Analyse et modélisation de la diversité des structures relationnelles à l'aide de graphes multipartis

Rémy Poulain

Sorbonne Université

Édite

Lip6/ Complex Networks

Analyse et modélisation de la diversité des structures relationnelles à l'aide de graphes multipartis.

Par Rémy Poulain

Thèse de doctorat d'informatique

Dirigée par Clémence Magnien et Fabien Tarissan

Les rapporteurs seront :

- Talel Abdessalem.
- Céline Robardet.

Table des matières

Chapitre

1

Introduction

Il n'est plus à prouver que le numérique, Internet et le web ont entraîné une révolution notamment dans la manière de s'informer. Comme toute révolution, elle est suivie par une série d'enjeux : égalité de traitement des utilisateurs et des fournisseurs, consommations écologiquement durable, liberté d'expression et censure etc... [1] Il est nécessaire que la recherche apporte une vision claire de ces enjeux. Pour introduire notre travail nous allons d'abord décrire quelques enjeux liés à l'information sur Internet, pour nous focaliser sur l'accès à l'information et plus précisément sa visibilité permise par les algorithmes facilitant la recherche d'information. Ceci nous permettra d'axer notre analyse sur la diversité induite par ces algorithmes que l'on peut analyser de l'intérieur (en étudiant directement l'algorithme), ou de l'extérieur (en analysant ses traces). Nous nous intéresserons plus particulièrement à l'analyse des traces et proposerons un formalisme adapté : les graphes n-parti. Ainsi nous pourrons définir le cadre de cette étude : analyser la *diversité* sur des graphes n-parti. Nous finirons par présenter rapidement la structure de la suite du manuscrit en explicitant mieux le titre complet de notre thèse.

1.1 L'INFORMATION SUR LE WEB : DES ENJEUX DIVERS

Aujourd'hui, quand on pense enjeux de l'information sur Internet, on pense naturellement aux *fake news*. On peut les définir comme des informations fabriquées qui imitent le contenu des médias d'information dans la forme mais pas dans leur organisation ni dans leur intention [53]. Elles ont une influence assez forte sur la façon de s'informer des gens et donc, par exemple, de voter [6]. Cependant comme le rappelle David Chavalarias dans un article dédié au phénomène [19] :"Le phénomène des fausses informations cache donc un phénomène de bien plus grande ampleur : un saut technologique dans les stratégies d'influence d'opinion issu de l'appropriation, par certains groupes coordonnés, des possibilités et des outils offerts par les technologies numériques au sens large." En d'autres termes il y a des entités (entreprises, États, groupes

1. Tout ceci sans compter les enjeux économiques liés à cette nouvelle forme d'information [14].

politiques) qui utilisent les nouvelles plateformes numériques pour avoir une influence sur la vie politique. Il y a donc une nécessité d'identifier l'impact de ces groupes. Afin de contribuer à cette question, nous allons nous intéresser dans cette thèse à l'une des formes que prend cet impact et qui est liée à la visibilité des informations partagées par ces groupes. Car pour avoir un impact, une information doit avant tout acquérir de la visibilité. Pour comprendre en quoi la visibilité est importante sur Internet, il suffit d'avoir une notion de l'ampleur du phénomène. Selon le site internetlivestats [2] il y a aujourd'hui plus de 1 750 millions de sites Internet, et $4,6$ milliards d'utilisateurs, ce qui rend l'analyse humaine de ce phénomène impossible. Les quatre milliards et demi d'utilisateurs ne vont donc pas directement fouiller page après page, mais ils utilisent des algorithmes facilitant la recherche d'information. Pour illustrer ce phénomène on décompte aussi, sur ce site, 4 milliards de recherches faites via le moteur de recherche Google par jour (ce qui représente 65% des recherches sur internet). Ainsi les algorithmes facilitant la recherche d'information ont une place primordiale dans la visibilité de l'information. C'est cette place que nous allons évoquer dans cette section.

1.1.1 *Les algorithmes facilitant la recherche d'information, des outils neutres ?*

Nous allons parler de deux phénomènes, que nous définissons dans le prochain paragraphe, qui sont parfois un peu confondus dans les médias : le phénomène de chambre d'écho et le phénomène de bulle de filtre. Le phénomène de bulle de filtre a été conceptualisé par Élie Pariser [71] et nous le distinguons du phénomène de chambre d'écho [98]. Le phénomène de chambre d'écho, qui fait référence à un phénomène acoustique, désigne le fait que des individus sont plus souvent exposés à des informations qui leur correspondent. En effet, il y a une corrélation entre le milieu d'un individu (ses amis, sa famille, son travail, ses voisins...) et les informations qu'il reçoit. De plus il y a une corrélation entre le milieu d'un individu et ses opinions (en général on fréquente des personnes qui nous ressemblent en terme d'opinion). Ce phénomène s'appelle l'homophilie [61] et il est très présent sur les réseaux sociaux (85% d'homophilie sur Twitter par exemple). Il y a donc de fait une corrélation entre nos opinions et les informations que l'on reçoit. Ce phénomène n'est pas nouveau et n'est pas seulement lié à l'avènement d'Internet, c'est un phénomène qui existe quand il y a une correlation entre les individus que l'on fréquente et notre façon de nous renseigner. On parle par exemple d'exposition sélective due à notre environnement [86]. Ainsi, connaître la structure de ce réseau, c'est à dire ici connaître les différents liens entres les entités, nous donne une grande information sur le milieu dans lequel les utilisateurs évoluent et la façon qu'ils ont de s'informer. C'est pour cela que la modélisation en terme de graphe nous parait adaptée. Décrit avec le vocabulaire des graphes, que nous définirons formellement plus tard, notre voisinage influence notre façon de nous informer et notre voisinage n'est évidemment pas un ensemble de nœuds pris au hasard.

Le phénomène de bulle de filtre est lui directement lié aux algorithmes de recommandation. Il a été formalisé en premier par Eli Pariser dans son livre [71]. Ce phénomène désigne le fait que les algorithmes facilitant la recherche d'information filtrent l'information en la sélectionnant,

2. `https://www.internetlivestats.com/`

ce qui relève de la nature même des algorithmes facilitant la recherche d'information. Il nous dit aussi que ce qui est plus préoccupant, c'est que l'utilisateur n'a pas connaissance de ce qui est invisibilisé par l'algorithme, ni même de l'existence d'un biais.

Sans entrer en détail dans le détail de ces deux phénomènes (ce que nous ferons dans la prochaine section), une étude récente permet d'ores et déjà de saisir les liens qu'entretiennent les phénomènes de bulles de filtre et de chambres d'écho. Il s'agit d'une étude réalisée en 2015 par Eytan Bakshy, Solomon Messing et Lada A Adamic [9] et qui nous permet de voir une première estimation du lien entre ces deux phénomènes (bulles de filtre et chambre d'écho), mais aussi en gardant en tête cette question (quel sont les liens entre deux phénomènes), qui sera abordées plus profondément dans la sous-section suivante. Les auteurs s'intéressent ici à la manière dont on accède à l'information sur Facebook. Pour ceci ils vont s'intéresser à un grand jeu de données de 10 millions de comptes Facebook et 7 millions d'URL partagées par ces comptes durant une durée de 6 mois entre 2014 et 2015. Ces comptes ont l'avantage d'avoir indiqué leur affiliation idéologique (libéral, conservateur ou neutre). De plus les auteurs ont extrait de leurs URL ce qu'ils appellent des "hard news", c'est à dire des publications "factuelles" ou "sérieuses" (celles qui ont une forte tendance à être des publications politiques). En regardant l'alignement moyen [3] des personnes qui partagent ces articles, ils peuvent alors donner un score d'alignement à ces liens. Cette base de données une fois constituée leur permet de voir le rapport entre des utilisateurs et les liens qu'ils partagent en ayant une vision de l'alignement de chaque utilisateur et de chaque lien. Remarquons un biais qu'il est important d'identifier ici (et que nous tenterons de limiter dans d'autres études) : l'alignement des articles est déterminé par l'alignement des personnes qui le partagent. Ce qui signifie que l'on n'a pas regardé (c'est plus difficile) le contenu des articles. Comme c'est aussi une dimension qui va être étudiée par les auteurs (comment ces articles sont partagés en fonction de l'alignement), il faut donc éviter de faire des observations trop tautologiques.

Tout d'abord les auteurs remarquent que nos amis ne sont pas alignés politiquement au hasard : il y a en moyenne seulement 20% des amis qui s'identifient comme de l'autre bord politique. De plus les auteurs s'intéressent aux informations *"cross-cutting"* c'est à dire aux informations qui viennent de l'autre bord, que l'on traduira par *transversales*. Ils remarquent que seulement 25% des informations partagées par les libéraux sont transversales, phénomène qui, bien qu'amplifié, est également faible pour les conservateurs (35 %). La question reste de savoir à quel point le NewsFeed (l'algorithme de recommandation de Facebook) a une influence sur cette polarisation des partages. Les auteurs estiment rapidement qu'après intervention du NewsFeed les utilisateurs sont légèrement moins exposés à des contenus transversaux.

Nous pouvons, sans avoir de réponse directe, avoir une meilleure vision de l'enjeu et de ce qu'il serait intéressant d'analyser. Le principe est plutôt bien résumé dans le titre de cet article du Monde du 18 avril 2018, " Facebook est un média au même titre que les autres éditeurs". C'est évidemment une question qui fait débat, ce qui est important à noter ici c'est que, même si la question de la responsabilité de ces choix n'est pas tranchée, les choix faits par l'algorithme ont des conséquences qui posent des questions éthiques et politiques connus

3. L'alignement n'est pas toujours regardé de la même manière soit ils classent les individus sur une échelle à l'aide d'un questionnaire, soit ils les classent directement dans les trois catégories libéral conservateur ou neutre.

des éditeurs.[4] L'algorithme de Facebook fait forcément des choix qui influencent l'exposition aux informations des utilisateurs. Une enquête faite par MediaPart a par exemple révélé l'influence que cet algorithme a eue sur l'audience de la gauche radicale[5]. Pour résumer, que l'algorithme soit le plus neutre possible (en supposant que ça soit possible) ou non, il fait des choix qui influencent l'exposition médiatique des utilisateurs. Ces choix peuvent donc placer les concepteurs et les gérants de ces algorithmes en face de questionnement éthiques qui ressemblent à ceux des éditeurs.

1.1.2 *Une implication politique directe*

Ce premier exemple illustre assez bien l'influence que l'algorithme de Facebook peut avoir sur notre façon de nous informer. Mais cela ne nous dit pas si cette influence a un impact dans notre vie sociale, collective. Pour cela, nous allons nous intéresser à une étude qui s'est précisément focalisée sur la mesure quantitative de l'impact que peut avoir un algorithme de classement dans le contexte d'une élection politique en condition réelle ou proche. Cette étude a été écrite par Robert Epstein et Ronald E. Robertson elle est parue en 2015 [27]. Nous l'avons choisie parce que les expériences qui y sont présentées sont parlantes, mais aussi parce que ses auteurs se sont appliqués à montrer en quoi ces expériences évaluent l'impact direct que ces algorithmes peuvent avoir.

Les auteurs étudient donc l'impact de l'algorithme facilitant la recherche d'information qui ressemble à Google, sur la façon de s'informer et la façon de voter des utilisateurs. L'idée est de faire utiliser à des participants un algorithme facilitant la recherche d'information biaisé, fabriqué pour l'expérience, pour ensuite voir la différence dans la manière de voter des sujets qui ont utilisé l'algorithme biaisé et de ceux qui se sont informés avec celui qui est neutre. Ce qui est important dans les différentes expériences présentées c'est que le contenu des articles n'est jamais changé, la seule chose qui est modifiée par les expérimentateurs est l'ordre facilitant la recherche d'information donné par l'algorithme pour des recherches spécifiques. C'est donc bien la vision et l'accès à l'information qui sont analysés et non la diversité des articles et des points de vue existant sur Internet.

Intéressons nous donc aux expériences proposées. Les auteurs ont mené cinq expériences qu'ils ont classées dans trois études. La première étude regroupe les trois premières expériences, les deux études suivantes sont basées sur une expérience chacune. L'étude 2 analyse l'expérience 4 et l'étude 3 l'expérience 5.

