HEC MONTRÉAL

Analyser les tendances : Une étude comparative du contenu de la revue Management international et des cahiers du GERAD

par

Gabrielle Alie

Gilles Caporossi
HEC Montréal
Directeur de recherche

Sciences de la gestion
(spécialisation Sciences des données et analytique d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Avril 2024

Résumé

L'accès rapide aux informations pertinentes est primordial dans le contexte actuel marqué par un flux constant de données, en particulier pour l'extraction d'informations à partir de données non structurées telles que le texte. Les techniques de modélisation thématique sont particulièrement utiles dans ce contexte. Elles fournissent des moyens concrets pour analyser ces données, permettant de découvrir de manière automatisée des structures et des thèmes cachés dans le contenu textuel. L'utilisation de ces techniques facilite la prise de décision, améliore la gestion du contenu et rend la recherche d'informations plus efficace.

Cette recherche est une étude comparative qui vise à évaluer la modélisation thématique grâce à l'application du modèle Allocation Latente de Dirichlet (LDA). Elle explore également la dynamique des communautés sur un corpus de 828 résumés de la revue Management international (Mi) juxtaposé à un corpus parallèle, les cahiers du GERAD. L'étude présente une analyse thématique de la revue et révèle des communautés distinctes qui reflètent non seulement l'identité de la revue, mais contribuent également à son évolution dynamique. Les cahiers du GERAD, avec son équipe diversifiée de spécialistes en sciences des données et de la décision, serviront de référence pour établir des comparaisons. Bien que l'objectif principal soit d'amplifier l'impact de Mi, l'analyse comparative est essentielle pour découvrir le potentiel de la revue. Cette juxtaposition permettra de fournir des recommandations concrètes pour améliorer les stratégies de curation de contenu et d'engagement, assurant ainsi la pertinence continue de Mi dans le domaine en constante évolution, celle du management international.

Mots clés : Traitement du langage, Extraction des thèmes, LDA, Données non structurées, Optimisation, Détection de communautés, Management international

Méthodes de recherche : Allocation Latente de Dirichlet, Détection de communauté grâce à un réseau

Abstract

Rapid access to relevant information is essential in today's context of constant data flow, particularly when it comes to extracting information from unstructured data such as text. Thematic modeling techniques are particularly useful in this context. They provide a practical means of analyzing such data, enabling the automated discovery of structures and themes hidden within textual content. The use of these techniques facilitates decision-making, improves content management and makes information retrieval more efficient.

This research is a comparative study that aims to evaluate thematic modeling through the application of Dirichlet's Latent Allocation (LDA) model. It also explores community dynamics on a corpus of 828 articles from Management International (Mi) juxtaposed with a parallel corpus, the cahiers du GERAD. The study presents a thematic analysis of the journal, revealing distinct communities that not only reflect the journal's identity, but also contribute to its dynamic evolution. Les cahiers du GERAD, with its diverse team of data and decision scientists, will serve as a benchmark for comparison. Although the main aim is to amplify Mi's impact, comparative analysis is essential to discover the journal's potential. This juxtaposition will provide concrete recommendations for improving content curation and engagement strategies, ensuring Mi's continued relevance in the ever-evolving field of international management.

Keywords: Language processing, Theme extraction, LDA, Unstructured data, Optimization, Community detection, International Management

Research methods: Latent Dirichlet Allocation, Network-based Community Detection

Table des matières

Liste des tableaux

Liste des Figures

Liste des abréviations

2D : deux dimensions

BoW : sac de mots

BTM : biterm topic model

CRF : Conditional Random Fields

DBSCAN : density-based spatial clustering of applications with noise

Doc2Vec : Document To Vector

DSBM : dynamic stochastic block model

Eps: epsilon

FD : fréquence des documents

GCN : graph convolutional networks

GERAD : Groupe Études et de Recherche en Analyse de Décision

HEC : Haute Étude Commerciale

HMM : hidden Markov model

IDF : fréquence inverse du document

k-NN : approche des k-voisins les plus proches

LDA : allocation latente de dirichlet

LPA : label propagation algorithm

LSA : analyse sémantique latente

Mi : revue Management International

MIT : Massachusetts Institute of Technology

MRF : Markov random field

NER: named entity recognition

NLP : traitement du langage naturel

NMF : factorisation de matrice non négative

NMI : Normalized mutual information

SBM : stochastic block model

SHRDLU : natural language understanding program created by Terry Winograd at the MIT Artificial Intelligence Laboratory between 1968 and 1969.

SNE : stochastic neighbor embedding

SVD : décomposition en valeur singulière

SVM : support vector machine

TF-IDF : fréquence du terme – fréquence inverse du document

TF : fréquence du terme

TOEFL : test of english as a foreign language

t-SNE : t-distributed stochastic neighbor embedding

VSM : vector space model

Remerciements

Je tiens d’abord à remercier ma famille pour le soutien constant qu'elle m'a apporté tout au long de cette aventure. Leur patience et leur compréhension ont été cruciales, en particulier lorsque les engagements académiques ont nécessité mon absence lors d'événements personnels importants. Je remercie également mon employeur pour la flexibilité et le soutien dont il a fait preuve en m'accordant le temps nécessaire pour les cours et le travail de mémoire.

Je dois également exprimer ma gratitude au professeur Franck Barès pour les discussions éclairantes qui ont élargi ma perspective de la revue Mi et de ses subtilités. Ces conversations ont non seulement renforcé mes connaissances, mais m'ont également permis d'apporter des idées sur l'orientation future de la revue.

Enfin, je remercie le professeur Gilles Caporossi, directeur de mémoire de m'avoir guidé et orienté tout au long de cette dernière année. Nos discussions ont toujours été productives et m'ont aidé à me concentrer sur les aspects essentiels de ma recherche.

1 Introduction

À l'ère numérique, l'internet est devenu un vaste dépôt de contenus divers, ce qui rend l'organisation et la gestion de données à grande échelle de plus en plus difficiles. Cela est particulièrement vrai pour le volume croissant de la littérature scientifique, le nombre de ces publications sont facilement accessibles via des bases de données bibliographiques telles que ResearchGate, arxiv (Cornell University) ou encore Google Scholar, ce qui présente un ensemble unique de défis pour une gestion efficace des données. Cette abondance d'informations souligne la nécessité de disposer de technologies et d'outils automatisés pour transformer habilement cette quantité de données brutes en informations exploitables [1].

Le principal défi consiste à extraire des informations significatives de données textuelles non structurées. Contrairement à la cognition humaine, qui interprète le monde à travers des mots et des récits, les systèmes informatiques fonctionnent principalement avec des chiffres et des algorithmes. Des technologies innovantes et des outils automatisés sont nécessaires pour transformer intelligemment les données en informations utilisables [2].

1.1 Contexte

Les revues comme Management international (Mi) et les centres de recherche comme le GERAD sont confrontés au défi de maintenir des normes de publication strictes tout en s'adaptant aux exigences technologiques de l'ère numérique. Cette étude aborde comment Mi gère la transition vers la communication numérique et, utilise les cahiers du GERAD comme référence comparative pour comprendre comment la revue pourrait maintenir sa mission principale et continuer à rassembler des esprits brillants issus de milieux linguistiques différents. L'abondance des données oblige Mi à gérer son contenu avec précision et à veiller à ce qu'elle atteigne le bon public, en engageant et en élargissant sa communauté de chercheurs.

La modélisation thématique et la détection de communautés sont des techniques qui pourraient être très utiles pour la revue. Dans cette étude, nous appliquerons ces outils à Mi et aux cahiers du GERAD, en établissant des comparaisons pour identifier les thèmes de recherche et les liens avec la communauté. Le résultat permettra non seulement d'identifier les stratégies de gestion de contenu de Management international, mais aussi de fournir un point de référence basé sur les pratiques du GERAD. Cela pourrait révéler de nouvelles façons de favoriser la collaboration et la discussion dans le domaine de la gestion internationale.

Mi et le GERAD sont étroitement affiliés et soutenus par HEC Montréal, mais il existe entre eux des différences notables qu'il est important de souligner. Le GERAD est un centre de recherche axé sur la mathématique et l'analyse des décisions avec des membres et leurs cahiers qui ne sont pas arbitrés, mais plutôt des traces du travail de tous les membres. En revanche, Mi opère dans le domaine de la gestion et soumet ses publications à un processus d'évaluation par les pairs.

1.2 Historique

Le GERAD, fondé en 1979, est un centre de recherche interuniversitaire qui rassemble des experts en science des données et de la décision, en informatique, en mathématiques appliquées et en ingénierie mathématique. L'objectif du centre est de développer la mathématique de la décision pour les systèmes complexes, à former du personnel hautement qualifié, à fédérer des chercheurs universitaires, des industries et des organisations autour de projets de recherche d'envergure et à contribuer au rayonnement international de la recherche québécoise et pour finir à créer un impact sociétal grâce à l'innovation scientifique, au transfert de connaissances et aux outils de décision. Le programme scientifique du GERAD se concentre sur quatre axes principaux : la valorisation des données pour la prise de décision, l'aide à la décision prise dans les systèmes complexes, l'aide à la décision en situation d'incertitude et l'aide à la décision en temps réel. La particularité du centre est de déployer ces méthodes dans des domaines clés tels que l'économie, la finance, l'énergie, l'environnement, les ressources naturelles, les infrastructures intelligentes, l'ingénierie, la logistique, le marketing et la santé.

La revue Mi, fondée en 1996, est une revue académique indépendante qui publie des articles en français, anglais et espagnol couvrant les divers aspects du management à l'échelle international. La mission de la revue est d'améliorer la compréhension globale et les pratiques dans ce domaine en présentant des recherches empiriques, des analyses théoriques et des opinions fondées émises par des chercheurs qui s'attaquent aux défis mondiaux du management. En outre, elle aide les nouveaux chercheurs dans le domaine du management à s'intégrer dans la communauté scientifique. Respectant les normes internationales les plus élevées, l'équipe éditoriale de Mi s'engage à respecter l'intégrité, la confidentialité et les processus scientifiques rigoureux, garantissant que les contributions sont à la fois significatives pour le domaine et fondées sur la littérature scientifique. Mi incarne un engagement à faire progresser les connaissances en matière de gestion internationale grâce à un riche éventail de perspectives, à stimuler le débat scientifique et à faire progresser la compréhension collective dans la discipline.

En 2020, une décision stratégique a été prise de confier la gestion de la revue à une nouvelle équipe de personnes engagées. Cette transition n'est pas seulement l'occasion d'affiner son orientation thématique, mais aussi d'intégrer les pratiques du GERAD, renforçant ainsi la cohérence et son contenu. Compte tenu de la volonté de Mi de revoir sa stratégie éditoriale, cette étude constitue un premier pas vers une exploration approfondie des diverses possibilités de son orientation future.

1.3 Énoncé du problème

La revue Management international se tourne dans un premier temps vers la modélisation thématique. Cette dernière s'avère une voie prometteuse vers la modernisation de l'accès et de la gestion des connaissances, en permettant aux chercheurs de suivre l'évolution rapide des données à l'ère numérique et aux rédacteurs en chef de prendre de la hauteur sur les éventuelles évolutions souhaitables de la ligne éditoriale en fonction des manuscrits reçus au fil des ans. En intégrant la modélisation thématique dans ses processus analytiques, Mi est en mesure d'analyser les vastes quantités d'informations numériques et d'identifier les tendances et les thèmes émergents. L'analyse comparative avec les cahiers du GERAD, fournira des pistes de directions à explorer. En second lieu, Mi souhaite explorer les méthodes de détection de communautés en s'appuyant sur

la structure communautaire des cahiers du GERAD. Cela permettra à Mi de comprendre en profondeur les interactions entre les auteurs, ce qui pourrait ultimement influencer les thèmes à publier. Cette approche comparative est essentielle pour comprendre la manière dont les communautés interagissent avec les thèmes du corpus et comment cette relation peut guider les orientations futures de la revue. Ces objectifs sont abordés dans le cadre d'une approche structurée, présentée dans cinq chapitres complets, chacun étant conçu pour s'appuyer progressivement sur les résultats du précédent.

1.4 Question de recherche

Compte tenu des objectifs, la question centrale de ce mémoire est la suivante : Quels sont les éléments qui découlent d'une analyse comparative de la modélisation thématique et des structures communautaires dans les publications de Mi par rapport à celles du GERAD, et comment ces thèmes et ces communautés sont-ils liés ? Trois objectifs de recherche découlent de cette question :

- Premièrement, il s'agit de fournir une représentation thématique d'un corpus de résumé de la revue Mi et des cahiers du GERAD afin d'obtenir des analyses comparatives.
- Deuxièmement, explorer les structures communautaires en comprenant les distinctions dues à leurs portées différentes et déterminer ce que Mi peut exploiter des stratégies d'engagement communautaire des cahiers du GERAD.
- Enfin, déterminer la relation entre le contenu thématique et les communautés au sein de Mi et des cahiers du GERAD.

Ce mémoire est structuré en cinq chapitres, chacun servant un objectif distinct. Le premier chapitre ouvre la voie en décrivant les défis numériques auxquels sont confrontées les revues académiques telles que Mi, en préparant le terrain pour une analyse comparative avec les cahiers du GERAD.
Le deuxième chapitre met en lumière les différentes techniques existantes sur l'analyse des données, la modélisation des thèmes et la détection des communautés, ce qui fournit une base théorique à l'étude. Le troisième chapitre quant à lui, décrit la conception et la méthodologie de la recherche, en mettant l'accent sur l'application comparative de l'analyse des données à Mi et au

GERAD. Les résultats et analyses sont présentés au chapitre quatre. Ce dernier met en valeur les résultats de l'analyse, discute des communautés thématiques identifiées et explore davantage leurs implications tout en abordant la collaboration des auteurs. Le dernier chapitre synthétise les résultats de la recherche, présente des conclusions répondant aux objectifs et aux questions de la recherche, et formule des recommandations principalement pour Mi, en s'appuyant sur l'analyse comparative avec le GERAD, tout en reconnaissant les limites de l'étude.

2 Revue de la littérature

La quantité croissante de données textuelles non structurées provenant de diverses plateformes représente un défi important pour l'analyse et la gestion des données. Ces données, bien que facilement compréhensibles par l'humain, posent des problèmes importants pour le traitement et l'interprétation automatisés. Il est donc nécessaire de disposer de méthodes et d'algorithmes efficaces pour traiter cette grande quantité de texte dans diverses applications. [3]. Dans cette revue de littérature, nous allons nous pencher sur les pratiques actuelles de la modélisation thématique et de la détection des communautés, dans le but d'appliquer ces techniques à notre étude. Nous allons nous attarder à comprendre comment ces méthodes peuvent être utilisées efficacement dans l'analyse de leurs données textuelles spécifiques.

2.1 Traitement automatique du langage naturel (NLP)

Le traitement du langage naturel (NLP) est devenu un outil essentiel pour gérer et interpréter la grande quantité de données textuelles présente dans de nombreux domaines. Initialement centré sur la traduction de textes entre différentes langues, le champ d'application du NLP s'est élargi, avec des applications variées, y compris en politique et en sciences sociales [4]. La science politique, par exemple, utilise l'analyse automatisée des textes pour exploiter de larges ensembles de textes politiques, une tâche complexe et volumineuse qui posait auparavant des défis importants [5]. Cependant, ces méthodes ne sont pas sans faille et reposent sur des modèles de langage qui peuvent ne pas être parfaits. Elles nécessitent donc une validation approfondie pour garantir la précision et la fiabilité de leur analyse [5].

Histoire et son évolution

Dans les années 1930, les premiers brevets pour la traduction automatique ont marqué les premiers pas du traitement du langage naturel. Les progrès dans le domaine ont pris une ampleur

considérable pendant la Seconde Guerre mondiale, principalement grâce au développement de la machine Enigma et aux travaux d'Alan Turing à Bletchley Park [4].

Ce n'est que quelques années plus tard, en 1950, que les grandes avancées ont pu être observées. Cette période a marqué la naissance du NLP par rapport à la recherche d'information textuelle, qui se concentrait sur l'indexation et la recherche de grands volumes de texte à l'aide de statistique [7]. Cependant, les gens œuvrant dans le domaine se sont rapidement trouvés confrontés à des défis. La complexité, la taille et l'ambiguïté du langage naturel ont montré que les règles élaborées étaient insuffisantes. C'est cette incompréhension qui a poussé ce domaine à s'orienter vers la sémantique, c'est-à-dire le sens du texte [7]. Cela a conduit à des efforts pour développer de nouvelles approches qui allaient au-delà de l'analyse syntaxique de base [8].

L'introduction des structures syntaxiques par Noam Chomsky en 1957 [156] a été un moment charnière pour la NLP. Elle a mis l'accent sur une théorie formalisée de la structure linguistique, bien que des critiques ultérieures aient mis en évidence ses limites dans l'utilisation pratique du langage [4].

La création de SHRDLU en 1960 par Terry Winograd au MIT a été un événement marquant. Ce programme combinait une analyse syntaxique avancée avec un système déductif, ce qui a fondamentalement changé la recherche sur l'intelligence artificielle et le NLP [8]. Il a toutefois, comme son prédécesseur, été heurté à l'ambiguïté et la complexité de la réalité. L'introduction des jetons par *Roger Schank* en 1969 a permis d'améliorer la compréhension des phrases en tenant compte des facteurs du monde réel [4]. Les réseaux de transition augmentés de *William Woods* en 1970 ont représenté une autre avancée significative, en abordant l'ambiguïté des phrases avec la récursivité et les informations partielles.

Les années 1980 ont été marquées par un changement important dans le domaine du NLP. Le domaine s'est orienté vers des approches plus simples et plus robustes, une évaluation rigoureuse et des méthodes probabilistes d'apprentissage automatique formées sur de vastes corpus de textes annotés [7]. Au fur et à mesure de son évolution, le NLP a englobé une série de tâches complexes telles que l'identification des erreurs orthographiques et grammaticales, la reconnaissance des

entités nommées et l'extraction de relations, chacune d'entre elles nécessitant des outils et des informations spécifiques [7].

L'ère actuelle du NLP est fortement influencée par l'apprentissage profond, qui résout les ambiguïtés linguistiques en utilisant de grands ensembles de données et une puissance de calcul avancée, sans règles fixes. Malgré ces avancées, l'analyse syntaxique de grammaires complexes reste un défi. Le NLP moderne comprend des composantes telles que l'analyse morphologique, syntaxique, sémantique, discursive et pragmatique, chacune abordant des aspects différents du traitement du langage, mais essentiels pour comprendre la structure et la signification du langage [4]. Ses applications pratiques vont du filtrage des pourriels, le traitement des documents numérique, la radiologie médicale, ce qui témoigne de sa polyvalence et son impact. La traduction automatique et les *chatbots* sont désormais des applications phares du NLP dans les moteurs de recherche et les sites web d'organisations, améliorant l'interaction avec l'utilisateur et le traitement de l'information [4].

2.1.1 Considérations générales

Vous êtes probablement tous familiers avec « Google, quelle est la température aujourd'hui ? ». Si ce n'est pas Google Home, peut-être qu'Alexa d'Amazon ou encore Siri d'Apple vous ont déjà guidé pour des demandes similaires. Nous parlons à ces assistants non pas dans un langage de programmation, mais dans notre langage naturel. Toutefois, les ordinateurs ne peuvent traiter les données qu'en binaire, c'est-à-dire 0 et 1. Alors, comment quantifier le texte?
Dans le processus d'analyse des documents dans le cadre du NLP, le prétraitement est une étape essentielle. Il commence par la tokenisation, c'est-à-dire la décomposition du texte en mots ou en phrases [6][3]. Vient ensuite le filtrage, qui consiste à supprimer les données non pertinentes telles que les caractères spéciaux ou les balises [6]. La lemmatisation réduit ensuite les mots à leur forme de base, ce qui garantit la cohérence de l'analyse. L'étape suivante consiste à supprimer les *stopwords*, c'est-à-dire les mots courants tels que "le", "est" et "et", qui n'ajoutent généralement pas de signification significative à l'analyse [6][3].

Une fois le prétraitement terminé, l'étape suivante consiste à définir le vocabulaire, c'est-à-dire l'ensemble des mots uniques du texte. Ce vocabulaire constitue la base du modèle de l'espace vectoriel, une méthode courante pour représenter le texte sous forme de vecteurs numériques. Ce modèle simplifie l'analyse du texte en se concentrant sur la fréquence des termes, sans tenir compte de l'ordre des mots [3].

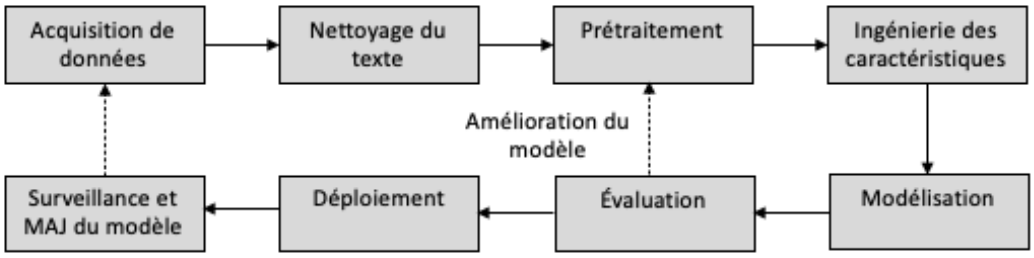


Figure 2.1 - Montre les principales composantes d'un pipeline générique pour le développement de systèmes de NLP modernes axés sur les données.

Tout au long de ces étapes, illustrées dans la figure 2.1, l'essentiel est de transformer les données textuelles dans un format qui peut être facilement analysé tout en préservant leur sens et leur structure essentiels. Il existe bien évidemment de multiple technique et méthodes pour traiter efficacement ces données textuelles non structurées. L'objectif n'est pas de les passer tous en revue, mais nous verrons quelques méthodes traditionnelles qui sont couramment utilisées.

Vectorisation

La vectorisation dans le traitement du langage naturel (NLP) est une technique fondamentale pour convertir les données textuelles en forme numérique, permettant aux modèles informatiques de traiter et d'analyser le langage [10][20]. L'objectif de la vectorisation est de transformer les informations textuelles en un modèle d'espace vectoriel (VSM), où les documents sont représentés sous forme de vecteurs dans un espace multidimensionnel, ce qui facilite les opérations telles que la mesure de la similarité et la classification [3]. La distance cosinus est largement utilisée pour mesurer la similarité entre les documents, en se concentrant sur la direction plutôt que sur la taille

des vecteurs. La distance euclidienne quantifie la dissimilarité directe entre deux documents, en mesurant la distance linéaire entre leurs représentations vectorielles [10]. La similarité de Jaccard quant à elle est définie comme la taille de l'intersection divisée par la taille de l'union de deux ensembles. Cette mesure permet de résoudre le problème de la similarité à travers l'ensemble ; lorsque le texte est relativement long, la similarité sera plus faible [146].

La méthode *Term Frequency-Inverse Document Frequency* (TF-IDF), introduite par *Salton et al.* en 1975 [157], est largement utilisée dans la vectorisation. Elle quantifie l'importance d'un mot dans un document par rapport à une collection, en utilisant la formule

$$W_{t,j} = tf_{t,d} * \log\left(\frac{N}{df_t}\right)$$

où $tf_{t,d}$ est la fréquence des termes dans le document t, N est le nombre total de documents, et df_t est le nombre de documents contenant le terme t [21].

Cette méthode met l'accent sur les mots qui sont fréquents dans un document, mais rare dans l'ensemble du corpus de documents, ce qui permet de saisir l'importance des mots dans le contexte du document [3][22].

Classification

La classification des textes fait l'objet de recherches approfondies dans des domaines tels que l'exploration de données, la gestion de bases de données, l'apprentissage automatique et la recherche d'informations, et est utilisée dans divers domaines tels que le traitement d'images et le diagnostic médical [3]. La classification des textes est une méthode qui attribue des classes prédéfinies aux documents textuels [9]. Quelle que soit la méthode spécifique employée, une tâche de classification par exploration de données commence par un ensemble d'apprentissages $D = (d1, \ldots, dn)$ de documents qui sont déjà étiquetés avec une classe $L \in \mathbb{L}$ (par exemple, gestion, ressources humaines). La tâche consiste ensuite à déterminer un modèle de classification

$$f : D \rightarrow \mathbb{L} \qquad f(d) = L$$

qui est capable d'attribuer la bonne classe à un nouveau document d du domaine [10].

Les performances du modèle de classification sont évaluées en définissant une fraction aléatoire de documents étiquetés comme ensemble de tests, en entraînant le classificateur, en classant l'ensemble de tests et en comparant les étiquettes estimées avec les vraies étiquettes, ce qui permet de mesurer les performances [3]. La proportion de documents correctement classés par rapport au nombre total de documents est appelée précision [10].

Classificateurs bayésiens

Les classificateurs bayésiens simples ont gagné en popularité ces derniers temps et se sont révélés étonnamment performants *(Friedman [158] ; Friedman et al. [159]; Sahami [160]; Langley et al. [161])* [11]. Le modèle utilise une approche probabiliste pour prédire la distribution des documents dans chaque classe, en supposant que les différentes distributions de termes sont indépendantes [3]. Deux modèles de classification de Bayes naïfs sont largement utilisés, chacun visant à déterminer la probabilité a posteriori d'une classe sur la base de la distribution des mots dans un document, l'un d'entre eux prenant en compte la fréquence des mots [3][11].

- Le modèle multivarié de Bernoulli est un système de représentation de documents qui utilise des caractéristiques binaires pour indiquer la présence ou l'absence de mots, sans tenir compte de leur fréquence [3][11][12].
- Le modèle multinomial de classification des textes prend en compte la fréquence des mots dans les documents, en les traitant comme une séquence ordonnée de mots issus d'un vocabulaire. Il suppose que la probabilité de chaque mot est indépendante de son contexte et de sa position, ce qui donne une représentation de type « sac de mots ». La probabilité d'un document en fonction de sa classe est calculée à l'aide de la distribution multinomiale, en tenant compte de l'occurrence de chaque mot. Ce modèle est particulièrement efficace dans les cas où les informations de fréquence sont utiles à la classification, par exemple pour différencier les articles de presse [11].

Les machines à vecteurs de support

Les machines à vecteurs de support (SVM) pour la classification des textes constituent une méthode puissante pour le traitement des données de haute dimension [10][13][14]. Elles se concentrent sur la recherche d'un hyperplan qui maximise la marge entre les différentes classes. La clé de l'efficacité des SVM réside dans l'utilisation de fonctions à noyau, qui permettent des séparations plus complexes que les limites linéaires [15]. La robustesse de la technique face à des caractéristiques de pertinence variable met en évidence son applicabilité dans divers scénarios de classification de textes [16]. La capacité des SVM à traiter efficacement des données textuelles éparses, à gérer le surajustement et à fournir une catégorisation précise même avec des données d'apprentissage limitées souligne leur importance dans les applications NLP. Les SVM sont efficaces dans le traitement des données textuelles en raison de leur capacité à traiter de grands ensembles de caractéristiques, ce qui les rend idéaux pour les tâches NLP avec des données peu nombreuses et de grande dimension, et pour fournir une classification fiable avec des données d'apprentissage limitées [14].

Les arbres de décisions

Les arbres de décision fournissent un modèle interprétable pour la classification des textes, en décomposant les décisions sur la base de caractéristiques [3][10]. Les arbres de décision dans la classification des textes sont reconnus pour leur simplicité et leur approche directe. Cet accent mis sur la simplicité et la clarté en fait un choix populaire pour les tâches nécessitant un processus de classification facilement compréhensible [16]. L'article [17] démontre l'application plus large des arbres de décision, en mettant en évidence leur utilité dans divers domaines tels que la classification de l'occupation du sol, soulignant ainsi leur adaptabilité et leur efficacité dans une série de tâches de classification.

k-NN

L'approche des k-voisins les plus proches (k-NN) pour la classification des textes fonctionne en identifiant les points de données les plus proches dans un espace multidimensionnel, une stratégie efficace pour catégoriser les textes sur la base de la ressemblance des caractéristiques [3].

L'adaptabilité de k-NN est encore affinée lorsque des poids sont introduits dans les attributs, ce qui améliore la différenciation entre les catégories de texte [18]. Au-delà du texte, l'applicabilité du k-NN dans la classification des données médicales démontre sa grande utilité dans toutes les disciplines, y compris son efficacité dans la classification de modèles complexes dans les ensembles de données cornéennes [19]. *Yang et Liu* [162] place k-NN parmi d'autres méthodes robustes, confirmant sa validité et son efficacité dans la classification des textes [16].

2.1.2 Analyse de regroupement

Le regroupement (aussi appelé clustering) est une méthode d'exploration de données largement utilisée qui a fait l'objet de recherches approfondies dans le contexte des données textuelles. Il englobe des techniques telles que le clustering hiérarchique, le clustering k-means, le clustering probabiliste et le clustering basé sur la densité comme DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Ces méthodes sont essentielles pour des tâches telles que la classification de textes, la recherche d'informations et la compréhension de la similarité des documents, car elles permettent de mieux comprendre la structure et les thèmes des grands ensembles de données [3][34][35][36][37].

Le regroupement hiérarchique est un élément essentiel de l'analyse des données, défini par des méthodes descendantes (aussi appelé divise) et ascendantes (aussi appelé agglomérative) d'organisation des données sous forme de dendrogrammes [3]. Il prend en charge l'analyse de la structure des données à plusieurs niveaux sans nécessiter un nombre prédéfini de grappes, ce qui est précieux dans divers domaines pour révéler les relations entre les données. Cette technique est adaptable et s'applique à des domaines différents, ce qui est essentiel pour extraire des informations de données complexes [38].

La méthode de regroupement K-means est une approche de partitionnement qui organise les données en un nombre prédéfini de grappes, visant à minimiser la variance à l'intérieur de chaque grappe tout en maximisant la variance entre les grappes [3].

ALGORITHM 1: *k*-means clustering algorithm

Input : Document set $\mathcal{D}$, similarity measure $\mathcal{S}$, number k of cluster

Output: Set of k clusters

initialization

Select randomly k data points as starting centroids.

while *not converged* **do**

  Assign documents to the centroids based on the closest similarity.

  Calculate the the cluster centroids for all the clusters.

end

return *k clusters*

Figure 2.2 - Algorithme de regroupement k-means [3]

L'algorithme, comme illustré à la figure 2.2 commence par sélectionner aléatoirement K centroïdes initiaux et itère ensuite sur deux étapes principales : l'affectation de chaque point de données au centroïde le plus proche sur la base de la métrique de la distance, et le recalcul des centroïdes comme la moyenne des points affectés. Ce processus itératif se poursuit jusqu'à ce que les centroïdes se stabilisent, ce qui indique que les grappes sont aussi distinctes et aussi similaires que possible [37][39][40]. Le choix de K est primordial et peut affecter de manière significative le résultat [3]. Les performances de l'algorithme sont sensibles à la sélection initiale des centroïdes.

La modélisation thématique est un algorithme de regroupement probabiliste populaire qui a fait l'objet d'une attention particulière ces derniers temps [3]. C'est un modèle probabiliste génératif conçu pour découvrir la structure thématique sous-jacente d'un corpus de documents en associant les documents à des thèmes [3][41][42][43]. Il part du principe que les documents sont des mélanges de thèmes, où un thème est caractérisé par une distribution de mots. Cette approche permet d'explorer la manière dont les documents sont liés les uns aux autres par des thèmes partagés, ce qui facilite une compréhension plus approfondie du contenu du corpus sans avoir à catégoriser manuellement chaque document [41]. Il existe plusieurs modèles, mais les plus populaires sont l'analyse sémantique latente (LSA) [46], l'allocation latente de Dirichlet (LDA) [42] et la matrice non négative de factorisation (NMF) [44]. Nous aurons l'occasion d'explorer chacun de ces modèles dans la section 3 de cette étude.

La méthode DBSCAN est un algorithme largement étudié qui construit des grappes à partir d'objets centraux avec des objets connectés en densité, nécessitant au moins MinPts (le plus petit nombre d'objets dans epsilon) voisin dans un rayon Eps (rayon de voisinage fixe) [144]. Il s'est avéré efficace dans la détection de formes et de tailles différentes, dans la gestion du bruit et ne nécessite pas que l'utilisateur connaisse à l'avance le nombre de grappes. Ses performances globales dépendent de deux paramètres : Eps et MinPts [145]. Les auteurs [144] ont développé une heuristique pour déterminer les paramètres Eps et MinPts de la grappe la plus « fine » de la base de données en se basant sur l'observation que le d-voisinage d'un point contient exactement $k + l$ points pour presque tous les points p.

Technique d'extraction

Les textes en langage naturel contiennent de nombreuses informations qui ne sont pas facilement accessibles par l'analyse informatique. Toutefois, les ordinateurs peuvent analyser efficacement de vastes volumes de textes, en identifiant des informations à partir de mots, de phrases ou de passages individuels [3]. L'exploration de texte utilise des algorithmes spécialisés pour traiter ces grands volumes de textes non structurés et les transformer en données structurées et exploitables [10]. L'extraction d'informations comporte deux tâches essentielles : la reconnaissance des entités nommées (NER) et l'extraction de relations, qui font toutes deux appels à des méthodes d'apprentissage statistique.

D'abord, la reconnaissance des entités nommées consiste à extraire les mots ou de l'information numérique du texte qui correspondent à des catégories prédéfinies telles que les personnes, les organisations, les dates, les lieux…[3][25][26]. Les origines de la NER remontent aux efforts déployés dans les années 1990 pour améliorer les systèmes d'extraction d'informations, soulignant la nécessité d'une reconnaissance précise des noms dans les textes [26]. La NER ne peut pas être entièrement réalisée en comparant des chaînes de caractères à un dictionnaire en raison de son caractère incomplet et des formes variables des entités nommées, qui dépendent souvent du contexte. Elle utilise plutôt des méthodes d'apprentissage statistique pour une identification précise [3].

En outre, la NER a progressé grâce à des techniques d'apprentissage automatique telles que les machines à vecteurs de support (SVM), qui sont plus performantes que les modèles traditionnels pour la classification des entités dans les documents [25]. Le modèle de Markov caché (HMM) est quant à lui, un modèle statistique qui permet d'identifier et de classer les éléments de la séquence de données. Il est particulièrement efficace en raison de sa capacité à prendre en compte le comportement dynamique temporel, qui est importante pour comprendre les données séquentielles telles que le texte [27]. Une autre technique est celle des champs aléatoires conditionnels (CRF) qui améliore la segmentation et l'étiquetage des données séquentielles en incorporant des caractéristiques interdépendantes complexes sans simplifier les hypothèses sur la distribution des données. Cette approche permet une compréhension et un traitement plus nuancés des données séquentielles [28].

L'extraction des relations, la deuxième tâche essentielle, est vitale pour comprendre les relations sémantiques dans le texte et améliorer les applications du NLP. On pense par exemple à l'extraction d'informations biomédicales ou encore la réponse aux questions. Les techniques pour l'extraction des relations incluent des systèmes supervisés, des connaissances de base et des ontologies hiérarchiques pour améliorer la précision. Des sources de connaissances externes telles que Wikipédia et les informations de coréférence sont utilisées pour affiner les processus d'extraction de relations, ce qui permet d'obtenir des modèles de NLP mieux informés et plus sensibles au contexte [29][30][31][32][33].

Application

Le NLP est largement utilisé dans différents domaines [4]. Cette section traite des applications réussies, en soulignant les caractéristiques spécifiques de chaque application lors de la sélection des méthodes appropriées. Tout d'abord, les agents conversationnels sont essentiels dans le domaine du traitement automatique du langage. Ils sont utilisés principalement dans les *chatbots* et les systèmes de dialogue. Ces derniers imitent la conversation humaine, soit par un apprentissage basé sur des règles, soit par un apprentissage basé sur un corpus [21]. Un autre exemple serait Twitter qui nécessite l'utilisation du NLP pour l'analyse des sentiments qui permet de classer les tweets en catégories positives, négatives et neutres pour une collecte de données efficace [4][24].

Une application un peu plus connue est celle du filtrage du pourriel qui a révolutionné le NLP en utilisant des algorithmes d'apprentissage automatique tels que l'arbre de décision, la régression logistique, la forêt aléatoire, *Bernoulli's Naive Bayes*, les SVM linéaires et gaussiens. Ces algorithmes visent à différencier les données réelles de celles qui sont fausses dans la relève du courriel [4]. Dans le domaine de la médecine, la reconnaissance des bio-entités est une méthode de biologie moléculaire qui identifie et classifie les termes techniques liés à des concepts biologiques, tels que les protéines, les gènes et leurs lieux d'activité. Cette technique est de plus en plus importante en raison des méthodes expérimentales à haut débit et peut être utilisée dans des tâches d'accès à l'information de plus haut niveau telles que l'extraction de relations et la réponse à des questions [10].

L'avenir du NLP dépend probablement du développement d'algorithmes qui traitent le langage aussi efficacement que le font les systèmes de recherche d'informations. Les réalisations d'IBM Watson illustrent le potentiel du NLP, mais elles soulignent également la nécessité de surmonter certains obstacles [23].

2.1.3 Modélisation des thèmes

Qu'est-ce qu'un thème? Selon le Larousse, un thème peut être interprété de 6 façons différentes, mais celle qui nous intéresse est la première soit : sujet, idée sur lesquels portent une réflexion, un discours, une œuvre, autour desquels s'organise une action. Pour interpréter un thème, *Duriau et al.* [163] parle d'analyse de contenu. Il se concentre sur les méthodes qui sont principalement utilisées pour la catégorisation et l'analyse de données textuelles [45]. L'analyse de contenu est présentée comme une technique systématique et reproductible qui simplifie le processus de catégorisation d'importants volumes de texte en un nombre réduit de catégories de contenu plus faciles à gérer. Une des méthodes que nous allons aborder dans cette section est la modélisation thématique. Cette méthode est mise de l'avant pour sa précision, fondée sur des règles de codage explicites, qui permet d'identifier des modèles, des tendances et des thèmes dans les données textuelles.

