorées, le graphe de similarité étudiant–étudiant et le réseau bipartite étudiant–ressource, permettent d'aborder les mêmes problématiques sous des angles complémentaires. Cela offre une compréhension plus fine des comportements d'apprentissage où chaque méthode révèlent des aspects que l'autre saisit moins directement.

Malgré leurs différences méthodologiques, les deux analyses convergent vers des conclusions cohérentes. Le graphe étudiant–étudiant a mis en évidence des communautés reflétant des profils d'usage distincts, par exemple un groupe fortement impliqué dans les activités collaboratives (communauté 4). De son côté, l'approche bipartite a confirmé la centralité du `forumng` et son lien étroit avec la réussite académique. Ainsi, même si les modèles diffèrent, ils traduisent une même logique d'analyse : les étudiants les plus engagés et collaboratifs sont aussi ceux qui réussissent le mieux.

Chacune des deux méthodes apporte cependant ses spécificités. Le graphe de similarité offre une vision structurelle des relations entre étudiants, utile pour détecter des communautés cohérentes, mais dont la pureté académique reste modérée ($ARI = 0.058$). L'approche bipartite, quant à elle plus proche des interactions réelles avec les ressources, ont en revanche une forte densité du réseau, rendant la détection de communautés plus difficile.

Ces limites ouvrent des pistes d'amélioration à noter :

- **Hybridation des modèles :** combiner les relations étudiant–étudiant et étudiant–ressource au sein d'un graphe multiplex pour croiser les deux perspectives ;
- **Intégration temporelle :** étudier l'évolution des comportements au fil du semestre afin de mieux comprendre les dynamiques d'engagement ;
- **Enrichissement contextuel :** inclure d'autre variable, tels que démographiques ou académiques pour affiner la caractérisation des profils.

En définitive, la complémentarité entre ces deux approches offre une vision plus globale et interprétable des mécanismes d'apprentissage. Des approches comme celle-ci ouvrent la voie à des dispositifs pédagogiques plus adaptatifs, capables de tenir compte à la fois des structures relationnelle et des pratiques réelles d'utilisation des ressources.

Références

[1] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1) :1–8, 2017.

[2] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset (oulad). *Scientific Data*, 4 :170171, 2017. Licence CC-BY 4.0.

Annexe

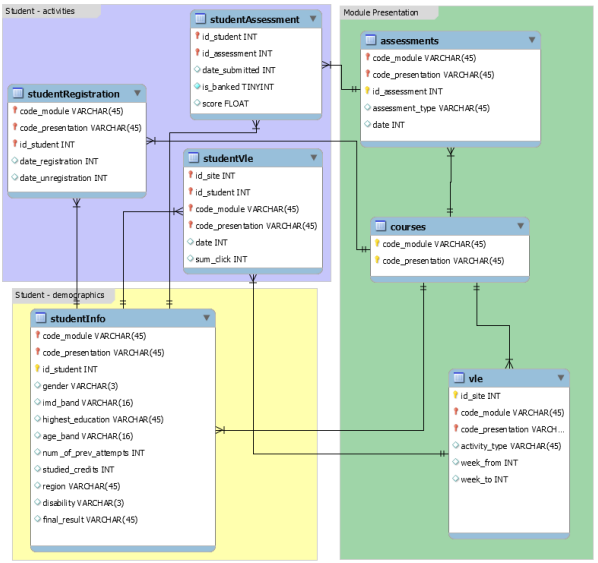

FIGURE 16 – Schéma de la base de données