Ces cinq expériences sont quasiment toutes basées sur la même méthode, nous allons donc d'abord la présenter globalement puis analyser brièvement ce qui les différencie. Les participants sont renseignés rapidement sur les candidats d'une élection (les auteurs parlent de la lecture d'une courte biographie). Ensuite ils remplissent un questionnaire pour savoir leur avis

4. https://www.lemonde.fr/idees/article/2018/04/18/facebook-est-un-media-au-meme-titre-
5287181_3232.html

5. https://www.mediapart.fr/journal/france/290819/facebook-aneantit-l-audience-d-une-p
onglet=full.

sur les différents candidats. Puis ils doivent se renseigner sur les candidats en utilisant l'algorithme facilitant la recherche d'information dédiée à l'expérience. Certains des sujets reçoivent pour toutes leurs recherches les mêmes classements qu'un algorithme de base, d'autres au contraire reçoivent des classements où un candidat a été mis en valeur. Plus précisément, un groupe de sujets va être associé à un candidat et pour toutes leurs recherches ils vont recevoir un classement où ce candidat sera valorisé : les articles pro vont être mis en premier et les articles contre en dernier. Une fois que les sujets se sont assez renseignés [6] ils remplissent une autre fois un questionnaire pour avoir leur avis sur les candidats. C'est alors que les auteurs analysent la différence de résultat des questionnaires suivant le groupe du sujet.

Pour ce faire, les auteurs définissent un indicateur explicite le *"vote manipulation power" (VMP)* que nous traduirons par intensité de la manipulation sur le vote. Le *VMP* est défini pour chaque candidat. On sélectionne le groupe *biaisé* pour ce candidat, c'est à dire le groupe de sujets qui a reçu un classement en faveur de ce candidat, pour lui appliquer cet indicateur. Si x est la proportion des sujets qui voteraient pour ce candidat avant d'utiliser l'algorithme et x' après : le VMP est défini ainsi $VMP = \frac{x'-x}{x}$. C'est une façon de quantifier l'impact qu'a l'algorithme sur le vote des électeurs.

Regardons maintenant plus précisément ces cinq expériences. Les trois premières expériences concernent une centaine de personnes chacune. Elles sont sélectionnées à San Diego en Californie pour élire le premier ministre australien parmi les deux candidat.e.s Tony Abbott et Julia Gillard. La quatrième expérience est similaire au trois premières mais avec une plus grande échelle : elle concerne 2100 américains choisis dans tous les Etats, pour la même élection que les précédentes études. La dernière expérience est une reproduction de la précédente dans un autre contexte : l'élection de la chambre basse du parlement indien en 2014. Cette fois-ci les sujets avaient le choix entre trois candidats (Gandhi, Kejriwal, Modi), et de plus ils étaient réellement concernés par l'élection puisque c'étaient des sujets indiens. Ce sont les deux points primordiaux qui différencient le contexte de la dernière expérience par rapport aux précédentes. Cette progression permet aux auteurs d'affiner leurs recherches et de tester leurs hypothèses dans des cas différents. [7]

Les résultats sont assez variables mais le VMP est toujours positif : il varie entre 36 % et 64% sur les quatre premières expériences mais est seulement à 10 % sur la dernière. Pour évaluer l'impact réel de cet effet les auteurs font une série de suppositions. Prenons en une : si 80% des électeurs ont accès à Internet et que 10% d'entre eux sont indécis avant l'élection et que l'on arrive à biaiser l'algorithme pour avoir un VPM de 25% ceci pourrait changer de 2% le résultat d'une élection. Ceci aurait pu par exemple changer le premier tour de la dernière élection française pour deux candidats François Fillon ou Jean Luc Mélenchon, qui auraient pu finir au second tour [8]. Dans ces cas, avec des suppositions qui ne sont pas aberrantes, on peut considérer qu'avec une stratégie similaire, si quelqu'un ayant le pouvoir d'intervenir sur l'algorithme facilitant la recherche d'information de Google (rappelons qu'il représente 65%

6. Il y a un temps fixé.

7. Par exemple l'avis des sujets n'est pas toujours demandé de la même façon (binaire, note entre -5 et 5...).

8. En pourcentage des inscrits François Fillon a obtenu $20,01\%$ et $19,58$ pour Jean Luc Mélenchon à moins de 2% d'une place au second tour (Marine Le Pen a obtenu$21,30\%$)

des recherches) et le désirait, il aurait pu changer le résultat de cette élection. C'est cet effet qui est appelé le *search engine manipulation effect (SEME)* soit l'effet de manipulation par les moteurs de recherches. Ce premier point est évidemment la preuve d'un enjeu majeur pour la société.

De plus, les auteurs ont utilisé les trois premières expériences pour se poser la question de la prise de conscience du biais de l'algorithme. Sur la première expérience 25% des participants ont, dans leur questionnaire final, fait état de suspicions de biais de l'algorithme. Dans la deuxième et la troisième expérience, les auteurs ont rajouté un *masque* aux classements retournés. Les classements biaisés étaient biaisés mais de façon un peu plus subtile en rajoutant des articles contre le candidat dans les premiers articles (mais jamais les tout premiers, ce qui est important pour avoir un fort impact). Ces expériences ont eu un résultat probant car au fil des expériences il y a de moins en moins de participants qui ont exprimé ce genre de ressenti et personne ne s'en est rendu compte dans la troisième expérience (a minima personne ne l'a exprimé dans les commentaires).

Donc, non seulement il est possible, selon certaines conditions, d'influencer le résultat d'une élection via l'algorithme facilitant la recherche d'information, mais en plus il est possible de le faire d'une façon imperceptible pour les utilisateurs.

1.2 UNE NÉCESSITÉ D'ÉVALUER LES ALGORITHMES FACILITANT LA RECHERCHE D'INFORMATION

Nous avons vu dans la précédente section que l'information sur Internet et plus précisément la visibilité de cette information était un enjeu majeur à analyser. Évidemment dans le cas précédent, nous supposions une mauvaise intention des concepteurs. Mais même même avec une apparente neutralité à leur égard, il est possible de montre que l'algorithme facilitant la recherche d'information de Google amplifie les torts actuels de la société comme le racisme [68] ou le sexisme [20] par exemple. Scientifiquement ceci nous pousse à chercher une manière d'analyser ces algorithmes. L'une des questions qui se pose alors est de savoir quelle notion est la mieux à même de décrire, caractériser, même quantifier l'impact de ces algorithmes. C'est dans cette perspective qu'une notion nous est apparue comme centrale : la diversité est une notion qui est assez opposée à la notion d'enfermement que représentent les chambres d'écho et les bulles de filtres. Nous avons choisi d'utiliser des mesures pour évaluer ce concept de *diversité*. Nous expliquerons plus formellement dans le chapitre suivant ce que représentent ces mesures. Pour reprendre l'un des exemples précédents, nous pouvons par exemple exprimer les effets négatifs des SEME par un manque de diversité, les classements biaisés pour un candidat ne contenant que des articles pour ce candidat (au moins dans le top 5), alors que le classement non biaisé aura une plus grosse diversité de points de vue. La diversité ainsi que d'autres notions (sérendipité, nouveauté, couverture) sont des notions bien installées dans la recherche sur les algorithmes facilitant la recherche d'information [45].

Nous allons voir comment évaluer la diversité produite par les algorithmes d'abord en considérant tout d'abord le point de vue de l'intérieur (c'est à dire en ce plaçant à l'intérieur de

l'algorithme et en modifiant le code afin d'observer le résultat) puis celui de l'extérieur (c'est à dire en analysant les traces laissées par ces algorithmes). Ceci nous permettra de dégager le cadre de ce manuscrit.

1.2.1 *Analyser la diversité de l'intérieur*

Le premier réflexe quand on a un algorithme qui ne prend pas en compte une dimension (ici la diversité) est de tenter de l'améliorer en lui rajoutant cette dimension. Si on assimile les algorithmes facilitant la recherche d'informations à des algorithmes d'optimisation, qui chercheraient à minimiser une fonction de coût, il suffirait d'ajouter à cette fonction de coût une évaluation de l'absence de diversité pour intégrer ceci dans l'algorithme. Cette approche a fait l'objet de nombreuses études [106], [107], [95],[101],[16],[8],[43],[75], dans lesquels est systématiquement soulevée la question du compromis entre augmenter la diversité et augmenter la précision (*accuracy*) du résultat.

Nous pouvons essayer de comprendre en quoi ces deux objectifs semblent opposés. Le but d'un algorithme facilitant la recherche d'information est de donner à l'utilisateur un résultat lié à une recherche, mesurer sa précision signifie donc mesurer à quel point les éléments du classement répondent à la question donnée (ou *requête*). Plus l'algorithme donnera des réponses adaptées à la requête plus il sera considéré (à première vue) comme précis. Nous pouvons imaginer que plus les réponses sont proches de la requête plus elles sont proches les unes des autres. La diversité d'un classement exprime, de façon simplifiée, la façon dont sont présents dans ce classement des articles différents.

Ainsi, par essence, la précision semble s'opposer à la diversité. Dans tous ces articles ce principe est remis en question en pratique, mais la question reste cependant ouverte. Qu'il soient opposés ou pas, le but de ces articles et d'arriver à gérer au mieux le compromis. Nous allons voir qu'une des études suivantes apporte une autre réponse.

Une autre façon d'analyser les algorithmes est de les simuler pour les comparer, approche adoptée par plusieurs études comme [18] et [62].Détaillons l'une d'entre elles, conduite par Judith Möller, Damian Trilling, Natali Helberger et Bram van Es et qui se concentre sur des recommandations d'articles issus de médias allemands. Les auteurs utilisent une bases de données contenant des médias allemands : 1000 articles publiés en septembre 2016. Pour chacun des articles les auteurs simulent sept différentes logiques de recommandation et récupèrent 3 autres articles qui seraient recommandés pour chacun des paradigmes. Sans entrer dans le détail de chacun des paradigmes [9], notons que les auteurs considèrent ici le filtrage collaboratif [10][84] et le filtrage sémantique [11] [4]. Pour résumer la différence faite ici entre ces deux

9. Notons que nous reprenons les noms et les définitions de ces paradigmes qui ont étés fixés par les auteurs.
10. Ce processus "automatise le processus de " bouche à oreille " : les articles sont recommandés à un utilisateur sur les valeurs attribuées par d'autres personnes aux goûts similaires.
11. Cette approche recommande des éléments qui correspondent à un ou des articles précédemment utilisés par le même utilisateur sur un certain nombre de critères prédéfinis. Dans le cas d'articles de presse, ces fonctionnalités peuvent être des mots ou des sujets, mais également l'auteur ou la source d'un article.

paradigmes nous pouvons dire que le filtrage sémantique désigne ce qui peut être fait par rapport au passé de l'individu cible alors que le filtrage collaboratif, lui, se base plus sur deux paradigmes pouvant être basés sur les objets ou les utilisateurs, ce qui donne donc 4 paradigmes différents. À cela les auteurs ajoutent (comme logique de recommandation), le choix d'éditeurs humains, une recommandation basée sur la popularité et une recommandation basée sur de l'aléatoire. Ensuite les auteurs veulent évaluer la "diversité" pour chacun des résultats de recommandation. Là encore, sans entrer dans les détails de ces indicateurs de diversités (qui feront l'objet des sections 2.2 et 2.3 de cette thèse), on peut néanmoins noter que les auteurs utilisent comme indicateur principal l'entropie de Shannon, que nous utiliserons également par la suite.

Cet article conclut qu'il n'y a pas de réduction de la diversité sur des algorithmes facilitant la recherche d'information, même s'il y a une invisibilisation des "petits" articles, c'est à dire des articles moins populaires. Malheureusement ces algorithmes simulés sont assez rudimentaires et ne reflètent pas la complexité des algorithmes réels qu'utilisent Google et Facebook par exemple. Cependant il est important de voir que les bases de ces algorithmes ne sont finalement pas réductrices de diversité en tant que telles, mais que ce serait leur utilisation répétée qui leur donnerait cette propriété de bulles filtrantes.

Il est important de noter que la question de la diversité dans les résultats des algorithmes n'est pas seulement une questions de performance du coté des concepteurs de ces algorithmes mais également (et surtout) un besoin exprimé par les utilisateurs, comme le montre une étude[15] mêlant simulation et enquêtes sur les utilisateurs. Dans cette étude conduite par Sylvain Castagnos, Armelle Brun et Anne Boyer, les auteurs ont ajouté à l'algorithme de base des coûts capturant la diversité. Il ressort des enquêtes sur les utilisateurs qu'ils perçoivent la diversité et qu'ils l'apprécient. Cette étude montre que les utilisateurs apprécient la diversité : c'est d'ailleurs pour cela aussi que les développeurs d'algorithmes tentent de prendre en compte la diversité dans leurs résultats.