La modélisation thématique est une méthodologie utilisée pour identifier efficacement les concepts cachés, les caractéristiques principales ou les variables latentes dans de grandes quantités de données générées par les progrès de l'informatique [47]. Il existe 2 modèles de classification. Le modèle probabiliste et le modèle non probabiliste (modèle algébrique). Les approches non probabilistes, telles que les approches algébriques de factorisation matricielle, sont apparues au début des années 1990 avec les concepts d'analyse sémantique latente [46] et de factorisation matricielle non négative [44]. Les modèles probabilistes ont été développés pour améliorer le modèle algébrique en ajoutant un sens de probabilité à l'aide d'approches de modèles génératifs tels que LDA [42] et PSLA [43].

Maintenant comment peut-on faire pour analyser ou plutôt quantifier ces données textuelles? Les modèles thématiques modélisent trois entités : les constructions, les collections et les thèmes. Les constructions forment des collections de mots, tandis que les thèmes décrivent des significations sémantiques. Un thème est une distribution de probabilité sur les constructions, et la plupart des modèles observent la cooccurrence des constructions dans les collections. Les mots sont caractérisés par leur contexte, ce qui les rend utiles pour l'analyse des données textuelles [48]. Une technique couramment utilisée pour quantifier ces données est celle du sac de mots, dans laquelle la sémantique et la signification des phrases ne sont pas évaluées. La méthode évalue plutôt la fréquence des mots. On part donc du principe que les mots les plus fréquents dans un thème présentent un intérêt pour ce thème [49].

Il existe plusieurs types d’algorithmes de modélisation thématique. Les plus populaires qui ont contribué à toutes les sphères de l'analyse de texte dans de multiples domaines comprennent l'analyse sémantique latente [46], la factorisation matricielle non négative [44], l'analyse sémantique latente dite probabiliste [43] et l'allocation de Dirichlet latente [42]. Nous aurons la chance d’explorer davantage ces techniques dans la prochaine section.

Il est important de souligner l’importance des textes courts et la façon dont les divers algorithmes les traitent. Les textes courts sont des documents qui ne contiennent que très peu de mots. De ce fait, il est plus rare d’observer deux mots qui sont corrélés à un même thème. Ça complexifie la détection des thèmes dans un corpus. Toutefois, l'intégration des techniques de modélisation thématique pour l'analyse des textes courts a connu des avancées significatives. Une nouvelle

approche de modélisation thématique a été proposée en 2013. Biterm (BTM) modélise directement les cooccurrences de paires de mots dans l'ensemble du corpus afin de remédier à la rareté des données dans les documents courts, démontrant une cohérence et une pertinence thématiques améliorées par rapport aux modèles traditionnels [53]. L'étude [51] met en évidence l'application de méthodes courantes telles que LDA et NMF à des textes courts, en montrant leur efficacité à découvrir des thèmes significatifs à partir de données de médias sociaux. Certains chercheurs n'étaient pas du même avis et en 2023, ils soulignent l'écart entre le développement de modèles thématiques et leur application pratique dans l'analyse des médias sociaux, suggérant un décalage avec les besoins des utilisateurs et appelant à des améliorations méthodologiques [52].

LDA

Latent Dirichlet Allocation (LDA) est un modèle probabiliste génératif largement reconnu introduit par *Blei et Jordan* [42], utilisé pour découvrir des informations thématiques latentes dans une grande collection de données discrètes, telles que les corpus de textes [60]. Ce modèle utilise l'approche connue du sac de mots (BoW), qui permet de traiter chaque document comme un vecteur de fréquences de mots. Il réduit efficacement la dimensionnalité du modèle BoW en représentant un document comme une combinaison de thèmes [60]. LDA suppose que les documents sont des mélanges de thèmes, où un thème est une distribution de mots. Cette méthode a été développée pour surmonter les limites des modèles de thèmes latents précédents [43] en fournissant un moyen d'approximer efficacement l'inférence et l'estimation des paramètres [42]. Des études récentes ont appliqué la méthode LDA à diverses analyses de contenu, telles que les pétitions électroniques, démontrant son efficacité dans l'identification de thèmes latents dans de grands volumes de texte et ses avantages par rapport aux méthodes manuelles d'analyse de contenu [66]. Malgré son application généralisée, des défis subsistent, notamment la détermination du nombre optimal de thèmes et la garantie de l'interopérabilité des résultats du modèle. Le nombre optimal de thèmes (K), selon *Quinn et al.* [164] doit être suffisamment grand pour générer des thèmes interprétables qui n'ont pas été trop agrégés et suffisamment petits pour être utilisable. De nouvelles techniques ont vu le jour afin de déterminer ce nombre [67] ainsi que l'interopérabilité des thèmes [68][69]. Une récente étude met en lumière le prétraitement des données textuelles, en examinant les effets de la suppression des mots peu fréquents sur la qualité des thèmes découverte

à travers la méthode LDA [70]. Une technique que nous mettrons de l'avant lors du chapitre de la méthodologie.

Avant d'aller plus loin, nous allons aborder quelques notations et terminologies qui nous aideront à comprendre la partie mathématique qui suit. Un *mot* est caractérisé comme l'unité de base des données discrètes. Il est défini comme un élément d'un vocabulaire indexé par $\{1, \ldots, V\}$. Les mots sont représentés à l'aide de vecteurs dont une seule composante est égale à un et toutes les autres composantes sont égales à zéro. Un *document* est une séquence de N mots désignés par $w = (w_1, w_2, \ldots, w_N)$, où w_n est le $n$ième mot de la séquence. Un corpus est une collection de M documents notés $D = \{w_1, w_2, \ldots w_M\}$. L'idée de base derrière le modèle LDA est que les documents sont représentés comme des mélanges aléatoires de thèmes latents, ou chaque thème est caractérisé par une distribution de mot [42]. Le modèle LDA suppose le processus génératif suivant pour chaque document w dans un corpus D :

1. Choisir $N \sim$ Poisson (ξ).
2. Choisir $\theta \sim$ Dir(α).
3. Pour chacun des N mots w_n :
   (a) Choisir un thème $z_n \sim$ Multinomial (θ).
   (b) Choisir un mot w_n à partir de $p(w_n | z_n, \beta)$, une probabilité multinomiale conditionnée par le thème z_n.

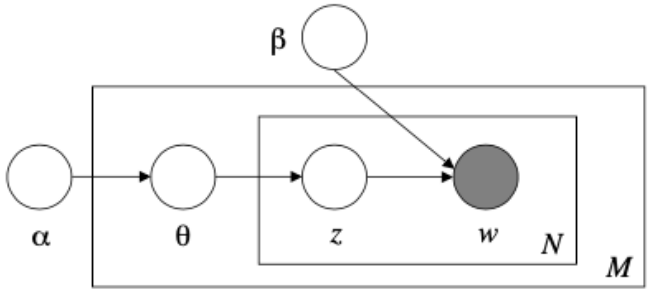


Figure 2.3 - Représentation graphique du modèle LDA [42]

Graphiquement, comme illustrée à la figure 2.3, la représentation LDA est composée de trois niveaux.

Les paramètres α et β sont les paramètres du corpus, qui devraient être échantillonné une fois dans le processus de génération d'un corpus. Les variables θ_d sont les variables du document, échantillonnées une fois par document. Les variables z_{dn} et w_{dn} sont les variables du mot et sont échantillonnées une fois pour chaque mot dans chaque document [42]. Mathématiquement, en prenant le produit des probabilités marginales des documents individuels, nous obtenons la probabilité d'un corpus suivant :

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(Z_n|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d.$$

Une des mesures d'interopérabilité des thèmes est d'évaluer la cohérence sémantique, qui est une mesure sommaire permettant de saisir la tendance des mots à forte probabilité d'un thème à cooccurrence dans le même document.

Les hyperparamètres α et β sont très spécifique au modèle LDA. L'hyperparamètre Alpha est un paramètre qui détermine les propriétés d'une distribution de Dirichlet, la distribution de probabilité préalable de la distribution thème-document. Il détermine la fréquence attendue des thèmes dans le corpus et la confiance en leur présence. Alpha représente l'hypothèse sur la distribution des données dans les documents avant l'inférence du modèle. Une valeur alpha plus élevée dans un document indique une plus grande probabilité qu'un thème apparaisse dans ce document, ce qui affecte la qualité du modèle [152]. L'hyperparamètre Bêta affecte la distribution des mots entre les thèmes. Une valeur plus faible signifie que chaque thème est susceptible d'être composé d'un ensemble plus restreint de mots et est donc plus spécifique. À l'inverse, une valeur plus élevée permet aux thèmes d'inclure un plus large éventail de mots, ce qui les rend moins distincts les uns des autres. Le choix des valeurs α et β implique généralement une combinaison de connaissances d'expert sur l'ensemble de données et de tests empiriques. Il est souvent utile d'utiliser des mesures telles que les scores de cohérence pour évaluer dans quelle mesure les sujets peuvent être interprétés par l'être humain.

Même si la méthode LDA existe depuis un certain temps, elle reste l'algorithme le plus utilisé et le plus populaire pour explorer des thèmes dans les articles de recherche [71]. Les recherches futures

se concentreront sur l'amélioration des méthodes d'évaluation des modèles, sur l'amélioration de l'interopérabilité et sur le développement de meilleurs outils de visualisation. De même que le développement de modèles qui gèrent mieux les changements de thèmes dans le temps et qui utilisent plus efficacement les métadonnées, dans le but d'accroître son applicabilité à travers divers ensembles de données et domaines [72].

NMF

La matrice non négative de factorisation est une technique d'analyse des données mise au point par *Lee et Seung* [61] pour décomposer les vecteurs à haute dimension en une représentation réduite. Elle repose sur le principe selon lequel une matrice de données peut être factorisée en deux matrices dont les entrées ne sont pas négatives. L'algorithme de la NMF met à jour ces matrices de manière itérative afin de minimiser la différence entre la matrice de données d'origine et le produit de ses composantes factorisées. Ce processus est essentiel pour trouver des modèles et des caractéristiques dans les ensembles de données dans divers domaines [44]. Visuellement, la NMF peut-être représentée de la manière suivante :

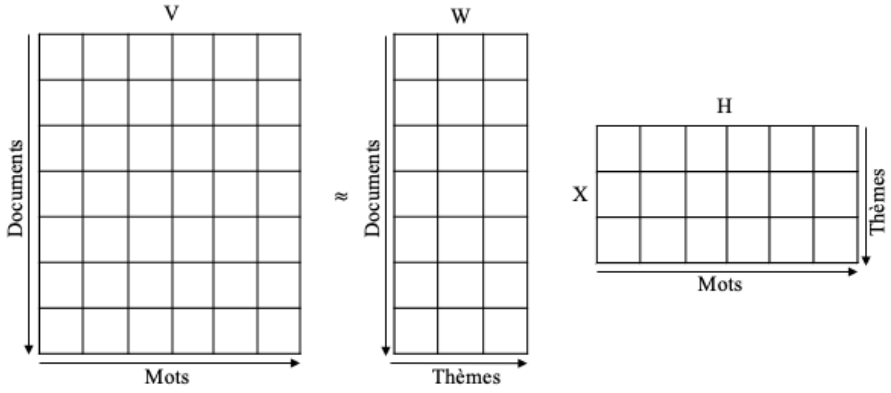


Figure 2.4 - Factorisation schématique des matrices non négatives ; la matrice originale V est décomposée en W et H.

Étant donné un ensemble de vecteurs de données multivariées à n dimensions, les vecteurs sont placés dans les colonnes n x m d'une matrice V où m est le nombre de mots dans l'ensemble de données. Cette matrice est ensuite approximativement factorisée en une n x r de matrice W et une r x m de matrice H. Habituellement, r est choisi pour être plus petit que n ou m, de sorte que W

et H sont plus petits que la matrice originale V. Il en résulte une version comprimée de la matrice de données originale [44].

Le développement de la NMF a donné lieu à diverses stratégies visant à améliorer ses performances, notamment une meilleure sélection des valeurs initiales, la définition de critères de convergence et la gestion de la rareté des données. Les méthodes itératives, en particulier celles qui utilisent des règles de mise à jour multiplicatives, sont remarquables pour l'amélioration incrémentale de l'approximation des données. Ces méthodes améliorent la représentation des données en identifiant et en soulignant efficacement les modèles sous-jacents [61]. Dans le traitement d'images la NMF, aide à identifier des parties d'objets, facilitant ainsi des tâches telles que la reconnaissance faciale et l'extraction de caractéristiques [62].

LSA

L'analyse sémantique latente (LSA) développée par *Landauer et Dumais* [165], est une technique d'extraction et de représentation de la signification de l'usage contextuel des mots par des calculs statistiques appliqués à un vaste corpus de texte. Cette méthode repose sur l'algèbre linéaire, en particulier la décomposition en valeurs singulières (SVD), qui décompose une matrice en trois autres matrices, révélant la structure sous-jacente en termes de relations entre les documents et les mots du corpus [64]. L'analyse factorielle simplifie les relations complexes entre les variables en les représentant par un nombre réduit de facteurs abstraits. Ces facteurs, indépendants les uns des autres, peuvent être combinés pour régénérer l'ensemble des données d'origine [65]. Ce processus implique la construction d'une matrice dont les lignes représentent les mots uniques, les colonnes les documents et les entrées les fréquences des mots, qui est ensuite transformée et réduite en dimensionnalité par SVD. La figure 2.5 permet d'offrir une représentation visuelle du schéma de la décomposition en valeurs singulières (SVD) d'une matrice (X) rectangulaire de mot (w) par contexte (c). La matrice originale est décomposée en trois matrices : W et C, qui sont orthonormées, et S, une matrice diagonale. Les m colonnes de W et les m lignes de C' sont linéairement indépendantes [63][65].

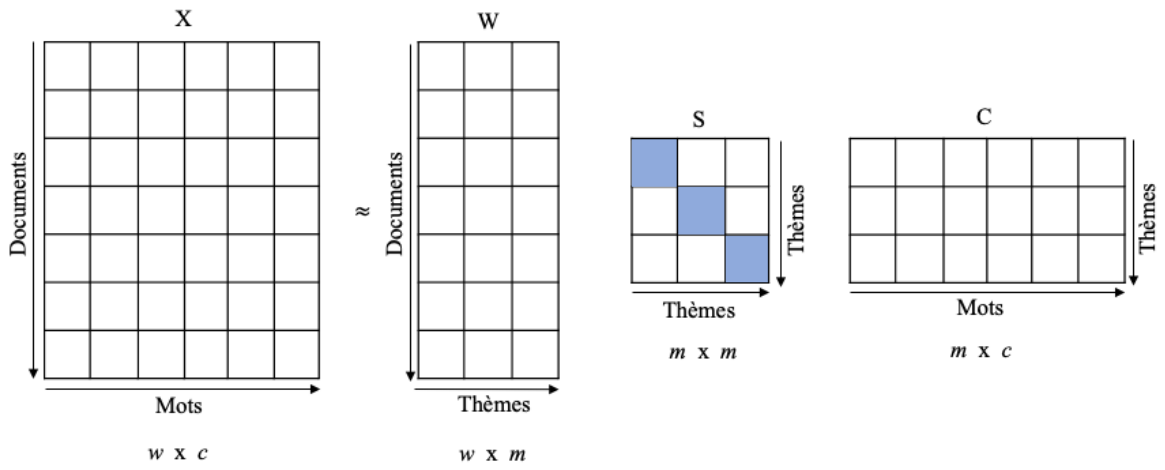


Figure 2.5 - Schéma de la décomposition en valeurs singulières

L'évaluation du modèle LSA de *Landauer et Dumais* [165] a été testée à l'aide d'items synonymes du TOEFL à travers quatre questions qui ont atteint une précision comparable. D'abord si un modèle linéaire simple peut acquérir une connaissance du sens des mots semblable à celle des humains à partir de grands volumes de textes. En deuxième lieu, l'impact de la dimensionnalité sur son succès. Ensuite, son taux d'acquisition de connaissances par rapport aux humains lisant le même texte. Pour finir, l'étendue de la connaissance dérivée des inférences indirectes par rapport à la contiguïté contextuelle directe [64].

Les nouvelles tendances consistent à améliorer leur capacité à traiter la polysémie (mots à sens multiples) et à améliorer leur évolutivité et leur efficacité avec des ensembles de données plus importants [43][46]. L'intégration de l'analyse sémantique avec d'autres techniques d'apprentissage automatique et de traitement du langage naturel est aussi nécessaire afin d'améliorer sa robustesse et son applicabilité. Des avancées telles que les modèles d'apprentissage profond et les modèles probabilistes de thèmes étendent les capacités du modèle [43][63].

Visualisation

Comme nous l'avons défini un peu plus tôt, la modélisation thématique est une technique populaire pour l'analyse de grands corpus de textes. Il est peu probable qu'un utilisateur dispose du temps nécessaire pour comprendre et exploiter les résultats bruts de la modélisation thématique pour l'analyse d'un corpus. Par conséquent, une visualisation intéressante et intuitive est nécessaire pour qu'un modèle thématique apporte une valeur ajoutée [54]. Plusieurs méthodes ont été développées afin de nous permettre de comprendre, d'explorer et de naviguer dans une collection de documents. Turbo topic, par exemple, introduit par *Blei et Lafferty* [56], est une approche qui intègre les n-grammes aux modèles traditionnels d'unigramme. Il fournit une compréhension intuitive des thèmes en visualisant les distributions sur les termes et en identifiant les expressions multimots significatives [56]. Termite, une autre technique, introduit par *Chuang et al.* [55] est un outil d'analyse visuelle qui aide à interpréter des modèles thématiques complexes dans des données textuelles. Il utilise des tableaux pour comparer les termes à l'intérieur et entre les thèmes, améliorant l'évaluation de la qualité du modèle grâce à des mesures de saillance et à des algorithmes de sériation [55]. Quoique ces méthodes nous permettent d'atteindre notre objectif de visualiser des données digestes, nous allons nous attarder sur les représentations suivantes : textflow, LDAvis ainsi que t-SNE.

Textflow

L'évolution des thèmes dans les grandes collections de textes est importante pour de nombreux individus, mais il n'est pas facile d'analyser comment et pourquoi les thèmes évoluent au fil du temps. De nombreuses méthodes ont été développées pour analyser l'évolution des thèmes, mais peu de travaux se sont concentrés sur l'étude des modèles de fusion et de division des thèmes [59]. La stratégie la plus répandue pour visualiser l'évolution des thèmes dans le temps est basée sur des graphiques empilés, mais elle ne présente pas l'évolution détaillée du contenu aux utilisateurs pour faciliter leur compréhension de l'information et leur prise de décision. TextFlow est un outil d'analyse visuelle interactive qui aide les utilisateurs à analyser comment et pourquoi les thèmes corrélés changent au fil du temps. Il intègre la visualisation interactive et les techniques de modélisation des thèmes pour aider les utilisateurs à découvrir les modèles d'évolution à différents

niveaux de détail. La méthode extrait des thèmes d'une collection de textes et modélise les relations d'évolution entre eux, telles que les événements critiques et les corrélations entre les mots clés. La visualisation, comme illustrer à la figure 2.6 aide les utilisateurs à comprendre les relations de fusion et de séparation entre des thèmes en évolution en utilisant une visualisation basée sur l'écoulement d'une rivière et de multiples indices visuels. Le graphique des flux thématiques représente l'évolution des thèmes dans le temps et peut être divisé en plusieurs branches ou fusionné avec plusieurs autres branches en un seul flux lorsque les thèmes correspondants fusionnent en un seul thème [59].

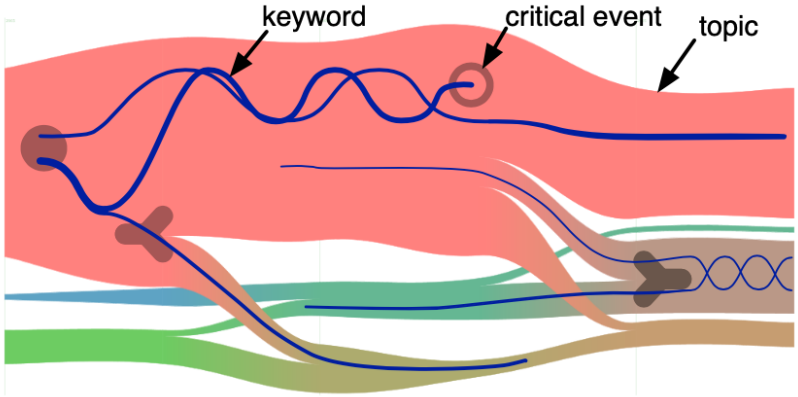


Figure 2.6 - La représentation visuelle de TextFlow, qui consiste en quatre flux thématiques, quatre événements critiques et cinq fils de mots clés [59].

LDAvis

LDAvis est un système de visualisation interactif qui tente de répondre à quelques questions de base sur un modèle thématique ajusté. Visuellement, tel qu'illustré à la figure 2.6, il est représenté avec une vue globale du modèle thématique sous forme de cercles dans un plan bidimensionnel et codifie la prévalence globale de chaque thème à l'aide des zones des cercles. Un diagramme à barres horizontal montre les mots les plus utiles pour interpréter le thème sélectionnés, et une paire de barres superposées montre la fréquence spécifique d'un mot donné. Les panneaux gauche et droit sont liés de telle sorte que la sélection d'un thème révèle les termes les plus utiles pour l'interprétation de ce thème, et la sélection d'un terme révèle la distribution conditionnelle des thèmes pour ce terme. Ce modèle propose une mesure de la pertinence des mots d'un thème qui permet aux utilisateurs de classer les mots par ordre d'utilité pour l'interprétation des thèmes [57].

Quoiqu'un certain nombre de systèmes de visualisation pour la modélisation thématique ont été développés ces dernières années, cette visualisation plus compacte offre une compréhension rapide et facile des thèmes individuels. Les travaux futurs porteront sur l'interprétation des thèmes dans les modèles LDA ajustés, y compris une comparaison de plusieurs méthodes, la visualisation des corrélations entre les thèmes et une solution au problème de la visualisation d'un grand nombre de thèmes d'une manière compacte [57].

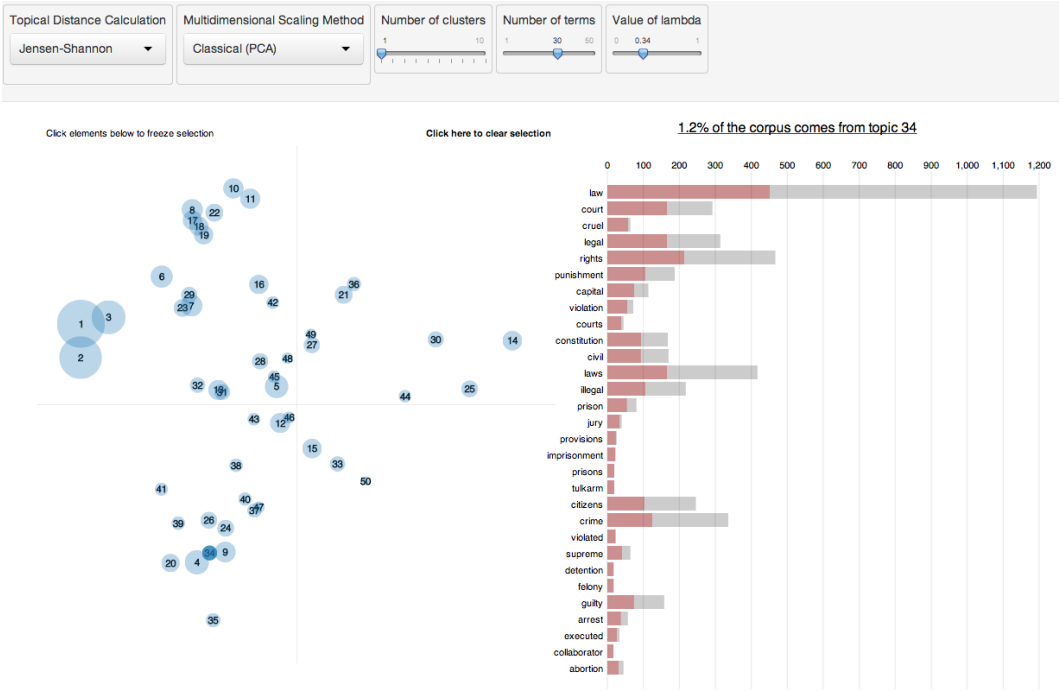


Figure 2.7 - La représentation visuelle de LDAvis [57]

T-SNE

La visualisation de données à haute dimension est un problème important dans de nombreux domaines différents, et diverses techniques ont été proposées, notamment des affichages iconographiques [166], des techniques basées sur les pixels [167] et des techniques qui représentent les dimensions des données comme les sommets d'un graphe [168]. Les méthodes de réduction de la dimensionnalité convertissent les ensembles de données à haute dimension en ensembles de données à deux ou trois dimensions qui peuvent être affichés dans un nuage de points, comme illustré à la figure 2.7. L'objectif de la réduction de la dimensionnalité est de préserver autant que

possible la structure significative des données à haute dimension dans la carte à basse dimension [58].

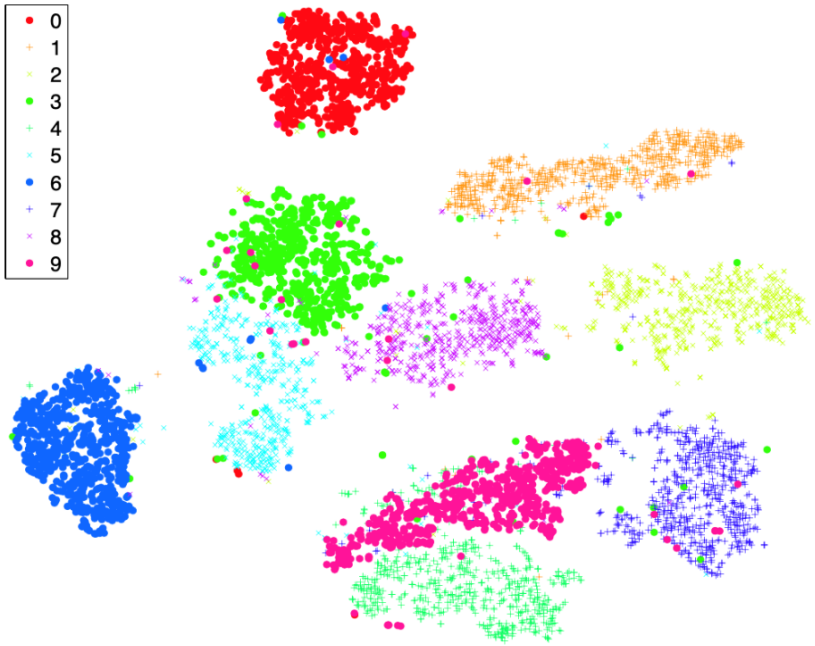


Figure 2.8 - La représentation visuelle de t-SNE [58]

Une nouvelle technique t-SNE [58], inspirée par SNE [169] utilise une distribution de Student-t au lieu d'une distribution gaussienne pour atténuer le problème de l'encombrement et les problèmes d'optimisation du SNE. Le t-SNE met l'accent sur la modélisation des points de données différents par de grandes distances par paire et sur la modélisation des points de données similaires par de petites distances par paire. Bien que le t-SNE se compare favorablement à d'autres techniques de visualisation de données, il présente trois faiblesses potentielles : (1) il est difficile de déterminer comment le t-SNE se comporte dans les tâches générales de réduction de la dimensionnalité. (2) La nature relativement locale du t-SNE le rend sensible à la malédiction de la dimensionnalité intrinsèque des données. (3) il n'est pas garanti que le t-SNE converge vers un optimum global de sa fonction de coût. Il est évident que des travaux futurs sont prévus afin d'optimiser le nombre de degrés de liberté de la distribution Student-t utilisée dans t-SNE, et de développer une version paramétrique de t-SNE qui permette une généralisation aux données de test retenues [58].

2.2 Les réseaux

Pour commencer, il est important de clarifier la terminologie qui sera utilisée au cours de cette section. Bien qu'il y ait une différenciation notable entre un graphe, composé de sommets et d'arêtes, et un réseau, caractérisé par des nœuds et des arcs, ce dernier représentant généralement une structure plus complexe et plus étendue, la terminologie sera utilisée de manière interchangeable. Cette approche est adoptée par souci de simplicité et pour éviter toute redondance dans la discussion, en reconnaissant que dans le contexte de ce travail, les termes sont fonctionnellement équivalents malgré leurs différences théoriques en termes d'échelle et de complexité.

Les réseaux, ou graphes constituent une méthode fondamentale d'analyse des systèmes complexes dans de nombreux contextes. Ils modélisent les réseaux de télécommunications tels que le World Wide Web, les systèmes de transport, y compris les réseaux ferroviaires et routiers, et les réseaux électriques. Les graphes sont également utilisés pour étudier les réseaux sociaux, tels que les structures d'entreprise de même que la relation de coauteurs. Ils sont utilisés dans les études biologiques, notamment pour cartographier les réseaux alimentaires et dans plusieurs autres domaines [75]. Les idées théoriques sur les graphes sont largement utilisées dans les applications informatiques, en particulier dans des domaines de recherche tels que l'exploration de données, la segmentation d'images, le regroupement, la capture d'images et la mise en réseau [73].

Un diagramme est une représentation visuelle de situations réelles, composée de points et de lignes reliant certaines paires de points. Par exemple des personnes représentées par des points, où des lignes relient des amis. L'objectif principal est de déterminer si deux points sont reliés par une ligne. Cette abstraction mathématique conduit au concept de graphe [76]. La théorie des graphes est un sous-ensemble des mathématiques discrètes qui met l'accent sur l'importance des graphes en tant que méthode d'expression d'informations picturales [74].

Un graphe $G = (V, E)$ est constitué d'un ensemble d'objets $V = \{v_1\ v_2, \ldots\}$ appelés sommets ou nœuds, et d'un autre ensemble $E = \{e_1, e_2, \ldots\}$, dont les éléments sont appelés arêtes. Chaque arête e_k est identifiée à une paire non ordonnée (v_i, v_j) de nœuds. La représentation la plus courante

d'un graphe est un diagramme, dans lequel les nœuds sont représentés par des points et chaque arête par un segment de ligne reliant ses extrémités [76][77][78].

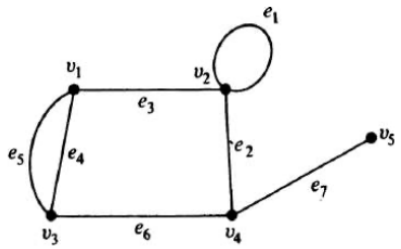


Figure 2.9 - La représentation visuelle d'un graphe [76]

Dans le contexte des systèmes spatiaux, les nœuds représentent des lieux ou des objets, tandis que les arêtes représentent des relations causales, statistiques, déductives ou spatiales ou des processus tels que les flux de matière ou d'énergie. Les arêtes d'un réseau peuvent être non dirigées ou dirigées, formant ainsi un graphe dirigé. Une séquence d'arêtes successives est une marche si aucun nœud n'est visité plusieurs fois. Les arêtes peuvent être pondérées en fonction de la force de la relation, de la distance ou de l'ampleur du flux. Un réseau possède une fonction qui attribue un poids non négatif à chaque arête [78].

2.1.1 Types de réseaux

La théorie des graphes, qui trouve son origine dans le problème du pont de Königsberg en 1735, a évolué au fil du temps pour inclure différents types de graphes dotés de propriétés uniques [73] [74][79]. Ces propriétés organisent les structures des sommets et des arêtes, faisant de la théorie des graphes un sujet vaste et complexe. Il existe plusieurs types de graphes qui ont chacun des caractéristiques uniques [74]. Un graphe non orienté n'a pas d'orientation, ses arêtes sont identiques les unes aux autres, et le nombre maximal d'arêtes sans boucle est $n(n\ -\ 1)/2$. Un graphe orienté est un graphe dont chaque arête est représentée par une paire ordonnée de deux sommets, par exemple (V_i, V_j) indiquant une arête allant du premier au deuxième sommet. Un graphe connecté est un graphe dans lequel chaque paire de sommets a un chemin. La matrice d'adjacence est une matrice binaire n x n associée à un graphe, indiquant la proximité des sommets adjacents. Elle peut

être représentée dans un tableau. Un graphe cyclique possède au moins un cycle, tandis qu'un graphe acyclique n'a pas de cycle [80]. Ceux-ci ne sont que des exemples, toutefois chacun d'entre eux nécessite l'utilisation de techniques spécifiques. Certaines de ces techniques et caractéristiques seront étudiées plus en détail, notamment les graphes épars, multigraphes, attributs, large ainsi que dynamique.

Épars

Un graphe dense est un graphe dont le nombre d'arêtes est proche du nombre maximal d'arêtes, reliant chaque paire de sommets par une arête. Un graphe épars comporte moins d'arêtes, ce qui indique une structure plus complexe.

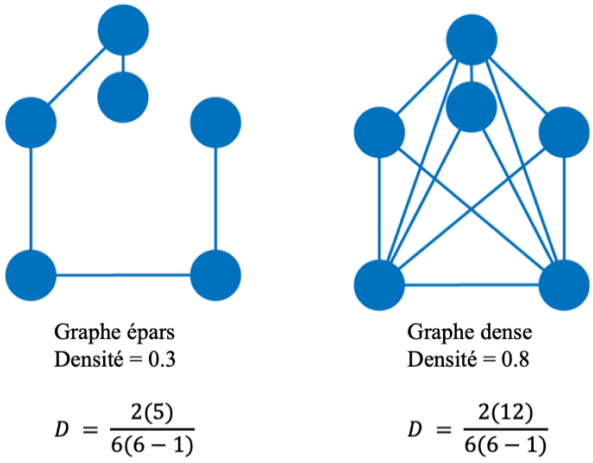


Figure 2.10 - La représentation visuelle d'un graphe épars et dense [143]

Étant donné que les graphes dans la vie courante sont souvent épars et incomplets, l'étude de la détection des communautés dans les graphes épars est très utile pour les applications pratiques et plusieurs chercheurs tendent vers cette exploration. Notamment *Chen et al.* [81] présente un nouvel algorithme de regroupement des graphes épars, plus performant que les méthodes précédentes. Il est conçu pour les grands graphes non pondérés et est utile pour la détection des communautés et la prédiction des liens. La méthode utilise la propriété de faible rang des matrices de regroupement et démontre son efficacité par des simulations. *Chin et al.* [82] présente un algorithme spectral optimisé à l'aide du modèle de blocs stochastiques. Il identifie les communautés de manière

optimale et traite de la détection des communautés dans les réseaux avec peu de connexions. L'algorithme gère des blocs multiples et fournit une base mathématique solide. Les résultats améliorent la compréhension de la détection des communautés dans les réseaux épars, ce qui est pertinent pour l'exploration théorique et l'analyse pratique des réseaux.

Multigraphes

Un multigraphe est composé de plusieurs graphes à couche unique. Un graphe à couche unique ne comporte qu'un seul type de nœud et d'arête. En reliant les nœuds des graphes à couche unique par des arêtes, on obtient un multigraphe. La figure 2.11 montre une représentation d'un graphe multigraphes. La définition dépend donc de celle d'un graphe à couche unique qui est représenté par un graphe pondéré (V, w) où V est un ensemble de sommets et w est un ensemble de poids des arêtes : $(V \times V) \rightarrow [0,1]$.

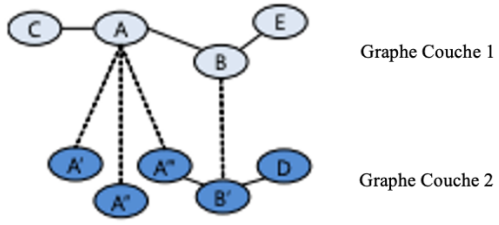


Figure 2.11 - La représentation visuelle d'un graphe multigraphes [83]

La compréhension de la correspondance entre les sommets d'un graphe et les sommets d'un autre est cruciale pour décrire ces graphes, qui sont composés de graphes interdépendants, et peut être formalisée à l'aide de la cartographie des nœuds. Un graphe multigraphes est un tuple $MLN = (L_1, \ldots, L_l, IM)$ où $L_i = (V_i, w_i), i \in 1, \ldots, l$ sont les couches du graphe et IM (Identity Mapping) est une matrice $l \times l$ de correspondances de nœuds, avec $IM_{i,j} : V_i \times V_j \rightarrow [0, 1]$ [83]. *Kim et Lee* [84] traitent des méthodes de maximisation de la modularité pour les réseaux multigraphes temporels et de l'heuristique de Louvain. Il démontre le compromis entre la structure statique de la communauté et la persistance à travers les couches, et applique ces méthodes aux réseaux de corrélation d'actifs financiers. *Vallès-Català et al.* [85] aborde le problème de

l'identification de la structure multigraphes des réseaux à partir de données agrégées en généralisant les modèles de blocs stochastiques (SBM) à couche unique et en tenant compte de différents mécanismes d'agrégation des couches. Il fournit une solution probabiliste pour identifier le SBM optimal et une approximation pour le calcul pratique, améliorant ainsi les capacités prédictives de l'analyse de la structure du réseau dans les systèmes complexes.