1.2.2 *Analyser la diversité de l'extérieur*

Ainsi les algorithmes peuvent être améliorés pour pouvoir augmenter la diversité des classements produite. Cependant, comme nous l'avons expliqué, les algorithmes utilisés pour établir ces résultats sont des version simplifiées des algorithmes utilisés en pratique. C'est d'ailleurs le problème principal : pour des raisons de confidentialité, il est impossible d'analyser l'algorithme de l'intérieur. Pour pouvoir répondre aux enjeux définis dans la première section, une autre approche consiste à analyser l'algorithme de l'extérieur, c'est à dire comme une boîte noire. Ceci est d'autant plus vrai quand il y a suspicion de mauvaise intention de la part du développeur. Cela signifie que l'on ne va pas analyser directement l'algorithme en tant que tel mais ses résultats : les traces de l'algorithme. Comme le soulignent J Kulshrestha, M Eslami et J Messias dans une étude portant sur l'évaluation des biais des algorithmes [48] "on a caractérisé le biais de l'algorithme facilitant la recherche d'information sur la plateforme Twitter, sans connaître ce qui se passe en interne".

Pour ce faire il faut donc raisonner sur des données récupérées *a posteriori*, comme par exemple la liste des classements des algorithmes facilitant la recherche d'information, ou alors des informations sur le comportement des utilisateurs. C'est dans ce dernier cadre que nous nous sommes placé dans cette thèse.

1.2.3 *Le cadre de notre étude : les graphe n-parti*

Nous nous éloignons un peu des algorithmes facilitant la recherche d'information pour définir un cadre général abstrait à notre étude. Nous verrons plus en détail le cadre formel dans le chapitre 2 qui lui est consacré et nous allons être plus intuitifs ici. Supposons que l'on ait accès aux traces de consommation [12] laissées par des utilisateurs. Nous avons donc déjà deux types d'entités : les utilisateurs et les objets consommés. Nous avons évidemment des liens entre ces deux entités quand un utilisateur consomme un objet. Maintenant comme nous allons vouloir comparer ces objets, il peut être intéressant de les classer dans des catégories. La catégorisation est parfois une information disponible car elle peut être utilisée aussi pour la recommandation notamment. Ceci crée un nouveau type d'entité de base : les catégories, qui ont des liens évidents avec les objets (un objet et une catégories sont liées si l'objet appartient à cette catégorie). Nous avons donc un réseau avec des liens entre différentes entités. Ce réseau contient une multitude d'information, une partie de l'information qui est contenue dans ce réseau est sa structure, c'est à dire, ici : quels sont les liens entre les différentes entités. C'est en ne gardant que cette information que l'on peut le modéliser sous forme de graphe. En fait, nous allons garder un peu plus d'informations ici car nous nous intéressons aux types des entités et nous allons baser la structure de notre graphe sur les entités. Nous avons par exemple trois types d'entités avec des liens relationnels entre elles, ce qui se représente très bien par des graphes tripartis. Si l'on complexifie un peu, on peut rajouter d'autres types d'entités (par exemple classifier les utilisateurs, avoir des sous-catégories...) ce qui peut rajouter des parties aux graphes pour en faire un graphe n-parti.

1.3 Présentation générale du manuscrit

Nous trouvons donc ici que le formalisme de graphe n-parti est une bonne manière de représenter une grande diversité de données. Même si nos premières données étudiées seront en lien avec les algorithmes de recommandation (consommation musicale ou achat d'article sur une plateforme) nous verrons au fil du manuscrit en quoi ce formalisme peut être adapté à d'autres types de données (utilisateurs politisés sur Twitter, invités d'émissions de télévision, installation d'ONG dans différents Etats...).

Il y a plusieurs objectifs dans cette étude :

— Définir mathématiquement des indicateurs de diversité sur les graphes n-partis.

12. consommation est à interpréter au sens large ici (achat de produits, mais aussi écoute de musique, lecture d'articles, visionnage de films, ...).

— Déterminer algorithmiquement comment les calculer.

— Implémenter ces algorithmes afin qu'ils soient utilisables en pratique sur des jeux de données de grande taille.

— Utiliser ces programmes sur des données assez variées et préciser les interprétations que l'on peut faire de nos différents indicateurs.

Dans ce manuscrit nous n'allons pas nous étendre sur l'implémentation informatique, nous nous focaliserons sur la définition mathématique des objets et leur utilisation sur les données.

Même si les motivations de ce travail sont de faire parler les données pour répondre à certains enjeux, nous nous focaliserons sur des questions de méthodologie. De plus, le coté général et appliqué est fondamental dans cette étude. C'est pour cela que nous avons essayé de toucher des domaines différents (sociologique, politique, géographique ...) et de construire un programme utilisable sans connaissance, exceptée le langage python (qui a lui aussi été choisi pour son côté transversal).

Nous avons donc une multitude de données différentes que nous allons regarder de la même manière. L'universalité de notre méthode vient du fait qu'on analyse la structure relationnelle entre les données que nous allons modéliser sous forme de graphes. Nous allons cependant analyser des données avec des entités de différentes natures, ce qui nous amène à travailler avec des graphes multipartis, chaque partie représentera un type d'entité particulière. Nous utiliserons cette structure de données pour pouvoir quantifier des scores qui seront interprétables pour pouvoir analyser la diversité des entités. Nous avons donc intitulé notre manuscrit : Analyse et modélisation de la diversité des structures relationnelles à l'aide de graphes multipartis.

Après cette introduction nous commencerons par décrire le formalisme mathématique nécessaire à notre étude (chapitre 2). Puis nous appliquerons notre objet mathématique à des exemples de base pour y voir toutes les possibilités que notre objet nous offre (chapitre 3). Ceci nous montrera l'importance de normaliser nos indicateurs car on va se rendre compte qu'il seront trop influencés par le volume des nœuds (notion proche du degré que l'on développera), et nous motivera à étudier une normalisation par l'aléatoire (chapitre 4). Enfin nous verrons une autre série d'exemples qui nous permettront d'aller plus loin sur nos indicateurs, en dépassant le coté statique et triparti pour aborder des graphes avec plus de couches et dépendant du temps (chapitre 5). Nous conclurons en montrant au passage d'autres façons d'utiliser nos indicateurs sur d'autres données.

Cette thèse est aussi basée sur trois articles que nous avons co-écrits. Le chapitre 2 est basé sur un article en phase de révision pour le journal Théorical Computer Science (version disponible ici [76]). Le chapitre 3 a été en partie publié dans des actes de conférences internationales avec comité de lecture, sous la référence [72]. Après cet article, nous avons co-écrits une version longue dans le journal *Information Processing & management* ([73]) dans lequel nous avons pu développer les modèles aléatoires que nous développons dans le chapitre 4. Le chapitre 5 est un chapitre d'ouverture qui traite de deux projets qui n'ont pas, pour l'instant, abouti a une publication scientifique.

Chapitre

2

Construction d'un formalisme pour évaluer la diversité.

Pour commencer nous avons besoin de définir formellement les objets que nous allons manipuler. Nous avons décidé de séparer les définitions formelles des applications, pour pouvoir donner au lecteur un chapitre de référence dans lequel les notions sont définies. Ce choix nous impose de ne pas traiter ensemble définitions et exemples applicatifs. C'est pourquoi nous avons agrémenté cette première partie d'intuitions et d'exemples simples mais abstraits, il faudra attendre le chapitre suivant pour avoir une réelle application de tous ces objets. Nous présenterons en préliminaire le formalisme de la thèse : l'objet analysé (un graphe n-parti) et nos indicateurs de diversités . Nous revenons sur le concept de diversité, pour en définir un cadre formel, puis pour en détacher des propriétés importantes . Ensuite nous définirons une façon d'utiliser nos indicateurs sur ces graphes, en calculant la diversité selon un chemin dans le graphe. Enfin nous regarderons comment étendre un peu cette définition pour que nos indicateurs collent au mieux avec la notion de diversité que nous voulons évaluer.

Ce travail consiste à faire un état de l'art de notions abordées dans beaucoup de domaines distincts à les unifier et à les reformuler dans notre cadre. La section 2.1 est une présentation des graphes et des objets associés, en insistant fortement sur la particularité n-parti. La section 2.2 est un état de l'art des indicateurs de diversité dans les différents domaines réécrit dans un formalisme simple. La section 2.3 est une analyse de la théorie connue sur ces indicateurs. Les deux sections suivantes sont des contributions, car nous inventons réellement des termes et des objets en utilisant évidemment ce qui existait. La section 2.4 utilise ces indicateurs et la marche aléatoire pour définir ce qu'est la diversité suivant un méta chemin dans un graphe n-parti. Enfin la section 2.5 présente des variations de ces indicateurs quand on ne s'intéresse plus à un nœud mais à la couche toute entière, ce qui nous permettra d'avoir un cadre plus expressif.

2.1 GRAPHE N-PARTI

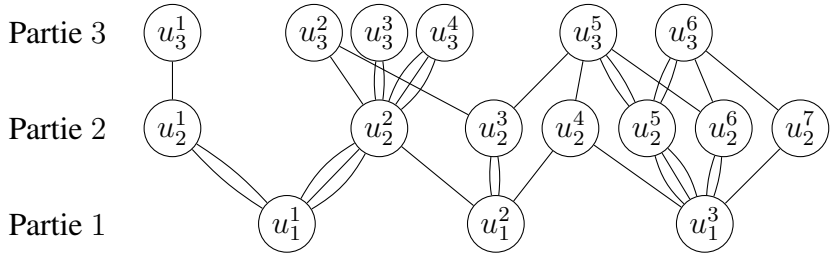


FIGURE 2.1 – Un exemple de graphe tri-parti (représenté comme un multi graphe)

Durant notre travail, nous avons modélisé nos données sous forme de graphes n-parti. Cette structure est assez générale pour s'appliquer à beaucoup de cas pratique et assez expressive pour que l'analyse structurelle de nos données révèle assez d'informations intéressantes. En effet un graphe est une abstraction essentielle qui peut représenter n'importe quel système dans le quel il y a des relations entre les objets.

Ici nous allons d'abord s'attacher à définir un cadre formel pour ces graphes, puis exprimer une intuition du sens que nous allons donner à cette représentation.

2.1.1 *Définir l'objet formel*

Nous allons utiliser des structures de données relationnelles c'est à dire que ce que nous étudions particulièrement ce sont les liens entre des entités. C'est pour ça que le formalisme de graphe est utile, en effet c'est une structure composée de *nœuds* qui représentent les entités et d'*arêtes* qui représentent les liens. Nous allons définir ici un concept de graphe appelé n*-parti* ainsi que plusieurs notations utilisées dans tout le manuscrit. Ce sont des notations classiques dans la littérature concernant les graphes. Ici les entités peuvent en plus avoir plusieurs natures différentes, ces graphes sont appelé graphe n-parti car leurs nœuds sont regroupés en couches (chaque couche accueille les nœuds de même nature). Cette particularité nécessite de redéfinir plusieurs notions classiques sous ce prisme (degrés, voisinages ...).

La notion de graphe n-parti Pour la suite n sera un entier supérieur à 2. Comme un graphe classique un **graphe n-parti** est défini par deux ensembles :

— $\boldsymbol{V}$ l'ensemble des **sommets** ou **nœuds** (vertex),

— $\boldsymbol{E}$ l'ensemble des **arêtes** ou **liens** (edges). Formellement $\boldsymbol{E}$ est un sous ensemble de $\boldsymbol{V} \times \boldsymbol{V}$.

On ajoute à cette définition de graphe, une partition n sous ensembles disjoints de V : $(\boldsymbol{V}_i)_{i\in[\![1,n]\!]} \in \mathcal{P}(V)$ tels que $\boldsymbol{V} = \bigcup_{i=1}^{n} \boldsymbol{V}_i$. Ces $(\boldsymbol{V}_i)$ sont appelées **parties** ou **couches**. On peut aussi définir grâce à cela une partition de $\boldsymbol{E}$ définissant l'ensemble des liens entre deux couches : pour

$i, j \in [\![1, n]\!]$, $E_{ij} = \{(a, b) \in \boldsymbol{E},\ a \in \boldsymbol{E}_i \wedge b \in \boldsymbol{E}_j\}$. Notons ici que nous nous plaçons dans un cadre général dans le quel il peut y avoir des liens entre les nœuds de deux couches mais aussi entre les nœuds d'une même couche. De plus, même si notre formalisme pourrait s'appliquer à des graphes orientés (c'est à dire dans lequel une arête (i, j) est distincte de l'arête (j, j)), nous nous restreignons dans notre cadre d'étude à des graphes non orientés.

La figure 2.1 est un exemple de graphe 3-parti, appelé **graphe tri-parti** (comme nous l'expliquons après, nous utilisons le formalisme de multi-graphe pour le représenter).