Attribut

Les graphes d'attributs sont des types de graphes dont les nœuds contiennent des informations supplémentaires. Ils sont associés à une collection d'attributs contenant m attributs différents, chaque nœud correspondant à un vecteur d'attribut. Les graphes complexes peuvent impliquer des attributs dans des arêtes avec des représentations similaires. Ces graphes sont largement utilisés dans des applications réelles telles que les réseaux sociaux et les réseaux de citations [99]. Cependant, elle nécessite une modélisation conjointe des structures des graphes et des attributs des nœuds, ce qui pose des problèmes. Les méthodes classiques de regroupement telles que les k-moyennes ne traitent que les caractéristiques des données, alors que de nombreuses méthodes de regroupement basées sur les graphes exploitent les modèles de connectivité des graphes [100]. Les approches bayésiennes et les représentations non négatives sont utilisées pour analyser ces graphes. *Jin et al.* [101] présente MRFasGCN, une méthode d'apprentissage profond qui combine les réseaux convolutionnels (GCN) et les champs aléatoires de Markov (MRF) pour améliorer la détection semi-supervisée des communautés dans les réseaux attribués. Cette approche exploite les informations structurelles et attributives des réseaux et s'attaque à leurs limites, surpassant les méthodes existantes en termes de précision et de temps d'exécution sur de grands réseaux. *Qin et al.* [102] présente une nouvelle méthode de détection des communautés qui intègre la topologie du réseau et les attributs. Elle utilise la factorisation matricielle non négative et un paramètre adaptatif pour gérer la contribution du contenu en fonction des incohérences. Cette méthode prend en charge la détection de communautés disjointes et chevauchantes, et démontre une robustesse et des performances accrues, en particulier dans les cas de non-concordance entre la topologie et le contenu.

Large

Les larges réseaux comportant des milliers ou des millions de nœuds sont courants dans divers domaines scientifiques [94]. Les premiers algorithmes se sont concentrés sur la structure globale, mais aujourd'hui, alors que nous explorons des réseaux comportant des milliards d'arêtes et des centaines de communautés, il est essentiel de se concentrer sur la structure microscopique. De plus en plus de travaux adoptent des méthodes d'expansion locale pour identifier les communautés à partir de membres exemplaires [95]. L'optimisation convexe est une méthode populaire pour la modélisation de problèmes dans divers domaines, mais son évolutivité est entravée par des ensembles de données plus vastes et plus complexes. Les méthodes classiques, souvent basées sur des méthodes de points intérieurs, échouent en raison de l'absence de structure connue du problème. Le défi consiste à développer des méthodes générales qui fonctionnent bien indépendamment de l'entrée et qui peuvent s'adapter à d'immenses ensembles de données [96]. *Wang et al.* [97] présente le concept de détection de communauté dans de larges réseaux, en mettant en évidence l'influence des utilisateurs et la disparité des comportements. Il propose deux algorithmes, GREEDY et WEBA, qui surpassent les méthodes de pointe dans l'identification des utilisateurs influents et la découverte des structures communautaires cachées. *De Meo et al.* [98] présente CONCLUDE, une méthode de détection des communautés dans les larges réseaux qui combine des approches globales et locales. Elle utilise la centralité $\kappa - path\ edge$ pour cartographier les sommets du réseau en points euclidiens et divise le réseau en grappes sur la base des distances. Cette méthode améliore les performances de regroupement sur des réseaux réels et synthétiques.

Dynamique

Les réseaux sont de plus en plus étudiés pour leur adaptabilité aux systèmes complexes, y compris ceux qui subissent des changements structurels dynamiques. Les réseaux sont modélisés comme des graphes, les nœuds représentant les objets individuels et les arêtes les interactions entre eux. Les communautés se forment au fur et à mesure que les individus interagissent et échangent des informations [86]. L'analyse du comportement dynamique et de l'évolution des réseaux dans le temps, est devenue essentielle.

Les réseaux dynamiques suivent l'évolution des interconnexions dans le temps, ce qui permet de suivre l'évolution de la structure du réseau à différents moments [86]. L'analyse traditionnelle des réseaux sociaux considère les réseaux comme des graphes statiques, soit en agrégeant des données au fil du temps, soit en prenant une image des données à un moment précis [87]. Cependant, des travaux récents se sont concentrés sur l'analyse des communautés et de leur évolution temporelle dans des réseaux dynamiques, mettant en lumière leur nature dynamique [88][89][90][91][92][93]. *Wang et al.* [93] propose une nouvelle structure appelée résumée des motifs locaux pondérés (LWEP) pour découvrir des communautés dynamiques dans des flux de graphes pondérés. L'approche est divisée en composantes en ligne et hors ligne, les statistiques étant maintenues pour préserver le maximum d'informations sur les voisins pondérés avec un stockage de mémoire limité. Les listes de voisins top-k et les listes de candidats top-k suivent les voisins top-k ayant les poids de lien les plus importants, tandis que le regroupement consolide les LWEP en grappes de haut niveau. *Yang et al.* [92] présente un modèle de blocs stochastiques dynamiques (DSBM) pour détecter les communautés et leur évolution dans les réseaux sociaux dynamiques à l'aide d'une approche bayésienne. Il capture la transition de la communauté en modélisant les changements d'appartenance des nœuds individuels au fil du temps. Contrairement aux méthodes traditionnelles d'estimation ponctuelle, le modèle DSBM utilise l'inférence bayésienne pour les distributions a posteriori, ce qui permet une gestion robuste du bruit et de l'incertitude des données.

2.2.2 Analyse des réseaux

La détection des communautés dans les réseaux est devenue une question fondamentale dans la science des réseaux [117]. Elle peut aider à identifier des sous-unités fonctionnelles et à découvrir des similitudes entre les sommets. Ils peuvent être classés en fonction de leur position structurelle au sein du groupe, qui peut être corrélée à leur rôle. Les nœuds centraux peuvent avoir des fonctions de contrôle et de stabilité, tandis que les nœuds frontières peuvent être des médiateurs entre différentes parties du graphe [118]. Ces communautés distinctes au sein des réseaux sont des sous-ensembles de nœuds dont les connexions sont plus denses que celles du reste du réseau [119][120]. La structure communautaire d'un réseau peut également constituer une puissante représentation

visuelle du système, permettant une description plus compacte et plus compréhensible du graphe dans son ensemble [118].

Des mesures ont été mises en place pour mesurer la densité interne et externe d'une communauté. Nous définissons la densité interne $\delta_{int}\ (C)$ du sous-graphe C comme le rapport entre le nombre d'arêtes internes de C et le nombre de toutes les arêtes internes possibles, c'est-à-dire

$$\delta_{int}\ (C) = \frac{\#\ \text{arêtes internes de C}}{\mathrm{n_c}(\mathrm{n_c}\ - 1)/2}$$

De même, la densité externe $\delta_{ext}\ (C)$ est le rapport entre le nombre d'arêtes allant des sommets de C au reste du graphe et le nombre maximum d'arêtes externes possibles, c'est-à-dire

$$\delta_{ext}\ (C) = \frac{\#\ \text{arêtes externes de C}}{\mathrm{n_c}(\mathrm{n_c}\ - 1)/2}$$

Pour qu'une communauté C soit une communauté, $\delta_{int}\ (C)$ doit être plus grand que la densité moyenne de liens $\delta(G)$ de G, qui est déterminée par le rapport entre le nombre d'arêtes de G et le nombre maximal d'arêtes possibles. De même, $\delta_{ext}\ (C)$ doit être plus petit que $\delta(G)$. La plupart des algorithmes de regroupement visent à trouver le meilleur compromis entre un grand $\delta_{int}\ (C)$ et un petit $\delta_{ext}\ (C)$, ce qui peut être réalisé en maximisant la somme des différences sur tous les regroupements [121].

Les communautés peuvent être étudiées selon deux approches : locale ou globale. Les communautés sont des entités distinctes au sein d'un graphe, dont les liens avec le système sont limités. Elles doivent être évaluées indépendamment du graphe dans son ensemble. Les définitions locales se concentrent sur le sous-graphe étudié, éventuellement sur son voisinage immédiat, en négligeant le reste du graphe [121]

Les communautés peuvent être définies sur la base du graphe dans son ensemble, en particulier lorsque les groupes sont cruciaux et ne peuvent être séparés sans affecter la fonctionnalité du système. La littérature fournit des critères globaux pour l'identification des communautés, souvent des définitions indirectes utilisant des propriétés globales du graphe. Cependant, il existe des

définitions appropriées basées sur l'idée qu'un graphe possède une structure communautaire s'il diffère d'un graphe aléatoire [121].

À la suite de ce survol de la détection des communautés locales et globales, il convient de se pencher sur les structures de réseaux spécialisées, telles que les réseaux bipartites. Les réseaux bipartites, qui consistent en deux ensembles de nœuds disjoints, sont souvent utilisés dans l'étude des réseaux complexes. Les réseaux de citations, les systèmes de recommandation, les réseaux d'interactions protéiques et les réseaux d'acteurs de cinéma en sont des exemples [103]. Les réseaux bipartites sont souvent étudiés en les projetant dans des réseaux unipartites ce qui peut entraîner une perte d'informations, une inflation des arêtes et d'autres inconvénients. Il devient important de développer des méthodes pour détecter les communautés sur les graphes bipartites originaux [104]. Un réseau bipartite est un triple $G = (\top, \bot, E)$, où $\top$ et $\bot$ sont des ensembles disjoints de nœuds, et $E \subseteq \top \times \bot$ est l'ensemble des liens du réseau. Il diffère des réseaux classiques, car les liens n'existent qu'entre les nœuds supérieurs et inférieurs [105].

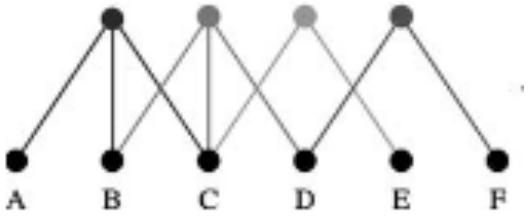


Figure 2.12 - La représentation visuelle d'un graphe multicouche [105]

Costa et Hansen [106] présentent une heuristique de division hiérarchique localement optimale pour la maximisation de la modularité dans les graphes bipartis, en se concentrant sur le regroupement précis des entités au sein des réseaux. L'heuristique atteint l'optimalité locale en résolvant le sous-problème de la bipartition d'un groupe. L'approche est validée par des expériences informatiques, qui soulignent son efficacité dans la compréhension de la structure des réseaux bipartis. *Ganji et al.* [107] présente une nouvelle méthode d'identification des modules dans les réseaux bipartites et unipartites dirigés. Cette méthode s'attaque aux algorithmes existants qui négligent souvent la directionnalité des connexions ou la nature bipartite des réseaux. Cette

approche innovante améliore la précision de la détection des structures modulaires, essentielles pour comprendre l'organisation et la fonction des réseaux.

Modularité

Comme de nombreux systèmes de notre quotidien peuvent être représentés par des réseaux, la recherche s'est principalement concentrée sur leur analyse afin d'identifier des modèles structurels sous-jacents [108]. Le critère de modularité, proposée pour la première fois par *Newman* [170] et *Newman et Girvan* [112], est une mesure de la structure de la communauté trouvée par un algorithme de détection de la communauté. Elle est considérée comme l'une des premières tentatives pour comprendre le problème du regroupement en mesurant la force de la structure de la communauté par rapport à un modèle nul [112][113]. La modularité mesure la densité d'un sous-graphe présentant une communauté qui ensuite comparée à un sous-graphe où les arêtes sont distribuées aléatoirement selon le degré des sommets [109]. Supposons que nous disposions d'un regroupement de sommets, $\{c_1, c_2, \ldots, c_n\}$, où c_i est la classe du sommet v_i. La modularité va mesurer la qualité du regroupement dans un réseau non dirigé et sa matrice d'adjacence W. Elle est définie par

$$Q = \frac{1}{2m}\sum_{ij}\left(w_{ij} - \frac{d_i d_j}{2m}\right)\delta(c_i, c_j),$$

où m est le nombre total d'arêtes, d_i est le degré du sommet v_i et $\delta(x, y)$ est le delta de Kronecker. La somme est calculée sur toutes les paires de sommets d'un même groupe. Elle mesure le nombre d'arêtes qui existent au sein des regroupements par rapport au nombre attendu si les arêtes sont distribuées aléatoirement en respect du degré des sommets. Par exemple, une arête spécifique du sommet v_i ayant un degré k_i peut être attachée à une extrémité de l'arête v_i et une autre à l'une des $2m$ extrémités de toutes les arêtes. La probabilité que l'autre extrémité de l'arête s'attache à l'une des d_j arêtes du sommet v_j est d_j. Le nombre total attendu de $2m$ est $d_i d_j$ puisque v_i a d_i arêtes [108][112][113]. La modularité continue de susciter de l'intérêt et un certain nombre d'études concernant les différentes applications et les ajustements possibles de la mesure ont vu le jour [108] [110][111][114][115][116].

Algorithmes de modularité

Il existe différentes méthodes pour détecter les communautés dans les réseaux, chacune ayant sa spécificité et s'appliquant à un type de graphe. Les premières approches se sont concentrées sur les graphes statiques avec des communautés hiérarchiques. La maximisation de la modularité est une méthode populaire pour détecter les communautés dans les graphes, car des valeurs élevées de modularité indiquent de bonnes partitions. Dans le cadre de ce mémoire, nous allons nous concentrer sur quelques algorithmes qui utilisent la modularité comme mesure de qualité.

L'optimisation de la modularité est un problème complexe encore aujourd'hui. En effet une optimisation exhaustive de Q [112] est impossible en raison du grand nombre de possibilités de partitionnement, même pour les petits graphes. Malgré cela, certains algorithmes peuvent trouver des approximations du maximum de modularité en un temps raisonnable [121].

Algorithme Greedy (glouton)

La méthode gloutonne de Newman [122] est une approche de regroupement hiérarchique agglomératif qui fusionne progressivement les sommets pour former des communautés plus larges, dans le but de maximiser la modularité. Elle commence avec autant de grappes que de sommets et ajoute des arêtes pour fusionner les grappes, maximisant ainsi la modularité. Avec une complexité de $O((m + n)n)$ ou $O(n)^2$ pour les graphes épars, elle permet l'analyse de grands réseaux [122].

Clauset et al. [124] ont amélioré la méthode de Newman en optimisant les structures de données pour les matrices éparses, réduisant la complexité à $O(md\ log\ n)$ et permettant l'analyse de grands graphes comportant jusqu'à 10^6 sommets [124]. Cependant, cette approche forme souvent de grandes communautés, ce qui peut entraîner des valeurs de modularité médiocres. Les modifications comprennent la normalisation de la variation de la modularité pour favoriser les petits groupes, la recherche de fusions pour équilibrer la taille de la communauté et le démarrage de l'agglomération hiérarchique à partir d'un état initial plus agrégé [121].

Blondel et al. [125] ont introduit une approche rapide et efficace pour les graphes pondérés, en se concentrant sur l'assignation séquentielle des sommets aux communautés pour une modularité maximale, capable d'analyser des graphes comportant jusqu'à 10^9 arêtes [125].

L'algorithme offre une faible complexité, un optimum local, de l'efficacité, et peut résoudre certains problèmes d'optimisation et être utilisé comme stratégie d'optimisation pour d'autres algorithmes. Cependant, il peut ne pas atteindre la solution optimale globale et nécessite que le problème soit doté d'une nature de sélection gloutonne. Il peut améliorer les résultats pour certains problèmes, mais pas pour les problèmes complexes comportant de nombreuses contraintes [123].

Algorithme LPA (propagation d'étiquettes)

La propagation d'étiquettes a été proposée par *Raghavan et al.* [130]. La méthode suppose qu'un nœud x ait des voisins $x_1, x_2, \ldots, x_k$ et que chaque voisin porte une étiquette indiquant la communauté à laquelle il appartient. Dans ce cas, x détermine sa communauté en fonction des étiquettes de ses voisins. Chaque nœud du réseau choisi rejoint la communauté qui compte le plus grand nombre de membres, ce qui garantit une égalité uniforme et égalitaire. Il y a ensuite initialisation de chaque nœud avec des étiquettes uniques et laisse les étiquettes se propager dans le réseau [130]. Au fur et à mesure que les étiquettes se propagent, les groupes de nœuds densément connectés parviennent rapidement à un consensus sur une étiquette unique si elle fait partie des plus fréquentes. Sinon une nouvelle étiquette est choisie au hasard parmi les plus fréquentes [131]. À la fin du processus de propagation, les nœuds ayant les mêmes étiquettes sont regroupés en une seule communauté [130].

Barber et Clark [172] ont introduit la modularité bipartite spécialisée LPA (LPAb) pour détecter les communautés dans les réseaux bipartis sur la base de l'optimisation de la modularité bipartite. Chaque nœud se voit initialement attribuer une étiquette unique, indiquant sa communauté. À chaque étape, les nœuds mettent à jour leurs étiquettes dans une séquence aléatoire afin de maximiser la modularité bipartite [132]. Ce n'est qu'un an plus tard que *Lui et Murata* [173] propose le modèle LPAb+. Un modèle qui échappe aux maxima locaux et détecte les divisions de communautés avec une modularité bipartite beaucoup plus élevée tout en conservant une vitesse rapide. C'est un algorithme hybride combinant LPAb modifié (LPAb') et l'algorithme agglomératif

glouton à plusieurs étapes (MSG). Il est divisé en deux phases itératives correspondant à LPAb' et MSG.

Algorithme de recuit simulé

La simulation de recuit est une technique d'optimisation globale basée sur la mécanique statistique. Il s'agit d'une forme avancée d'optimisation locale ou d'amélioration itérative qui permet des mouvements « ascendants » occasionnels, en évitant le piège des solutions sous-optimales. Cette méthode polyvalente peut être appliquée à divers problèmes sans nécessiter de connaissances spécifiques approfondies. Son principal avantage réside dans le fait qu'elle contourne les optima locaux médiocres, ce qui permet d'obtenir des résultats nettement meilleurs [128].

Le recuit simulé est une procédure probabiliste d'optimisation globale qui explore l'espace des états possibles afin de potentiellement trouver l'optimum global d'une fonction F. Les transitions se produisent avec une probabilité de 1 $si\ F$ augmente après un changement, et avec une probabilité $exp(\beta\Delta F)$, où ΔF est la diminution de la fonction et β est un indice de bruit stochastique. Le bruit réduit le risque que le système soit piégé dans des optima locaux. À un certain stade, le système converge vers un état stable, qui peut être une approximation arbitrairement bonne du maximum de F [121].

Il a été utilisé pour la première fois pour l'optimisation de la modularité par *Guimerà et al.* [129] et combine des mouvements locaux et globaux. La meilleure performance est obtenue en optimisant la modularité d'une bipartition de la grappe, ce qui est fait à nouveau avec le recuit simulé, en ne considérant que les mouvements individuels des sommets et en diminuant la température jusqu'à ce qu'elle atteigne la valeur courante pour l'optimisation globale [121]. Les mouvements globaux réduisent le risque d'être piégé dans des minima locaux et ont prouvé qu'ils conduisaient à de meilleurs optima qu'en utilisant simplement des mouvements locaux [113].

2.2.3 Mesure des distances dans les réseaux et les textes

Pour comprendre et classer les modèles de graphes, nous devons les comparer et comprendre ce qui rend deux graphes similaires ou différents [133]. Il existe plusieurs approches de la similarité. Nous n'en couvrirons que quelques-unes dans le cadre de ce mémoire.

Matrice d'adjacence

Une matrice d'adjacence est une méthode utilisée pour déterminer les sommets adjacents d'un graphe. Pour un graphe fini G à n sommets, la matrice d'adjacence est la matrice $n \times n$, où l'entrée non diagonale a_{ij} représente le nombre d'arêtes du sommet i au sommet j et l'entrée diagonale a_{ii} est soit une fois, soit deux fois le nombre d'arêtes du sommet i à lui-même. Les graphes non dirigés utilisent souvent la première convention, tandis que les graphes dirigés utilisent généralement la seconde. Il existe une matrice unique pour chaque graphe, qui n'est la matrice d'adjacence d'aucun autre graphe. Dans un graphe simple fini, elle est une matrice $(0, 1)$ avec des zéros sur sa diagonale. Si G est un graphe non orienté, elle est symétrique. Si G est un graphe multiple, elle est la matrice $m \times m$, $A = (a_{ij})$, définie en fixant a_{ij} au nombre d'arêtes, entre v_i et v_j [134]. Les matrices adjacences sont la mesure par défaut de nombreux algorithmes de regroupement, y compris le regroupement spectral [135].

Jaccard

L'indice de Jaccard, introduit par le botaniste *Paul Jaccard en 1901*, mesure la similarité entre des ensembles en comparant les tailles d'intersection et d'union [136]. Cette mesure est le rapport entre le nombre de paires de sommets classées dans le même groupe dans les deux partitions et le nombre de paires de sommets classées dans le même groupe dans au moins une partition [121]. Cette similarité peut être définie comme suit :

$$J_{ij} = \frac{|a_i \cap a_j|}{|a_i \cup a_j|}$$

où i et j sont considérer comme deux objets décrits respectivement par les ensembles d'attributs a_i et a_j [137]. La valeur de l'indice de Jaccard se situe toujours entre 0 et 1. Plus l'indice est élevé, plus la similarité est grande.

Cosine

La similarité cosinus est couramment utilisée pour mesurer la proximité entre deux documents vectorisés par des méthodes telles que TF-IDF ou Doc2Vec. Cependant, son utilisation s'étend également à l'analyse des graphes. La similarité cosinus est calculée par l'équation suivante.

$$cos(v_u, v_i) = \frac{v_u \cdot v_i}{\| v_u \| + \| v_i \|}$$

où v_u et v_i sont des vecteurs du graphe. Ces éléments peuvent être des attributs de nœuds ou des vecteurs. Dans l'équation ci-dessus, la proximité de l'angle formé par les vecteurs est exprimée en prenant une valeur de 0 à 1. Une valeur plus proche de 1 signifie que la similarité entre eux est plus élevée [138]. *Troussas et al.* [139] démontre l'utilisation de la similarité cosinus dans le contexte du graphe de connaissances. Elle améliore la sophistication du système de recommandation en tenant compte des relations multidimensionnelles et des préférences des apprenants et des entités éducatives.

Communicabilité

La communicabilité est une mesure de la façon dont deux nœuds d'un réseau sont liés l'un à l'autre. Elle est généralement déterminée entre deux nœuds d'un réseau par le chemin le plus court qui les relie [140]. La communicabilité est définie entre une paire de nœuds p et q comme une somme pondérée des moments $\mu_k\ (p, q)$. Le poids est donné de manière que les chemins les plus courts reçoivent plus de poids que les plus longs. Le poids $1\ /\ k$! a été utilisé pour le moment k. Ce qui a permis d'exprimer la communicabilité $G_{p,q}$ des nœuds p et q en termes de paramètres spectraux du graphe par la formule suivante :

$$G_{p,q} = \sum_{k=0}^{\infty} \frac{\mu_k(p,q)}{k!} = \sum_{k=0}^{\infty} \frac{(A^k)_{pq}}{k!} = \sum_{j=1}^{n} \phi_j(p)\, \phi_j(q) e^{\lambda_j}$$

où $\phi_j(p)$est la p ième composante du j ième vecteur propre de la matrice d'adjacente A, qui est associé à la valeur propre λ_j [141].

Le concept de communicabilité vise à fournir une meilleure approximation de la communication entre les nœuds du réseau en utilisant des chemins aléatoires sans mémoire. Il compte le nombre de chemins de différentes longueurs entre les nœuds, en utilisant une fonction de pondération différente et la matrice d'adjacente A [142] .

Résistance

Introduite par *Klein et Randic* [152], cette mesure propose une fonction de distance basée sur la théorie des réseaux électriques, où une résistance fixe est imaginée sur chaque arête. L'idée est de considérer un graphe où chaque arête agit comme une résistance égale. La distance entre deux sommets est définie comme la résistance entre les deux nœuds. Les distances entre d'autres paires de sommets sont obtenues comme résistances effectives [152].

Sur un graphe G, la distance de résistante $\Omega_{i,j}$ entre 2 sommets V_i et V_j est défini comme suit :

$$\Omega_{i,j} = \Gamma_{i,i} + \Gamma_{j,j} - \Gamma_{i,j} - \Gamma_{j,i}$$

$$\text{Ou } \Gamma = \left(L + \frac{1}{|V|}\Phi\right)^{+},$$

Où

$^{+}$ désigne l'inverse de Moore-Penrose

L la matrice de Laplacien de G

$|V|$ est le nombre de sommets dans G

O est la matrice $|V|$ x $|V|$ contenant que le chiffre 1

Donc pour un graphe non dirigé si $i = j$ donc $\Omega_{i,j} = 0$ on obtient la formule mathématique suivante :

$$\Omega_{i,j} = \Omega_{j,i} = \Gamma_{i,i} + \Gamma_{j,j} - 2\Gamma_{i,j}$$

La formule calcule la distance de résistance, aussi interprétée comme une mesure de la distance du réseau entre deux nœuds dans un graphe, en tenant compte des chemins directs et indirects. La distance de résistance est une mesure de la théorie des graphes utilisée majoritairement dans l'analyse de la robustesse des réseaux, l'analyse de la marche aléatoire et la conception de circuits.

2.2.4 Mesure de centralité

Dans les réseaux complexes, les nœuds ont des caractéristiques uniques qui déterminent leur importance en fonction du contexte de l'application. L'importance d'un nœud change selon le contexte de l'application. Les nœuds à haut degré peuvent être influents de façon locale, mais ne peuvent pas rendre l'information virale au niveau global sans se connecter à d'autres nœuds influents. Ces caractéristiques peuvent être identifiées à l'aide de diverses mesures de centralité telle que le *degré* qui est calculé à partir d'informations locales sur le nœud. D'autres utilisent des informations globales sur le réseau, comme la *closeness centrality*, la centralité de Katz, etc…[148]. Les recherches se concentrent sur le développement de mesures de centralité permettant d'identifier efficacement ces nœuds influents dans des réseaux complexes [147]. *Bloch et al.* [149] révèle que les mesures de centralité dans l'analyse des réseaux sont basées sur des traitements additivement séparables et linéaires des statistiques qui capturent la position d'un nœud dans le réseau. Cela permet d'établir une taxonomie des mesures de centralité variant selon deux dimensions : les informations qu'elles utilisent sur les positions des nœuds et la manière dont elles sont pondérées en fonction de la distance. Dans le cadre de ce mémoire, nous allons nous pencher sur les mesures de centralité les plus observées.

Degree centrality

Le *degree centrality* mesure le nombre d'arêtes du nœud i, $d_i(g)$, et peut être normalisée par le degré maximal possible, $n - 1$, pour obtenir un nombre compris entre 0 et 1 [149]. La normalisation de cette mesure à l'aide du nombre total de connexions possibles améliore l'analyse et permet de mieux comparer deux nœuds de réseaux différents, quelle que soit la taille du réseau [148]. Elle peut être exprimée comme suit :

$$C_i^{deg}(g) = \frac{d_i(g)}{n-1}$$

Cette mesure est la plus fondamentale dans la science des réseaux. Elle donne un aperçu de la connectivité du nœud i, mais néglige l'architecture du réseau et la position du nœud [148].

Closeness centrality

Dans les applications réelles, les informations empruntent les chemins les plus courts, ce qui fait qu'un nœud a une grande influence s'il se trouve à une distance plus courte des autres nœuds. La *closeness centrality* capture cette propriété du réseau, en indiquant la proximité d'un nœud dans le réseau, qui est inversement proportionnelle à sa force. *Bavelas* [174] *and Sabidussi* [175], ont défini cette mesure normalisée comme suit [148].

$$C_i^{cls}(G) = \frac{n-1}{\sum_{j \neq i} \rho_G(i,j)}$$

Closeness centrality mesure la distance du réseau entre les nœuds et étend la centralité de degré en prenant en compte les voisinages de tous les rayons [149].

Betweenness centrality

Dans les réseaux complexes, l'unicité d'un nœud est déterminée par son importance dans le flux d'informations, ce qui est reflété par la *betweenness centrality* du nœud. Elle tient compte du

nombre de chemins les plus courts passant par un nœud, ce qui explique l'importance d'un nœud en ce qui concerne le flux d'informations [148].

La *betweenness centrality* est une mesure proposée par *Freeman* [176] qui prend en compte toutes les topographiques entre deux nœuds j, k différents de i qui passent par i. Elle mesure l'importance d'un nœud dans la connexion d'autres nœuds dans le réseau. Elle reflète le rôle d'un agent en tant qu'intermédiaire dans la transmission d'informations ou de ressources entre d'autres agents du réseau. Mathématiquement, cette mesure est :

$$c_i^{bet}(G) = \frac{2}{(n-1)(n-2)} \sum_{(j,k), j \neq i, k \neq i} \frac{v_G(i:j,k)}{v_G(j,k)}$$

qui pondère toutes les topographies de la même manière, indépendamment de la distance entre les nœuds ou du nombre d'autres moyens de se rejoindre [149]. L'interdépendance est une propriété unique qui augmente considérablement pour les nœuds qui forment des ponts entre les nœuds de différentes communautés ou groupes liés [142].

Katz centrality

Katz et Bonacich ont proposé en 1953 [148] une mesure de centralité basée sur le nombre de marches à partir d'un nœud i. Cette mesure est une extension de la mesure traditionnelle *degree centrality*, avec une différence notable, elle prend en considération les connexions indirectes. En raison de la longueur illimitée des marches dans un graphe, un facteur d'actualisation δ entre 0 et 1 est utilisé pour calculer la somme actualisée des marches [149]. L'algorithme attribue des poids différents aux chemins les plus courts en fonction de leur longueur, les chemins les plus courts étant considérés comme plus cruciaux pour le flux d'informations [148] [142]. La centralité de Katz peut être définie comme suit :

$$c^{KB}(A, \delta) = \sum_{i=1}^{\infty} \delta^k \sum_{j \epsilon V} A_{ij}^k$$

où A est une matrice d'adjacence

V est l'ensemble des nœuds du réseau

A_{ij} est un élément de la matrice A, indiquant la présence d'une arête entre les nœuds i et j.

Le score de centralité du nœud i est basé sur le comptage du nombre total de marches entre ce nœud et d'autres nœuds [149], avec une atténuation exponentielle en fonction de la longueur du chemin afin de garantir que les chemins plus courts ont un impact plus important sur la mesure de centralité que les chemins plus longs.

Eigenvector centrality

Eigenvector centrality est une mesure basée sur l'idée que la centralité du nœud i est liée à ses voisins. Elle est calculée en supposant que la centralité du nœud i est proportionnelle à la somme des centralités de ses voisins. La *eigenvector centrality* d'un nœud est donc autoréférentielle, mais possède un point fixe bien défini [149]. Elle est une mesure de l'influence d'un nœud au sein du réseau.

Pour un graphe G avec sa matrice d'adjacence A, la mesure de centralité x de chaque nœud est déterminé par vecteur propre correspondant à la plus grande valeur propre de λ_1 de A est dans sa forme la plus simple, donnée comme suit:

$$A\mathrm{x} = \lambda\mathrm{x}$$

où A est une matrice d'adjacence

x est le vecteur propre de A correspondant à la plus grande valeur propre λ_1

$\lambda_1$est la plus grande valeur propre de A.

Le score de centralité **x**, qui représente la plus grande valeur propre, est essentiel pour déterminer la centralité des vecteurs propres, car il garantit une solution stable proportionnelle à la somme des scores des voisins.

3 Méthodologie

Cette section sur la méthodologie présente une approche structurée, en commençant par une description détaillée de nos deux ensembles de données. Elle couvre ensuite les étapes de prétraitement prises en compte pour préparer les données à l'analyse, suivies de la modélisation thématique qui permettra d'identifier les thèmes sous-jacents. Enfin, nous introduisons une technique de détection des communautés pour explorer les relations et les regroupements au sein des données. Ce processus est essentiel pour atteindre les objectifs de l'étude et répondre aux questions de recherche posées qui se veut à analyser le contenu et la collaboration afin de croiser les résultats pour comprendre ces deux dimensions.

3.1 Descriptions des données

L'analyse comparative menée dans le cadre de cette étude repose sur deux ensembles de données distincts. Dans un premier temps, les données présentées au tableau 3.1 proviennent de la revue Management international (Mi), fournies par le rédacteur en chef. Cet ensemble de données se compose de résumés concis de tous les articles publiés dans la revue entre 2009 et 2023, formatés en Excel. Le corpus, qui comprend 829 documents, présente une diversité linguistique avec des textes en français, en anglais et en espagnol.

Une première évaluation a confirmé l'exhaustivité de l'ensemble des données. Un examen préliminaire de l'ensemble des colonnes a été réalisé, identifiant les attributs de chacune et leur pertinence pour l'étude. Cette procédure a garanti la disponibilité de toutes les informations nécessaires et confirmé que la structure des données convenait à l'analyse projetée.

Titre	Désignation de la colonne des données reçues
Article_id	Numéro unique permettant d'identifier l'article
issue_year	Année de la publication
article_title_FR	Titre de l'article en français
article_authors	Résumé de l'article en français
article_abstract_fr	Auteurs ayant contribué à la rédaction de l'article

Tableau 3.1 - Ensemble de données de la revue Management international (colonnes pertinentes pour la recherche)

Une vérification sommaire de l'intégrité des données a ensuite révélé l'absence de plusieurs documents français, probablement due à l'absence de traductions originales. Pour résoudre ce problème, des traductions ont été obtenues via le logiciel DeepL, assurant ainsi la cohérence et l'exhaustivité des données. Subséquemment, l'attention s'est portée sur les auteurs, importants pour l'identification des communautés. Des différences dans la saisie des noms d'auteurs ont été identifiées, risquant de biaiser les résultats. Une correction a été apportée pour uniformiser les noms d'auteurs, assurant ainsi l'exactitude des analyses futures.

L'accès aux textes intégraux de la revue aurait représenté un avantage notable; toutefois, il a été établi que leurs formats n'étaient pas compatibles avec un traitement efficace dans le cadre de cette étude. Par ailleurs, l'absence de version numérisée des archives remontant à la fondation de la revue en 1995 a restreint l'envergure historique de l'analyse aux documents publiés après 2009.

Pour permettre notre analyse comparative, un second ensemble de données, les cahiers du GERAD, a été intégré grâce à la générosité de deux membres du GERAD. Ce corpus, formaté en Excel, comprend tous les numéros publiés depuis leur création en 1981, soit un total de 3 022 documents. Afin d'harmoniser la portée temporelle de l'analyse comparative, les entrées antérieures à 2009 ont été exclues, ce qui a permis d'obtenir un corpus affiné de 1 441 documents. Ce corpus affiné est destiné à fonctionner comme un ensemble de données parallèles dans le cadre de l'étude.

En parallèle des mesures préliminaires adoptées pour le traitement de l'ensemble de données de Mi, une analyse initiale des données des Cahiers du GERAD a été réalisée afin d'évaluer le contenu

des colonnes pertinentes, comme représenté au tableau 3.2. La qualité des données s'est révélée exceptionnellement élevée, attribuable vraisemblablement aux efforts des recherches antérieures, facilitant ainsi la transition vers l'étape subséquente de prétraitement des données.

Titre	Désignation de la colonne des données reçues
numero_cahier	Numéro unique permettant d'identifier l'article
titre	Titre de l'article
date_publication	Date de la publication
résumé (FR)	Résumé de l'article en français
résumé (EN)	Résumé de l'article en anglais
résumé	Résumé de l'article traduit en anglais qui sera utilisé pour l'étude.
auteur_1	Auteur ayant contribué à la rédaction de l'article
auteur_2	Auteur ayant contribué à la rédaction de l'article
…	
auteur_12	Auteur ayant contribué à la rédaction de l'article

Tableau 3.2 - Ensemble de données des cahiers du GERAD (colonnes pertinentes pour la recherche)

La décision de fusionner les titres des documents avec leurs résumés a été prise pour augmenter le contexte de l'analyse, compte tenu de la brièveté des résumés et de leur omission potentielle de termes clés. Cette intégration a permis d'obtenir un texte plus complet pour chaque entrée de l'ensemble de données.

3.2 Prétraitement des données

Dans les études quantitatives, le prétraitement est un processus qui convertit le texte en chiffres, dans le but de simplifier les données d'entrée sans affecter l'interopérabilité du modèle. Ce compromis entre des données plus simples et une perte d'information minimale est une question complexe [6]. Dans le cadre de cette étude, les données textuelles ont été préparées selon des

procédures conçues pour affiner et normaliser l'ensemble des données. Ces dernières permettent de garantir des données de qualité pour le modèle et de fournir des informations significatives.

Le prétraitement a commencé par la mise en place de modèles linguistiques spécifiques à la langue de l'ensemble de données, à savoir le français. Le modèle « fr_core_news_sm » de spaCy a été utilisé pour une analyse linguistique sophistiquée. Ce modèle a fourni les outils nécessaires à une analyse précise et à la compréhension des nuances de la langue française, essentielles pour un traitement efficace.

Normalisation

Pour normaliser l'ensemble des données, une série d'étapes de nettoyage du texte a été mise en œuvre. Dans un premier temps, les signes de ponctuation ont été supprimés, en partant du principe que, malgré leur importance structurelle, ils ne contribuent pas de manière substantielle au contenu sémantique du texte. Ensuite, le texte a été transformé en minuscules afin de garantir un traitement uniforme des variations de mots. Par exemple, « Gestion » et « gestion » sont considérées comme équivalents, ce qui permet de réduire la redondance et de maintenir la cohérence de l'ensemble des données. Parallèlement, un autre aspect essentiel de la normalisation des textes a consisté à éliminer les balises HTML à l'aide de librairies Python. Cette étape était impérative, car ces balises n'améliorent pas la valeur sémantique du texte et peuvent perturber les performances des modèles analytiques.

Stopwords

La présence de mots qui reviennent fréquemment, mais qui n'ont qu'une signification sémantique minimale, souvent appelés *stopwords*, est un phénomène notable. Ces mots, bien que cruciaux pour la structure de la phrase, n'améliorent généralement pas la compréhension du sujet du contenu. Il s'agit par exemple des prépositions, des conjonctions et des articles tels que « le », « la », « l' » et « un ». En utilisant la liste prédéfinie dans spaCy pour le français, ces mots ont été systématiquement identifiés et supprimés du texte. Bien qu'il existe des listes prédéfinies pour de nombreuses langues, il est possible de les personnaliser et la décision d'omettre ces mots dépend

des objectifs de l'analyse. La langue se caractérise par sa multidimensionnalité, reflétée dans le contenu et la forme de l'expression. Identifier et isoler les aspects pertinents de ces dimensions est essentiel, tandis que les éléments non pertinents constituent un bruit superflu. Dans le cadre de cette étude, nous avons personnalisé une liste de *stopwords* tels que « article », « étude », « recherche », « cas », etc., que l'on trouve couramment dans de les résumés de textes plus importants, mais qui n'apportent pas de signification majeure au texte lui-même. L'élimination de ces mots est une pratique courante dans le traitement du langage naturel (NLP). Ce processus permet de rationaliser l'ensemble des données, en diminuant le nombre de mots à analyser sans altérer sensiblement le sens général du texte. Souvent, ce processus d'élagage réduit considérablement la taille de l'ensemble de données, dans certains cas jusqu'à la moitié.

Lemmatisation

La lemmatisation regroupe différentes formes de mots en un seul élément. Elle améliore l'uniformité des données et simplifie l'ensemble des données en transformant les formes flexionnelles et les formes dérivées occasionnelles en leur forme de base. Dans cette étude, la lemmatisation a été utilisée à la place du stemming. Ce choix a été fait pour maintenir l'exactitude des formes de mots et de leurs fonctions grammaticales, une caractéristique particulièrement vitale dans les langues à conjugaison étendue comme le français. Le tableau 3.3 nous permet de comprendre ces transformations réalisées avec la lemmatisation, mettant en évidence la façon dont les mots sont réduits à leurs formes de base.

Management international		**Cahiers du GERAD**	
concerne	concerner	**based**	base
est	être	**demonstrated**	demonstrate
attendus	attendre	**Promising**	promise
fondée	fonder	**Have had**	have
estimés	estimer	**was**	be

Tableau 3.3 - Exemple de changement dans le corpus

Analyse de la distribution des fréquences

Une analyse de la distribution des fréquences a été effectuée pour évaluer l'importance de chaque mot dans le corpus, dans le but de l'affiner pour une représentation précise du contenu. L'analyse a commencé par le calcul de la fréquence des documents (FD) pour chaque mot afin d'évaluer sa prévalence dans l'ensemble des documents, à la différence de la fréquence des mots qui mesure la densité au sein d'un seul document. Pour affiner le corpus, des seuils spécifiques ont été établis par des tests empiriques, identifiant les mots apparaissant dans plus de 70 % des documents comme non discriminatoires pour l'analyse thématique. De même, les mots présents dans moins de quatre documents ont été jugés trop rares, indiquant une utilisation spécialisée ne reflétant pas le discours général. Cette méthode a établi des seuils pertinents pour la fréquence des termes. Cela a rendu possible l'élaboration d'un corpus où les mots peu informatifs, qui ne contribuaient pas à la compréhension claire des thèmes centraux, ont été écartés.

3.3 Modélisation des thèmes

Malgré les progrès réalisés dans le domaine de l'analyse de contenu, les chercheurs manquent encore de méthodologies sophistiquées pour utiliser les données massives stockées dans des formats textuels [47]. *Grimmer et Stewart* [5] proposent diverses méthodes informatiques de traitement des données textuelles, qui varient principalement en fonction du type de données et des paramètres des algorithmes [5]. La modélisation thématique est une approche utilisée pour identifier les thèmes dans l'analyse de contenu des données textuelles [47]. L'utilisation de la modélisation thématique pour les articles publiés de la revue Mi ainsi que les cahiers du GERAD permet d'identifier les thèmes de recherche contenus dans les articles textuels, tout en minimisant les risques de biais.

Comme mentionné un peu plus tôt dans la revue de littérature, il existe plusieurs types d’algorithmes de modélisation thématique. Dans le cadre de cette étude, trois algorithmes importants ont été pris en compte : LDA, NMF et LSA. À la suite d'une expérimentation préliminaire, la méthode LDA a été choisie comme approche principale en raison de sa capacité supérieure à générer des thèmes cohérents et interprétables pour l'ensemble de données. Cette

décision a été prise après une évaluation qualitative de l'interopérabilité des thèmes, mais surtout aux discussions approfondies avec le rédacteur en chef de Mi et les experts du GERAD.

Le corpus initial, que ce soit pour Mi ou les cahiers du GERAD comprenait un nombre important de mots qui après un prétraitement a été réduit de manière significative. Le prétraitement comprend la tokenisation, la normalisation (balise HTML, mot en minuscule…), l'élimination des mots dans valeur ajoutée, la lemmatisation ainsi que la suppression des termes rares et courants. Cette phase est essentielle pour affiner l'ensemble des données afin de garantir que la modélisation thématique soit effectuée sur des données qui sont pertinentes et significatives [6].

	Management international	**Cahiers du GERAD**
Taille du vocabulaire original	14380	20700
Taille du vocabulaire après prétraitement	6309	9693
Taille du vocabulaire après le filtre par fréquence des mots	1715	2881

Tableau 3.4 - Taille du vocabulaire avant et après prétraitement

Il est nécessaire de définir à nouveau les termes de base qui composent le vocabulaire de la modélisation thématique. Un mot est un élément de vocabulaire, un document est une séquence de mots et un corpus est une collection de documents [42].

Les algorithmes de modélisation thématique, tels que LDA, utilisent le traitement du langage naturel et l'apprentissage automatique. Ils permettent d'identifier et d'extraire automatiquement des thèmes clés à partir de l'occurrence des mots des documents [60]. Les paramètres du modèle sont le nombre de thèmes, α et β. Le nombre optimal de thèmes, K, doit être suffisamment important pour générer des catégories interprétables et suffisamment petit pour être utilisable, car un nombre trop élevé peut entraîner des résultats ininterprétables [66]. Les hyperparamètres du modèle α et β représentent respectivement la densité document-thème et la densité thème-mot [60], tel que décrit à la section 2.1.3. Une valeur alpha plus faible se traduit par des documents contenant moins de thèmes dominants, tandis qu'une valeur alpha plus élevée conduit à une distribution plus uniforme

des thèmes dans les documents. Inversement, une valeur bêta plus faible se traduit par des thèmes constitués d'un plus petit ensemble de mots, ce qui renforce leur caractère distinctif, tandis qu'une valeur bêta plus élevée permet une répartition plus uniforme des mots entre les thèmes.
Afin de déterminer un nombre approprié de thèmes pour notre ensemble de données, une approche itérative a été utilisée, en exécutant le modèle LDA plusieurs fois, en faisant varier les paramètres afin d'observer leur impact sur la cohérence. L'objectif final était d'atteindre un équilibre dans lequel les thèmes étaient à la fois significatifs et représentatifs du corpus, produisant des structures thématiques pertinentes pour les objectifs de la recherche.

	Management international	**Cahiers du GERAD**
Nombre de thèmes optimal	7	7
$\boldsymbol{\alpha}$	0,5	0,5
$\boldsymbol{\beta}$	0,2	0,2

Tableau 3.5 - Hyperparamètres final du modèle LDA

Les résumés du corpus de Mi et des cahiers du GERAD sont concis, avec environ 100 à 120 mots chacun. Cette concision peut conduire à l'omission d'informations détaillées ou d'aspects secondaires essentiels pour une compréhension complète du thème. Ces brefs résumés d'un contenu complexe peuvent entraîner une simplification excessive, ce qui pourrait affecter les résultats de l'analyse de modélisation thématique. Dans l'étude [53], les auteurs proposent une nouvelle méthode de modélisation des thèmes dans les textes courts, appelée modèle thématique biterm (BTM). Cette méthode permet d'apprendre les thèmes en modélisant directement la génération de motifs de cooccurrence de mots (c'est-à-dire les bitermes) dans l'ensemble du corpus. Bien que l'étude ait montré des résultats très concluants, la méthode utilisée dans cette étude a suivi une approche thématique classique. Cependant, il est à noter que l'approche par bitermes reste une avenue pour de futures recherches nécessitant une analyse textuelle plus approfondie.

3.4 Détection de communautés

L'objectif principal de la modélisation des réseaux est de répondre à la question de recherche, à savoir l'exploration des différentes structures communautaires et déterminer ce que Mi peut exploiter des stratégies d'engagement communautaire des cahiers du GERAD. Cette phase, bien que complémentaire à ce mémoire sera exploitée de manière préliminaire. Elle servira de tremplin pour établir et croiser les résultats de la modélisation thématique et les communautés au sein des deux ensembles de données.

Avant de poursuivre, il est essentiel de préciser que, pour les deux ensembles de données, une étape de prétraitement a été nécessaire en ce qui concerne les données des auteurs. À l'origine, tous les auteurs ont été regroupés sur une seule ligne pour chaque résumé, souvent accompagnés d'informations supplémentaires telles que les affiliations institutionnelles ou d'autres titres importants. Pour représenter correctement le réseau, il était nécessaire de s'assurer que chaque entrée dans l'ensemble de données correspondait à un seul auteur par ligne. Par conséquent, cette modification a entraîné une expansion des ensembles de données, puisqu'un résumé peut désormais être reproduit jusqu'à douze fois s'il a douze auteurs contributeurs.

Les données ont par la suite été classées dans des intervalles prédéfinis de trois ans ce qui a permis d'établir une structure pour l'analyse temporelle. Le choix d'utiliser des périodes fixes plutôt qu'une approche de détection de communautés dynamique était délibéré. L'utilisation de périodes fixes permet d'analyser les tendances, les modèles et les changements au sein des données sur des intervalles de temps uniforme. Il devient possible de suivre l'émergence et l'évolution de phénomènes spécifiques, de comparer différentes tranches de temps et de tirer des conclusions sur la progression ou la régression de certaines variables dans le temps ce qui est aligné avec l'objectif de ce mémoire.

L'analyse utilise un graphe bipartite composé de deux sous-ensembles de sommets, soit les auteurs et les résumés qui seront reliés ensemble par des arêtes. Par définition, un graphe bipartite doit être constitué de deux ensembles de nœuds qui ne se chevauchent pas et dont les arêtes relient les sommets des différents ensembles [104]. Visuellement le graphe sera représenté comme suit dans

sa forme simplifiée, ou les chiffres sont représentés par les résumés et les lettres par les auteurs des résumés en question. Cette structure est idéale pour détecter les collaborations et les contributions des auteurs.

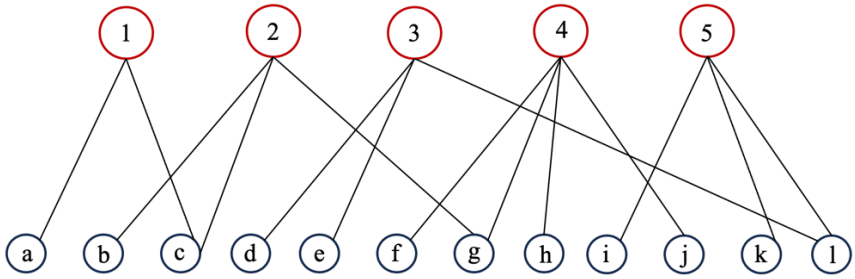


Figure 3.1 - Graphe bipartite des données à l'étude

La mesure de modularité, introduite par *Newman et Girvan* [112], évalue la force des communautés détectées en comparant la densité des liens intra-communautés aux liens inter-communautés. La mesure de modularité est particulièrement avantageuse, car elle est indépendante de l'algorithme et fournit une base de comparaison cohérente entre les différentes méthodes de détection des communautés. Elle permet d'évaluer la qualité de la structure de la communauté produite par des algorithmes. Comme mentionné dans la section 2.2, une multitude d'algorithmes existent afin de détecter des communautés dans un réseau. Toutefois, certains pourraient ne pas être adaptés à un graphe en particulier. Ils ont tous des particularités inhérentes aux graphes. L'algorithme glouton semblait le plus approprié dans le cas présent. Il vise à maximiser la modularité en fusionnant progressivement les communautés afin d'améliorer le score de modularité. Cette méthode est sensible à la structure locale du réseau, ce qui la rend efficace pour détecter à la fois les petites et les grandes structures communautaires, ce qui est essentiel pour les réseaux dont la taille des communautés est potentiellement variée, tout comme l'ensemble de nos données.

Cet algorithme a été implémenté pas l'intermédiaire de la bibliothèque NetworkX de l'environnement Python. Lorsque l'algorithme glouton est invoqué et appliqué à un objet graphique, il fusionne itérativement les communautés afin de maximiser le score de modularité global. Le résultat est une liste dont chaque élément est un ensemble de nœuds représentant une

communauté. Son utilisation a permis d'atteindre un niveau de modularité élevé comme présenté dans le tableau 3.6.

	Mi		GERAD	
	# communautés	Modularité	# communautés	Modularité
2021-2023	193	0,9360	33	0,8788
2018-2020	153	0,9127	39	0,8578
2015-2017	130	0,8126	35	0,8799
2012-2014	140	0,9298	31	0,8917
2009-2011	80	0,8444	24	0,8707

Tableau 3.6 - Comparatif des résultats obtenu de la détection des communautés pour Mi et les cahiers du GERAD pour chacune des périodes

Une valeur élevée de modularité indique une structure communautaire forte, où les nœuds d'un même groupe ont plus de connexions entre eux qu'avec des nœuds de groupes différents. Cependant, la taille des communautés peut avoir un impact sur l'interprétation des résultats. Nous aurons la chance d'explorer davantage dans la section suivante.

4 Résultats et analyses

Dans ce chapitre, nous effectuons une analyse comparative de deux ensembles de données à travers trois sections distinctes, chacune conçue pour répondre à une question de recherche spécifique. Dans un premier temps, nous allons utiliser des techniques d'analyse textuelle telles que les nuages de mots, PyLDAvis, t-SNE et les scores de cohérence. L'objectif est d'évaluer quantitativement et visuellement les thèmes dans chaque ensemble de données et de comparer leurs structures et leur contenu. Deuxièmement, nous explorons les communautés au sein des ensembles de données, en nous concentrant sur les caractéristiques des graphes, l'évolution et la taille des communautés à travers les différentes périodes. Cette analyse cherche à identifier les variations attribuables à leurs différentes portées et déterminer ce que Mi peut exploiter des stratégies d'engagement communautaire observées dans les cahiers du GERAD. Pour finir, nous combinerons les résultats des deux questions de recherche précédentes pour identifier les corrélations entre les thèmes et les structures des communautés dans les ensembles de données, afin de comprendre comment ces éléments interagissent et affectent chaque ensemble de données.

4.1 Modélisation des thèmes

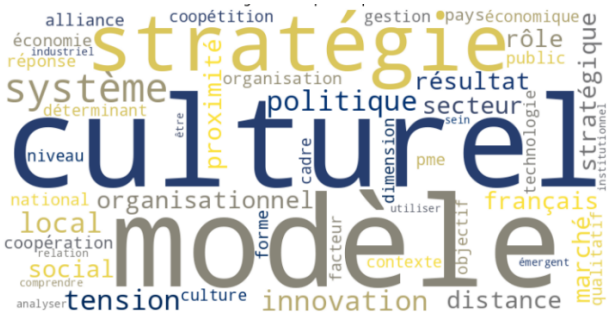
Notre approche consiste à utiliser des nuages de mots pour une représentation visuelle de la fréquence des mots, puis à analyser le poids des mots pour déterminer l'importance des mots dans le corpus. En utilisant PyLDAvis, nous visons à interpréter les modèles thématiques de manière interactive. L'algorithme t-SNE aide à visualiser les données à haute dimension, facilitant ainsi le discernement des regroupements de données. Enfin, nous appliquons des scores de cohérence pour valider la pertinence et le caractère distinctif des thèmes identifiés dans notre analyse textuelle.

4.1.1 Visualisation

Nuages de mots

Pour faciliter la compréhension, le modèle LDA a généré une liste de mots clés pondérés, indiquant leur pertinence au sein du corpus. Cette liste a ensuite été utilisée pour créer des nuages de mots, qui offrent une représentation visuelle intuitive des thèmes prédominants. Les mots de plus grande importance sont mis en évidence par une taille de police plus imposante, mettant en évidence les mots plus pertinents. La création d'un nuage de mots offre une représentation visuelle améliorée de l'importance et de la fréquence des différents mots qui définissent la structure thématique de la revue Management international (figure 4.1) et des cahiers du GERAD (figure 4.2). Nous avons opté pour cette méthode initiale car elle permet de saisir immédiatement les principaux thèmes. Le nom des thèmes et les détails de notre approche seront développés plus en profondeur dans la section 4.1.2.

Nuage de mot du thème 1

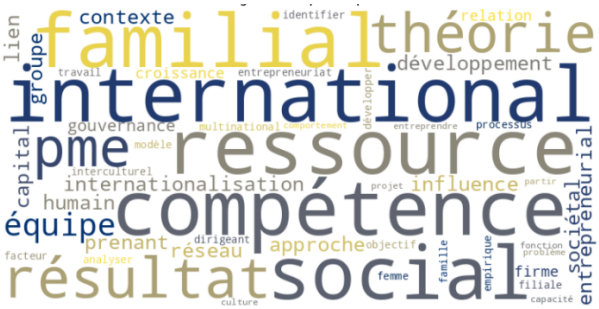


Nuage de mot du thème 2

Nuage de mot du thème 3

Nuage de mot du thème 4

Nuage de mot du thème 5

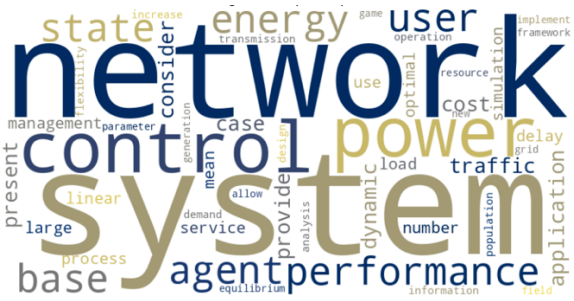


Nuage de mot du thème 6

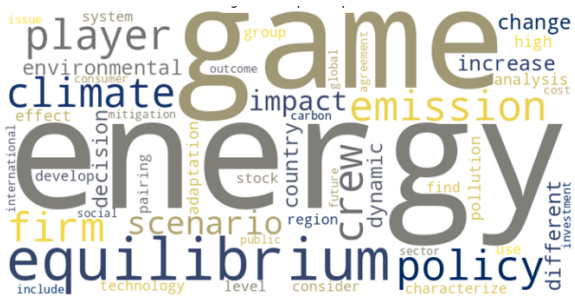
Nuage de mot du thème 7

Figure 4.1 - Nuage de mot de la revue Mi pour les différents thèmes découvert avec le modèle LDA

Nuage de mot du thème 1



Nuage de mot du thème 2



Nuage de mot du thème 3

Nuage de mot du thème 4

Nuage de mot du thème 5

Nuage de mot du thème 6

Nuage de mot du thème 7

Figure 4.2 - Nuage de mot des cahiers du GERAD pour les différents thèmes découverts avec le modèle LDA

Les nuages de mots des deux corpus présentent des caractéristiques linguistiques et thématiques distinctes. Le nuage de mots de la revue Mi englobe des termes socioculturels et de gestion stratégique plus large tels que « culture », « stratégie » et « pays », ce qui suggère une focalisation sur les relations internationales et stratégiques. En revanche, le nuage de mots des cahiers du GERAD présente des termes plus spécialisés tels que « network », « stochastic », « heuristic » et « demand » qui s'alignent sur des domaines de recherche technique tels que la recherche opérationnelle et l'optimisation des réseaux suggérant un accent sur l'analyse quantitative et la résolution de problèmes dans des domaines tels que la logistique et l'ingénierie. Ce contraste est le reflet des différents contextes ou Mi s'adresse à un public de management axé sur un contexte de discussion et d'analyse plus large tandis que les cahiers du GERAD ciblent une communauté de chercheurs spécialisés centré sur la résolution de problème et l'analyse technique.

LDAvis

LDAvis est un outil de visualisation interactif qui capture les relations thématiques et les distributions modélisées par le modèle LDA. Il est illustré comme un plan bidimensionnel ou chaque cercle représente un thème distinct, la taille du cercle indiquant sa prévalence du thème dans le corpus. La proximité entre les cercles indique une similarité thématique tandis que ceux qui se chevauchent peuvent indiquer que le modèle rencontre des difficultés à distinguer les thèmes. La distribution équilibrée des thèmes dans le modèle à sept thèmes, comme le montre la figure 4.3 et la figure 4.4 indique un contexte cohérent, chaque thème occupant un espace discursif distinct.

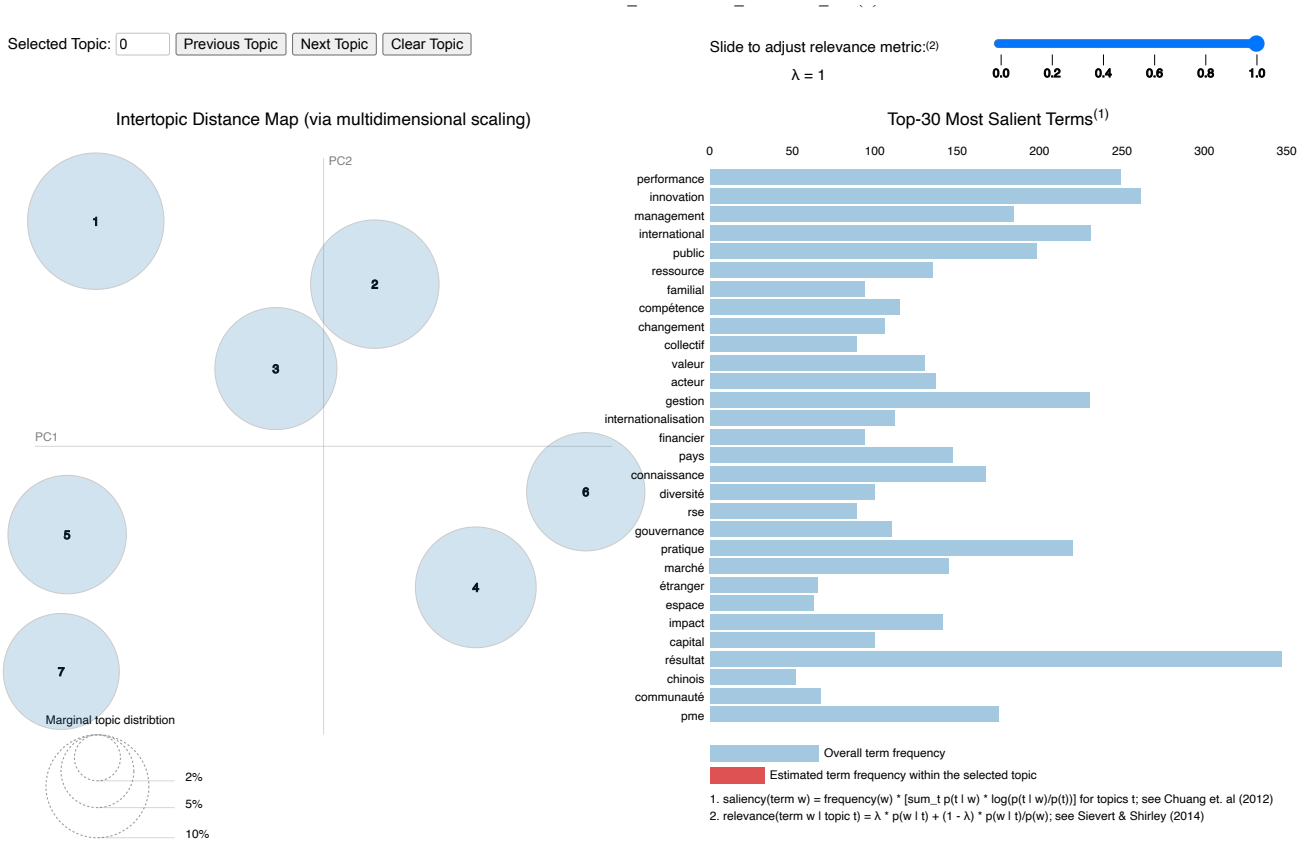


Figure 4.3 - Visualisation PyLDAvis de la revue Mi

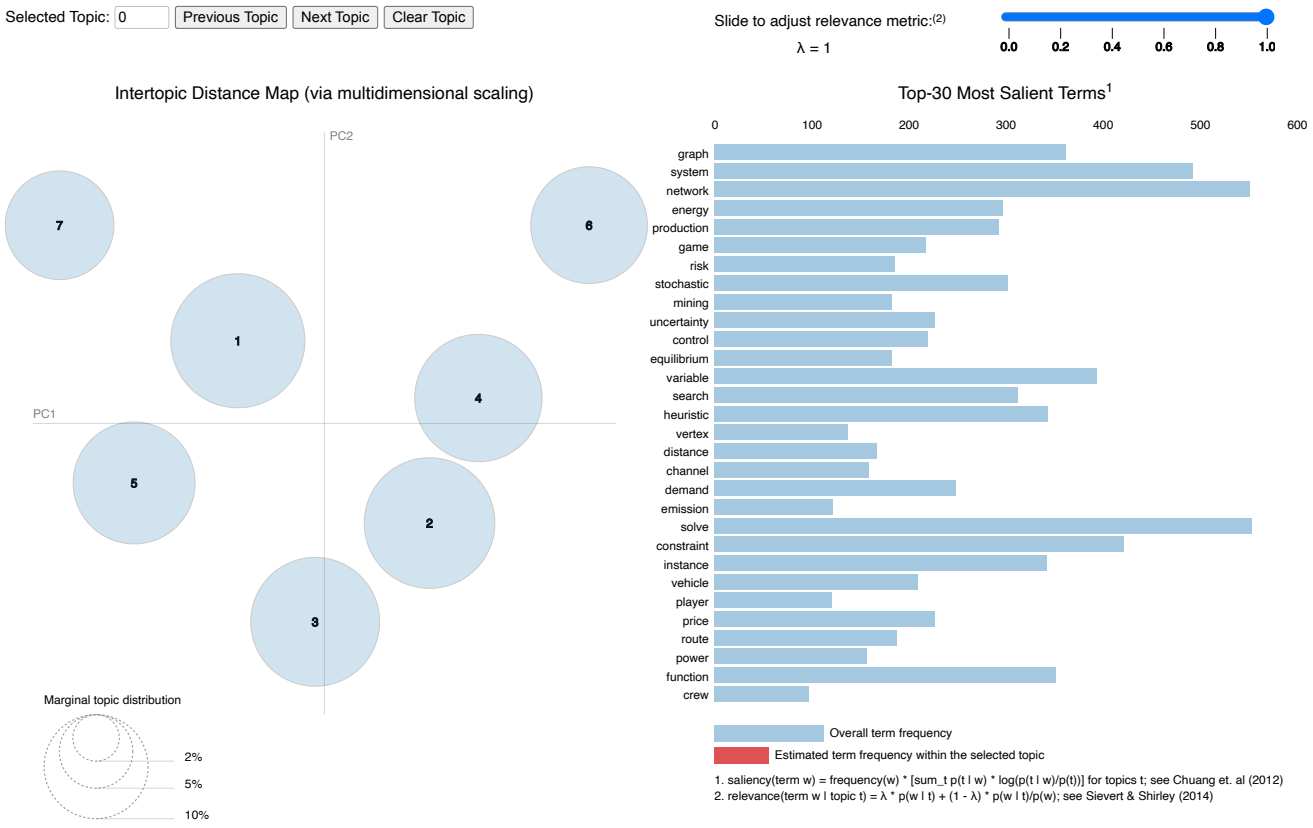


Figure 4.4 - Visualisation PyLDAvis des cahiers du GERAD

La visualisation des cahiers du GERAD valide nos observations précédentes sur la prévalence des mots techniques et quantitatifs dans les documents. La revue Mi se concentre sur les termes associés à la gestion et à l'entrepreneuriat. Cette visualisation globale permet de comprendre rapidement nos ensembles de données et, surtout, de déterminer si le nombre de thèmes choisi est approprié. Dans ce cas, le terme *approprié* se rapporte à la taille des thèmes et à leur éventuel chevauchement. Nous visons un équilibre thématique qui, après avoir décidé du nombre de thèmes de façon itérative, semble être le plus logique pour les deux ensembles de données.

t-SNE

t-SNE est une technique de réduction de dimensionnalité qui permet une représentation visuelle plus compréhensible. Pour visualiser la distribution des mots entre les thèmes dans nos modèles LDA, nous avons extrait la matrice de distribution mot-thème de chaque modèle. Nous avons normalisé cette distribution pour que la somme des distributions de chaque mot soit égale à un, afin de maintenir l'influence proportionnelle de chaque thème par mot. Le modèle t-SNE a ensuite été initialisé. Chaque point de ces représentations 2D symbolise le thème dominant pour chaque mot en identifiant la valeur la plus élevée dans la distribution mot-thème pour ce mot. Les points situés à proximité suggèrent une petite distance locale dans l'espace multidimensionnel original, ce qui implique une forte probabilité de vocabulaire partagé. Tandis que regroupement qui sont plus dispersés, suggère une plus grande variation au sein du thème.

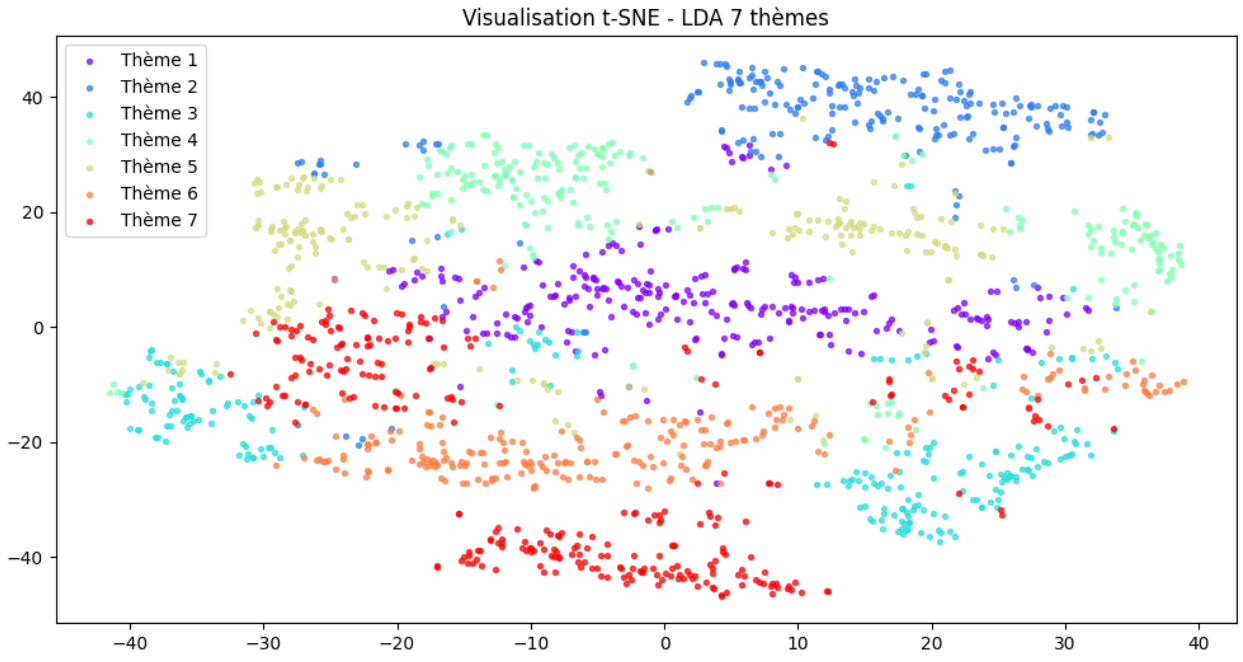


Figure 4.5 - Représentation t-sne de la revue Mi pour un modèle à 7 thèmes

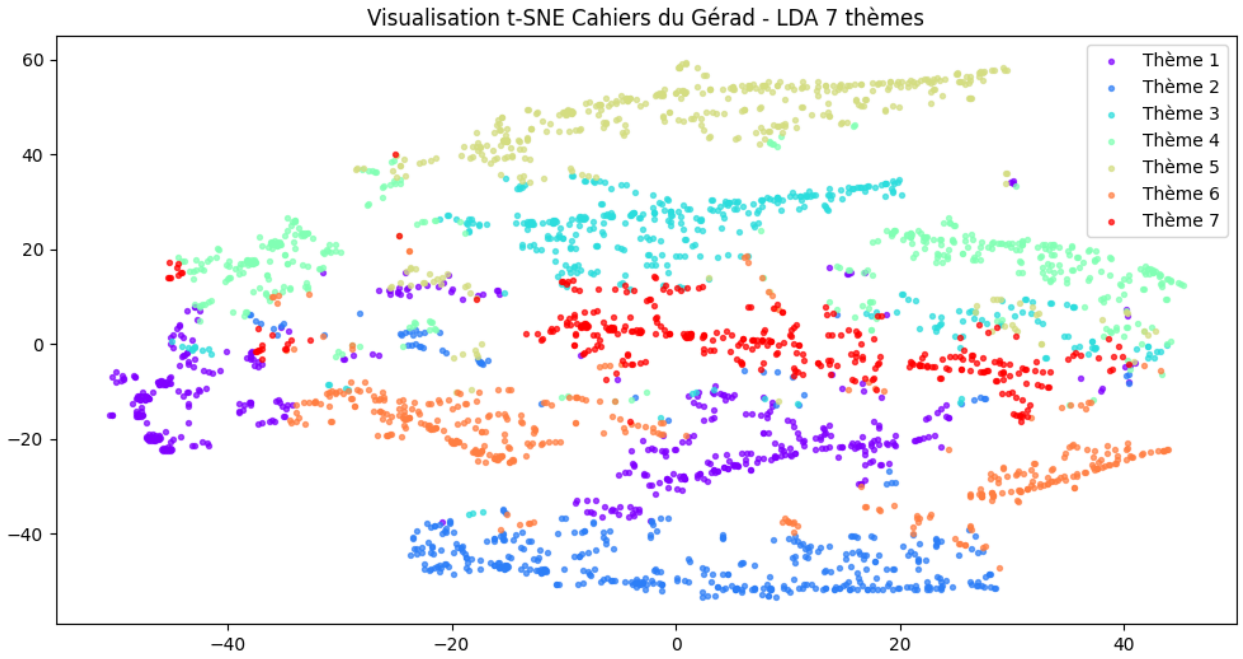


Figure 4.6 - Représentation t-sne des cahiers du GERAD pour un modèle à 7 thèmes

La visualisation de la revue Mi présente un modèle de regroupement dispersé, ce qui indique un ensemble de données caractérisé par une plus grande diversité thématique au sein de ses thèmes. En revanche, les cahiers du GERAD présentent des regroupements plus denses, ce qui suggère un ensemble de données présentant un degré plus élevé de cohérence thématique et un champ d'application plus étroit au sein de ses thèmes. Le chevauchement plus important des regroupements de la revue Mi indique que les thèmes sont moins distinctement séparés, alors que les cahiers du GERAD montrent des regroupements plus délimités, ce qui implique des thématiques plus claires. La compacité des regroupements dans les cahiers du GERAD suggère des thèmes avec des mots étroitement liés, tandis que les regroupements dispersés de la revue Mi reflètent des thèmes plus larges et des associations de mots plus hétérogènes. Cette comparaison met en évidence le fait que les caractéristiques inhérentes à chaque ensemble de données influencent considérablement la structure et l'interprétation des thèmes dérivés de la modélisation LDA.

4.1.2 Évaluation du modèle

Scores de coherence

Une fois que le modèle a accompli ces tâches, il est nécessaire de l'évaluer. La perplexité est une mesure prédictive utilisée pour évaluer les modèles probabilistes, en évaluant leur capacité à se généraliser globalement à des données non observées. Une faible valeur de perplexité valide un bon modèle génératif, mais ne garantit pas l'interopérabilité et ne permet pas de filtrer les thèmes incohérents ou redondants. L'étude de *Chang et al.* [177] suggère que l'évaluation des modèles devrait se concentrer sur la performance des tâches réelles, y compris l'annotation humaine, car la perplexité est souvent en corrélation négative avec la qualité des thèmes. En raison de ces limites, des mesures de cohérence ont été développées et ces c'est celles-ci que nous allons évaluer. Elles reposent essentiellement sur le fait que les paires de mots descripteurs de thèmes qui cooccurrent fréquemment ou qui sont proches les uns des autres dans un espace sémantique sont susceptibles de contribuer à des niveaux de cohérence plus élevés.