Graphe pondéré Nous travaillerons ici avec des graphes qui peuvent être **pondérés**. Formellement, en plus du couple $(\boldsymbol{V}, \boldsymbol{E})$, on ajoute une fonction de poids w définie sur $\boldsymbol{E}$ et à valeur dans $\mathbb{R}^+$. Notons que nous pouvons étendre w sur $\boldsymbol{V} \times \boldsymbol{V}$ par une convention simple : $\forall (a, b) \in \boldsymbol{V} \times \boldsymbol{V}$, $w(a, b) = 0$ si $(a, b) \notin \boldsymbol{E}$.

Multi-graphe On parle de **multi-graphe** quand l'ensemble des arêtes $\boldsymbol{E}$ n'est plus un ensemble mais un multi-ensemble. Ainsi il peut y avoir plusieurs arêtes entre deux nœuds.

Lemme 1. *À partir d'un multi-graphe on peut construire un graphe pondéré.*

Démonstration. Soit un multi-graphe $G = (\boldsymbol{V}, \boldsymbol{E})$ on peut construire un graphe pondéré $G' = (\boldsymbol{V}', \boldsymbol{E}', w')$ ainsi : Tout d'abord on fixe $a, b \in \boldsymbol{V}$ on fixe $w(a, b)$ étant nombre de liens entre a et b. Nous pouvons ainsi construire notre triplet :

— $V' = V$,

— $E' = \{(a, b) \in V \mid w(a, b) \neq 0\}$,

— $w = w'$.

□

La construction réciproque n'est pas forcément possible, en effet la pondération des liens dans un graphe pondérée n'est pas forcément un nombre entier mais peut être un nombre réel. C'est pour ne pas perdre cette généralité que nous décrivons notre formalisme dans le cadre des graphes pondérés. Nous pourrions définir un formalisme de multi graphe pondéré, mais ce n'est pas nécessaire ici, puisque nous n'avons, comme nous venons de le voir, pas de perte de généralité.

Cependant, dans un souci de rendre les figures les plus lisibles possibles, nous utiliserons souvent le formalisme de multi-graphe.

Dans notre exemple (figure 2.1) on a ainsi $w(u_2^1, u_1^1) = 2$ et $w(u_2^1, u_3^1) = 1$.

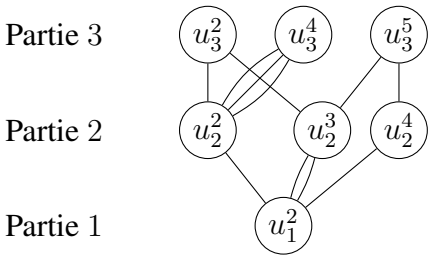


FIGURE 2.2 – Exemple de sous graphe.

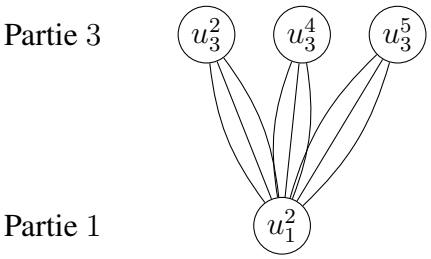


FIGURE 2.3 – Projeté de ce sous graphe

Degrés et degrés pondérés Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.
Soit $i, j \in [\![1, n]\!]$ et $a \in \boldsymbol{V}_i$ un nœud, on définit le **degré pondéré** (i, j) **de** a (noté $d_{ij}(a)$ comme la somme des pondérations des liens partant de a et allant vers la couche j. Formellement $d_{ij}(a) = \sum_{b \in V_j} w(a, b)$. De même on peut définir le **degré** (i, j) de a par le nombre de liens sortant de a et arrivant sur un nœud de la couche j. Formellement $d^*_{ij} = \sum_{b \in V_j} \mathbf{1}_{w(a,b)=0}$.
Où $\mathbf{1}_{w(a,b)}$ est égal à 0 si $w(a, b) = 0$ et 1 sinon.

Dans notre exemple (figure 2.1) on a ainsi $d_{21}(u_2^2) = 5$, $d_{23}(u_2^2) = 6$ et $d^*_{21}(u_2^2) = 3$.

Voisinage Avec les notations précédentes le degré est donc le cardinal de l'ensemble des nœuds de la couche j liés à a. Cet ensemble est appelé le **voisinage** ij **de** a (noté $\boldsymbol{E}_{ij}(a)$) et ses éléments sont appelés ses **voisins**. Formellement $\boldsymbol{E}_{ij}(a) = \{b \in \boldsymbol{V}_j \mid (a, b) \in \boldsymbol{E}_{ij}\}$.

Dans notre exemple (figure 2.1) : $\boldsymbol{E}_{21}(u_2^2) = \{u_1^1, u_1^2\}$ et $\boldsymbol{E}_{23}(u_2^2) = \{u_3^2, u_3^3, u_3^4\}$.

Sous graphe Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.
On appelle **sous graphe** un triplet $(\boldsymbol{V}', \boldsymbol{E}', w')$ où V' est un sous ensemble de V et E' et w' définie par rapport à V' : formellement :
$\boldsymbol{V}' \subseteq \boldsymbol{V}$, $\boldsymbol{E}' = \{(a, b) \in \boldsymbol{E} \mid (a, b) \in V'^2\}$ et $w' = w_{|\boldsymbol{E}'}$. La figure 2.2 est un exemple de sous graphe. Notons qu'il suffit que V' contienne au moins un nœud de chaque couche pour garder un graphe n-parti.

Graphe projeté Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.
Soit $k \in [\![1, n]\!]$, On appelle **graphe projeté** par rapport à la couche k, le graphe $(\boldsymbol{V}', \boldsymbol{E}', w')$ dans le quel l'on retire la couche k et on relie les nœuds pour lesquels les nœuds de la couche k jouaient le rôle de relai. Formellement : formellement :
$\boldsymbol{V}' = V \setminus \boldsymbol{V}_k$ et pour $i, j \in [\![1, n]\!] \setminus \{k\}$:

$$\boldsymbol{E}'_{ij} = \boldsymbol{E}_{ij} \cup \{(a, b) \in V_i \times V_j \mid \exists c \in V_k \mid (a, c) \in \boldsymbol{E}_{ik} \wedge \boldsymbol{E}_{kj}\}$$

De plus pour $a, b \in \boldsymbol{E}'_{ij}$

$$w'_{ij}(a, b) = w_{ij}(a, b) + \sum_{c \in \boldsymbol{V}_k} w_{ik}(a, c) w_{kj}(c, b)$$

La figure 2.3 est un projeté par rapport à la couche 2, du sous graphe de la figure 2.2. nous voyons ici qu'on perd une grande partie de l'information et de la complexité de la structure. Cependant, grâce au projeté il est facile de voir d'un seul coup d'œil qu'à partir de u_1^2, on peut atteindre chacun des nœuds de la couche 3 de trois façon différentes.

2.1.2 *Graphe triparti utilisateurs, objets, catégories.*

Nos définitions ici sont formelles et générales. Cependant le lecteur doit avoir une intuition de ce que l'on fait dans un cas plus particulier, qui sera le cas majoritaire ici. Nous allons étudier, en grande majorité des graphes tripartis. De plus ces graphes sont en général sous une forme particulière : La couche 1 représente des utilisateurs (consommateurs, reviewers, lecteurs...), la couche 2 représente des objets (musiques, articles de presse, achats...) et la couche 3 représente des catégories dans lesquelles classer ces objets (genre musical, sources d'un article, genre d'un article, auteur de l'article, chanteur...).

Ainsi dans ce graphe particulier, même si on regarde des nœuds et des liens, les couches d'où ils proviennent les caractérisent beaucoup mieux. Un lien entre la couche 1 et la couche 2 signifie qu'un utilisateur a consommé un objet alors qu'un lien pris dans la couche 2 et 3 signifie que l'objet est caractérisé d'une certaine manière. Si le sens du lien change, le sens de la pondération aussi évidemment. De même le degré ($d*_{1,2}$ d'un nœud désigne le nombre d'articles distincts consommés par un utilisateur. Son degré pondéré lui désigne le nombre de fois où il a consommé un objet.

2.2 La diversité, un concept à formaliser

Maintenant que le cadre de représentation de nos données est fixé, nous pouvons nous intéresser au concept que nous allons aborder : la **diversité**. Ce concept est aussi commun que difficile à définir. Cette section va nous permettre de formaliser ce qu'est un indicateur de diversité. Nous avons décidé de séparer cette section de la section suivante parce que nous souhaitons séparer la notion de diversité qui est associé à des indicateurs déjà existants (section 2.2) et la théorie qui nous a amené à en choisir un sous ensemble (section 2.3).

Nous allons tout d'abord commencer par définir le concept pour en avoir une première formalisation et enfin regarder grâce au formalisme les différents indicateurs qui proviennent des différentes disciplines scientifiques.

Étant donné que les premiers exemples applicatif que nous montrerons seront sur des écoutes musicales, nous allons souvent prendre l'exemple d'une *playlist* pour illustrer nos concepts.

2.2.1 *Le concept de diversité*

La diversité est un concept utilisé dans la littérature scientifique comme dans le discours commun. Par exemple essayons d'évaluer la diversité d'une *playlist* musicale. Pour ce faire nous

pourrions regarder les différences et les ressemblances entre les musiques de cette *playlist*. Il arrive assez vite une question : qu'est ce qui fait que deux musiques sont différentes, ou ressemblantes ? Cette question est certes intéressante, mais elle nous sort un peu de notre cadre général. En effet elle ne se pose pas du tout de la même manière si l'on parle de la ressemblance entre des musiques, des invités politiques ou des articles de journaux. Analyser les différences entre les musiques est donc quelque chose de trop fin à analyser pour nous. A contrario, nous pourrions considérer qu'une *playlist* est diversifiée simplement en comptant le nombre de musiques. Évidemment, cette solution n'est pas satisfaisante car elle supprime toute l'information des musiques. Pour trouver un compromis nous pouvons choisir des catégories qui définissent ces musiques. Ici on peut penser au genre musical (rock, rap, classique, techno, reggae...). Pour le reformuler dans le cadre le plus général, nous pouvons dire que la diversité est une propriété d'un système dans le quel chaque objet possède certaines caractéristiques. Ici notre système est une *playlist*, les musiques sont les objets et les caractéristiques sont les genres musicaux.

Dans cet exemple plus la *playlist* contiendra de genres musicaux, plus ces genres musicaux seront différents les uns des autres, plus nous pourrons considérer que la *playlist* est diverse. Au contraire si la *playlist* est composée de musiques d'un genre unique, nous considérerons que cette *playlist* n'est pas diverse. Évidemment il est intéressant d'analyser combien de genres il y a dans cette *playlist*, mais aussi combien de musiques de chaque genre, et quelle proportion de chaque genre l'utilisateur écoute.

Ici, dans l'exemple, nous voyons que nous n'analysons pas toutes les caractéristiques de l'objet, mais qu'il est plus simple de les classer dans des catégories. En effet nous ne regardons pas à quel point les musiques sont différentes mais nous cherchons d'abord à les classer dans des genres musicaux. Nous perdons évidemment de l'information, mais c'est l'axe que nous avons exploré dans ce travail. De même nous ne nous sommes pas intéressés à la différence entre les genres, ce qui ici aussi réduit l'information, mais nous laisse quand même un cadre d'étude assez fourni.

La diversité est donc une façon d'évaluer à quel point les différentes caractéristiques sont présentes dans le système.

Ainsi ce simple modèle d'objets classés dans des catégories est assez général pour l'usage de la diversité dans plusieurs domaines de recherches. Par exemple comment des unités de richesses catégorisées appartiennent à différentes personnes (en économie), ou le nombre d'individus catégorisés par leurs différentes espèces (en écologie), ou les unités produites d'un objet classées par leurs producteur (en droit de la concurrence).

Dans la sous-section suivante nous allons utiliser cette intuition pour définir un cadre formel pour calculer la diversité. Pour rester proche de la littérature nous ne parlerons plus de catégories mais de types.

2.2.2 *Objets, types et mesure de diversité*

Considérons un système composé d'un ensemble d'**objets** (items) I, d'un ensemble de **types** T et d'une relation d'appartenance $\tau \subseteq I \times T$. On dit d'un objet $i \in I$ qu'il est de type $t \in T$ (ou appartient au type t) lorsque $(i,t) \in \tau$. Cette fonction peut aussi être appelée **classification**.

Une mesure de diversité est une fonction $D : \tau \mapsto d \in \mathbb{R}^+$ qui s'applique à un système pour lui donner une valeur.

Dans la suite de cette étude, nous étudierons des systèmes dans lesquels les objets, les types, et la classification sont données. Évidemment une des grandes problématiques est de fixer cette classification, mais ce n'est pas l'objet de l'étude ici. Il sera donc nécessaire de la compléter avec une analyse fine des données et l'objet de cette thèse est aussi d'améliorer la possibilité d'interface entre notre analyse et celle ci plus fine.

Comme nous l'avons pressenti au préalable, l'importance pour évaluer la diversité est de quantifier la présence de chaque type. Nous définissons **l'abondance** d'un type $t \in T$, notée $a_\tau(t)$, comme le nombre d'objets de ce type : $a_\tau(t) = |\{i \in I : (i,t) \in \tau\}|$, et **l'abondance proportionnelle** comme $p_\tau(t) = \frac{a_\tau(t)}{|\tau|}$.

Ainsi nous pouvons nous intéresser à des mesures de diversité particulières qui ne s'appliquent pas globalement au système τ mais plus spécifiquement aux abondances propositionnelles. La définition précédente associe l'abondance proportionnelle provenant d'une classification à une valeur réelle positive $D(\tau)$ devient donc $D(p_\tau(t_1), \ldots, p_\tau(t_k))$ avec $k = |T|$.

Nous avons donc des mesures qui s'appliquent à un ensemble de distribution. Pour simplifier les notations nous allons définir cet ensemble. Tout d'abord, pour un entier k positif définissons l'ensemble

$$\Delta^k = \left\{(p_1, \ldots, p_k) \in \mathbb{R}^k : \forall i \leq k,\ p_i \in [0,1] \text{ et } \textstyle\sum_{i \leq k} p_i = 1\right\}.$$

Puis prenons l'union de tous ces ensembles : $\Delta^* = \bigcup_{i=1}^{\infty} \Delta^k$. Nous appellerons **distribution** un élément de Δ^*. Formellement une distribution est un n-upplet. Pour les identifier nous les noterons $\langle p \rangle$ ou $\langle p_1, \cdots, p_n \rangle$.

Ainsi, à partir de maintenant une **mesure de diversité** D est une fonction allant de Δ^* à $\mathbb{R}^+$.

2.2.3 *La diversité des mesures de diversité*

Comme nous l'avons dit plus haut, le terme *diversité*, est utilisé pour désigner une grande variété de propriétés de dissimilarité dans plusieurs domaines comme l'écologie, [39, 70, 37, 13, 60, 57, 59], les sciences du vivant [93], l'économie [28, 80], les politiques publiques [30, 24, 91, 69], la théorie de l'information [3, 79], les études des médias et d'internet [67, 49], la

physique [90, 85], les sciences sociales [33], l'étude des systèmes complexes [7, 25, 99, 81], et des dynamiques d'opinion [105]. Dans tous ces domaines on analyse la diversité d'un système dans le quel les objets sont classifiés sous différents types. Et, dans tous ces cas, les mesures de diversité sont des fonctions qui assignent à chaque système une valeur de diversité dans l'intention de mesurer différentes propriétés. Nous regardons ici quelles sont ces propriétés, pour nous servir de notre formalisme puis présenter ces différents indicateurs et enfin analyser comment ces indicateurs respectent ces propriétés.

Ces propriétés comportent trois composantes identifiées dans Selon [97], la **variété (variety)**, **l'équilibre (balance)**, et **la disparité (disparity)** que l'on reformulera ainsi : Pour une distribution $\langle p \rangle$

— **La variété** est le nombre de types dans lesquels les objets peuvent être classifiés : ce serait donc simplement le nombre d'éléments de $\langle p \rangle$

— **L'équilibre** mesure la manière dont notre distribution est répartie. Plus notre distribution est proche de la distribution uniforme plus nous dirons qu'elle est équilibrée. Plus les p_i sont proches les uns des autres plus notre distribution sera dite équilibrée

— et **la disparité** est une mesure du *dégré de différences* entre les types, qui nécessiterait une métrique sur T. Plus les types présent sont différents plus la distribution sera dite disparate.

Le lecteur peut aussi lire [96] pour approfondir ces propriétés.

Analysons maintenant plus particulièrement le concept de la diversité dans des exemples précis de différents domaines.

Pour ceci fixons une distribution : $p = \langle p_1, p_2, \ldots, p_{|T|} \rangle$. Chacun des éléments représente une abondance proportionnelle provenant de la classification d'objets appartenant à I en types appartenant à T.

La Richness [56, 32] est une mesure commune de diversité qui exprime simplement la propriété de *varieté*.

Souvent utilisée en écologie, elle mesure le nombre de types utilisés effectivement (c'est à dire au moins une fois) pour classifier des objets. Si une *playlist* contient du rock du rap et du métal elle sera égal à 3 quelque soit leur proportion.

$$R(p) = |\{t \in \{1, 2, ..., |T|\} : p_t > 0\}|.$$

La Richness est la mesure la plus basique de la diversité. Nous pouvons même l'utiliser pour normaliser d'autres mesures (en prenant le ratio entre cette autre mesure et la richness par exemple)([70, Section 9]).

L'entropie de Shannon (Shannon entropy) [89, 88], elle,x est liée à la *variété* et à l'*équilibre*. C'est une mesure qui est utilisée dans plusieurs domaines mais principalement en théorie de l'information. Elle correspond à la quantité d'information contenue ou délivrée par une source d'information.

$$H(p) = -\sum_{i=1}^{|T|} p_i \log_2 p_i.$$

C'est aussi l'espérance d'une fonction $x \mapsto -log(x)$. Par exemple, dans une *playlist*, si nous connaissons l'abondance proportionnelle de chaque type il nous faudrait $H(p)$ question en moyenne pour déterminer le genre d'une musique prise au hasard.

C'est un indicateur qui mesure en même temps **l'équilibre** et la **variété** .

L'indice Herfindahl-Hirschman [80] est lui plus utilisé en droit de la concurrence ou dans la régulation antitrust et plus généralement en économie. Il correspond à la somme des carrés des abondances proportionnelles.

$$\text{HHI}(p) = \sum_{i=1}^{|T|} p_i^2.$$

C'est un autre indice lié à la **varieté** et **l'équilibre** et aussi connu comme l'**indice Simpson** [92], il fut premièrement introduit par Hirschman [40] puis par Herfindahl [38] dans une étude de concentration de la production industrielle.

Il mesure le degré de concentration d'objets dans les types. Ici l'abondance proportionnelle correspond à la part de marché d'une entreprise. Le but est de repérer si la part de marché est trop concentré dans un trop petit nombre d'entreprises. Dans ce cas **l'indice Herfindahl-Hirschman** sera grand (petite diversité).

Pour l'exprimer aussi de façon probabiliste : en connaissant l'abondance proportionnelle des genres musicaux d'une *playlist*, cet indice donne la probabilité de tomber sur le même genre en prenant 2 musiques au hasard. Évidemment, plus les genres seront distribués de façon uniforme (grande diversité), moins il sera fréquent de tomber, au hasard, sur le même (petit indice).

Une autre de ces mesures : **l'Indice de Gini-Simpson** [31], aussi appellée **indice Gibbs-Martin** en sociologie et en psychologie [29] et encore **Population Heterozygosity** en génétique [64], est un autre exemple de mesure prenant en compte la **varieté** et l'**équilibre**. C'est en fait 1 moins l'indice de **Herfhindhal**.

$$\text{GS}(p) = 1 - \sum_{i=1}^{|T|} p_i^2 = 1 - HHI(p)$$

Ce qui peut être vu, de manière probabiliste, comme la probabilité dans une *playlist* de choisir deux genres différents quand on choisit deux musiques au hasard (i.e la probabilité associée à l'événement complémentaire mesurée par **Herfindahl-Hirschman**).

C'est exactement ce qui est regardé en écologie ([42]). En effet on regarde "the probability of interspecific encounter", la probabilité de tomber sur des espèces distinctes.

Il ne faut pas confondre cet indice avec l'**indice de Gini** (ou le **coefficient de Gini** [87]). Celui là est plus utilisé en économie. Une des formulations de ce coefficient est donnée par l'équation :

$$\text{Gini}(p) = \frac{1}{2|T|} \sum_{i=1}^{|T|} \sum_{j=1}^{|T|} |p_i - p_j|.$$

Nous pouvons considérer qu'il ne mesure que l'équilibre, puisque contrairement aux autres il ne prend pas vraiment en compte les abondances proportionnelles mais la distance entre elles. Contrairement aux trois autres, si nous prenons une solution totalement équilibrée ($\forall i, p_i = \frac{1}{T}$), ce coefficient ne dépendra pas de T (il sera égal à 0 dans tous les cas).

L'indice de Berger-Parker est autre mesure de diversité qui est seulement focalisée sur l'*équilibre*. C'est aussi un indice utilisé en écologie. Il mesure juste l'abondance proportionnelle du type le plus abondant :

$$\text{BPI}(p) = \max_{i \in \{1,2,\ldots,|T|\}} p_i.$$

Par exemple s'il y a 90% de rock dans une *playlist*, cet indice sera égal à 0.9, quelle que soit la répartition des autres 10%. C'est pour cela qu'on peut dire qu'il ne mesure que l'équilibre, puisque même si ces 10% sont constitués d'autant de genres que l'on veut, l'indice ne variera pas.

Il existe d'autres mesures qui évaluent la *disparité*. Pour ce faire, il nous faudrait une fonction de distance sur T, ou de façon plus générale, une fonction partant de T^2 pour arriver dans un espace exprimant la disparité [94, 103]. Même si c'est une dimension de la diversité importante que l'on trouve en paléontologie [104], en économie [66], en biologie [83], nous ne nous intéresserons pas à cette dimension dans la suite de cette étude.

La raison principale est liée à la visée empirique du travail qui est entrepris ici. Dans l'ensemble des jeux de données dont nous avons disposé, nous n'obtenons de l'information qu'à partir de mesures dévoilant les relations structurelles entre les entités du réseau mais sans information sur les distances entre les différentes de catégories. Ceci est vrai sur les données liées à l'écoute musicale sur laquelle nous avons travaillé.

2.3 UNE THÉORIE UNIFIANT LES MESURES DE DIVERSITÉ

Maintenant que l'on a vu les principaux indicateurs et la manière dont ils étaient énoncés dans les différents domaines, nous allons analyser les propriétés mathématiques que peuvent avoir ces différents indicateurs. En théorie de l'information, il y a plusieurs théories axiomatiques qui permettent de définir l'entropie et d'autres mesures de diversités (cf. [77, 21, 102, 2]). À partir de ces théories nous avons choisi quelques unes des propriétés qui nous paraissent importantes. Ceci va nous permettre de classifier les mesures et de définir une famille d'indicateurs (α**-diversité**) que nous allons garder pour la suite de notre étude.

D'abord nous allons définir 4 principes qui nous paraissent importants pour une mesure de diversité : **la symétrie**, **l'extensibilité**, **le principe de transfert** et **la normalisation**. Ensuite nous choisissons une famille connue ayant une forme explicite, c'est à dire que nous pouvons écrire une formule qui la décrit (**les moyennes quasilinéaires**), ainsi nous allons pouvoir la modifier pour qu'elle réponde à ces axiomes. Nous ajouterons à ce moment un dernier axiome, **l'axiome de réplication**, pour se rapprocher encore plus de notre intuition de la diversité. Ensuite en verra en quoi les indicateurs réunis section 2.2.3, sont fortement liés à des éléments de notre famille. Enfin nous verrons en quoi "l'ordre" c'est à dire le paramètre α de nos indicateurs exprime une continuité entre variété et équilibre.

2.3.1 *Les propriétés notables des mesures de diversité*

On fixe, $k \in N^*$ un entier strictement positif, une mesure de diversité comme une fonction $D : \Delta^k \to \mathbb{R}^+$ et une distribution de probabilité $p = \langle p_1, ..., p_k \rangle \in \Delta^k$. Ce qui va nous permettre de présenter des axiomes définissant ce que doit être une "bonne" mesure de diversité. [3].

La première propriété est la **symétrie** (ou **anonymat**), et qui désigne pour notre mesure de diversité le fait d'être invariable par permutation sur les types.

Par exemple, une *playlist* avec 75% de métal et 25% de musique classique doit obtenir la même diversité qu'une *playlist* avec 75% de rock et 25% de rap.

Pour le dire autrement la diversité ne s'intéresse pas à la nature des types (ce qui correspond bien au sens de l'anonymat), mais au rapport de proportion entre les types. Notons que nous validons aussi avec cet axiome, le choix de ne pas prendre en compte la disparité puisque la notion de distance entre les types ne peut plus être prise en compte..