Newman et al [150] ont évalué les méthodes de classement automatique des thèmes en comparant les évaluations humaines aux meilleurs ensembles de mots. L'étude s'est faite selon une échelle de pertinence a 3 niveaux: bonne, neutre, mauvaise. Ils ont constaté que les mesures de cohérence, basées sur l'information mutuelle ponctuelle (PMI), présentaient la corrélation la plus élevée avec les évaluations humaines.

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} .$$

Les probabilités de p(w_i, w_j) sont calculées à l'aide des bases de connaissances externes et ϵ sert de constante de lissage.

Mimno et al. [178] ont proposé une probabilité conditionnelle lissée comme mesure de confirmation asymétrique pour les paires de mots supérieures, en incorporant la cohérence UMass pour tenir compte de l'ordre des mots [150].

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j + \epsilon)}{P(w_j)} .$$

Les probabilités des mots sont calculées en estimant la fréquence des documents originaux utilisés pour l'apprentissage des thèmes.

Aletras et Stevenson [179] ont introduit la cohérence thématique en utilisant des vecteurs de contexte pour les mots les plus importants. Les comptes de cooccurrence des mots sont utilisés pour créer des vecteurs de contexte, qui sont constitués de ± 5 jetons autour du mot. La corrélation la plus forte avec les évaluations humaines de la cohérence thématique est observée lorsque ces vecteurs sont définis comme des PMI normalisés (NPMI). La restriction des cooccurrences de mots aux mêmes thèmes donne les meilleurs résultats [150].

$$v_{ij} = NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-log(P(w_i, w_j) + \epsilon} \right)^{\gamma}$$

Röder et al. [180] ont introduit un cadre unificateur pour les mesures de cohérence, appelée cohérence Cv, visant à améliorer la corrélation avec les jugements humains par le biais d'une composition de parties [151]. Le nouveau cadre unificateur des mesures de cohérence se compose de quatre dimensions : S, M, P et Σ. La première dimension est le type de segmentation utilisé pour diviser un ensemble de mots en morceaux plus petits, S. La deuxième dimension est la mesure de confirmation, M, qui évalue la concordance d'une paire donnée, P. La troisième dimension est la méthode d'estimation des probabilités des mots, P. La quatrième dimension est la méthode d'agrégation des valeurs scalaires calculées par la mesure de confirmation, Σ. Le cadre définit un espace de configurations qui est le produit croisé de ces quatre ensembles [150]. Dans le cadre de

ce mémoire, c'est cette approche qui a été appliquée et les résultats se retrouvent dans le tableau 4.1.

	Score de cohérence	
Nombre de Thèmes	**Mi**	**Cahier du GERAD**
4	0,2347	0,3067
5	0,2398	0,3037
6	0,2293	0,3295
7	**0,2404**	**0,3294**
8	0,2340	0,3306
9	0,2350	0,3191
10	0,2282	0,3165
12	0,2329	0,3148
16	0,2323	0,3225
20	0,2290	0,3199

Tableau 4.1 - Score de cohérence des modèles selon le nombre de thèmes pour Mi et les cahiers du GERAD

L'évaluation de la cohérence d'un ensemble de mots sans expertise dans le domaine présente des défis. Afin de discerner la bonne structure thématique, plusieurs modèles LDA ont été implémentés avec un nombre différent de thèmes allant de 4 à 20. La sélection du nombre de thèmes a été déterminée sur la base du contenu existant sur les sites web de la revue Mi et des cahiers du GERAD. Cette approche visait à la fois à identifier les thèmes pertinents et à rationaliser le nombre de sous-catégories. Un maximum de 20 thèmes a été fixé pour éviter les thèmes trop vastes et maintenir une segmentation claire. Un minimum de quatre thèmes garantit que la catégorisation reste suffisamment détaillée pour que les chercheurs et les scientifiques puissent l'associer avec précision à leurs articles. Les résultats démontrent que les cahiers du GERAD présentent une plus grande cohérence que ceux du Mi. La cohérence de Mi étant de sept thèmes, ce qui correspond à nos analyses visuelles antérieures. La collaboration d'experts a permis d'affiner l'identification des thèmes. Malgré des scores de cohérence plus élevés pour six et huit thèmes pour les cahiers du GERAD, la validation des experts et l'analyse thématique ont favorisé le modèle à sept thèmes. Toutefois, on peut observer que dans les deux ensembles de données, la variation des scores de cohérence est subtile, ce qui peut suggérer une structure thématique qui n'est pas clairement définie.

Nombre de documents par thèmes

S'appuyant sur les résultats des visualisations de la modélisation des thèmes et des évaluations de la cohérence, les graphiques à barres suivants représentent quantitativement la répartition des documents par thème. Ils mettent en lumière la prédominance thématique au sein de chaque ensemble de données.

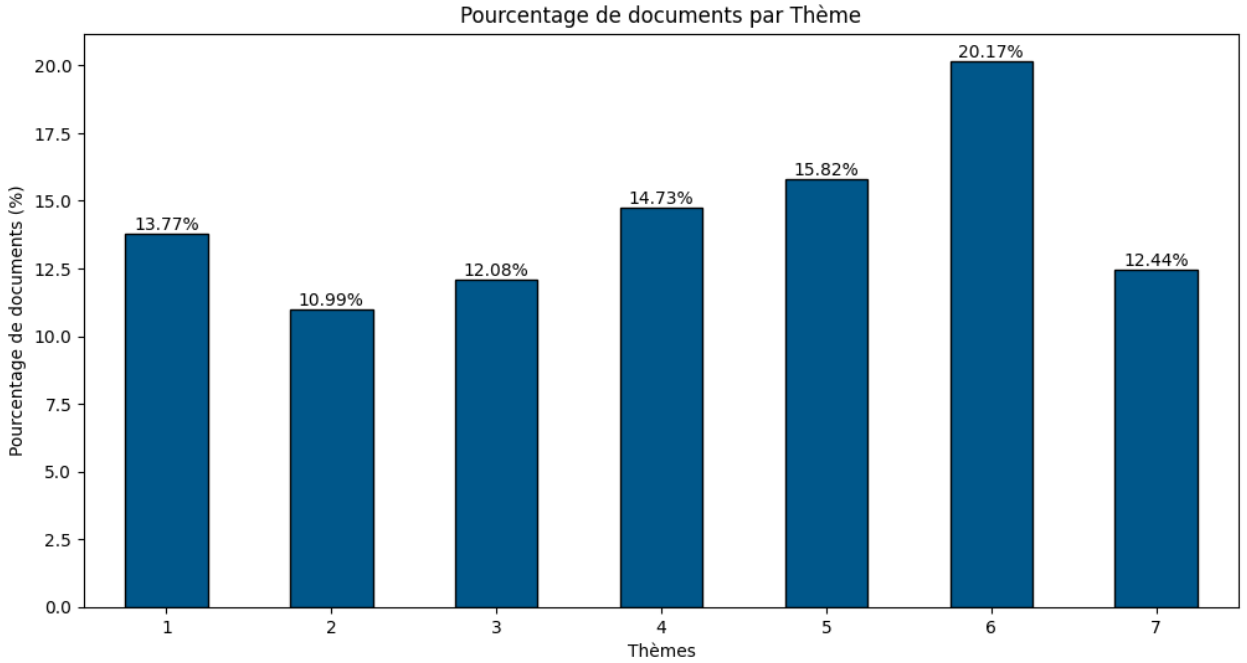


Figure 4.7 – Répartition des pourcentages de documents pour chaque thème de la revue Mi

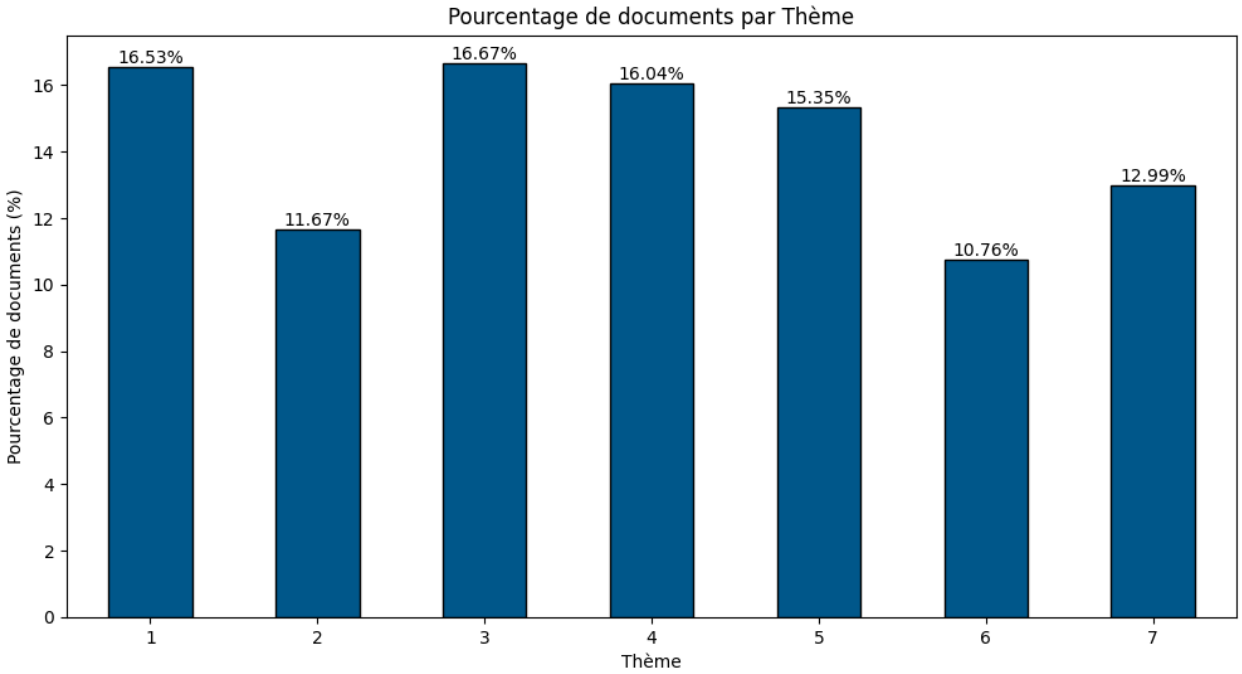


Figure 4.8 – Répartition des pourcentages de documents pour chaque thème des cahiers du GERAD

La distribution des documents de la revue Mi suggère une répartition équitable entre sept thèmes, le sixième thème apparaissant comme le plus prédominant. Inversement, le deuxième thème affiche le pourcentage le plus faible, ce qui signifie un thème moins développé. Dans les cahiers du GERAD, la répartition est uniformément comparable, mais les thèmes un et trois sont plus prononcés, ce qui peut mettre en évidence des domaines d'intérêt potentiels. Le sixième thème est le moins représenté, ce qui suggère un champ d'application plus restreint. La variation des thématiques entre ces ensembles de données, qui proviennent de domaines totalement différents, nous informe sur la profondeur et la concentration thématique inhérentes à chaque corpus.

Thèmes découverts

Sur la base de nos observations, nous sommes maintenant en mesure de finaliser cette analyse en présentant une représentation thématique des corpus, avec des thèmes bien définis pour la revue Mi et les cahiers du GERAD. Il reste à évaluer si les thèmes identifiés correspondent au contenu des articles ou à l'orientation thématique prévue.

Thèmes pour la revue Mi

Pour le thème 1, les termes *modèle* et *culturel* apparaissent avec le plus de poids, ancrant le thème dans un discours de stratégie d'entreprise tout en se concentrant sur la gestion de la diversité au sein des industries et organisations. Le rapprochement d'entreprise et la stratégie-concurrence font partie intégrante de ce thème. Ainsi, le thème qui émerge est celui de la **Management stratégique.**

Pour le thème 2, les termes *performance, résultat et pays*, faisant référence aux grandes entreprises ainsi qu'aux PME ont visuellement le poids le plus important. Ceci permet d'illustrer l'ampleur du terrain de jeu mondial dans lequel ces entreprises sont confrontées, que ce soit au niveau des risques ou encore des stratégies associés. Dans ce contexte, la communication-négociation est donc intrinsèquement liée aux stratégies marketing qui permettent de naviguer et tirer parti de l'internationalisation. **Marketing et consommation** pourrait donc être un thème approprié.

Pour le thème 3, **Développement durable** apparaît comme étant un thème probable, étant dominé par le mot *gestion privé*/*public et éthique* faisant référence aux aspects plutôt consultatifs, réglementaires et environnementaux inhérents au thème. Le poids important de ces termes confirme leur centralité dans les discussions axées sur les relations avec les parties prenantes au sein des organisations.

Le thème 4, **Finance et gouvernance** est caractérisé par les mots *résultats, valeur et financier*, soulignant l'accent thématique sur l'analyse financière et l'évaluation complexe de ces derniers. Ce thème met de l'avant le rôle des données dans l'élaboration des stratégies d'investissement. Il englobe les politiques économiques, les tendances du marché et les informations financières, qui constituent la pierre angulaire des processus de prise de décision.

Le thème 5, axé sur l'**Entrepreneuriat et PME,** met en évidence le rôle central des mots *compétences intra- entrepreneuriales et international*, tandis que *ressource et capital social* suggèrent une convergence vers l'importance des réseaux de soutien et des relations interpersonnelles dans le succès et entrepreneurial.

Le thème 6 est celui de **l'Innovation, communauté et digitalisation**. Les mots, *processus*, *dynamique, communautés* et *innovation* au premier plan, indiquant une attention particulière à la gestion des connaissances et à l'apprentissage organisationnel en tant que moteurs clés du développement économique. Les technologies de l'information jouent un rôle crucial, servant de catalyseur au processus de l'innovation.

Enfin, le thème 7, ponctué par les mots *management, changement et diversité* soulignent le besoin pour les organisations de s'adapter et d'évoluer en intégrant de nouvelles perspectives et des pratiques diverses. Ce thème met l'accent sur le besoin de résilience pour favoriser la croissance des organisations et s'adapter au changement. Ainsi le thème qui émerge est celui de **Management et GRH**

Thèmes pour les cahiers du GERAD

Le thème 1, **Gestion des systèmes et des infrastructures**, englobe des mots tels que *système*, *réseau*, *contrôle* et *énergie*. Il se concentre sur les aspects techniques de la gestion des réseaux et de l'énergie au sein des systèmes d'infrastructure. Il met l'accent sur la nécessité de mesures de contrôle et sur l'optimisation des performances grâce à une gestion efficace des systèmes.
Le thème 2, **Environnement**, où des mots tels que *jeu, énergie, équilibre* et *émission* renvoient à la prise de décision stratégique dans les secteurs de l'environnement et de l'énergie. Il traite de l'interaction entre les questions environnementales et les stratégies économiques et politiques, en mettant l'accent sur leur impact sur le climat et la dynamique de l'utilisation de l'énergie.

Le thème 3, **Optimisation opérationnelle**, soulignée par les mots tel que *contrainte, résoudre, optimisation* et *variable,* indique un thème centré sur la résolution de problèmes opérationnels et l'amélioration de l'efficacité. Ce thème est probablement axé sur l'application de techniques d'optimisation et de méthodes informatiques pour améliorer les processus logistiques tels que la chaîne d'approvisionnement.

Le thème 4, **Méthodes heuristiques**, avec des mots tels qu'*heuristique, test, optimisation* et *recherche*, qui se rapportent à la résolution de problèmes dans des environnements informatiques.

L'accent est mis sur le développement et le test de solutions algorithmiques dans le domaine des heuristiques.

Le thème 5, **Théorie des graphes**, soulignés par les mots tels que *graphe, nombre, distance* et *sommet*, se concentre sur l'étude des structures mathématiques des graphes. Ce thème met l'accent sur l'utilisation de méthodes numériques et de la théorie des graphes dans diverses applications.

Le thème 6, **Économie**, est caractérisé par des mots tels que *stochastique, production, risque* et *exploitation minière*, qui renvoient à la gestion des risques et à l'incertitude dans les contextes économiques et de production. L'accent est mis sur l'incertitude dans des scénarios de prise de décision complexes au sein de ces systèmes.

Le thème 7, **Logistique**, est centré sur des mots tels que *demande, coût, canal* et *réseau,* soulignant les subtilités de la logistique et de l'allocation des ressources. Ce thème semble se concentrer sur l'économie de la distribution des ressources, en abordant des éléments tels que la prévision de la demande, l'analyse des coûts, les canaux de distribution et l'ordonnancement dans le secteur de la logistique.

L'analyse par LDA a permis de délimiter des thèmes qui correspondent aux domaines spécialisés de chaque corpus. La revue Mi se concentre sur la gestion internationale, avec une expertise dans des domaines tels que la gouvernance, le leadership et les ressources humaines. En revanche, les cahiers du GERAD se concentrent sur les mathématiques appliquées. Cette différenciation thématique confirme l'efficacité du modèle pour l'identification de thèmes dans des disciplines distinctes.

4.2 Détection des communautés

Dans la section suivante, nous explorerons les caractéristiques définissant les communautés au sein du réseau. Nous évaluerons la structure globale du graphe en utilisant des mesures de centralité pour comprendre l'importance des nœuds, et nous utiliserons la modularité pour évaluer la force et la clarté des divisions de la communauté. Nous nous pencherons sur les caractéristiques des auteurs,

en analysant comment leurs contributions et leur centralité au sein du réseau soulignent leur influence et le flux d'informations. Chacun de ces aspects sera abordé par le biais d'une analyse quantitative, en s'appuyant sur les méthodologies vues précédemment.

4.2.1 Caractéristiques des communautés

L'analyse de la détection des communautés au sein d'un graphe bipartite révèle des informations importantes sur la dynamique temporelle et l'évolution des structures des communautés. La revue Mi a commencé avec 80 différentes communautés en 2009 tandis que les Cahiers du GERAD ont commencé avec 28 communautés. La variabilité du nombre de communautés au cours de chaque période démontre leur nature dynamique.

Le nombre de communautés de la revue Mi a considérablement augmenté, particulièrement après la période 2025-2017, tel que nous pouvons l'observer sur la figure 4.9. Ce qui indique un réseau en expansion. D'autre part, le nombre stable de communautés au sein des cahiers du GERAD suggère que les résultats de la recherche et les efforts de collaboration restent cohérents. Cette disparité suggère un critère d'inclusion plus large dans la revue Mi, par opposition à la taille plus constante de la communauté dans les cahiers du GERAD, qui peut être influencée par ses critères de publication basés sur l'adhésion d'au moins un membre.

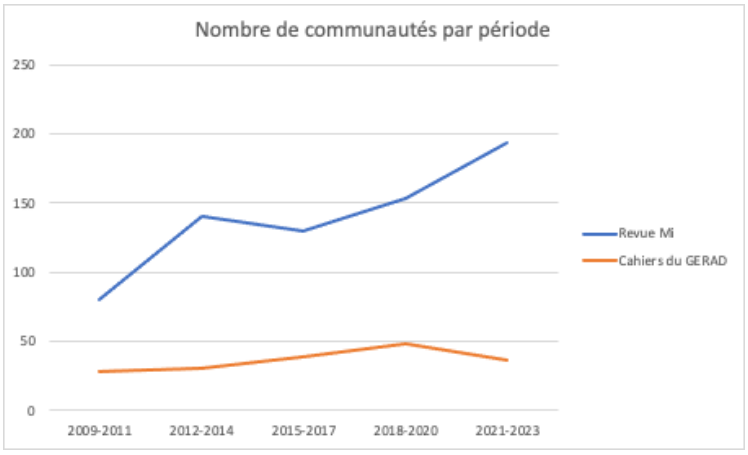


Figure 4.9 - Nombre de communautés par période pour la revue Mi et les cahiers du GERAD

À la figure 4.10 correspondant à la revue Mi, nous observons que les communautés de moins de trois auteurs constituent la majorité. Cette tendance persiste à travers toutes les périodes, suggérant une prédominance de petits groupes d'auteurs au sein de la communauté. La présence de communautés comptant de 4 à 6 et de 7 à 9 auteurs est notable, mais reste secondaire. Les communautés de plus de dix auteurs sont les moins représentées, ce qui indique que les grands groupes de collaboration sont relativement rares. Cette répartition indique une culture de recherche qui penche vers des efforts de recherche individuels ou en petits groupes plutôt que vers des réseaux de collaboration étendus. La nature libre d'accès de la revue peut attirer un large éventail de contributeurs individuels ou de petites équipes qui travaillent de manière indépendante.

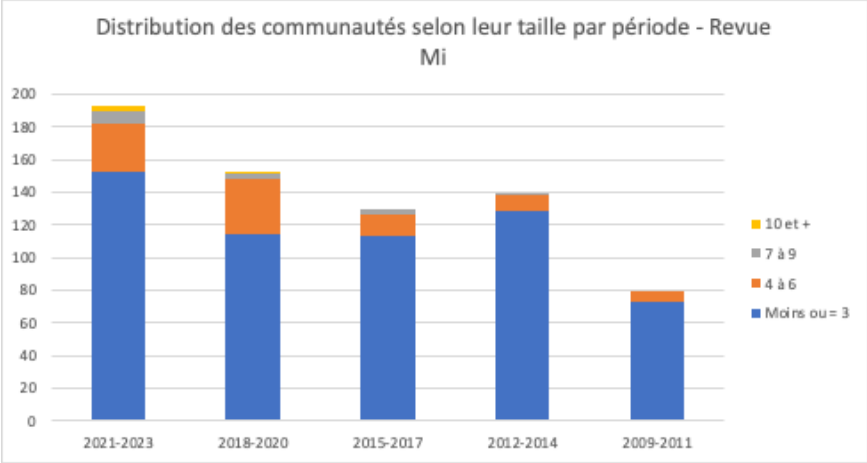


Figure 4.10 - Distribution des communautés selon leur taille par période pour la revue Mi

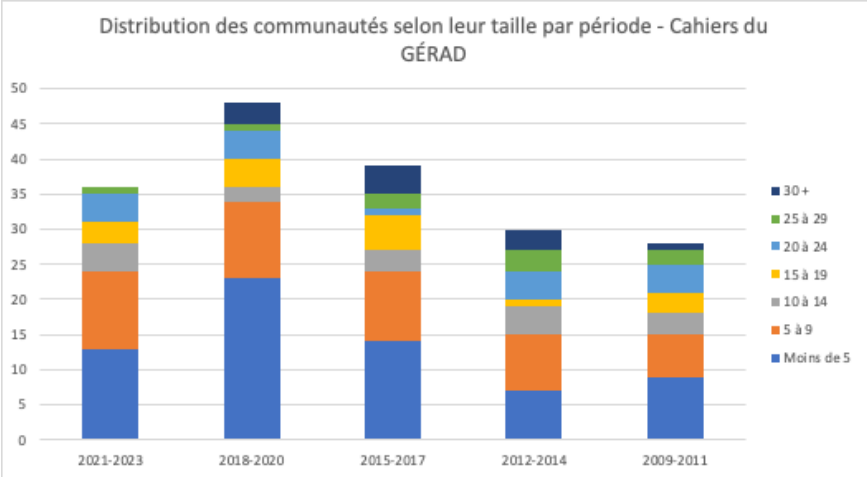


Figure 4.11 - Distribution des communautés selon leur taille par période pour les cahiers du GERAD

En revanche, la figure 4.11 des Cahiers du GERAD présente un schéma différent. Il révèle une distribution plus diversifiée de la taille des communautés, avec une présence notable de communautés comptant plus de 10 auteurs. La présence de groupes aussi importants suggère une tendance à une collaboration plus étendue. La diversité de la taille des communautés, de moins de cinq auteurs à plus de trente, suggère un environnement de collaboration dynamique où l'on trouve à la fois des collaborations à petite et à grande échelle. La prévalence de communautés plus larges dans les Cahiers du GERAD est directement liée aux exigences de collaboration de la recherche mathématique et à la politique de soumission. Cette politique exige que les contributeurs externes soient co-auteurs avec les membres du GERAD, ce qui favorise une collaboration étendue. De telles collaborations sont typiques en mathématiques en raison de la complexité du thème, ce qui donne lieu à des communautés de recherche plus grandes et plus intégrées, axé sur des problèmes mathématiques avancés.

Ces analyses permettent de comprendre le comportement des différentes communautés au sein de nos deux ensembles de données sur des périodes différentes. Des distinctions notables sont observées à la fois dans le nombre de communautés et, plus significativement, dans leur taille. Ces différences peuvent être attribuées principalement à leurs domaines de recherche distincts.

4.2.2 Caractéristiques des auteurs

Il est intéressant d’étudier les auteurs afin de mieux comprendre leur regroupement en communautés. À la figure 4.12, on peut voir le nombre de communautés d’appartenance des différents auteurs pour toutes périodes confondus pour la revue Mi. La politique d’évaluation par ses pairs de la revue Mi et son orientation plus large vers la gestion pourraient attirer un plus grand nombre d'auteurs issus de diverses sous-disciplines de la gestion. Cela pourrait expliquer le pourcentage plus élevé (82%) d'auteurs qui n'appartiennent qu'à une seule communauté. L'étendue du domaine pourrait se traduire par un ensemble plus diversifié d'intérêts de recherche et moins de chevauchement entre les communautés. La gestion peut également impliquer des méthodologies et des approches de recherche distinctes, propres à chaque sous-discipline, ce qui renforce les frontières entre les communautés.

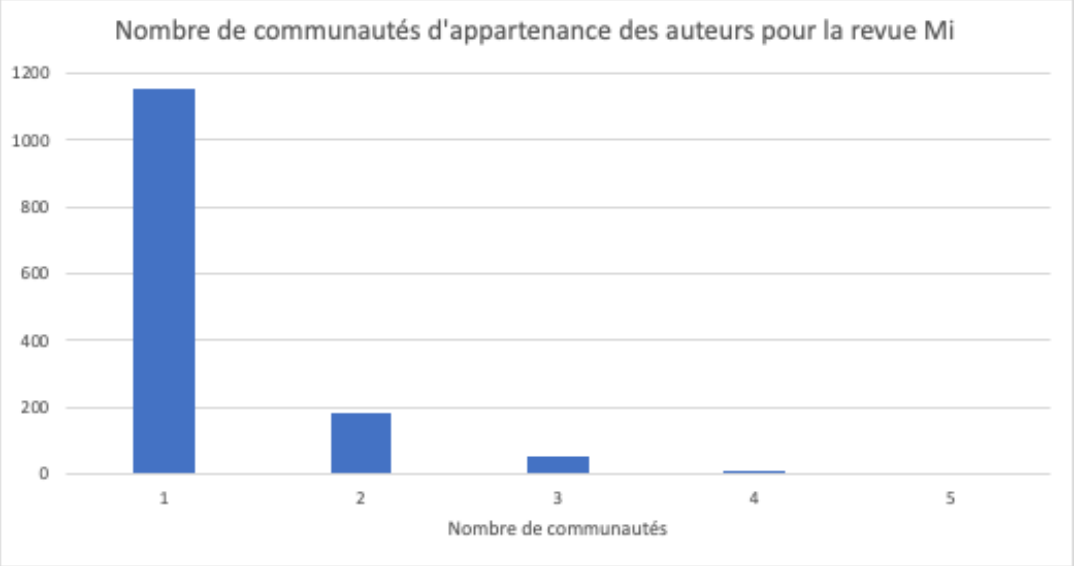


Figure 4.12 - Nombre de communautés d'appartenance des auteurs pour la revue Mi – nonobstant les périodes

La politique des cahiers du GERAD selon laquelle au moins un auteur doit être membre pourrait créer une communauté d'auteurs plus insulaire, car cette exigence limite naturellement la qualité d'auteur à ceux qui font partie d'un certain groupe. Cela pourrait conduire à un degré plus élevé d'interconnexion au sein de la communauté, ce qui pourrait expliquer la raison d'un pourcentage plus faible (70 %) d'auteurs appartenant à une seule communauté, tel qu'observé à la figure 4.13. L'environnement collaboratif nécessaire à la recherche mathématique, qui implique souvent la résolution de problèmes complexes et des travaux théoriques complexes, pourrait contribuer davantage à cette interconnexion.

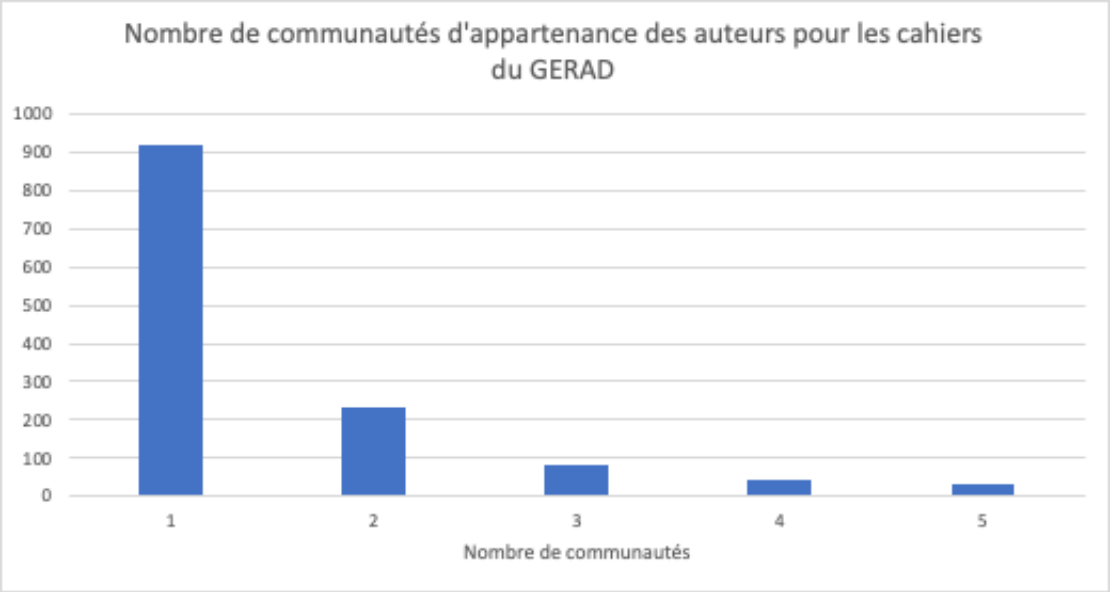


Figure 4.13 - Nombre de communautés d'appartenance des auteurs pour les cahiers du GERAD – nonobstant les périodes

Pour la revue Mi, la figure 4.14 indique une augmentation importante du nombre de nouveaux auteurs pour la période 2012-2014 suivie de quelques fluctuations au fil du temps. Les données suggèrent que Mi a réussi à attirer initialement un grand nombre de nouveaux auteurs, peut-être en raison de son accessibilité et de son orientation plus large vers la gestion, qui pourrait attirer un large public. L'augmentation du nombre d'auteurs qui reviennent au cours des périodes intermédiaires pourrait refléter le développement d'une base d'auteurs fidèles. Toutefois le taux de retour est particulièrement faible. Plusieurs facteurs pourraient influencer cette tendance, tels que l'évolution des intérêts dans le domaine de la gestion pour les auteurs, l'accessibilité, ou même la dynamique plus large de la revue.

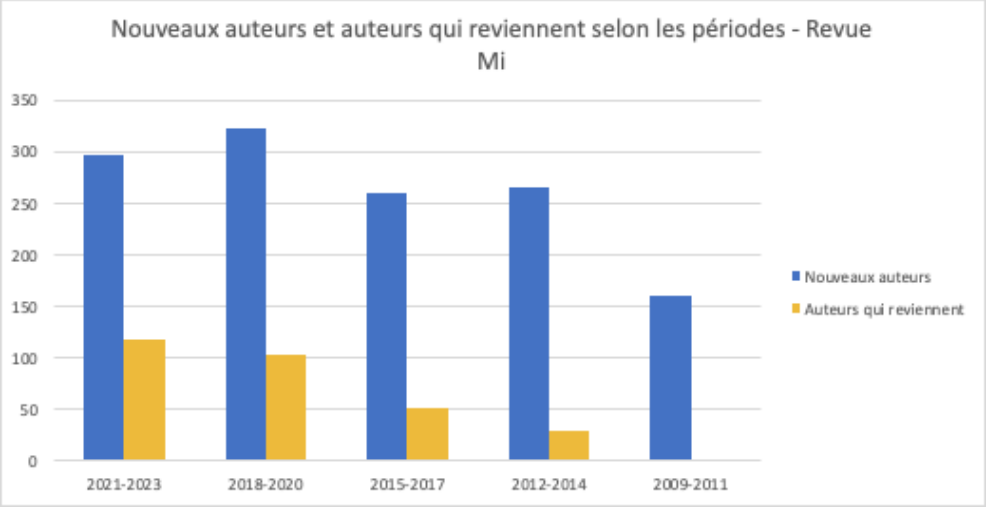


Figure 4.14 - Nombre d'auteurs qui reviennent et nombre de nouveaux auteurs pour la revue Mi

À l'inverse, la figure 4.15 des cahiers du GERAD présente un nombre relativement stable de nouveaux auteurs à travers les périodes, avec une diminution notable dans l'intervalle de 2021-2023. Pour contrer cette baisse, les auteurs qui reviennent viennent dépasser le nombre de nouveaux auteurs pour cette même période. Les auteurs qui reviennent présentent une plus grande variabilité, mais affichent une proportion quasi constante en lien avec les nouveaux auteurs. Compte tenu de la nature spécialisée et de l'obligation d'être membre pour publier, ces données suggèrent un intérêt constant pour les contributions au sein de la communauté. Le pic d'auteurs revenant dans la période la plus récente pourrait être attribué à une culture réussie d'une communauté scientifique engagée et récurrente ou à une augmentation possible du nombre de travaux collaboratifs qui impliquent des contributions répétées de la part de membres existants.

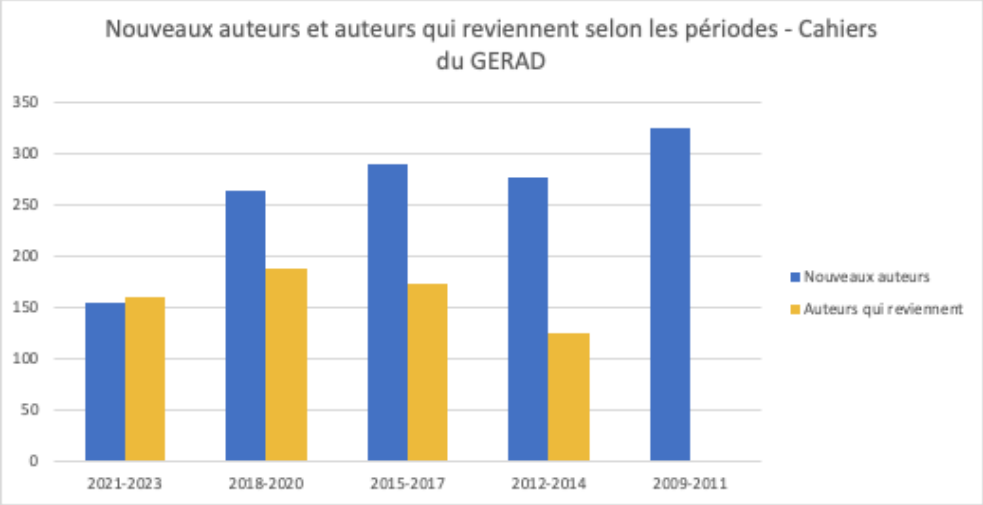


Figure 4.15 - Nombre d'auteurs qui reviennent et nombre de nouveaux auteurs pour les cahiers du GERAD

Pour la revue Mi, les données de la figure 4.16 montrent une tendance cohérente à travers les périodes : les résumés avec un, deux ou trois auteurs sont les plus fréquents, avec un pic à deux auteurs. Cette prévalence indique un niveau modéré de collaboration, ce qui est typique de la recherche en gestion où de petites équipes ou des chercheurs individuels entreprennent souvent des études. Il y a une diminution notable des résumés avec un plus grand nombre d'auteurs, ce qui suggère que les efforts de collaboration importants sont moins fréquents.

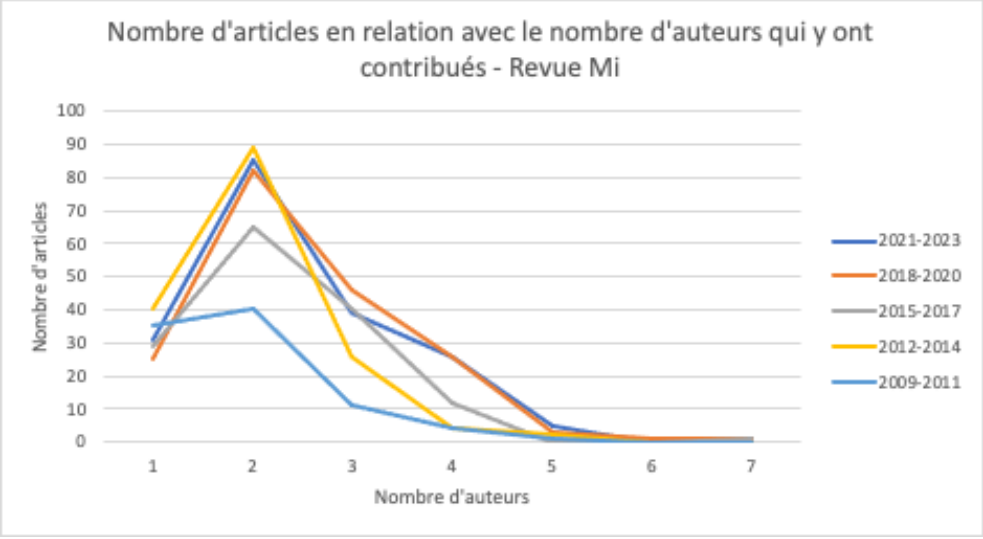


Figure 4.16 - Nombre de résumés en relation avec le nombre d'auteurs qui y ont contribué pour la revue Mi

La figure 4.17 des cahiers du GERAD est particulièrement différente, reflétant un degré de collaboration plus élevé. Le pic de résumés rédigés par deux à quatre auteurs au cours de différentes périodes souligne une tendance à la collaboration conforme à la nature spécialisée et souvent complexe de la recherche mathématique, qui nécessite fréquemment un travail d'équipe. Il est à noter qu'il existe également un nombre petit, mais non négligeable de résumés rédigés par six auteurs ou plus, ce qui indique l'existence de projets de collaboration importants.