Axiome 1 (Symétrie)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
*La mesure de diversité D est **symétrique** lorsque pour toute permutation σ sur l'ensemble $[\![1, k]\!]$(formellement : $\sigma \in \mathfrak{S}_k$) et pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$:*

$$D(p_1, p_2, ..., p_k) = D(p_{\sigma(1)}, p_{\sigma(2)}, ..., p_{\sigma(k)}).$$

(Où $\mathfrak{S}_k$ désigne l'ensemble des permutations sur $[\![1, k]\!]$)

Ensuite nous ajoutons à notre mesure le fait qu'elle soit **extensible**, c'est à dire que notre mesure ne doit pas être influencée par le fait de rajouter des types "absent" (i.e dont leur abondance proportionnelle est nulle). Nous pouvons aussi appeler cela **invariance par type absent**

Si une *playlist* contient 50% de rock et 50% de rap,il est équivalent de l'analyser en considérant seulement les catégories rock et rap ou en considérant le rock et le rap, ou en regardant le rock, le rap et le métal.

Axiome 2 (Extensibilité)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
*La mesure de diversité D est **extensible** lorsque, pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$:*

$$D(\underbrace{p_1, p_2, ..., p_k}_{k \text{ entrées}}) = D(\underbrace{p_1, p_2, ..., p_k, 0}_{k+1 \text{ entrées}}).$$

Maintenant nous désirons capturer dans un axiome la notion d'équilibre. Prenons l'exemple d'une *playlist* contenant 15% de rock et 30% de classique. Nous considérerons évidement qu'elle est strictement plus équilibrée si nous ajoutons du rock et supprimons du classique pour passer à 20% de rock, et 25% de classique. Ce principe s'appelle le **principe de Pigou-Dalton** , ou **principe de transfert** [22].

Axiome 3 (Principe de transfert)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
*La mesure de diversité D valide le **principe de transfert** lorsque pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$ et pour tout $i, j \in [\![1, k]\!]$, si $p_i > p_j$, alors*

$$\forall \epsilon \leq \frac{p_i - p_j}{2}, \quad D(\underbrace{\ldots, p_i - \epsilon, \ldots, p_j + \epsilon, \ldots}_{k \text{ entrées}}) \geq D(\underbrace{\ldots, p_i, \ldots, p_j, \ldots}_{k \text{ entrées}}).$$

En appliquant à la suite ces trois axiomes . On peut vérifier facilement le principe de **fusion** donné ci dessous.

Théorème 1 (Fusion)**.** *La mesure de diversité D qui satisfait les axiomes 1, 2 & 3 vérifie **le principe de fusion** :*

$$D(\underbrace{\ldots, p_i, p_{i+1}, \ldots}_{k \text{ entrées}}) \quad \geq \quad D(\underbrace{\ldots, p_i + p_{i+1}, \ldots}_{k-1 \text{ entrées}}).$$

Démonstration. Fixons $k \in N^*$ un entier positif et $\langle p_1 \cdots p_k \rangle \in \Delta^k$ une distribution.
D'abord par extensibilité :

$$D(\underbrace{\ldots, p_i + p_{i+1}, \ldots}_{k-1 \text{ entrées}}) = D(\underbrace{\ldots, p_i + p_{i+1}, \ldots}_{k-1 \text{ entrées}}, 0).$$

Puis on applique une symétrie pour avoir

$$D(\ldots, p_i + p_{i+1}, \ldots, 0) = D(\ldots, p_i + p_{i+1}, 0, \ldots)$$

Ensuite il nous reste à appliquer le principe de transfert :
Nous voulons l'appliquer pour passer de $p_i + p_{i+1}, 0$ a p_i, p_{i+1}. Faisons une disjonction de cas pour choisir ϵ si $p_{i+1} < p_i$ dans ce cas on choisit $\epsilon = p_{i+1}$ ainsi $\epsilon \leq \frac{p_{i+1}+p_i+0}{2}$.
Nous avons donc :

$$D(\ldots, p_i + p_{i+1}, 0, \ldots) \leq D(\ldots, p_i + p_{i+1} - p_{i+1}, 0 + p_{i+1}, \ldots).$$

Ainsi :

$$D(\ldots, p_i + p_{i+1}, 0, \ldots) \leq D(\ldots, p_i, p_{i+1}, \ldots).$$

Dans l'autre cas ($p_i \leq p_{i+1}$) on choisit $\epsilon = p_i$. Ainsi on a avec le même raisonnement :

$$D(\ldots, p_i + p_{i+1}, 0, \ldots) \leq D(\ldots, p_{i+1}, p_i, \ldots).$$

En utilisant le principe de symétrie on obtient :

$$D(\ldots, p_i + p_{i+1}, 0, \ldots) \leq D(\ldots, p_i, p_{i+1}, \ldots).$$

Dans tous les cas nous avons donc bien

$$D(\underbrace{\ldots, p_i, p_{i+1}, \ldots}_{k \text{ entrées}}) \quad \geq \quad D(\underbrace{\ldots, p_i + p_{i+1}, \ldots}_{k-1 \text{ entrées}}).$$

□

Ce principe signifie que si nous fusionnons deux types en ajoutant leurs abondances proportionnelles, la diversité de notre distribution diminue. Par exemple une *playlist* contenant 70% de techno 20% de rock et 10% de reggae, est moins diversifiée qu'une *playlist* contenant 70% de techno et 30% de reggae-rock (on a fusionné reggae et rock).

Ici aussi nous voyons que la classification est importante, car si nous avons une classification qui fusionne deux genres musicaux nous n'aurons pas le même résultat que si nous les prenons séparément. Par exemple : devons-nous considérer que le black métal et le hard métal sont deux genres différents, ou un seul genre (le métal) ?

Nous pouvons nous intéresser au maximum et au minimum de nos indicateurs. Il est assez simple (par récurrence) de montrer que si une mesure de diversité suit nos trois axiomes 1(2 puis 3, alors elle est bornée comme décrit ci dessous :

Théorème 2 (Majoration et minoration pour les mesures de diversité.)**.** *Une mesure de diversité D qui satisfait les axiomes de symétrie, d'extensibilité et de transfert est bornée comme suit :*

$$D(1) \quad \leq \quad D(p_1, p_2, \ldots, p_k) \quad \leq \quad D(\underbrace{1/k, 1/k, \ldots, 1/k}_{k \text{ entrées}}).$$

Ceci signifie que la *playlist* la moins diversifiée est la *playlist* qui ne contient qu'un seul genre, et que celle qui est la plus diversifiée est celle dans laquelle tous les genres sont présents de façon équivalente (ils ont tous la même abondance proportionnelle).

Dans le but de normaliser nos indicateurs, nous avons décidé de choisir une valeur pour ce maximum [17]. Nous choisissons comme borne maximum le nombre de types présents.

Par exemple la *playlist* contenant 25% de rock, 25% de rap, 25% de reggae et 25% de techno obtiendra une note de diversité de 4.

Axiome 4 (Normalisation)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
La mesure de diversité D satisfait l'axiome de ***normalisation*** *lorsque pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$:*

$$D(\underbrace{1/k, ..., 1/k}_{k \text{ entrées}}) = k.$$

D'évidence (en prenannt $k = 1$) on a $D(1) = 1$. Ce qui nous donne (en reformulant le théorème 2) :

Théorème 3 (Bornes pour des mesures de diversités)**.** *Une mesure de diversité D qui satisfait les axiomes de symétrie, d'extensibilité, de transfert et de normalisation est bornée comme suit :*
Pour tout $p \in \Delta^{k-1}$, on a $1 \leq D(p) \leq k$.

Notons que ce théorème est simplement un corollaire du précédent, nous le nommons théorème aussi car les deux sont tout autant fondamentaux.

2.3.2 *Vers la α-diversité : les moyennes quasilinéaires autopondérées*

Maintenant que l'on a déterminé une série d'axiomes qualifiant les mesures de diversités, nous pouvons regarder quelles sont les familles d'indicateurs qui les respectent . Il est pratique d'avoir une formule explicite pour pouvoir les étudier, et les calculer. Notons tout d'abord que, pour suivre l'état de l'art [1] nous ne définissons pas directement la famille qui nous intéresse mais leur inverse. En effet, ces familles prennent une valeur basse lorsque la distribution est équilibrée et haute lorsque elle est inégale. Nous pourrons alors parler d'indicateur de concentration. C'est l'inverse de ce que nous voulons. Nous les dénoterons par un I, pour ne pas les confondre avec nos indicateurs que l'on dénote avec un D.

Cette famille est centrale en quantification de l'information dans la théorie de l'information [3] et elles est de la forme suivante :

Définition 1 (Moyenne quasilinéaire)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
Une fonction $I : \Delta^* \to \mathbb{R}^+$ est une ***moyenne quasilinéaire*** *lorsqu'il existe $w \in \Delta^k$ une distribution et ϕ une fonction continue strictement croissante telles que, pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$:*

$$I(p) = \phi^{-1}\left(\sum_{i=1}^{k} w_i \phi(p_i)\right),$$

Nous allons prendre une sous-famille de celle là, car il est facile de voir que, pour l'instant, cette famille ne correspond pas aux axiomes de symétrie, à moins que les w_i soient tous égaux. En fait, nous n'allons pas les prendre égaux entre eux mais égaux aux p_i ce qui va permettre aussi la symétrie, ce qui définit la famille suivante.

1. Nous regarderons une famille générale (notament compatible avec les axiomes de probabilités[41]), qui est la famille des **moyennes quasilinéaires** (quasilinear means) (elle vient de Kolmogorov [46] et Nagumo [63]).

Définition 2 (Moyennes autopondérées quasilinéaires (self-weighted quasilinear means [74]))**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
*Une fonction $I : \Delta^* \to \mathbb{R}^+$ est une **moyenne autopondérée quasilinéaire** lorsqu'il existe $\phi : \mathbb{R} \mapsto \mathbb{R}^{+*}$ une fonction continue strictement croissante telle que, pour toute distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$:*

$$I(p) = \phi^{-1}\left(\sum_{i=1}^{k} p_i \phi(p_i)\right),$$

Notons que ϕ est une fonction réelle continue et strictement croissante donc par le théorème de la bijection (ou le théorème des fonctions réciproques) elle est inversible ce qui nous permet de parler de ϕ^{-1}. De plus ϕ^{-1} est, elle aussi, continue strictement croissante.

Ensuite nous la restreignons encore un petit peu pour pouvoir enfin respecter nos axiomes. On a

Théorème 4 (Moyennes autopondérées quasilinéaires concaves (concaves self-weighted quasilinear means))**.** *Soit I,telle que ϕ est concave ou convexe (avec ϕ provenant de la Définition 2). Dans ce cas en posant $D : p \mapsto \frac{1}{I(p)}$, D satisfait les Axiomes 1, 2, 3 & 4.*

Démonstration. Soit $\phi : \mathbb{R} \mapsto \mathbb{R}^{+*}$ une fonction continue strictement croissante concave ou convexe. Notons $I : p \mapsto \phi^{-1}\left(\sum_{i=1}^{k} p_i \phi(p_i)\right)$. Et $D : p \mapsto \frac{1}{I(p)}$. Montrons que D satisfait les axiomes 1, 2, 3 & 4.
Soit $k \in \mathbb{N}^*$ un entier strictement positif.
Soit une distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$.

Tout d'abord l'axiome de symétrie.
Soit $\sigma \in \mathfrak{S}_k$, notons $p_\sigma = \langle p_{\sigma(1)}, ..., p_{\sigma(k)} \rangle$.
Dans ce cas $I(p_\sigma) = \phi^{-1}\left(\sum_{i=1}^{k} p_{\sigma(i)} \phi(p_{\sigma(i)})\right)$.
En ré-indexant la somme on a $I(p_\sigma) = I(p)$ donc $D(p_\sigma) = D(p)$.
D vérifie donc l'axiome de symétrie.

Ensuite l'axiome d'extensibilité.
$I(\underbrace{p_1, p_2, ..., p_k, 0}_{k+1 \text{ entrées}}) = \phi^{-1}\left(\sum_{i=1}^{k} p_{\sigma(i)} \phi(p_{\sigma(i)}) + 0 * \phi(0)\right) = I(p)$. Donc $D(p_\sigma) = D(p)$.
Ainsi D vérifie l'axiome d'extensibilité.

Puis le principe de transfert.
Soit $i, j \in [\![1, k]\!]$ tels que $p_i > p_j$ et $\epsilon \leq \frac{p_i - p_j}{2}$. Notons $p' = \langle \ldots, p_i - \epsilon, \ldots, p_j + \epsilon \ldots \rangle$.
$I(p') = \phi^{-1}\left(\sum_{l=1}^{k} p'_l \phi(p'_l)\right) = \phi^{-1}\left(\sum_{l \neq i \wedge l \neq j} p'_l \phi(p'_l) + \epsilon\left(\phi(p_j + \epsilon) - \phi(p_i - \epsilon)\right)\right)$
Par définition de ϵ : $p_j + \epsilon \leq p_i - \epsilon$. Ainsi par croissance de ϕ : $\phi(p_j + \epsilon) - \phi(p_i - \epsilon) \leq 0$. Donc par

croissance de ϕ^{-1} : $\phi^{-1}\left(\sum_{l\neq i\wedge l\neq j} p'_l\phi(p'_l) + \epsilon\left(\phi(p_j+\epsilon) - \phi(p_i-\epsilon)\right)\right) \leq \phi^{-1}\left(\sum_{l\leq k} p'_l\phi(p'_l)\right)$
Donc $I(p') \leq I(p)$, et ce sont, par définition de ϕ et donc de I, deux nombres strictement positifs donc $D(p') \geq D(p)$.
Ainsi D vérifie le principe de transfert.

Enfin l'axiome de normalisation.
Notons $p = \langle \frac{1}{k} \cdots \frac{1}{k} \rangle$.
$I(p) = \phi^{-1}\left(\sum_{i=1}^{k} \frac{1}{k}\phi(\frac{1}{k})\right)$. Si ϕ est convexe, par convexité ϕ, $\sum_{i=1}^{k} \frac{1}{k}\phi(\frac{1}{k}) \geq \phi(\sum_{i=1}^{k} \frac{1}{k^2})$.
Notons que $\sum_{i=1}^{k} \frac{1}{k^2} = \frac{1}{k}$. Par croissance de la fonction ϕ^{-1}, $I(p) \geq \phi^{-1}(\phi(\frac{1}{k})) = \frac{1}{k}$. Si ϕ est concave, ϕ^{-1} est convexe, de la même manière on a : $\phi^{-1}\left(\sum_{i=1}^{k} \frac{1}{k}\phi(\frac{1}{k})\right) \geq \sum_{i=1}^{k} \frac{1}{k}\phi^{-1}(\phi(\frac{1}{k}))$.
Ainsi dans tous les cas $I(p) \geq \frac{1}{k}$ Donc $D(p) \leq k$. Ce qui vérifie le dernier axiome de normalisation. □

Notons que les **moyennes autopondérées quasilinéaires concaves** sont des mesures de diversité venant de la théorie [17, 26].

Attention, certes ces mesures de diversité satisfont nos axiomes et les axiomes de probabilités mais ce ne sont plus des moyenne quasilinéaires (car la pondération dépend de l'entrée) (e.g. the Hall-Tideman Index [34]).

2.3.3 *La famille d'indicateurs choisie : α-diversité*

Pour finir, et arriver à une famille satisfaisante, nous allons ajouter un dernier axiome. Cet axiome nous permettra d'instaurer une hiérarchie entre les distributions. Car, pour l'instant, à part pour les deux cas extrêmes (ceux du théorème 3), nous n'avons pas introduit d'inégalités strictes dans nos hypothèses. Ce qui signifie que nos indicateurs, peuvent être incapables de distinguer beaucoup de cas. Ils doivent seulement avoir une valeur pour toute les distributions uniformes : ($\langle 1\rangle$,$\langle \frac{1}{2}, \frac{1}{2}\rangle$,$\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\rangle$...). Par exemple, pour l'instant avec notre axiome de normalisation, nous pouvons dire qu'une *playlist* contenant seulement de la techno est moins diversifiée qu'une *playlist* contenant 50% de rap et 50% de rock. (Car $D(1) = 1$ et $D(\frac{1}{2}, \frac{1}{2}) = 2$). Fondamentalement, ici ce que nous avons fait, c'est prendre une distribution et la "répartir" en deux distributions identiques : Nous avons pris la techno qu'on a séparé en deux parts égales : rap et rock. Ceci a eu comme effet de multiplier la diversité par 2. Nous allons généraliser ce principe ainsi :

Axiome 5 (Le principe de réplication)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
*La mesure de diversité D valide le **principe de réplication** lorsque pour toute distribution*

$p = \langle p_1, ..., p_k \rangle \in \Delta^k$:

$$D\left(\underbrace{\frac{p_1}{m}, \frac{p_2}{m}, ..., \frac{p_k}{m}}_{1^{st}\ copie}, \underbrace{\frac{p_1}{m}, \frac{p_2}{m}, ..., \frac{p_k}{m}}_{2^{nd}\ copie}, ..., \underbrace{\frac{p_1}{m}, \frac{p_2}{m}, ..., \frac{p_k}{m}}_{m^{th}\ copie}\right) = m\, D(p_1, ..., p_k).$$

Ce principe a été décrit par Chakravarty [17]. Par exemple si une *playlist* comporte 70% de reggae et 30% de rap, nos indicateurs donneront une note deux fois plus forte à une *playlist* comportant 35% de classique 35% de reggae 15 % de rock et 15% de techno.

Ce principe de réplication est nécessaire pour éviter des résultats paradoxaux applicatifs qui se retrouvent par exemple en écologie [23], mais qui sont assez éloignés de ce que nous allons étudier.

Cet axiome restreint encore notre famille d'indicateurs pour arriver à la définition suivante.

Définition 3 (α-diversité)**.** *Soit $\alpha \in \mathbb{R}^+$ un réel positif différent de 1. La α-diversité, notée D_α, est une fonction $D_\alpha : \Delta^* \to \mathbb{R}^+$,telle que, pour $p = (p_1, \ldots, p_k) \in \Delta^*$:*

$$D_\alpha(p) = \left(\sum_{i=1}^{k} p_i^\alpha\right)^{\frac{1}{1-\alpha}}$$

Notons que cette définition correspond à l'inverse d'une moyenne autopondérée quasilinéaire avec $\phi : x \mapsto x^{\alpha-1}$. Selon le paramètre α cette fonction est concave ou convexe.

Cette définition ne permet pas de considérer le cas $\alpha = 1$ mais nous pouvons néanmoins regarder sa limite ou définir directement avec $\phi = log$:

Définition 4 (1-diversité)**.** *La 1-diversité, notée D_1, est une fonction $D_1 : \Delta^* \to \mathbb{R}^+$,telle que, pour $p = (p_1, \ldots, p_k) \in \Delta^*$:*

$$D_1(p) = 2^{-\sum\limits_{i=1}^{k} p_i log(p_i)}$$

Que l'on peut réécrire $D_1 = \left(\prod\limits_{\substack{i=1 \\ p_i \neq 0}}^{k} p_i^{p_i}\right)^{-1}$

De même nous pouvons regarder la limite de cet indicateur en ∞ ou définir directement :

Définition 5 (∞-diversité)**.** *La ∞-diversité, notée D_∞, est une fonction $D_\infty : \Delta^* \to \mathbb{R}^+$,telle que, pour $p = (p_1, \ldots, p_k) \in \Delta^*$:*

$$D_\infty(p) = \left(\max_{i \in \{1,\ldots,k\}} \{p_i\}\right)^{-1}$$

Notons que des variantes de la α-diversité (introduite pour la première fois comme le nombre de Hill [39] puis nommée *true diversity* [44]) éxistent dans differents domaines. L'indice de concentration de Hannah-Kay (The Hannah-Kay concentration index) d'ordre α [35] est l'inverse de D_α. En théorie de l'information l'entropie de Rényi [79] d'ordre α, notée H_α, est le logarithme de D_α : $H_\alpha(p) = \log(D_\alpha(p))$.

Comme annoncé, grâce au théorème 4, ces mesures respectent bien nos axiomes.

Nous pouvons voir l'axiomatique developpée Chakravaratry[17] pour en avoir une preuve.

Nous pouvons maintenant voir, en quoi notre famille est fortement liée aux indicateurs que l'on avait exprimés dans la section 2.2.3.

Fixons $p \in \Delta^*$. La *Richness* est exactement $D_0(p)$. Nous remarquons, qu'avec la convention peu classique $0^0 = 0$, $D_0(p)$ est réellement définie avec notre définition de α-diversité. Si nous voulons nous passer de cette convention nous pouvons vérifier que c'est la limite de notre famille quand α tant vers 0. Cependant elle n'est pas directement une moyenne auto-pondérée. Comme précisé précédemment cela revient simplement à compter le nombre de types présents. Nous l'appellerons par la suite **Diversité de variété**.

$D_1(p)$, est liée à l'entropie de Shannon ($H(p)$), c'est simplement son logarithme :
$D_1(p) = 2^{H(p)}$. Nous l'appellerons **Diversité de Shannon**.

$D_2(p)$, est l'inverse de l'indice d'Herfindhal-Hirschman : $D_2(p) = 1/\text{HHI}(p)$. Nous l'appellerons **Diversité de Herfindhal**.

Enfin $D_\infty(p)$ est l'inverse de l'indice de Berger-Parker $D_\infty(p) = 1/\text{BPI}(p)$.
Nous l'appellerons **Diversité de Berger**.

Nous avons regroupé ces relations dans le tableau 2.4.

Encore une fois, il n'est pas étonnant au final que deux de nos indices soient l'inverse de deux indices classiques de la littérature puisque ce sont des indices de concentration, et que nous voulons exprimer la diversité.

Maintenant que nous avons unifié nos mesures de diversité, nous pouvons nous interroger sur l'influence qu'a α sur nos indicateurs. Nous pouvons commencer à avoir une intuition grâce aux deux valeurs extrêmes $\alpha = 0$, $\alpha = \infty$ associées respectivement à la diversité de variété et la diversité de Berger. Nous avons vu que la diversité de variété ne mesure que la variété et que la diversité de Berger ne mesure que l'équilibre. Intuitivement nous pouvons imaginer qu'un grand α donnera un indicateur plus basé sur l'équilibre et un petit α donnera un indicateur plus basé sur la variété. Dit d'une autre manière plus l'ordre augmente plus les distributions inégales sont pénalisées. Ainsi le paramètre α nous permet de formaliser une continuité entre variété et équilibre. Nous pourrons mieux visualiser ce rapport entre les indicateurs dans le premier exemple, particulièrement à la section 3.4.

Table 2.4 – Résumé de la α-diversitéd'ordre 0, 1, 2, and ∞, et leur relation avec les mesures classiques.

Order (α)	Name	True diversity	Expression	Relation to other diversity measures
0	diversité de variété	$D_0(p)$	$\lvert\{i \in \{1,\dots,k\} : p_i > 0 \in\}\rvert$	Richness[56, 32].
1	diversité de Shannon	$D_1(p)$	$\left(\prod_{\substack{i=1 \\ p_i \neq 0}}^{k} p_i^{p_i}\right)^{-1}$	Exponentiélle de l'entropie de Shannon [89, 88] : $H(p) = \log_2(D_1(p))$
2	diversité de Herfindhal	$D_2(p)$	$\left(\sum_{i=1}^{k} p_i^2\right)^{-1}$	Inverse de l'indice d'Herfindahl-Hirschman [80] : $\text{HHI}(p) = 1/D_2(p)$.
∞	diversité de Berger	$D_\infty(p)$	$\left(\max_{i\in\{1,\dots,k\}}\{p_i\}\right)^{-1}$	Inverse de l'indice Berger-Parker [11] : $\text{BPI}(p) = 1/D_\infty(p)$.

2.4 DIVERSITÉ SELON UN CHEMIN DANS UN GRAPHE n-PARTI

Nous avons vu jusqu'à présent quel formalisme nous allons utiliser pour décrire les données (les graphes n-parti, cf section 2.1) et comment quantifier la notion de diversité à partir d'une distribution (la α-diversité , section 2.3). Cette section vise à formaliser le lien entre les deux afin de quantifier la diversité d'un nœud au regard des liens qu'il entretient (directement ou indirectement) avec les autres nœuds du réseau). Comme nous l'avons vu précédemment pour pouvoir appliquer nos mesures de diversité il nous faut d'abord obtenir une distribution de probabilité. Le principe est simple. On part d'un nœud et on cherche à savoir la distribution de probabilité d'atteindre les nœuds d'une couche donnée à partir de ce nœud de départ.

Ici nous allons utiliser la marche aléatoire suivant un méta-chemin pour obtenir une distribution de probabilité. Nous ne nous intéresserons pas du tout au comportement à l'infini, mais au comportement après un nombre d'étapes fixées, ce qui est un tout autre domaine d'étude.