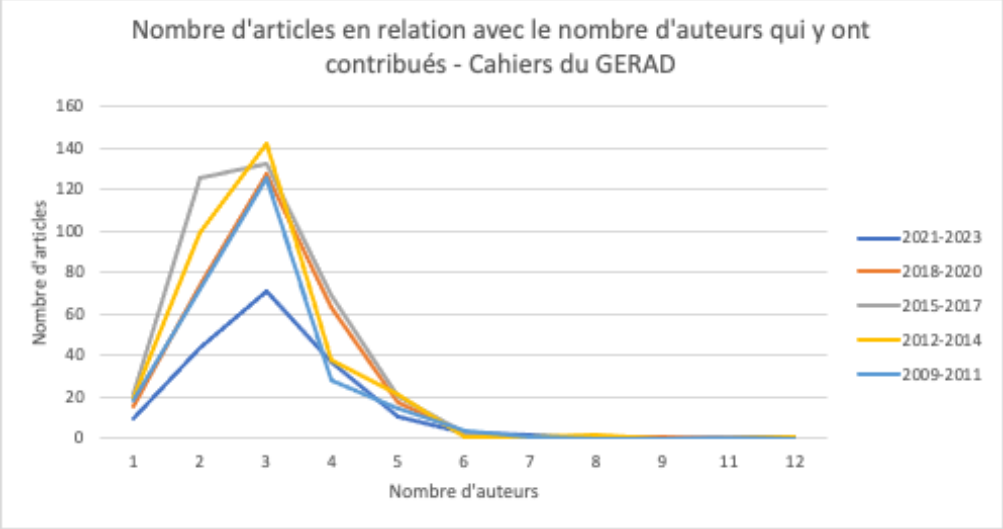


Figure 4.17 - Nombre de résumés en relation avec le nombre d'auteurs qui y ont contribué pour les cahiers du GERAD

Ces tendances doivent prendre en considération les différents champs d'application des deux ensemble de données, les disciplines auxquelles elles s'adressent et les barrières potentielles à l'entrée pour les nouveaux auteurs. En outre, le processus de soumission et les politiques éditoriales sont susceptibles d'affecter la dynamique de la paternité des articles. Processus qui est particulièrement plus difficile pour les cahiers du GERAD que pour la revue Mi et qui est reflétée dans les analyses. Toutefois, nos observations confirment que les cahiers du GERAD devraient présenter davantage de cas de collaboration que la revue Mi. Cette compréhension est cohérente avec les disciplines qu'ils représentent et leurs politiques de publication.

4.2.3 Évaluation des réseaux

Nous avons brièvement présenté à la section 3.2 les caractéristiques des différents réseaux. Sur cette base nous allons maintenant nous attarder sur les nœuds et les arêtes qui les composent et trouver les liens qui les unissent. Dans le contexte des données, les nœuds représentent des entités individuelles au sein du réseau, qui peuvent être des résumés ou des auteurs. Tandis que les arêtes signifient les relations entre ces nœuds. La relation entre les nœuds et les arêtes donne une idée de la complexité et de la connectivité du réseau.

	Mi				**GERAD**			
	# noeuds	# arêtes	# résumés	# auteurs	# noeuds	# arêtes	# résumés	# auteurs
2021-2023	782	596	245	537	493	563	178	315
2018-2020	610	459	184	426	752	916	301	451
2015-2017	460	334	147	313	834	1082	373	461
2012-2014	456	317	160	296	725	941	324	401
2009-2011	252	174	92	160	587	753	263	324

Tableau 4.2 - Attribut de la structure du graphe bipartite de la revue Mi et des cahiers du GERAD

Les graphes à l'étude sont répartis selon les périodes de trois ans. On se retrouve donc avec cinq graphes pour la revue Mi et cinq graphes pour les cahiers du GERAD. Ce dernier présente une expansion significative, avec des nœuds passant de 587 en 2009-2011 à 834 en 2015-2017, et des arêtes passant de 743 à 1082 au cours des mêmes périodes. En revanche, le réseau de Mi a connu une croissance plus modeste, avec des nœuds passant de 252 à 460 et des arêtes passant de 174 à 334 pour les mêmes périodes.

Nous observons que Mi avait 610 nœuds et 459 arêtes en 2018-2020, cela suggère que le réseau de résumés ou d'auteurs était relativement interconnecté, avec un nombre de connexions proche du nombre de nœud dans le graphe. Cependant, le rapport entre les arêtes et les nœuds est inférieur à 1, ce qui peut impliquer que chaque nœud n'est pas connecté à tous les autres nœuds. Un indicateur d'un réseau de collaboration plus sélectif ou moins dense reflétant la nature appliquée de la recherche en gestion. C'est le cas de tous les graphes pour chacune des périodes.

En revanche, les cahiers du GERAD, qui se concentre sur les mathématiques et la science de la décision, présente un schéma différent. Par exemple, en 2015-2017, on compte 834 nœuds et 1 082 arêtes, ce qui indique un nombre plus élevé de connexions par rapport aux nœuds. Cela suggère un réseau plus dense avec un degré plus élevé d'interconnectivité entre les nœuds. Cela reflète probablement un environnement collaboratif fort où de nombreuses entités sont interconnectées, ce qui est courant dans la recherche mathématique où les problèmes sont souvent résolus grâce à des efforts de collaboration. Pour chacune des périodes, le nombre d'arêtes est plus élevé que le nombre de nœuds des graphes.

Comparativement, les données de chacun des graphes par période suggèrent que le réseau des cahiers du GERAD est plus interconnecté et plus complexe que celui de Mi comme nous pouvons le voir dans les figures 4.18 à 4.22 suivantes.

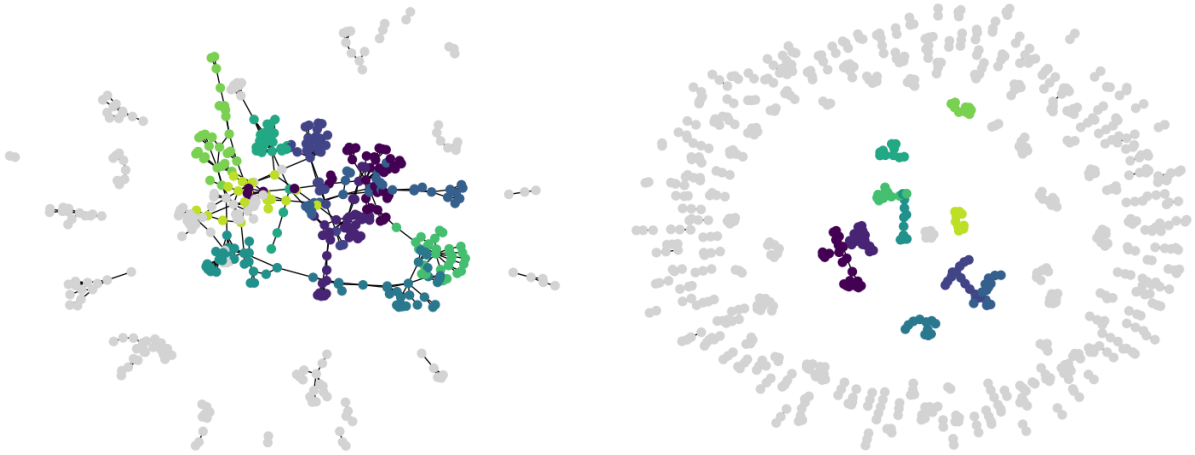


Figure 4.18 - À droite les 10 communautés les plus importantes en termes de taille de la revue Mi et à gauche les 10 communautés les plus importantes en termes de taille des cahiers du GERAD pour la période 2021-2023

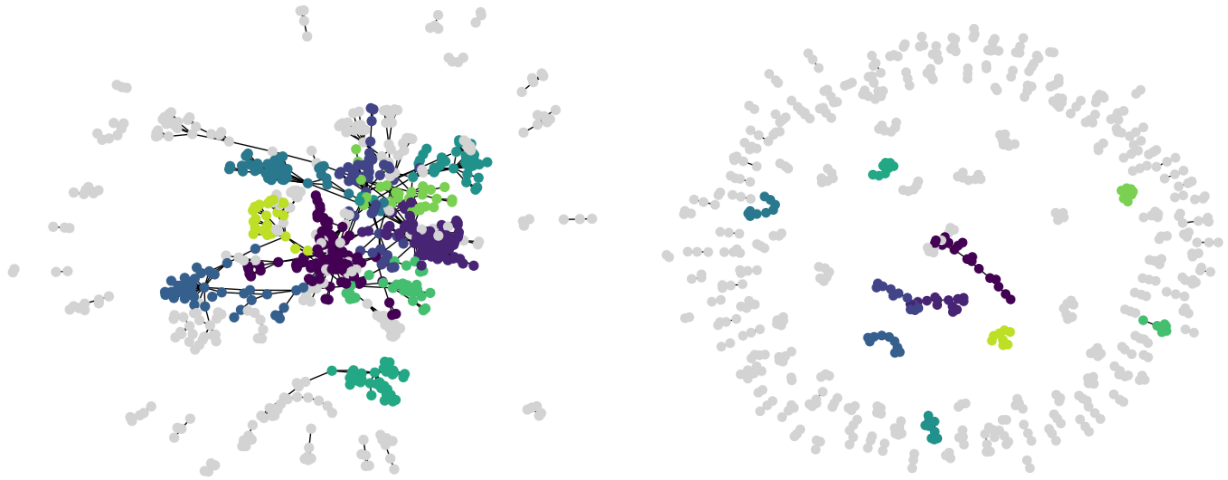


Figure 4.19 - À droite les 10 communautés les plus importantes en termes de taille de la revue Mi et à gauche les 10 communautés les plus importantes en termes de taille des cahiers du GERAD pour la période 2018-2020

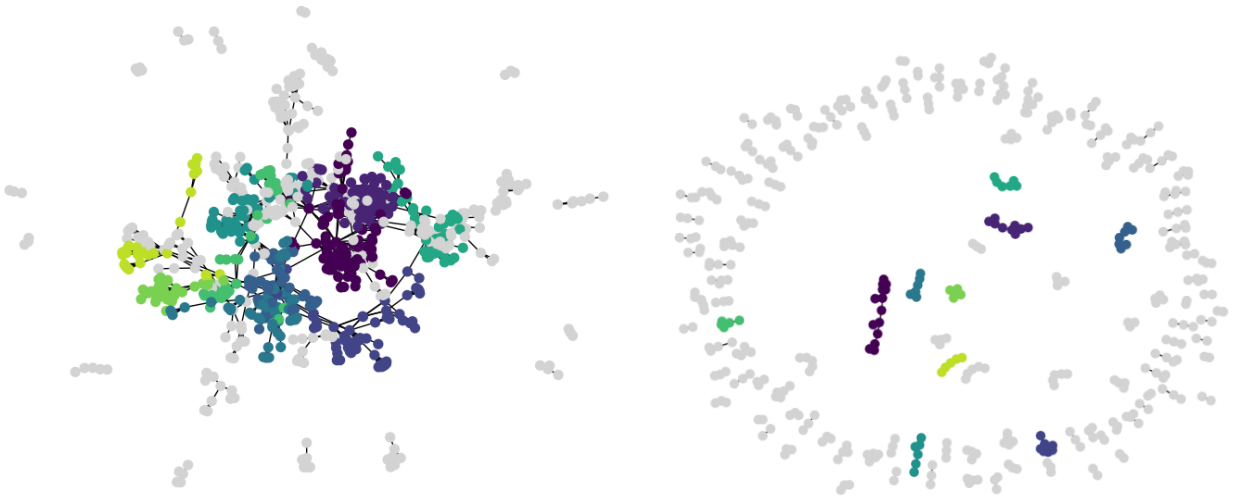


Figure 4.20 - À droite les 10 communautés les plus importantes en termes de taille de la revue Mi et à gauche les 10 communautés les plus importantes en termes de taille des cahiers du GERAD pour la période 2015-2017

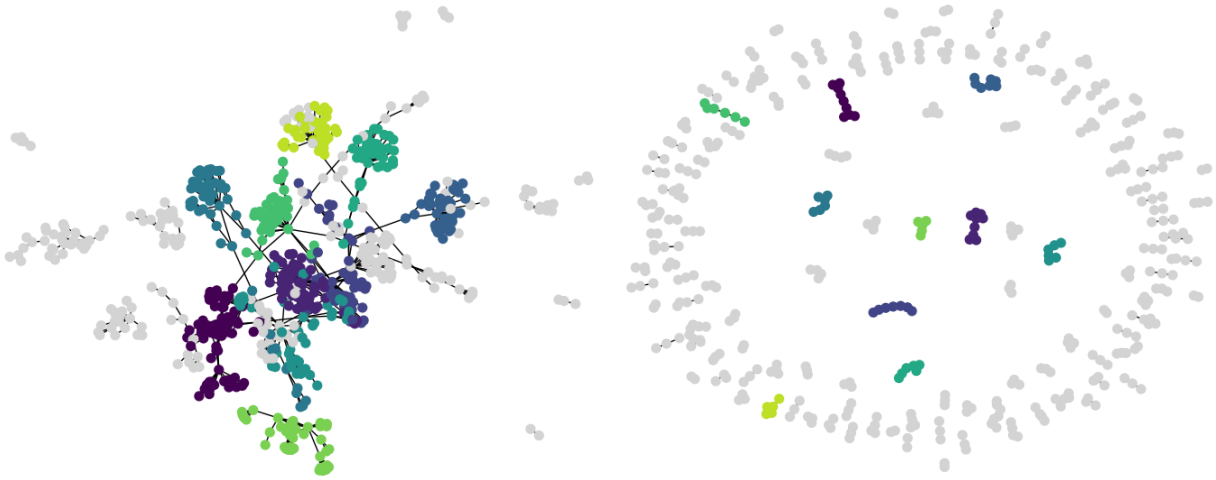


Figure 4.21 - À droite les 10 communautés les plus importantes en termes de taille de la revue Mi et à gauche les 10 communautés les plus importantes en termes de taille des cahiers du GERAD pour la période 2012-2014

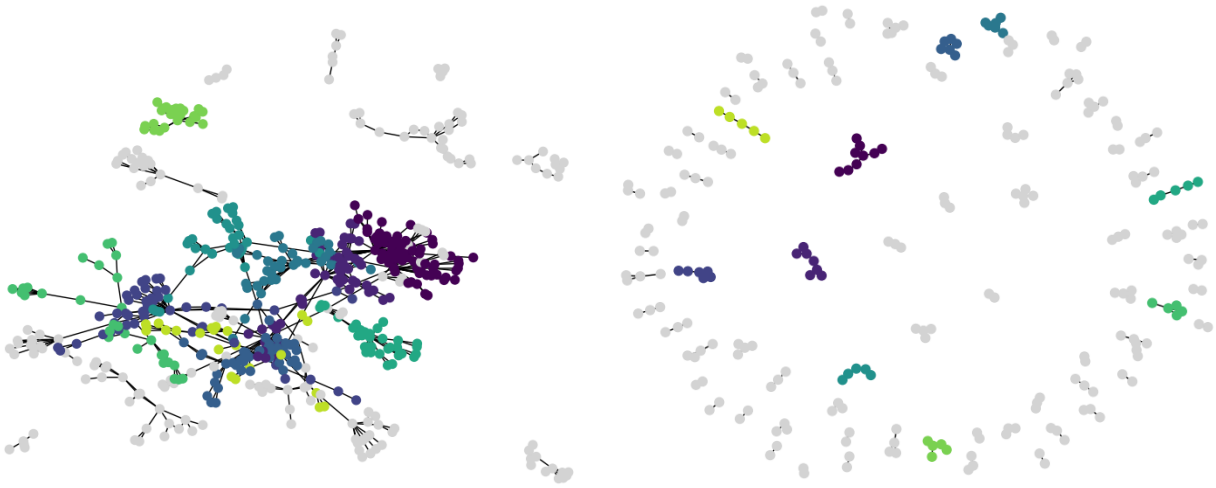


Figure 4.22 - À droite les 10 communautés les plus importantes en termes de taille de la revue Mi et à gauche les 10 communautés les plus importantes en termes de taille des cahiers du GERAD pour la période 2009-2011

Les 10 premières communautés fournies sont basées sur leur taille, déterminée par le nombre de nœuds qu'elles contiennent, et ce pour chaque période. La figure 4.18 pour les cahiers du GERAD montre une forte densité de connexions avec une structure communautaire claire, indiquant un réseau étroitement lié où les membres sont fortement interconnectés. La présence de plusieurs communautés distinctes peut indiquer des groupes spécialisés ou des domaines de recherche au sein du GERAD. En revanche, le réseau de Mi est plus clairsemé, avec moins d'interconnexions entre les nœuds. Cela peut refléter une plus grande variété de thèmes ou un environnement moins collaboratif. La figure 4.19 reste dense pour les cahiers du GERAD, mais semble avoir plus de nœuds isolés que la période précédente, ce qui suggère une légère évolution vers un travail individualiste ou l'inclusion de nouveaux membres qui ne sont pas encore intégrés. Le réseau de Mi présente un petit groupe de nœuds étroitement connectés, ce qui pourrait être le signe d'un petit nombre de contributeurs dominants à ce moment-là. Pour les années subséquentes, c'est sensiblement le même scénario qui se répète. On observe un groupe central dominant entouré de nœuds satellites pour les cahiers du GERAD. Cette structure de communauté plus groupée indique un environnement de recherche collaborative avec des interactions interdisciplinaires potentielles. En revanche, Mi continue de présenter des groupes isolés aux liens faibles. Toutefois, une comparaison entre les périodes 2009 et 2023 révèle une augmentation modeste des liens, ce qui indique un développement progressif des efforts de collaboration.

Si on explore davantage les structures communautaires, on peut tenter de voir quels sont les auteurs les plus prolifiques en fonction des différentes périodes selon leur degré de centralité Katz-Bonacich. Contrairement aux mesures qui ne prennent en compte que les liens directs, la centralité de Katz tient compte de l'ensemble de la structure du réseau en incorporant tous les chemins par le biais d'une atténuation paramétrée. Cette mesure est particulièrement avantageuse, car elle reconnaît non seulement les auteurs prolifiques, mais aussi ceux qui jouent un rôle central dans la connectivité du réseau. Cela nous permet d'en apprendre davantage sur la communauté et des individus qui la compose.

	Auteurs	Mesure de centralité	Contribution
2021-2023	Faten Lakhal	0,046912386	4
	Ulrike Mayrhofer	0,046229274	4
	Jean-Michel Sahut	0,043455131	3
	Eric Braune	0,043283352	3
	Lubica Hikkerova	0,04323916	3
2018-2020	Foued Cheriet	0,052401697	4
	Frédéric Teulon	0,048360587	3
	Caroline Mothe	0,047100683	3
	Alfredo Jiménez	0,043972634	2
	Hanane Beddi	0,043750473	2
2015-2017	Florence Charue-Duboc	0,056659796	3
	Sihem Ben Mahmoud-Jouini	0,056659796	3
	Thierry Burger-Helmchen	0,050895581	2
	Pascal Lièvre	0,050651575	2
	Mathias Guérineau	0,05061781	2
2012-2014	Eric Persais	0,053672896	3
	Sébastien Point	0,051861903	2
	Stéphanie Loup	0,050904303	2
	Raluca Mogos	0,050481681	2
	Björn Walliser	0,050481681	2
2009-2011	Ulrike Mayrhofer	0,074176536	3
	Patrick Cohendet	0,068566788	2
	Anne Loubès	0,068426156	2
	Julien Pénin	0,067955487	2
	Laurent Simon	0,067955487	2

Tableau 4.3 - Représente les 5 auteurs les plus prolifiques en fonction du nombre d'articles qu'ils ont écrits de même que leur mesure de centralité pour la revue Mi

	Auteurs	Mesure de centralité	Contribution
2021-2023	Dominique Orban	0,116245343	16
	Erick Delage	0,099029121	13
	Roussos Dimitrakopoulos	0,098597696	14
	Issmail El Hallaoui	0,09022727	10
	Charles Audet	0,087632857	11
2018-2020	Guy Desaulniers	0,163277259	27
	F. Miguel Anjos	0,125501764	21
	Sébastien Le Digabel	0,123136441	19
	Issmail El Hallaoui	0,110926464	18
	Charles Audet	0,110073813	18
2015-2017	Roussos Dimitrakopoulos	0,20369097	40
	Nenad Mladenović	0,123939831	22
	Guy Desaulniers	0,106627168	21
	Gilles Caporossi	0,098865971	17
	Alain Hertz	0,09100582	17
2012-2014	Brunilde Sansò	0,16662281	27
	Georges Zaccour	0,156780282	30
	François Soumis	0,120195729	21
	Pierre Hansen	0,117305585	22
	Roussos Dimitrakopoulos	0,108415968	23
2009-2011	Pierre Hansen	0,320404742	43
	Gilles Caporossi	0,134447006	19
	Guy Desaulniers	0,104861358	19
	Leo Liberti	0,100329398	10
	Charles Audet	0,097107748	16

Tableau 4.5 - Représente les 5 auteurs les plus prolifiques en fonction du nombre d'articles qu'ils ont écrits, de même que leur mesure de centralité pour les cahiers du GERAD

En examinant séparément le tableau 4.4 pour la revue Mi et le tableau 4.5 pour les cahiers du GERAD, nous observons différents modèles de collaboration et d'influence dans les deux réseaux. Les cahiers du GERAD, présente des scores élevés de centralité associés à un nombre considérable de contributions, ce qui suggère un réseau composé d'un petit nombre d'auteurs très influents et productifs. Dominique Orban, par exemple, au cours de la période 2021-2023, contribue non seulement à 16 articles, mais a également un score de centralité d'environ 0,116, ce qui indique un rôle central dans le réseau. Certaines périodes où des auteurs spécifiques, tel que Pierre Hansen en

2009-2011, a exercé une influence dominante, comme en témoignent son score de centralité élevé de 0,32. et sa contribution d'article au nombre de 43. À travers les différentes périodes, les auteurs ayant la centralité la plus élevée n'ont pas toujours le plus grand nombre de contributions, ce qui illustre la capacité de la centralité de Katz a capturé non seulement les influences directes, mais aussi indirectes à travers le réseau.

Le réseau de la revue Mi présente un modèle d'influence plus uniformément réparti entre ses principaux contributeurs, Faten Lakhal étant en tête pour la période la plus récente avec un score de centralité d'environ 0,047 et quatre contributions. Cela suggère un environnement collaboratif avec un équilibre entre la productivité des auteurs et leur centralité dans le réseau. Ulrike Mayrhofer apparaît systématiquement comme une personne influente au cours de deux périodes distinctes. Au cours de la période 2009-2011, elle a contribué à trois articles, et ses contributions sont passées à quatre articles au cours de la période 2021-2023. Cela indique qu'elle joue un rôle soutenu dans le réseau au fil des ans. La diminution progressive des scores de centralité au fil du temps pourrait impliquer un réseau en expansion avec une distribution plus équitable des liens de collaboration entre ses membres.

Compte tenu des différences de portée et d'exigences de soumission des cahiers du GERAD et de la revue Mi, l'analyse de leurs structures communautaires révèle des contrastes intéressants. Les mesures de centralité plus élevées du GERAD indiquent un réseau étroitement connecté, probablement en raison de l'exigence d'implication des membres du GERAD dans les publications et de sa spécialisation dans les mathématiques et la science de la décision. Dans ce contexte, les travaux de chaque membre sont inévitablement liés à ceux des autres, ce qui favorise la création d'un réseau très dense en connexions qui amplifie la centralité de ses membres. Ce modèle est bénéfique pour assurer la cohérence de la qualité de la recherche et maintenir une communauté scientifique étroitement liée qui peut tirer parti de son expertise collective de manière efficace.

Le réseau de la revue Mi, présente une structure plus dispersée avec des scores de centralité plus faibles. Cela est probablement dû au fait qu'elle se concentre davantage sur le management international et qu'elle applique une politique d'évaluation par les pairs, ce qui permet d'élargir et

de diversifier l'éventail des contributeurs. Il en résulte une distribution moins concentrée de l'influence à travers son réseau.

En termes pratiques, Mi pourrait bénéficier de certaines des stratégies d'engagement communautaire du GERAD pour stimuler la collaboration. Il pourrait s'agir d'encourager les co-auteurs ou de créer des groupes d'intérêt spéciaux au sein de la communauté Mi afin d'accroître l'interconnectivité entre les auteurs. Mi pourrait potentiellement créer un réseau plus fort et plus collaboratif au sein de sa propre communauté, sans compromettre son approche plus large et plus inclusive. Cela pourrait conduire à une augmentation des résultats de centralité des auteurs au fil du temps, reflétant une communauté scientifique plus interconnectée.

4.3 Relation entre modélisation des thèmes et communautés

Évolution des thèmes

Dans un premier temps, il est essentiel d'analyser l'évolution des thèmes à travers les différentes périodes qui ont été établis dans la deuxième partie de l'étude. Cette analyse nous permettra d'identifier les changements d'orientation ou les tendances émergentes dans le domaine. Rappelons les thèmes découverts pour chaque ensemble de données, qui sont présentés dans le tableau 4.6.1 et 4.6.2 ci-dessous. Il est également important de souligner que, contre toute attente, nous avons identifié le même nombre de thèmes malgré les différences notables entre Mi et des cahiers du GERAD.

	Mi
Thème 1	Management stratégique
Thème 2	Marketing et consommation
Thème 3	Développement durable
Thème 4	Finance et gouvernance
Thème 5	Entrepreneuriat et PME
Thème 6	Innovation, communauté et digitalisation
Thème 7	Management et GRH

Tableau 4.6.1 – Thèmes découverts lors de la modélisation des thèmes avec la méthode LDA pour la revue Mi

	Cahiers du GERAD
Thème 1	Gestion des systèmes et des infrastructures
Thème 2	Environnement
Thème 3	Optimisation opérationnelle
Thème 4	Méthodes heuristiques
Thème 5	Théorie des graphes
Thème 6	Économie
Thème 7	Logistique

Tableau 4.6.2 – Thèmes découverts lors de la modélisation des thèmes avec la méthode LDA pour les cahiers du GERAD

En observant la figure 4.23, de la revue Mi, on constate que les thèmes forts de la période 2009-2011 sont le *Management stratégique*, *Entreprenariat et PME* ainsi *qu'Innovation, communauté et digitalisation.* Les conséquences de la crise financière de 2008 auraient pu influencer cette orientation thématique. Ces thèmes correspondaient à la réorientation stratégique et aux initiatives entrepreneuriales durant la phase de sortie de crise, ainsi qu'à l'importance croissante de l'innovation numérique. Pour la période 2012-2014, une augmentation globale de tous les thèmes a été observée, à l'exception de *Management Stratégique*. En 2015, la plupart des thèmes sont restés assez constants à l'exception *d'Entrepreneuriat et PME* et *Management et GRH* qui ont tous deux diminué. Potentiellement dû à une stabilisation du marché ou une attention redirigée vers des questions mondiales émergentes. Malgré la signature de l'accord de Paris en 2015, aucune augmentation du thème *développement durable* n'est observée. Ce n'est que plus tard, lors de la période 2021-2023 qu'une hausse est constatée.

En 2018, on remarque une augmentation des thèmes *Management Stratégique* et *Marketing et Consommation*. Possiblement liée à l'évolution des stratégies de marketing numérique et à une attention renouvelée pour la planification stratégique dans le contexte des changements économiques mondiaux. Le thème *Entrepreneuriat et PME* a également connu une croissance, probablement en raison l'émergence de plateformes numériques favorisant de nouveaux modèles d'entreprise. En 2023, des augmentations notables ont été observées dans trois thèmes non seulement basé sur la période précédente, mais aussi en relation au nombre de documents qui représentent 53% du total de document publié pour cette période. L'augmentation de *Finance et*

Gouvernance pourrait être attribuée à la complexité croissante de la dynamique du commerce mondial et à l'importance accrue accordée à la gouvernance d'entreprise. La croissance d'*Innovation, Communauté et Digitalisation* reflète probablement le passage rapide aux modalités numériques déclenché par la pandémie de COVID-19, qui nécessite une transformation numérique rapide. L'augmentation d'*Entrepreneuriat et PME* suggère une expansion des nouvelles entreprises à mesure que l'économie se rétablit de la pandémie, les entreprises innovant pour s'adapter à l'évolution des préférences des consommateurs dans un environnement post-pandémique axé sur le numérique.

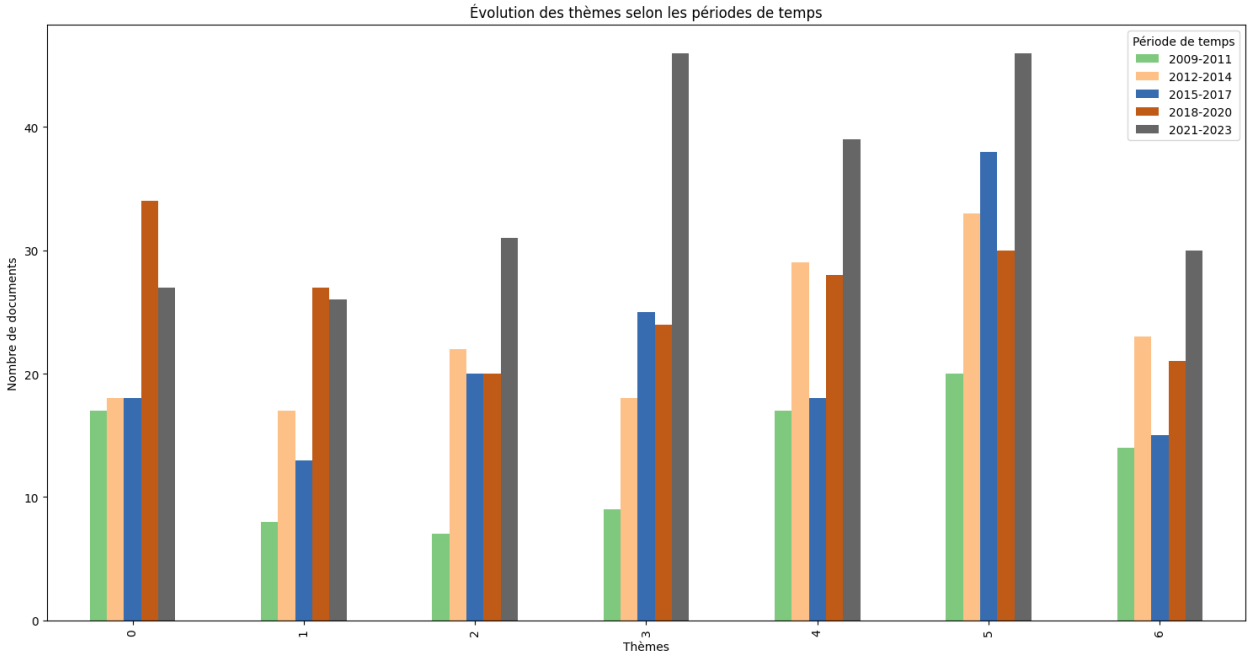


Figure 4.23 - Représente l'évolution des 7 thèmes au cours des différentes périodes pour la revue Mi

En observant la figure 4.24 des cahiers du GERAD, on constate une faible importance accordée aux thèmes *Gestion des systèmes et des infrastructures* et *Économie,* qui ne représentent que 11,4% des articles publiés au cours de la période 2009-2011. Malgré la crise économique de cette période, l'accent a peut-être été mis sur des domaines plus directement liés à la réponse à la crise plutôt que sur des stratégies économiques et d'infrastructure à long terme. Au cours de la période 2012-2014, on observe une augmentation substantielle (de plus de 100%) de ces thèmes, reflétant probablement une phase de reprise post-crise où les investissements dans l'infrastructure informatique et l'analyse

économique sont devenus plus prononcés en raison de l'amélioration des conditions financières et de la nécessité d'une prise de décision stratégique dans une économie en voie de stabilisation. De 2015 à 2017, on observe une augmentation marginale du thème *Environnement,* malgré les initiatives politiques mondiales en matière d'environnement, ce qui indique un possible décalage. L'augmentation de 65% du thème *Méthodes heuristiques* suggère une application accrue des méthodes de résolution de problèmes dans des scénarios complexes, tandis que la réduction drastique du thème *Logistique* pourrait indiquer un point de saturation dans le développement ou la recherche du thème.

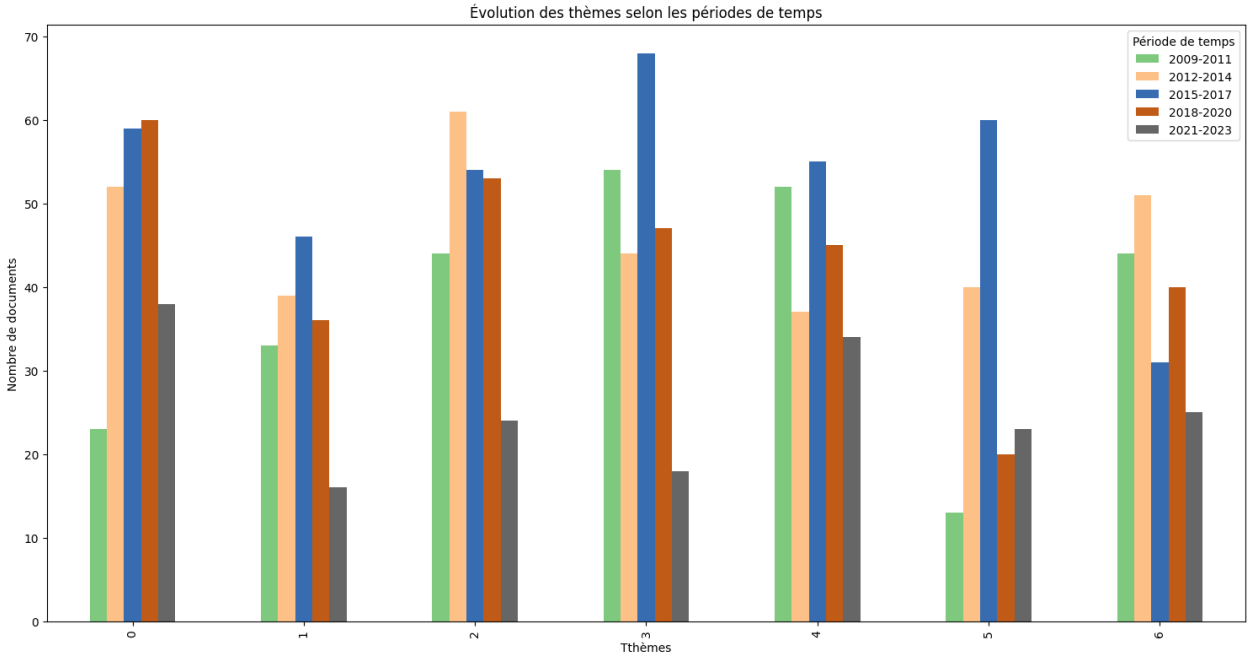


Figure 4.24 - Représente l'évolution des 7 thèmes au cours des différentes périodes pour les cahiers du GERAD

La période 2018-2020 a connu une diminution des thèmes *Environnement, Méthodes heuristiques* et *Économie.* L'Environnement pourrait être dû à plafonnement des nouvelles études à la suite de l'essor post-Accord de Paris. En ce qui concerne les Méthodes heuristiques, l'intégration de ces méthodes dans la pratique peut avoir réduit le volume de la recherche fondamentale. Dans le même temps, une période économique relativement stable peut avoir détourné l'attention vers des thèmes plus spécialisés ou émergents. Au cours de la dernière période, 2021-2023, on observe une diminution notable dans plusieurs thèmes, notamment *Gestion des systèmes et des infrastructures*, *Environnement*, *Optimisation opérationnelle* et *Méthodes heuristiques*. Cette tendance peut être

attribuée à un changement des priorités de recherche dû à l'influence continue de la pandémie sur les programmes de recherche mondiaux.

Les tendances générales observées sur l'ensemble des périodes des deux ensembles de données montrent comment des événements majeurs, tels que les ralentissements économiques et les crises sanitaires mondiales, influencent les thèmes de recherche, entraînant un changement d'orientation qui s'aligne sur l'évolution du contexte mondial.

Liens entre thèmes et communautés par période

Examiner l'alignement entre les communautés détectées et les thèmes dérivés de LDA, pourrait indiquer si les communautés sont centrées sur des thèmes spécifiques ou si elles ont une approche plus généraliste. Cette analyse servira à identifier les tendances ou les points communs entre les principales communautés afin de comprendre ce qui contribue au succès d'une communauté au sein de ces réseaux. Afin d'éviter l'encombrement et d'améliorer la visibilité des données, il a été décidé que seules les cinq plus grandes communautés de chaque période étaient nécessaires à l'analyse.

La figure 4.25 montre la relation entre le nombre d'articles publiés, la période et la répartition des communautés pour la revue Mi. Les couleurs indiquent le thème de l'article sur la base des résultats du modèle LDA décrits à la section 4.1.

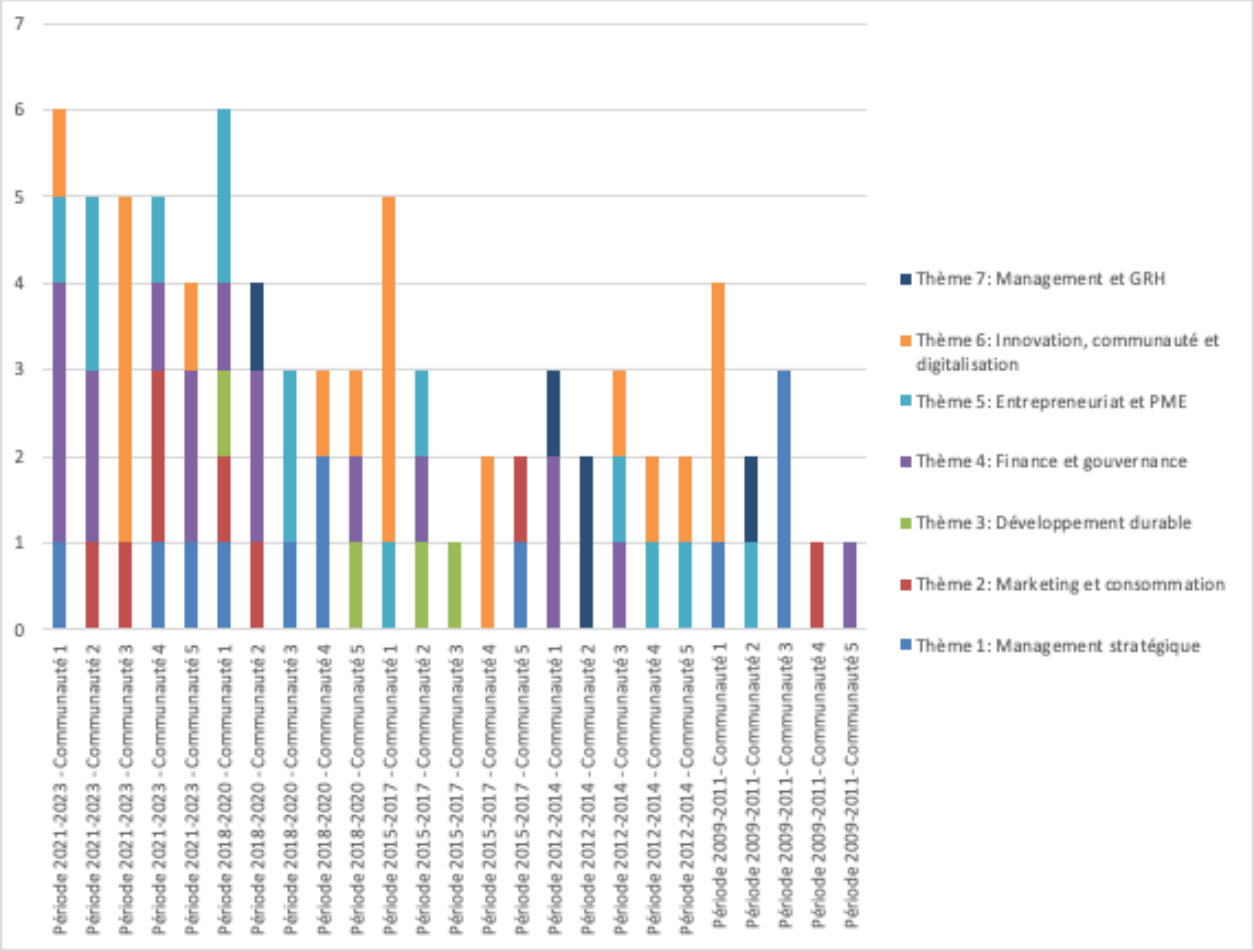


Figure 4.25 - Représentation des 7 thèmes à travers les 5 communautés les plus affluentes par période pour la revue Mi

Une tendance notable est l'accent mis sur le thème *Innovation, Communauté et digitalisation,* en particulier par la Communauté 3 au cours de la période 2021-2023. Ce thème met en évidence l'alignement de la communauté sur les changements numériques contemporains. Au cours des quatorze années d'analyse, ce thème a été récurrent ce qui indique sa pertinence soutenue dans ce domaine. Si les auteurs de la communauté 1 qui ont écrit sur le thème *Innovation, Communauté et digitalisation* entre 2009-2011 contribuent à la communauté 4 et la communauté 1 entre 2015 et 2017 et contribuent également à la communauté 3 en 2021-2023, cela peut indiquer un intérêt soutenu qui pourrait conduire à des collaborations intercommunautaires.

La Communauté 3 a manifesté un intérêt prononcé pour le *Management stratégique* au cours de la période 2009-2011. Ce thème absent au cours de la période 2011-2014 est réapparu au cours des périodes suivantes, bien qu'avec une attention moindre. Il serait utile de poursuivre les recherches sur la composition de ces communautés afin de déterminer si elles sont composées des mêmes individus ou s'il existe un potentiel de collaboration basé sur des intérêts communs.

Le *développement durable* a été mis de l'avant de façon très modeste par les communautés à l'étude; il ne s'agit pas d'un thème fréquemment abordé. Son émergence a été notée au cours de la période 2015-2017, avec un seul article par la communauté 3. Par conséquent, il serait utile d'examiner les trois autres communautés pour lesquels ce thème est abordé. Ceci aidera à déterminer le potentiel de collaboration visant à accroître l'accent mis sur le développement durable pour les périodes à venir. Une telle collaboration pourrait contribuer à la formation d'un collectif plus large avec un intérêt partagé dans ce domaine.

Le thème 4, *Finance et gouvernance*, affiche une présence relativement constante tout au long des périodes observées. Notamment, ce thème représente 70 % du contenu de 2018 à 2023 pour les 5 communautés prédominantes. Cette prévalence suggère un potentiel d'échange d'idées et d'efforts de collaboration dans ce domaine.

Pour les Mi, cette analyse offre une compréhension fondamentale de l'alignement et de la progression des intérêts de recherche de 2009 à 2023. L'étape suivante consiste à examiner en détail les stratégies de la ligne éditoriale que le comité Mi pourrait mettre en œuvre. Les options comprennent le recrutement d'un membre éminent de la communauté en tant que rédacteur en chef associé ou l'identification d'un auteur prolifique sur un thème spécifique et la proposition d'une co-écriture avec un autre auteur partageant les mêmes intérêts pour les collaborations futures. Bien qu'il s'agisse d'un défi de taille, l'exploitation de ces informations peut sans aucun doute renforcer le potentiel de la revue en impliquant les membres les plus influents de chaque communauté.

La figure 4.26 détaille les articles publiés des cahiers du GERAD au sein de cinq communautés distinctes entre 2009 et 2023, réparties en sept thèmes spécifiques. À première vue, les données révèlent un engagement variable à l'égard de thèmes au cours des différentes périodes, ce qui indique des changements d'orientation et d'intérêt de la recherche au fur et à mesure que les communautés évoluent.

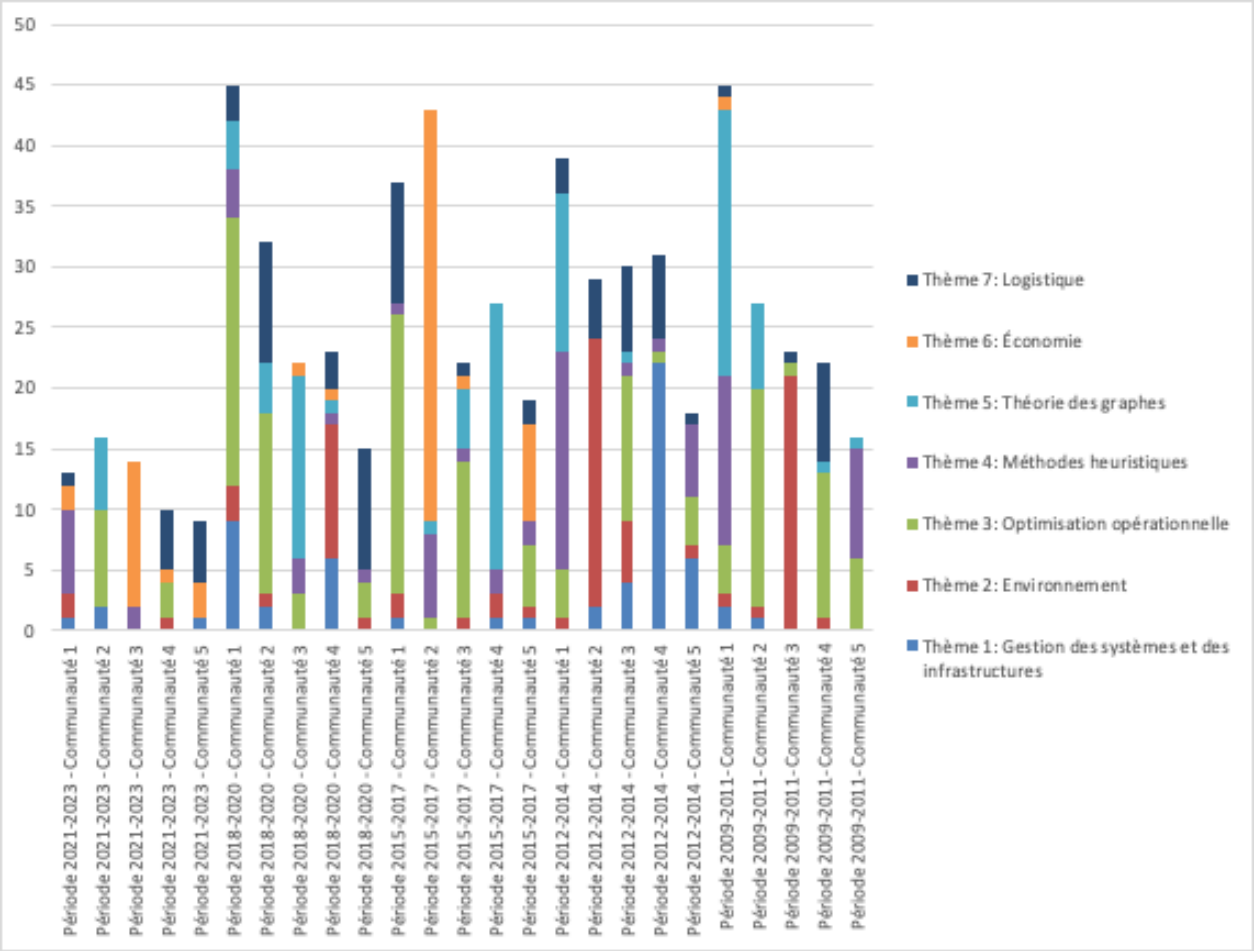


Figure 4.26 - Représentation des 7 thèmes à travers les 5 communautés les plus affluentes par période pour les cahiers du GERAD

Un volume constant de publications sur *l'Optimisation opérationnelle* est observé d'une année sur l'autre, ce qui indique une attention constante portée à ce domaine dans la recherche universitaire. En revanche, le thème de l'*environnement* connaît des fluctuations, avec des augmentations notables en 2009-2011, 2012-2014 et à nouveau en 2018-2020, ce qui pourrait refléter l'attention accrue portée aux questions environnementales à l'échelle mondiale au cours de ces années.

Malgré la variété des thèmes présents dans les différentes communautés, qui témoigne de la diversité des intérêts des membres, un thème se distingue souvent par sa prévalence. Par exemple, le thème 5, *Théorie des graphes*, lorsqu'il est présent dans une communauté, a tendance à avoir un grand nombre d'articles publié. Il s'agit également d'un thème qui apparaît systématiquement à chaque période, ce qui démontre son développement continu dans le domaine.

Le thème 6, qui porte sur l'*économie*, a occupé une place prépondérante dans la communauté 2 au cours de la période 2015-2017. Ce thème n'était pas présent au cours des périodes précédentes et est réapparu au cours de la récente période 2021-2023. L'*économie* présente une fluctuation qui suggère une réactivité aux changements économiques mondiaux, qui peuvent correspondre à des périodes de crise ou de croissance économique, soulignant ainsi l'engagement des cahiers du GERAD dans les questions d'actualité.

L'analyse des cahiers du GERAD en termes de communauté et de composition thématique vise à établir des comparaisons avec la revue Mi afin d'identifier les principales différences. Les cahiers du GERAD révèlent une structure qui soutient les communautés les plus larges, exigeant des auteurs qu'ils collaborent avec un membre du GERAD. Une des raisons pour laquelle les cahiers du GERAD présente une plus grande variété à la fois dans les communautés et les thèmes. Cela contraste beaucoup avec la revue Mi. La tendance récente à la diversification des intérêts au sein des communautés du GERAD indique une évolution vers la recherche interdisciplinaire, avec des communautés qui élargissent leur champ académique au-delà de leur expertise principale pour adopter une approche de recherche plus globale. Cette tendance s'éloigne des communautés plus homogènes que l'on observe actuellement dans la revue Mi et pourrait suggérer une orientation stratégique à prendre en considération.

Liens entre thèmes et communautés sans période

Cette analyse explore une dimension supplémentaire en procédant à la détection des communautés sans délimiter de périodes. L'objectif est d'examiner en détail la relation entre la taille des communautés et leurs mots thématiques respectifs. Plus précisément, l'étude se concentre sur les

cinq plus grandes communautés afin de déterminer si leur composition lexicale correspond aux nuages de mots présentés dans la section 4.1.1. Cette comparaison est essentielle pour établir la cohérence thématique entre les communautés identifiées. En outre, l'étude calcule l'information mutuelle normalisée (NMI) afin d'évaluer quantitativement la cohérence entre les structures des communautés.

Le réseau représentant l'ensemble des données sans période Mi comprend 2 403 nœuds et 1 880 arêtes. Cette structure indique un rapport plus faible entre les arêtes et les nœuds, ce qui implique un réseau avec moins de connexions par article. Le réseau comprend 828 résumés et 1 575 auteurs, ce qui révèle que le nombre d'auteurs dépasse le nombre de résumés, ce qui peut suggérer une tendance au travail collaboratif avec plusieurs auteurs par résumé. En revanche, le réseau des cahiers du GERAD contient 2 746 nœuds et 4 255 arêtes, ce qui reflète un ratio plus élevé d'arêtes par rapport aux nœuds et donc une plus grande connectivité au sein du réseau. Il y a 1 439 résumés associés à 1 307 auteurs dans le réseau GERAD, ce qui indique que la plupart des auteurs sont susceptibles d'avoir contribué à au moins un résumé.

Mi				**GERAD**			
# noeuds	# arêtes	# résumés	# auteurs	**# noeuds**	# arêtes	# résumés	# auteurs
2403	1880	828	1575	**2746**	4255	1439	1307

Tableau 4.7 - Attribut de la structure du graphe bipartite de la revue Mi et des cahiers du GERAD sans périodes

Ces différences dans la structure du réseau suggèrent que l'ensemble de données des cahiers du GERAD est caractérisé par des connexions denses et un degré élevé de collaboration au sein d'un groupe restreint d'auteurs, tandis que l'ensemble de données de la revue Mi est caractérisé par un plus grand nombre d'auteurs avec moins d'interconnectivité entre les articles. Cela peut indiquer des sujets de recherche plus variés de la revue Mi et un domaine d'étude plus ciblé dans les cahiers du GERAD.

La visualisation des réseaux est essentielle pour comprendre la dynamique des communautés, mais elle peut s'avérer difficile lorsqu'il s'agit de communautés de grande taille ou densément regroupées. Le graphe de l'ensemble de données de Mi, avec 557 communautés, rend difficile

l'identification des relations. Le réseau des cahiers du GERAD, avec 51 communautés, manque de clarté visuelle en raison d'un regroupement dense, qui masque les détails de l'interconnectivité et rend l'analyse insuffisante pour comprendre les collaborations complexes.

La complexité et la densité des réseaux nécessitent l'utilisation de techniques analytiques telles que des algorithmes de détection des communautés et des mesures de distance pour comprendre les connexions et la force des liens, ce qui permet d'obtenir des informations significatives.

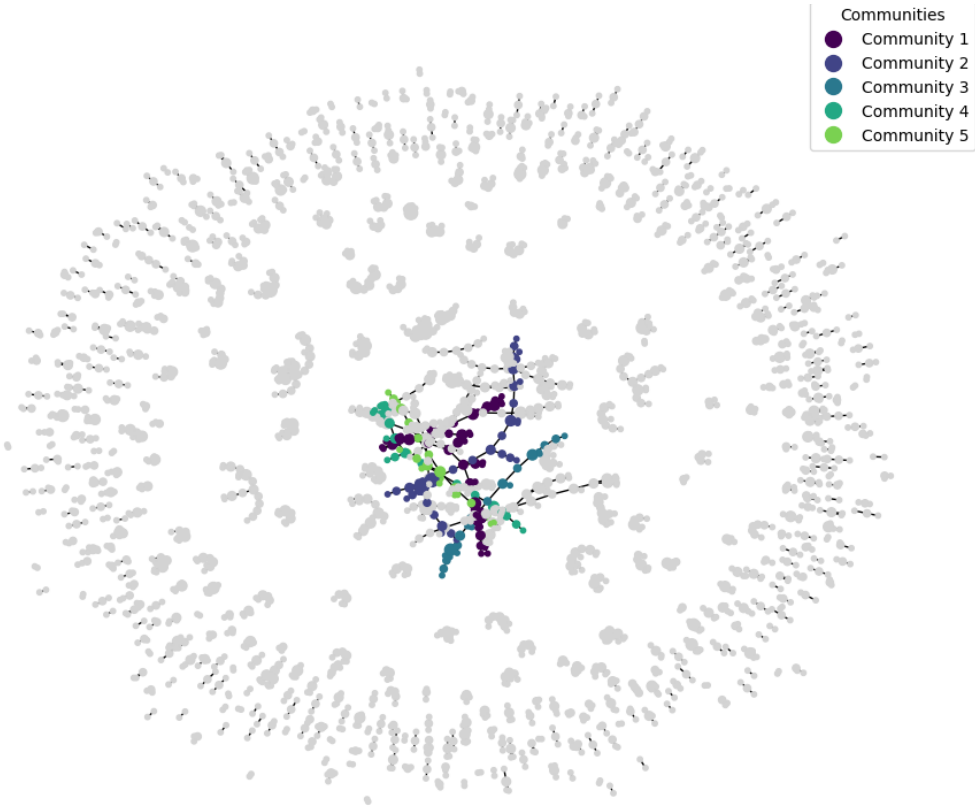


Figure 4.27 – Représentation visuelle des cinq communautés les plus significatives pour la revue Mi

Le graphe de la revue Mi, tel que présenté à la figure 4.27, indique un réseau fragmenté avec de nombreux petits groupes. Cela indique des intérêts de recherche divers et de faibles connexions interdisciplinaires. La collaboration se fait souvent au sein de poches isolées et il est difficile de discerner des liens clairs. En revanche, le graphe des cahiers du GERAD, à la figure 4.28, présente

un contraste avec celui de Mi. Moins de communautés indiquent une plus grande intégration et de connexions avec les membres ce qui suggère un domaine de recherche plus ciblée et collaboration fréquentes des auteurs.

Figure 4.28 – Représentation visuelle des cinq communautés les plus significatives des cahiers du GERAD

En explorant davantage la structure des cinq communautés les plus significatives, nous observons que les communautés des cahiers du GERAD sont significativement plus grandes et plus productives, comme identifié au tableau 4.8. Par exemple, la plus grande communauté de Mi de taille 34 est composée de 34 auteurs et de 18 résumés, tandis que la communauté des cahiers du GERAD correspondante est de taille 296, composée de 136 auteurs et 160 résumés. Les communautés de la revue Mi varient en termes de taille de 22 et 52, et le nombre d'auteurs entre 13 et 34. Le nombre de résumés par communauté est proche ou parfois supérieur au nombre d'auteurs, ce qui suggère des efforts de collaboration. Toutefois, la taille réduite des communautés

suggère que ces efforts sont plus isolés et c'est ce que nous observons à la figure 4.28. Les communautés des cahiers du GERAD ont une taille plus importante, allant de 178 à 296, ce qui indique des réseaux plus étendus avec davantage d'interactions et de collaborations. Le nombre de résumés est élevé, dépassant le nombre d'auteurs dans chaque communauté, ce qui suggère un travail d'auteur prolifique où les auteurs individuels peuvent contribuer à plusieurs travaux. Ces résultats soulignent l'importance de l'engagement de la communauté et de la productivité dans la recherche.

	Mi			**GERAD**	
Taille de la communauté	# auteurs	# résumés	**Taille de la communauté**	# auteurs	# résumés
52	34	18	**296**	136	160
40	21	19	**213**	79	134
24	17	7	**209**	95	114
23	15	8	**199**	79	120
22	13	9	**178**	71	107

Tableau 4.8 – Taille des cinq communautés les plus prolifiques pour la revue Mi et les cahiers du GERAD ainsi que le nombre d'auteurs et les nombres de résumés dans la communauté

Mi pourrait améliorer la dynamique de son réseau en exploitant des stratégies d'engagement des cahiers du GERAD. Cela impliquerait d'encourager une plus grande collaboration, de se concentrer sur la concentration thématique, de renforcer les groupes d'auteurs principaux et de tirer parti des nœuds centraux. Mi pourrait également promouvoir des groupes de recherche autour de thèmes spécifiques afin de développer des explorations approfondies. L'identification et l'exploitation des nœuds centraux au sein de Mi pourraient améliorer le flux d'informations et créer des communautés plus cohésives. Mi pourrait mettre en œuvre des changements stratégiques dans son propre réseau afin de favoriser un environnement de recherche plus interconnecté et plus productif.

L'objectif de cette section est d'identifier les connexions entre les communautés et les thèmes identifiés précédemment. En raison du manque de corrélation entre les données capturées par le réseau et les résultats du modèle LDA, nous avons choisi d'utiliser des nuages de mots pour la visualisation. Cela nous permet d'évaluer si les mots correspondent à ceux générés par le modèle LDA. Nous avons appliqué les mêmes méthodes de prétraitement des données que celles utilisées

pour les thèmes afin de garantir la cohérence de notre représentation. Les figures 4.29 et 4.30 illustrent les communautés les plus significatives et leurs mots clés tels qu'ils apparaissent dans les résumés. En comparant ces figures aux nuages de mots de la section 4.1.1, nous pouvons évaluer les similitudes et éventuellement les associer à des thèmes correspondants.

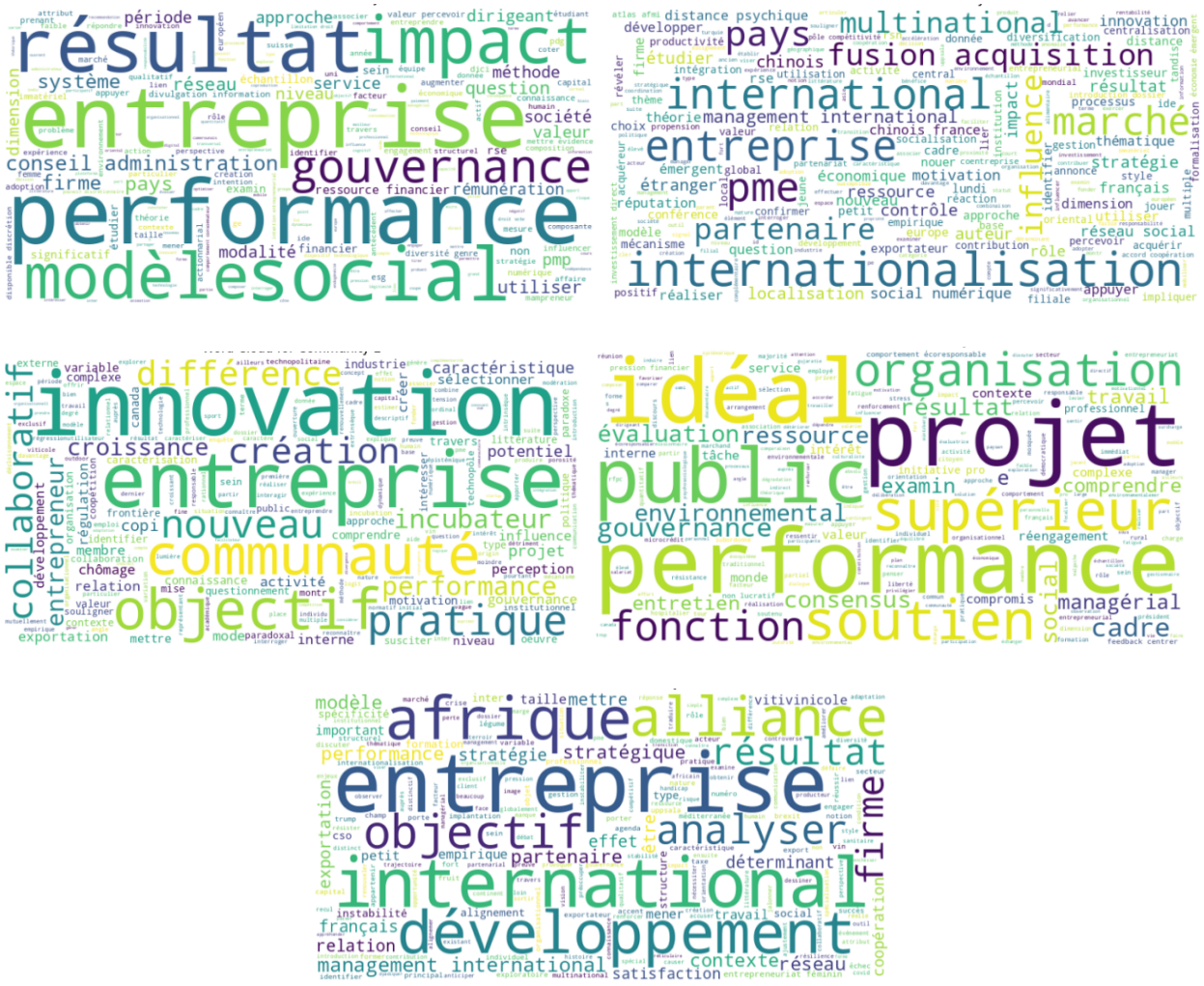


Figure 4.29 – Nuages de mots des 5 communautés les plus importantes dans le réseau de la revue Mi. Le nuage de mot de la communauté 1 (la plus importante) est en haut à gauche. La communauté 2 en haut à droite et ainsi de suite.

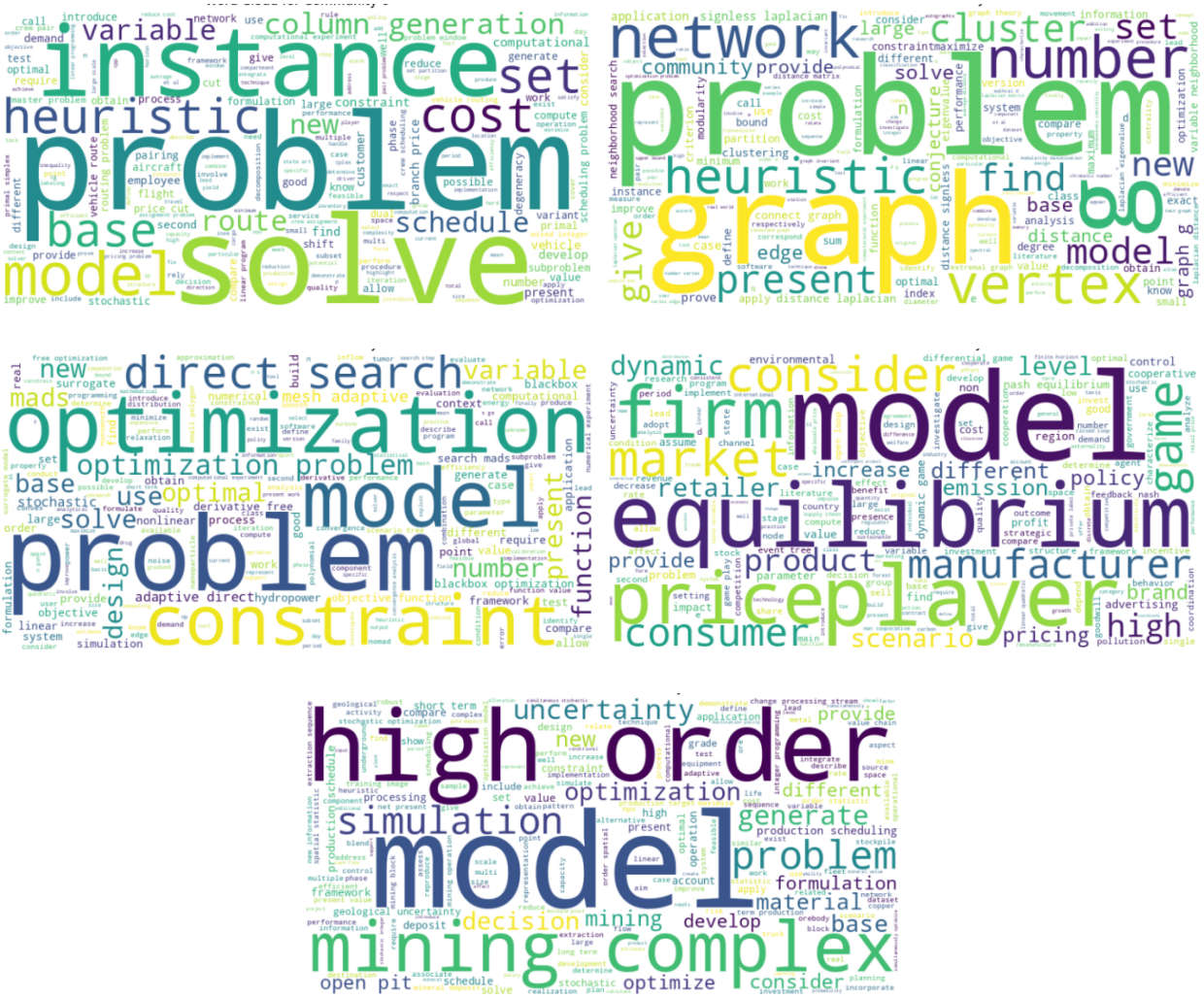


Figure 4.30 – Nuages de mots des 5 communautés les plus importantes dans le réseau des cahiers du GERAD. Le nuage de mot de la communauté 1 (la plus importante) est en haut à gauche. La communauté 2 en haut à droite et ainsi de suite.

L'analyse individuelle des mots est un processus difficile. Pour simplifier, nous avons sélectionné les 50 mots les plus importants en termes de poids pour les comparer à nos nuages de mots. Notre analyse a consisté à vérifier l'existence de mots communs. Nous avons choisi une matrice pour visualiser efficacement les résultats de la comparaison. Les tableaux 4.11 et 4.12 illustrent le pourcentage de chevauchement des mots entre les communautés et les thèmes issus de la modélisation thématique. Une diagonale rouge a été introduite pour souligner les pourcentages les plus élevés dans chaque communauté, facilitant ainsi l'identification des thèmes prédominants. En outre, l'ordre des thèmes a été ajusté pour révéler les similitudes dans la composition structurelle

des communautés Le tableau 4.11 montre que les communautés 2 et 3 ont une représentation élevée des mots associés au thème 2. En outre, la communauté 3 présente un lien fort avec le thème 6. Dans l'analyse du contenu de la revue Mi, aucun pourcentage ne dépasse 50 %. Toutefois, cela ne signifie pas qu'il n'y a pas de lien ; un lien significatif peut exister même si le pourcentage de chevauchement est à peine plus élevé que ce que l'on attendrait, soit 14,29 % pour une distribution uniforme. En fin de compte, notre objectif est de faire des liens entre les communautés et les thèmes afin de comprendre comment Mi peut utiliser ces informations pour sa stratégie éditoriale.

	Thème 2	Thème 4	Thème 1	Thème 5	Thème 3	Thème 7	Thème 6
Communauté 2	42%	14%	28%	22%	16%	12%	16%
Communauté 1	36%	32%	18%	24%	20%	8%	18%
Communauté 5	24%	20%	28%	26%	12%	12%	20%
Communauté 4	14%	26%	20%	20%	28%	16%	16%
Communauté 3	16%	10%	10%	16%	12%	8%	32%

Tableau 4.11 – Distribution en pourcentage des liens entre les communautés de la revue Mi et les thèmes, dérivée uniquement des nuages de mots composés de 50 mots

Le tableau 4.12 des cahiers GERAD indique des préférences distinctes pour des thèmes spécifiques. Cela suggère un intérêt prononcé pour certains thèmes plutôt que d'autres. Par exemple, la communauté 4 montre un lien clair avec le thème 2, alors que les autres communautés ne montrent aucun intérêt. Contrairement à la revue Mi, chaque communauté des cahiers du GERAD semble se spécialiser à des thèmes différents. Les communautés 1 et 3 ont des intérêts avec le thème 3, la communauté 2 s'aligne sur le thème 5, la communauté 4 sur le thème 2 et la communauté 5 sur le thème 6. Les thèmes 1 et 7 semblent toutefois sous-utilisés, bien que les communautés 1 et 5 manifestent un certain intérêt. Cet intérêt indique des thèmes communs possibles ou des opportunités d'interaction entre les communautés. Dans l'ensemble, les cahiers du GERAD présentent des préférences plus définies, reflétant une concentration d'expertise et des domaines de discussion ciblés.

	Thème 6	Thème 5	Thème 3	Thème 7	Thème 2	Thème 1	Thème 4
Communauté 5	58%	22%	24%	30%	16%	28%	26%
Communauté 2	16%	50%	28%	28%	8%	24%	38%
Communauté 1	26%	24%	50%	44%	12%	22%	36%
Communauté 4	22%	8%	12%	18%	46%	24%	12%
Communauté 3	28%	36%	42%	28%	6%	32%	38%

Tableau 4.12 – Distribution en pourcentage des liens entre les communautés des cahiers du GERAD et les thèmes, dérivée uniquement des nuages de mots composés de 50 mots

Les communautés dont certains thèmes se recoupent largement peuvent bénéficier d'un contenu plus important ou de discussions centrées sur ces thèmes, car ils représentent des domaines d'intérêt ou d'expertise importants. Pour les communautés dont les intérêts se répartissent équitablement entre les thèmes, il serait pertinent de proposer un contenu varié afin de maintenir l'engagement. Il serait important d'identifier et encourager les intérêts uniques, afin de renforcer l'identité et l'engagement de la communauté. S'intéresser aux thèmes sous-représentés, s'ils présentent des pourcentages systématiquement faibles dans toutes les communautés, afin de déterminer s'ils doivent faire l'objet d'une plus grande attention. Encourager l'interaction entre les communautés, si certaines d'entre elles s'intéressent fortement à certains thèmes tandis que d'autres possèdent une expertise. Revoir la stratégie de contenu si l'objectif est d'obtenir une répartition plus équilibrée des thèmes. Suivre régulièrement l'évolution de la communauté pour comprendre comment les intérêts évoluent, ce qui permet de procéder à des ajustements proactifs.

Nous allons maintenant appliquer une mesure de cohérence pour évaluer la force des liens entre les communautés et les thèmes. L'information mutuelle normalisée (NMI) présentée par [154] est une mesure utilisée pour évaluer les performances des algorithmes de détection des communautés. Il va de 0 à 1, indiquant la similarité entre deux communautés. Une valeur plus élevée signifie un plus grand degré de similitude entre deux communautés. NMI nécessite l'existence d'étiquettes de classe pour les calculs, ce qui implique la nécessité d'une hypothèse de base [155]. Cette mesure peut être exprimée comme suit :

$$I(A,B) = \frac{-2\ \sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}N}{N_i N_j}\right)}{\sum_{i=1}^{C_A} N_i \log\left(\frac{N_i}{N}\right) + \sum_{j=1}^{C_B} N_j \log\left(\frac{N_j}{N}\right)} \quad .$$

La matrice de confusion N est utilisée pour représenter le nombre de nœuds d'une communauté réelle i qui apparaissent dans une communauté trouvée j, les lignes représentant les communautés réelles et les colonnes les communautés trouvées. Où le nombre de communautés réelles est représenté par C_A et le nombre de communautés trouvées est noté C_B, la somme sur la ligne i de la matrice N_{ij} est notée N_i et la somme sur la colonne j est notée N_j [154].

En ce qui concerne nos ensembles de données, la liste des communautés représente les étiquettes de communautés détectées par l'algorithme. Ces étiquettes correspondent à un regroupement attribué à un résumé, sur la base des caractéristiques ou des similitudes des données. Dans notre modélisation, les thèmes représentent les étiquettes attribuées à chaque résumé, qui représente les valeurs réelles. La NMI trouvée pour la revue Mi est de **0,3717** alors que pour les cahiers du GERAD, le résultat est de **0,3372**. Les valeurs vont de 0 à 1, ce qui indique une absence d'information mutuelle ou une concordance parfaite. Les valeurs de 0,37 et 0,33 indiquent une information mutuelle modérée, ce qui suggère une similitude dans les structures des communautés, mais des différences substantielles subsistent.

Les résultats de la revue Mi et des cahiers du GERAD sont influencés par le nombre de communautés dans chaque ensemble de données. Le grand nombre de communautés de Mi (556 communautés) suggère une couverture thématique plus large ou des limites de communautés moins distinctes. Le nombre moins élevé de communautés de GERAD (51 communautés) indique une distribution plus concentrée des thèmes, ce qui suggère une discussion plus ciblée. La sensibilité de la mesure à la taille des communautés est évidente, puisqu'un résultat similaire dans Mi pourrait indiquer une structure thématique différente. Les résultats doivent être considérés dans le contexte des spécificités de chaque ensemble de données.

Similarité de Jaccard

Après avoir exploré NMI en tant que mesure de la cohérence, nous allons nous intéresser à une autre mesure : l'analyse entre les paires d'articles, connue sous le nom de similarité de Jaccard. Les résultats présentés au tableau 4.9, nous montre la similarité de Jaccard entre les résumées 1085036ar et 1085037ar qui a été calculée à 0,382, ce qui signifie un niveau de similarité modéré. En revanche, les résumés 1086417ar et 1060031ar ont affiché un score de similarité plus faible de 0,25. Les scores suivants sont très similaires oscillant de 0,2391 à 0,2245.

Article 1	Article 2	Similarité Jaccard
1085036ar	1085037ar	0,3823
1086417ar	1060031ar	0,25
1074363ar	1069098ar	0,2391
1027870ar	1015405ar	0,2333
1060060ar	1026028ar	0,2245

Tableau 4.9 - Résultat des 5 plus grands résultats de Jaccard pour la revue Mi

L'évaluation de ces similitudes indique qu'à mesure que le pourcentage diminue, la prévisibilité des résultats diminue. Plus précisément, les paires de résumés dont le pourcentage de similitude est proche de 20 % affichent des résultats susceptibles d'être influencés par seulement quelques mots. Une analyse des paires dont la similarité est 0 a donné un taux de précision de 89 %, ce qui implique que les mots identifiés dans ces résumés sont distincts les uns des autres et le modèle LDA a été en mesure de les distinguer. Pour les valeurs de similarité inférieures à 10 %, qui indiquent une correspondance minimale, la cohérence des résultats entre les paires de résumés s'élève à 85 %. Cela démontre un accord prédominant entre le modèle LDA et la mesure de similarité de Jaccard. Toutefois, une dispersion significative des résultats a été observée pour les scores de similarité supérieurs à 10 %, avec un taux de correspondance de 33 % seulement.

Pour les paires de résumés présentant un score de similarité supérieur à 20 %, ce qui n'englobe que treize paires dans l'ensemble de données, la corrélation a atteint 70 %. Ce schéma implique une forte corrélation entre les thèmes de chaque paire de résumés à mesure que le score de similarité augmente. Malgré ces résultats, la rareté générale des mots communs entre les paires de résumés rend l'identification des thèmes difficile.

Pour les cahiers du GERAD, une importante corrélation thématique entre les résumés est évidente, avec des résultats dépassant de manière significative ceux de Mi, comme présenté au tableau 4.10. Par exemple, les résumés G-2012-68 et G2011-49 présentent une similarité impressionnante de 0,9556. Ces différences indiquent que les cahiers du GERAD ont tendance à contenir des résumés plus étroitement liés les uns aux autres, contrairement à Mi où les résumés sont généralement plus variés et couvrent un plus large éventail de sujets.

Article 1	Article 2	Similarité Jaccard
G-2012-68	G-2011-49	0,9556
G-2014-26	G-2011-82	0,9333
G-2016-72	G-2016-09	0,9324
G-2018-58	G-2014-31	0,8814
G-2017-16	G-2014-31	0,8814

Tableau 4.10 - Résultat des 5 plus grands résultats de Jaccard pour les cahiers du GERAD

Il convient donc de s'interroger sur les facteurs susceptibles d'expliquer les écarts observés dans les mesures de similarité. L'analyse des taux de similarité supérieurs à 20 %, le modèle LDA et la similarité de Jaccard a atteint un taux de précision de 90 %. Cette précision atteint 96% pour tous les résultats qui sont supérieurs à 50%. Seulement une paire de résumés n'a pas bien été identifiée sur un total de 28. Ce qui indique que des pourcentages de similarité plus élevés sont en corrélation avec des résultats plus fiables. Pour les similitudes inférieures ou égales à 10 %, la précision du modèle LDA dans l'identification des thèmes est de 86 %. Avec une similarité de Jaccard de 0, la précision de l'identification des thèmes est de 90 %. Cela indique que le modèle LDA est capable de discerner des mots parmi les résumés dont la similarité de Jaccard est faible.

L'ensemble de données a révélé une anomalie avec un score de similarité parfait (1) entre deux paires de résumés qui ne sont pas intrinsèquement liées. Cette anomalie s'est avérée être un cas de deux résumés distincts partageant des résumés identiques dans l'ensemble de données, en raison d'une duplication de la saisie des données. Cette paire de doublons ne sera pas prise en compte dans l'analyse.

En résumé, les mesures de similarité de Jaccard suggèrent que les résumés des cahiers du GERAD sont plus étroitement corrélés que celles de Mi. Les cahiers du GERAD présentent un alignement thématique plus strict et plus clair, ce qui se traduit par une corrélation plus forte entre les termes en raison de l'adhésion à des structures thématiques bien établies. En revanche, les directives de publication de la revue Mi semblent être plus souples en ce qui concerne le contenu. Ce que Mi peut tirer de cette analyse est qu'en révisant son cadre thématique - en s'alignant plus étroitement sur l'un des sept thèmes que la revue cherche à représenter et en refusant les documents de recherche qui ne correspondent pas à ces thèmes - elle peut obtenir une plus grande similitude de mots dans ses résumés.

5 Conclusion

Cette étude a fourni une analyse comparative détaillée de la modélisation thématique et des structures communautaires dans les publications de Management international (Mi) et des cahiers du GERAD. La recherche a révélé que le champ thématique de Mi est large, englobant un large éventail de thèmes de gestion internationale, alors que les thèmes du GERAD sont plus concentrés autour des sciences de la décision et des mathématiques. Cette divergence reflète les priorités académiques et opérationnelles distinctes des deux entités.

L'analyse de la communauté a montré que le GERAD a une communauté plus cohésive et interconnectée, probablement en raison de son domaine de recherche spécialisé. En revanche, la communauté du Mi est plus diversifiée et dispersée, ce qui peut être attribué à sa portée thématique plus large et à son envergure internationale. Cette diversité, bien qu'elle soit une force, présente également des défis dans la création d'une communauté de recherche très soudée.

En mettant en relation les thèmes et les communautés découverts des deux ensembles de données, ce mémoire représente une analyse exploratoire sur les liens qui les unissent. Malgré les liens subtils entre les communautés et les thèmes, l'étude a mis en évidence des schémas perceptibles dans ces liens. L'une des principales conclusions de cette étude est l'identification d'une relation modérée entre les structures des communautés et les sujets, comme le montrent les indices de Jaccard et NMI. Bien qu'une forte corrélation n'ait pas été établie, la recherche a permis de découvrir des modèles dans les données qui ont une valeur significative pour la compréhension de l'engagement de la communauté et du développement des thèmes pour la revue Mi. L'objectif principal n'était pas de découvrir toutes les connexions possibles, mais de mettre en évidence les interactions les plus pertinentes. Ce faisant, cette étude présente des recommandations pratiques pour Management international, en se concentrant sur le développement et la capitalisation des sept thèmes distincts identifiés. Ces thèmes devraient être exploités de manière stratégique pour consolider la position de Mi dans des secteurs spécifiques de la gestion internationale, en s'alignant sur les tendances mondiales actuelles et sur l'expertise spécialisée de ses contributeurs.

En conclusion, ce mémoire fournit des informations qui ne sont pas seulement d'un intérêt académique, mais aussi d'une utilité pratique pour la direction éditoriale de *Management*

international. Elle suggère qu'en adoptant une approche axée sur les données pour analyser les tendances de publication et la dynamique de la communauté, Mi peut améliorer ses stratégies de gestion du contenu. Les recommandations formulées ici visent à éclairer les décisions stratégiques concernant la préservation du contenu, l'orientation éditoriale et les tactiques d'engagement communautaire, favorisant ainsi l'intégrité thématique de la revue et son engagement envers sa communauté.

5.1 Limites

L'étude a rencontré des limites en raison des différences inhérentes entre les ensembles de données de Mi et les cahiers du GERAD. Le contenu de Mi est principalement ancré dans la gestion, la gestion des ressources humaines et la prise de décision stratégique, tandis que les cahiers du GERAD se concentrent sur la science de la décision et les mathématiques. Malgré ces différences, la décision a été prise de poursuivre l'analyse comparative, tout en sachant que la comparabilité directe pourrait être limitée.

La brièveté des textes, souvent inférieurs à 100 mots, a potentiellement influencé l'efficacité de l'analyse LDA. Des textes aussi courts peuvent ne pas fournir suffisamment de données contextuelles pour une modélisation thématique précise. La détection de communautés pour compléter les résultats de l’analyse primaire a été introduite en cours d’étude. Elle a été réalisée avec un certain degré de minimalisme, dans le but d'identifier les structures communautaires de base au sein des ensembles de données qui pourraient fournir des informations sur leur dynamique structurelle sans se lancer dans la création de modèles complexes.

Il est essentiel de noter que bien que Mi et le GERAD aient plus de trois décennies de contributions scientifiques, les données historiques complètes n'étaient pas accessibles au moment de cette étude de même que leurs textes intégraux. Par conséquent, les résultats présentés ne reflètent pas toute l'étendue de leur production académique. Cette limitation réduit la nature exhaustive des résultats, suggérant que les conclusions tirées peuvent représenter un instantané plutôt qu'un portrait complet.

5.2 Recommandation pour les recherches futures

Les recherches futures devraient tenir compte de ces limites en intégrant des méthodes de détection des communautés dynamiques afin de mieux comprendre la nature évolutive des thèmes et des structures communautaires. L'élargissement de l'ensemble de données aux textes intégraux et non seulement aux résumés des articles publiés renforcerait l'analyse. En outre, l'utilisation de techniques NLP avancées telles que l'analyse sémantique et l'exploration des sentiments pourrait fournir des informations plus approfondies sur les thèmes. La modélisation BTM serait aussi une méthode envisageable qui permet de découvrir la structure thématique de textes courts. Comparativement à la méthode utilisée dans ce mémoire, BTM compte la cooccurrence des paires de mots dans l'ensemble du corpus, et pas seulement dans les documents individuels.

Des études comparatives avec d'autres revues universitaires similaires à Mi pourraient également révéler des tendances plus larges en matière de publication universitaire et d'engagement communautaire, offrant ainsi une vision plus intégrale du domaine de la gestion internationale.

En conclusion, ce mémoire souligne l'importance de l'orientation thématique stratégique et de l'engagement communautaire pour les revues universitaires telles que Mi. En tirant parti des idées et des recommandations présentées ici, Mi peut améliorer sa cohérence thématique, renforcer sa communauté de chercheurs et consolider sa position dans le domaine de la gestion internationale.

Bibliographie

[1] Han, J. & Kamber, M. (2006) *Data Mining: Concepts and Techniques*. Massachusetts, 2nd edn. Burlington: Morgan Kaufmann.

[2] Kantardzic, M. (2011) *Data Mining: Concepts, Models, Methods, and Algorithms.* Online library, John Wiley & Sons, Ltd.

[3] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B. & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering andExtraction Techniques. *Journal arXiv e-prints: 1707.02919v2 [cs.CL]*

[4] Johri, P., Khatri, S.K., Al-Taani, A. & Sabharwal, M. (2021). Natural language Processing : History, Evolution, Application, and Future Work. *Proceedings of 3rd International Conference on Computing Informatics and Networks (pp.365-375)*

[5] Grimmer, J. & Stewart, B. M. (2013). Text as Data : The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(03), 267-297.

[6] Denny, M. & Spirling A. (2017). Text Preprocessing For Unsupervised Learning: Why it Matters, When it Misleads, And What To Do About It. *Available at SSRN: https://ssrn.com/abstract=2849145, September 27, 2017*

[7] Nadkarni, P.M., Ohno-Machado, L. & Chapman, W.W. (2011) Natural language processing: an introduction. *J Am Med Inform Assoc 2011, Sep-Oct;18(5):544-551. doi:10.1136/amiajnl-2011-000464*

[8] Winograd, T. (1980). What does it mean to understand language? *Cognitive Science, Volume 4, issue 3, 209-241. https://doi.org/10.1016/S0364-0213(80)80003-6.*

[9] Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill Science/Engineering/Math; (March 1, 1997)

[10] Hotho, A., Nürnberger A. & Paass, G. (2005). A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics 20(1):19-62*

[11] McCallum, A. & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *Learning for Text Categorization: Papers from the 1998 AAAI Workshop, page 41--48.*

[12] Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, vol 1398. Springer, Berlin, Heidelberg.*

[13] Joachims, T. (2001). A statistical learning learning model of text classifica-

tion for support vector machines. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, September 2001, pages 128-136.*

[14] Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, vol 1398. Springer, Berlin, Heidelberg.

[15] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning 20, 3 (1995), 273–297.*

[16] Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. August 1999, pages 42-49.*

[17] Friedl, M. A. & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment 61, 3 (1997), 399–409.*

[18] Han, E. H., Karypis, G. & Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification. *In: Cheung, D., Williams, G.J., Li, Q. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2001. Lecture Notes in Computer Science(), vol 2035. Springer, Berlin, Heidelberg.*

[19] Rezaeiye, P. P., Bazrafkan, M. & al (2014). Use HMM and KNN for classifying corneal data. *Journal arXiv:1401.7486 [cs.CV]*

[20] Wang, B., Wang, A., Chen, F., Wang, Y. & Jay Kuo, C.-C. (2019). Evaluating Word Embedding Models: Methods and Experimental Results, *Journal arXiv: 1901.09785v2 [cs.CL]*

[21] Jurafsky, D. & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall New Jersey, 2000.

[22] Catellier, R., Vaiter, S. & Garreau, D. (2023). On the Robustness of Text Vectorizers, *Journal arXiv: 2303.07203v2 [cs.CL]*

[23] Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA 18 5 (2011): 544-51.*

[24] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *In Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon. Association for Computational Linguistics.*

[25] Isozaki, H. & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. *In Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1-7.*

[26] Leong Chieu, H. & Tou Ng, H. (2003). Named Entity Recognition with a Maximum Entropy Approach. *In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 160–163.*

[27] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE 77, 2 (1989), 257–286.*

[28] Lafferty, J. D., McCallum, A. & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Published in International Conference on machine learning.*

[29] Seng Chan, Y. & Roth, D. (2010). Exploiting background knowledge for relation extraction. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 152–160, Beijing, China. Coling 2010 Organizing Committee.*

[30] Seng Chan, Y & Roth, D. (2011). Exploiting syntactico-semantic structures for relation extraction. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 551–560, Portland, Oregon, USA. Association for Computational Linguistics*

[31] Guo Dong, Z., Jian, S., Jie, Z. & Min, Z. (2005). Exploring various knowledge in relation extraction. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.*

[32] Jiang, J. & Zhai, C. (2007). A Systematic Exploration of the Feature Space for Relation Extraction. *In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 113–120, Rochester, New York. Association for Computational Linguistics.*

[33] Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.*

[34] Douglas Baker, L. & Kachites McCallum, A. (1998). Distributional clustering of words for text classification. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. August 1998, Pages 96-103*

[35] Bekkerman, R., El-Yaniv, R., Tishby, N. & Winter, Y. (2001). On feature

distributional clustering for text categorization. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. September 2001, pages 146–153.*

[36] Cadez, I., Heckerman, D., Meek, C., Smyth, P. & White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery 7, (2003), 399–424.*

[37] Steinbach, M., Karypis, G. & Kumar, V. (2000). A comparison of document clustering techniques. *Proceedings of the International KDD Workshop on Text Mining, Boston, 525–526.*

[38] Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management, volume 24, issue 5, 1988, pages 577–597.*

[39] Alsabti, K., Ranka, S. & Singh, V. (2000). An efficient k-means clustering algorithm. *Proc First Workshop High Performance Data Mining.*

[40] Kanungo, T. Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, July 2002.*

[41] Griffiths, T. L. & Steyvers, M. (2002). A probabilistic approach to semantic representation. *In Proceedings of the 24th annual conference of the cognitive science society, p. 381–386.*

[42] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research* ***3****(Jan):* 993-1022.

[43] Hofmann, T. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 50–57.*

[44] Lee, D. & Seung, S. (2001). Algorithms for Non-negative Matrix Factorization. *Adv. Neural Inform. Process. Syst. 13.*

[45] Duriau, V., Reger, R. & Pfarrer, M. (2007). A Content Analysis of the Content Analysis Literature in Organization Studies: Research Themes, Data Sources, and Methodological Refinements. *Article in Organizational Research Methods, 10. 5-34.*

[46] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American society for information science, 41(6): 391.*

[47] Piepenbrink, A. & Gaur, A. S. (2017). Topic models as a novel approach to identify themes in content analysis. *Conference Paper in Academy of Management Proceedings, August 2017.*

[48] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W. & Hassan, A. (2022). Topic modeling algorithms and applications: A survey. *Information Systems, Volume 112, 2023, 102131, ISSN 0306-4379,*

[49] Asmussen, C. B. & Moller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Asmussen and Møller J Big Data (2019) 6:93*

[50] Asmussen, C. B. & Moller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Asmussen and Møller J Big Data (2019) 6:93*

[51] Albalawi, R., Yeap, T. H. & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell. 2020, Jul 14; 3:42. doi: 10.3389/frai.2020.00042*

[52] Poet Laureate, C. D., Buntine, W & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis*. Artificial Intelligence Review (2023) 56:14223–14255 https://doi.org/10.1007/s10462-023-10471-x*

[53] Yan, X, Guo, J., Lan, Y, & Cheng, X (2013). A biterm topic model for short texts. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web. 1445-1456. 10.1145/2488388.2488514.*

[54] Smith, A, Chuang, J., Hu, Y, Boyd-Graber, Y. & Findlater, L. (2014). Concurrent Visualization of Relationships between Words and Topics in Topic Models. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 79–82, Baltimore, Maryland, USA, June 27, 2014.*

[55] Chuang, J., Manning, C. D. & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *In Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12). Association for Computing Machinery, New York, NY, USA, 74–77.*

[56] Blei, D. M. & Lafferty, J. D. (2009). Visualizing Topics with Multi-Word Expressions. *Journal arXiv: arXiv:0907.1013v1 [stat.ML].*

[57] Sievert, C & Shirley, K. (2014). LDAvis : A method for visualizing and interpreting topics. *In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.*

[58] Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579-2605.

[59] Cui, W. et al., (2011). TextFlow: Towards Better Understanding of Evolving Topics in Text. *In IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2412-2421, Dec. 2011, doi: 10.1109/TVCG.2011.239.*

[60] Krishnan, A. (2023) Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. *Journal arXiv: arXiv:2308.11520v1 [cs.CL].*

[61] Lee, D. D & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 1999, Oct 21*; 401(6755): 788- 791.

[62] Pauca, V. P., Piper, J. & Plemmons, R. (2005). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications, volume 416, Issue 1, 2006, Pages 29-47, ISSN 0024-3795, https://doi.org/10.1016/j.laa.2005.06.025.*

[63] Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning*, 427. 424-440.

[64] Landauer, T. K., Foltz, P. & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes. 25. 259-284. 10.1080/01638539809545028.*

[65] Landauer, T. K. & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211*

[66] Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management, volume 54, issue 6, 2018, pages 1292-1307, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2018.05.006.*

[67] Arun, R., Suresh, V., Madhavan, C. & Murthy, M. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *In: Advances in Knowledge Discovery and Data Mining (pp. 391–402). Springer Berlin Heidelberg.*

[68] Allahyari, M. & Kochut, K. J. (2015). Automatic topic labeling using ontology- based topic models. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 259-264.*

[69] Aletras, N., Stevenson, M. (2013). Evaluating Topic Coherence Using Distributional Semantics. *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.*

[70] Bystrov, V., Naboka-Krell, V., Strazewska-Bystrova, A. & Winker, P. (2023). Analysing the Impact of Removing Infrequent Words on Topic Quality in LDA Models. *Journal arXiv: 2311.14505 [cs.CL]*

[71] Parlina, P. & Kusumarani, R. (2023) A Latent Dirichlet Allocation – Based bibliometric exploration of top-3 journals in management information system. *Article in Jurnal Studi Komunikas dan Media, 27(1), 77-92*

[72] Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, volume *55, issu (4):77-84, https://doi.org/10.1145/2133806.2133826*

[73] Shirinivas, S. G., Vetrivel, S. & Elango, N. (2010). Applications of graph theory in computer science an overview. *International Journal of Engineering Science and Technology, 2(9), pp. 4610-4621.*

[74] Mondal, B. & De, K. (2017). An Overview Applications of Graph Theory in Real Field. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2017, Volume 2, Issue 5, ISSN : 2456-3307*

[75] Cafieri, S., Hansen, P. & Liberti, L. (2010). Edge ratio and community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics. 81. 026105. 10.1103/PhysRevE.81.026105.*

[76] Bondy, J. A. & Murty, U. S. R. (1976). *Graph theory and applications.* Elsevier Science Publishing Co., Inc.

[77] Deo, N. (2014). Graph Theory with Applications to Engineering and Computer Science. *Prentice Hall of India, Delhi, India.*

[78] Heckmann, T., Schwanghart, W. & Phillips, J. D. (2015). Graph theory—Recent developments of its application in geomorphology. *Geomorphology, Volume 243, 2015, Pages 130-146, ISSN 0169-555X, https://doi.org/10.1016/j.geomorph.2014.12.024.*

[79] Majeed, A. & Rauf, I. (2020). Graph Theory: A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks. *Inventions 2020, 5, 10. https://doi.org/10.3390/inventions5010010*

[80] Patel, P. & Patel, C. (2013). Various Graphs and Their Applications in Real World. *International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 12, December - 2013*

[81] Chen, Y., Sanghavi, S. & Xu, H. (2012). Clustering sparse graphs. *In: Advances in neural information processing systems. 2012, pp. 2204–2212.*

[82] Chin, P., Rao, A. & Vu, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. *In: Conference on Learning Theory. 2015, pp. 391–423.*

[83] Kim, J. & Lee, J-G. (2015). Community Detection in Multi-Layer Graphs: A Survey. *In: ACM SIGMOD Record 44.3 (2015), pp. 37–48.*

[84] Bazzi, M. et al. (2016). Community detection in temporal multilayer networks, with an application to correlation networks. *In: Multiscale Modeling & Simulation: A SIAM Interdisciplinary Journal, Vol. 14, No. 1, pp. 1– 41.*

[85] Vallès-Català, T. et al. (2016). Multilayer stochastic block models reveal the multilayer structure of complex networks. *In: Physical Review X 6.1 (2016), p. 11036.*

[86] Folino, F. & Pizzuti, C. (2013). An evolutionary multiobjective approach for community discovery in dynamic networks. *In: IEEE Transactions on Knowledge and Data Engineering 26.8 (2013), pp. 1838–1852.*

[87] Lin, Y., Chi, Y., Zhu, S., Sundaram, H. & Tseng, B. (2009). Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 3, no. 2, p. 8, 2009.*

[88] Cazabet, R., Amblard, F. & Hanachi, C. (2010). Detection of overlapping communities in dynamical social networks. *In Social Computing (SocialCom), 2010 IEEE Second International Conference on, p. 309–314, IEEE, 2010.*

[89] Chen, Y., Kawadia, V. & Urgaonkar, R. (2013). Detecting overlapping temporal community structure in time-evolving networks. *In journal arXiv:1303.7226 (2013).*

[90] Xu, K. (2015). Stochastic block transition models for dynamic networks. *In: Artificial Intelligence and Statistics. 2015, pp. 1079–1087.*

[91] Ghasemian, A. et al. (2016). Detectability thresholds and optimal algorithms for community structure in dynamic networks. *In: Physical Review X 6.3 (2016), p. 31005.*

[92] Yang, T., Chi, Y., Zhu, S., Gong, Y. & Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Mach Learn (2011) 82: 157–189 DOI 10.1007/s10994-010-5214-7*

[93] Wang, C. D., Lai, J. H. & Yu, P. (2013). Dynamic Community Detection in Weighted Graph Streams. *SDM, 2013: 10.1137/1.9781611972832.17.*

[94] Harenberg S. et al. (2014). Community detection in large-scale networks: a survey and empirical evaluation. *In: Wiley Interdisciplinary Reviews: Computational Statistics 6.6 (2014), pp. 426–439.*

[95] Li, Y., He, K., Bindel, D. & Hopcroft, J. E. (2015). Uncovering the small community structure in large networks: A local spectral approach. *In: Proceedings of the 24th international conference on world wide web. 2015, pp. 658–668.*

[96] Hallac, D., Leskovec, J. & Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. *In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM. 2015, pp. 387–396.*

[97] Wang, L., Lou, T., Tang, J. & Hopcroft, J. E. (2011). Detecting community kernels in large social networks. *In: Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE. 2011, pp. 784–793.*

[98] De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. (2014). Mixing local and global information for community detection in large networks. *In: Journal of Computer and System Sciences 80.1 (2014), pp. 72–87.*

[99] Huang, X., Cheng, H. & Xu Yu, J. (2015). Dense community detection in multi-valued attributed networks. *In: Information Sciences, volume 314, pp. 77–99.*

[100] Zhang, X., Liu, H., Li, Q. & Wu, X-M. (2019). Attributed graph clustering via adaptive graph convolution. *In: journal arXiv :1906.01210 (2019).*

[101] Jin, D., Liu, Z., Li, W., He, D. & Zhang, W. (2019). Graph convolutional networks meet Markov random fields: Semi-supervised community detection in attribute networks. *In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019, pp. 152–159.*

[102] Qin, M., Jin, D., Lei, K., Gabrys, B. & Musial-Gabrys, K. (2018). Adaptive community detection incorporating topology and content in social networks. *In: Knowledge-Based Systems 161 (2018), pp. 342–356.*

[103] Alzahrani, T., Horadam, K. J. (2016). Community Detection in Bipartite Networks: Algorithms and Case studies. *In Complex Systems and Networks: Dynamics, Controls and Applications, pages 25–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.*

[104] Zhang, P., Wang, J., Li, X., Di, Z. & Fan, Y. (2008). The clustering coefficient and community structure of bipartite networks. *Physica A : Statistical Mechanics and its Applications, vol. 387, no. 27, p. 6869–6875, 2008.*

[105] Guillaume, J. L. & Latapy, M. (2004). Bipartite structure of all complex networks. *Information Processing Letters, Volume 90, Issue 5,2004,Pages 215-221,ISSN 0020-0190, https://doi.org/10.1016/j.ipl.2004.03.007.*

[106] Costa, A. & Hansen, P. (2014). A locally optimal hierarchical divisive heuristic for bipartite modularity maximization. *Optimization Letters, 8(3):903–917, 2014.*

[107] Guimerà, R., Sales-Pardo, M. & Amaral, L. (2007). Module identification in bipartite and directed networks. *Physical review. E, Statistical, nonlinear, and soft matter physics. 76. 036102. 10.1103/PhysRevE.76.036102.*

[108] Ganji, M., Seifi, A., Alizadeh, H., Bailey, J. & Stuckey, P.J. (2015). Generalized Modularity for Community Detection. *9285. 655-670. 10.1007/978-3-319-23525-7_40.*

[109] Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E, vol. 76, no. 6, p. 066102, 2007.*

[110] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, K. & Wagner, D. (2006). Maximizing modularity is hard. *Journal arXiv: physics/0608255*

[111] Guimerà, R., Sales-Pardo, M. & Nunes Amaral, L. A. (2007). Module identification in bipartite and directed networks. *Physical Review E, 76(3):036102, 2007.*

[112] Newman, M. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, vol. 69, no. 2, p. 026113, 2004.*

[113] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.*

[114] Boutalbi, R., Ait-Saada, M., Iurshina, A., Staab, S. & Nadif, M. (2022). Tensor-based Graph modularity for text data clustering. *2227-2231. 10.1145/3477495.3531834.*

[115] Sun, H., Guyon, I. (2023). Modularity in Deep Learning: a Survey. *Journal arXiv:2310.01154 [cs.LG]*

[116] Esfahlani, F. Z., Jo, Y., Puxeddu, M. G., Merritt, H., Tanner, J. C., Greenwell, S., Patel, R., Faskowitz, J. & Betzel, R. F. (2021). Modularity maximization as a flexible and generic framework for brain network exploratory analysis. *NeuroImage,Volume 244, 2021, 118607, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2021.118607.*

[117] Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E, vol. 80, no. 5, p. 056117, 2009.*

[118] Lancichinetti, A., Radicchi, F., Ramasco, J. & Fortunato, S. (2011). Finding statistically significant communities in networks. *Plos One, vol. 6, no. 4, p. e18961, 2011.*

[119] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 9, p. 2658– 2663, 2004.*

[120] Danon, L., Diaz-Guilera, J., Duch, J. & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment, vol. 2005, no. 09, p. P09008, 2005.*

[121] Fortunato, S. (2010). Community detection in graphs. *Physics Reports, vol. 486, no. 3, p. 75–174, 2010.*

[122] Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, vol. 69, no. 6, p. 066133, 2004.*

[123] Wang, Y. (2023). Review on greedy algorithm. *In Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics DOI: 10.54254/2753-8818/14/20241041*

[124] Clauset, A., Newman, M. & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70(6):066111, 2004.*

[125] Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008, 2008.*

[126] Mahoney, M.W., Orecchia, L. & Vishnoi, N. K. (2012). A Local Spectral Method for Graphs: With Applications to Improving Graph Partitions and Exploring Data Graphs Locally. *Journal of Machine Learning Research 13 (2012) 2339-2365*

[127] Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008.*

[128] Johnson, D. S., Aragon, C. R. L., McGeoch, A. & Schevon, C. (1989). Optimizartion by simulated annealing: an experimental evaluation; part 1, graph partitioning. *Operation Research, Vol. 37, No. 6. November-December 1989.*

[129] Guimerà, R., Sales Pardo, M. & Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys.Rev.E70(2)(2004)025101 (R).*

[130] Raghavan, U. N., Albert, R. & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E, 76(3):036106, 2007.*

[131] Liu, X. & Murata, T. (2009). Community Detection in Large-Scale Bipartite Networks. *In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, volume 1, pages 50–57, 2009.*

[132] Barber, M.J. & Clark, J. W. (2009). Detecting network communities by propagating labels under constraints. *Physical Review E, 80(2):026129, 2009.*

[133] Gervens, T. & Grohe, M. (2022). Graph Similarity Based on Matrix Norms. *In journal arXiv:2207.00090v1 [cs.DM] 30 Jun 2022*

[134] Singh, H., Sharma, R. (2012). Role of Adjacency Matrix & Adjacency List in Graph Theory. *Article in International journal of computers & technology · August 2012 DOI: 10.24297/ijct.v3i1c.2775*

[135] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing, 17(4):395–416, 2007.*

[136] Bouchard, M., Jousselme, A.-L. & Doré, P.-E. (2013). A proof for the positive definiteness of the Jaccard index matrix. *In International Journal of Approximate Reasoning 54 (2013) 615–626*

[137] Caporossi, G., Camby, E. (2017). The extended jaccard distance in complex networks. *Les Cahiers du GERAD G–2017–77*

[138] Takano, Y., Iikima, Y., Kobayashi, K., Sakuta, H., Sakaji, H., Kohana, M. & Kobayashi, A. (2020). Improving Document Similarity Calculation Using Cosine-Similarity Graphs. *10.1007/978-3-030-15032-7_43.*

[139] Troussas, C., Krouska, A., Tselenti, P., Kardaras, D. K. & Barbounaki, S. (2023). Enhancing Personalized Educational Content Recommendation through Cosine Similarity-Based Knowledge Graphs and Contextual Signals. *Information 2023,14,505. https://doi.org/ 10.3390/info14090505*

[140] Estrada E. & Hatano, N. (2008). Communicability in complex networks. *Phys. Rev. E, 77:036111, Mar 2008.*

[141] Estrada E. & Hatano, N. (2009). Communicability Graph and Community Structures in Complex Networks. *arXiv:0905.4103 [physics.soc-ph]*

[142] Caporossi, G., Camby, E. (2023). Complex Networks Analysis. *HEC Montreal, 2023.*

[143] Needham, M. & Hodler, A. (2019). Graph Algorithms. Published by O'Reilly Media, Inc.

[144] Ester, M., Kriegel, H-P., Sander, J., Xu, X. (1996). A Density-Based algorithm for discovering Clusters in a large spatial databases with noise. *From: KDD-96 Proceedings. 1996, AAAI (www.aaai.org)*

[145] Fahim, A. M.(2023). Adaptive Density-Based Spatial Clustering of Applications with Noise (ADBSCAN) for Clusters of Different Densities. *Computers, Materials & Continua, 2023. 75. 3695-3712. 10.32604/cmc.2023.036820.*

[146] Wang, J. & Dong, Y. (2020). Measurement of text similarity: A survey. *Information 2020, 11, 421; doi:10.3390/info11090421*

[147] Gupta, M., Singh, A. & Cherifi, H. (2016). Centrality Measures for Networks with Community Structure. *Journal arXiv:1601.07108 [cs.SI]*

[148] Saxena, A., Iyengar S. (2020). Centrality Measures in Complex Networks: A Survey. *arXiv:2011.07190 [cs.SI]*

[149] Bloch, F., Jackson, M. O. & Tebaldi, P. (2021). Centrality Measures in Networks. *Journal arXiv:1608.05845 [physics.soc-ph]*

[150] Roder, M., Both, A. & Hinneburg A. (2015). Exploring the Space of Topic Coherence Measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 399-408. 10.1145/2684822.2685324.*

[151] O'Callaghan, D., Greene, D., Carthy, J. & Cunningham P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications, Volume 42, Issue 13, 2015, Pages 5645-5657, ISSN 0957-4174.*

[152] Klein, D.J. & Randic, M. (1993). Resistance distance. Journal of Mathematical Chemistry. 12. 81-95. 10.1007/BF01164627.

[153] Du, K. (2002). Evaluating Hyperparameter Alpha of LDA Topic Modeling. Conference: DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022)

[154] Danon, L., Duch, J., Diaz-Guilera, A. & Arenas, A. (2005). Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment. 2005. 10.1088/1742-5468/2005/09/P09008.

[155] Mahmoudi, A. & Jemielniak, D. (2024). Proof of biased behavior of Normalized Mutual Information. Sci Rep 14, 9021 (2024). https://doi.org/10.1038/s41598-024-59073-9

[156] Chomsky, N. (2002). *Syntactic Structures.* Second Edition, Mouton de Gruyter, Berlin, Germany.

[157] Salton, G., Wong, A. & Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 18, 613-620.*

[158] Friedman, R. (1997). Towards a (Bayesian) Convergence? *Int'l J. Evidence & Proof 1 (1997): 348-53.*

[159] Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning. 29. 131-163. 10.1023/A:1007465528199.*

[160] Sahami. M. (1996). Learning limited dependence Bayesian classifiers. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 335–338.*

[161] Langley, P. Iba, W. & Thompson, K. (1998). An Analysis of Bayesian Classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence. 90.

[162] Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. *In Proceedings 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley (pp. 42–49).*

[163] Duriau, V., Reger, R. & Pfarrer, M. (2007). A Content Analysis of the Content Analysis Literature in Organization Studies: Research Themes, Data Sources, and Methodological Refinements. *Organizational Research Methods. 10. 5-34. 10.1177/1094428106289252.*

[164] Quinn, K. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54(1):209–28.*

[165] Landauer, T., Foltz, P. & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes. 25. 259-284. 10.1080/01638539809545028.*

[166] Chernoff, H. (1971). The use of faces to represent points in n-dimensional space graphically. *Technical Report 71, Department of Statistics, Stanford University.*

[167] Keim, D. (2000). Designing Pixel-oriented Visualization Techniques : Theory and Applications. *First publ. in: IEEE transactions on visualization and computer graphics 6 (2000), 1, pp. 59-78. 6. 10.1109/2945.841121.*

[168] Battista, G., Eades, P., Tamassia, R. & Tollis, I. (1994). Algorithms for Drawing Graphs: An Annotated Bibliography. *Computational Geometry. 4. 235-282. 10.1016/0925-7721(94)00014-X.*

[169] Hinton, G. & Roweis, S. (2002). Stochastic neighbor embedding. *In Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS'02). MIT Press, Cambridge, MA, USA, 857–864.*

[170] Newman, M.E.J. (2003) The Structure and Function of Complex Networks. *SIAM Review, 45, 167-256. https://doi.org/10.1137/S003614450342480*

[172] Barber, M. & Clark, J. (2009). Detecting Network Communities by Propagating Labels under Constraints. *Physical review. E, Statistical, nonlinear, and soft matter physics. 80. 026129. 10.1103/PhysRevE.80.026129.*

[173] Liu, X. & Murata, T. (2010). An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks. *JACIII. 14. 408-415. 10.20965/jaciii.2010.p0408.*

[174] Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *Journal of the Acoustical Society of America, 22, 725-730. http://dx.doi.org/10.1121/1.1906679*

[175] Sabidussi G. (1966). The centrality of a graph. *Psychometrika. 1966 Dec;31(4):581-603. doi: 10.1007/BF02289527. PMID: 5232444.*

[176] Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. Sociometry, 40(1), 35–41. https://doi.org/10.2307/3033543

[177] Chang, J., Boyd-Graber, J., Gerrish, S. Wang, C. & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems. 32. 288-296.*

[178] *Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011).* Optimizing Semantic Coherence in Topic Models. *In Proceedings of the 2011 Conference on Empirical*

Methods in Natural Language Processing, pages 262–272, Edinburgh, Scotland, UK.. Association for Computational Linguistics.

[179] *Aletras, N. & Stevenson, M. (2013)* Evaluating Topic Coherence Using Distributional Semantics. *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.*

[180] Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 399-408. 10.1145/2684822.2685324.*