Pour comprendre en quoi la marche aléatoire est une bonne façon d'obtenir des distributions intéressantes il faut revenir à l'intuition de la sous section 2.1.2. En effet si on a un graphe tri-parti utilisateurs /objets/catégories, faire une marche aléatoire à partir d'un nœud de la couche 1 (utilisateur) vers les nœuds de la couche 3 (catégories), nous donnera une distribution donnant pour chaque catégorie la probabilité de l'atteindre à partir de cet utilisateur. Si un utilisateur atteint équitablement chaque catégorie, la distribution obtenue sera équilibrée, si au contraire il ne consomme que des objets d'une catégorie, la distribution sera inégale.

Nous définirons d'abord la probabilité de transition (section 2.4.1) entre deux nœuds, puis la marche aléatoire (section 2.4.2) selon un chemin, puis enfin la diversité selon un chemin (section 2.4.3).

Pour pouvoir suivre les différentes étapes de la marche aléatoire nous nous reporterons, dans

toute la suite de cette section aux quatre figures regroupées dans la figure 2.9 .

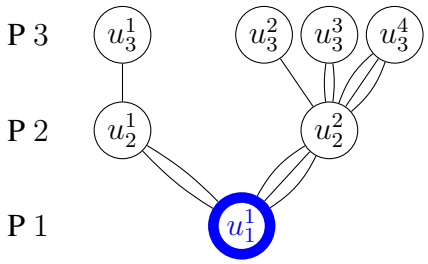


FIGURE 2.5 – Graphe initial

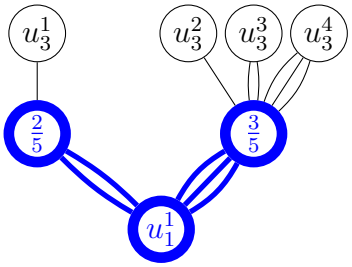


FIGURE 2.6 – Première étape

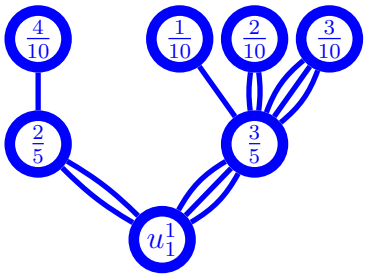


FIGURE 2.7 – Dernière étape

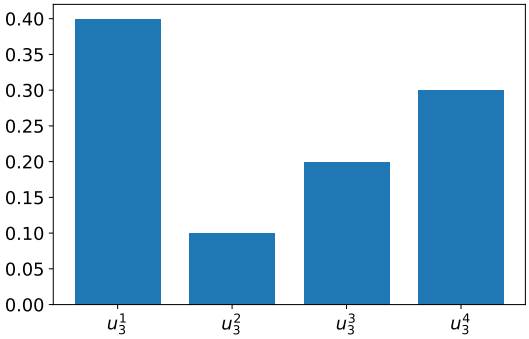


FIGURE 2.8 – Distribution obtenue

FIGURE 2.9 – Marche aléatoire suivant le chemin 1,2,3 partant de u_1^1

Pour la suite on fixera n un entier, et $G = (\boldsymbol{E}, \boldsymbol{V}, w)$ un graphe n-parti pondéré. De plus nous fixerons $i, j \in [\![1, n]\!]$, deux entiers qui nous permettrons d'identifier deux couches. Comme nous l'avons dit précédemment, en pratique ces couches seront distinctes, mais le formalisme reste valable pour $i = j$.

2.4.1 *Probabilité de transition*

Pour commencer il nous faut définir la probabilité de transition entre deux nœuds :

Définition 6 (Probabilité de transition entre deux nœuds)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
Soit $v_i \in V_i$ et $v_j \in V_j$, , la probabilité de transition entre v_i et v_j, notée $p_{\boldsymbol{E}_{ij}}(v_j \,|\, v_i)$ est définie par

- *si $d_{ij}(v_i) \neq 0$, $p_{\boldsymbol{E}_{ij}}(v_j \mid v_i) = \frac{w(v_i, v_j)}{d_{ij}(v_i)}$,*
- *sinon la probabilité n'est pas définie.*

Par exemple sur les figures 2.5 et 2.6, nous voyons que : $p_{\boldsymbol{E}_{12}}(u_2^1 \,|\, u_1^1) = \frac{2}{5}$ et $p_{\boldsymbol{E}_{12}}(u_2^2 \,|\, u_1^1) = \frac{3}{5}$.

Notons que ces définitions sont semblables aux définitions des *processus markoviens*, c'est pour cela que nous n'écrivons pas la probabilité comme une fonction à deux arguments $p(v_i, v_j)$ mais que nous gardons une écriture de probabilité conditionnelle. En effet c'est bien la probabilité d'arriver sur v_i en sachant que l'on part de v_j (et qu'on suit un chemin E_{ij}). On a donc la notation qui en découle :

Notation 1 (Transition aléatoire entre deux nœuds)**.** *On note la transition à partir d'un nœud aléatoire $X_i \in \boldsymbol{V}_i$ vers un nœud aléatoire d'arrivée $X_j \in \boldsymbol{V}_j$ suivant la distribution de probabilité $p_{\boldsymbol{E}_{ij}}$ ainsi : $X_i \xrightarrow{E_{ij}} X_j$.*

Nous utiliserons l'abréviation $p_{ij} = p_{\boldsymbol{E}_{ij}}$ et même p_j, ou p lorsqu'il n'y aura pas d'ambiguïté. De même nous noterons la précédente transition ainsi $X_i \xrightarrow{ij} X_j$ ou même $X_i \xrightarrow{j} X_j$.

Grâce à la définition 6 nous avons le lemme suivant :

Lemme 2 (distribution de probabilité)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti. $\forall v_i \in \boldsymbol{V}_i,\ p_{E_{ij}}(. \mid v_i) : \boldsymbol{V}_j \to \mathbb{R}^+$ est une **distribution de probabilité** sur $\boldsymbol{V}_j$*

En effet pour tout $v_j \in \boldsymbol{V}_j$ on a bien $p_{E_{ij}}(v_j \mid v_i) \in [0, 1]$ et $\sum_{v_j \in \boldsymbol{V}_j} p_{E_{ij}}(v_j \mid v_i) = 1$.

Le cas où $d_{ij}(v_i) = 0$ est un cas problématique, il signifie qu'il n'y a pas de lien partant de v_i allant dans la couche j. Nous appellerons ces cas **impasses**, il y a plusieurs façons de considérer ces cas, nous les verrons dans la suite (voir, par exemple, les sections 3.1.1 et 3.1.2). Cependant, nous préférons considérer que ces cas ne peuvent pas arriver (formellement que les probabilités suivantes ne sont pas définies), car en pratique, ce genre de cas peut signifier des choses de natures différentes sur les données. Nous préférerons donc, en général, traiter ce cas dans le prétraitement des données en fonction de la signification de cette absence d'information. Nous supprimerons donc les nœuds de degré nul.

2.4.2 *Méta-chemin et marche aléatoire*

Nous désirons définir une marche aléatoire suivant un chemin précis. Nous allons utiliser la structure de nos graphes et surtout le fait qu'ils soient n-partis. Nous allons donc définir notre marche suivant un chemin défini par les couches. Pour différencier le chemin entre les couches d'un chemin classique entre les nœuds, nous appellerons le premier **méta-chemin**.

Définition 7 (Méta-chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
Soit $N \in \mathbb{N}^*$,
*On appelle **méta-chemin** la suite donnée $\Pi = (i_k)_{k \in [\![0,N]\!]} \in [\![1, n]\!]^{N+1}$ qui est une suite de couches d'un graphe n-parti. On pourra le noter $(i_0, \cdots, i_k)$ ou $i_1 \to i_2 \to \cdots \to i_k$. N est appelée la **taille** du méta-chemin.*

Attention, nous avons ici considéré que N était la taille du chemin parce que c'est le nombre de transitions, nous aurions pu considérer le nombre de couches (qui aurait été $N+1$). Remarquons aussi que nous pouvons choisir n'importe quel chemin dans le graphe, par exemple on

peut repasser deux fois par la même couche. N défini ici n'a donc rien à voir avec le nombre de couches du graphe (n).

Ceci nous permet de définir une marche aléatoire suivant un méta-chemin.

Définition 8 (Marche aléatoire suivant un méta-chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
*On fixe $N \in \mathbb{N}^*$ un entier strictement positif, $\Pi = (i_k)_{k\in[\![1,N]\!]} \in [\![1, n]\!]^N$ un méta-chemin de taille N et une variable aléatoire $X_0 \in \boldsymbol{V}_{i_0}$ représentant la position de départ de la marche aléatoire. La **marche aléatoire suivant le méta-chemin** Π est une suite de $N + 1$ variables aléatoires $(X_0, X_1, \ldots, X_N)$ provenant de la suite de transition aléatoire entre deux nœuds (cf Définition 6) :*

$$X_0 \xrightarrow{i_1} X_1 \xrightarrow{i_2} X_2 \xrightarrow{i_3} \cdots \xrightarrow{i_N} X_k,$$

où $\forall k \in [\![0, N]\!],\ X_k \in \boldsymbol{V}_k$.

Ceci ressemble à la définition d'une marche aléatoire contrainte par un chemin (path-constrained random walk) que l'on peut retrouver dans les articles de Lao [51],[50].

Il en découle, avec les notations précédentes, le lemme suivant

Lemme 3. *Chaine de Markov. Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
Une marche aléatoire suivant ce méta-chemin Π est une chaine de Markov avec des transitions définies comme suit. $\forall k \in [\![1, N]\!]$

$$\Pr(X_k = v_k \mid X_{k-1} = v_{k-1}) = p_{k-1k}(v_k \mid v_{k-1}),$$

avec $v_{k-1} \in \boldsymbol{V}_{k-1}$ and $v_k \in \boldsymbol{V}_k$.

Pour les deux définitions qui suivent on considère un entier strictement positif N et un méta-chemin $\Pi = (i_k)_{k\in[\![0,N]\!]} \in [\![1, n]\!]^{N+1}$ de taille N et sa marche aléatoire associée $(X_0, X_1, \ldots, X_k)$.

Par construction nous avons les deux définitions suivantes

Définition 9 (Probabilité conditionnelle suivant un méta-chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
On fixe $X_N \in \boldsymbol{V}_{i_N}$, la variable aléatoire représentant le nœud d'arrivée de la marche aléatoire, et $v_0 \in V_{i_0}$ un nœud de départ. La probabilité conditionnelle suivant un méta-chemin Π partant de $v_0 \in V_{\Pi_0}$ (i.e. $X_0 = v_0$) est notée pour tout $v_N \in \boldsymbol{V}_{i_N}$: $p_\Pi(v_N \mid v_0)$, et peut être calculée récursivement ainsi :

- *Si $N = 1$: $p_\Pi(v_N \mid v_0)$ est donnée par la définition 6*
- *Sinon on fixe $\Pi_2 = (i_k)_{k\in[\![2,N]\!]}$ et :*

$$\begin{aligned} p_\Pi(v_N \mid v_0) &= \Pr(X_N = v_N \mid X_0 = v_0) \\ &= \sum_{v_1 \in \boldsymbol{V} i_1} p_{\Pi_2}(v_N \mid v_1)\ p_{i_0 i_1}(v_1 | v_0). \end{aligned}$$

Notons que Π_0 *est l'identifiant de la première couche du chemin de* V_{Π_0}.

Par exemple sur la figure 2.7 nous voyons que $p_{1,2,3}(u_3^1 \mid u_1^1) = \frac{4}{10}$. Nous avons défini la probabilité conditionnelle suivant un méta chemin, si nous l'appliquons à tous les nœuds de la couche d'arrivée nous obtenons une distribution.

Définition 10 (Distribution conditionnelle suivant un méta-chemin)**.** *On appelle distribution conditionnelle suivant un méta-chemin la distribution suivante (que l'on notera par abus de notation* $p_\Pi(v_0)$*) :* $p_\Pi(v_0) = (p_\Pi(v \mid v_0))_{v \in V_N}$

La figure 2.8 est une représentation de la distribution $p_{1,2,3}(u_1^1) = \langle \frac{4}{10}, \frac{1}{10}, \frac{2}{10}, \frac{3}{10} \rangle$.

2.4.3 *Diversité selon un méta-chemin*

Dans la sous-section 2.3.3 nous avons défini des indicateurs de diversité particuliers (appartenant à la α-diversité) ils s'appliquent à partir d'une distribution. Dans la sous-section précédente nous avons défini la marche aléatoire selon un chemin qui nous donne une distribution de probabilité. Il nous reste à utiliser les deux successivement pour définir la diversité selon un chemin.

Définition 11 (α-diversité suivant un méta-chemin)**.** *Soit* $n > 2$ *un entier et* $G = (V, E)$ *un graphe* n*-parti.*
Soit $\alpha \in \mathbb{R}^+$ *un réel positif,* Π *un méta chemin et* $u_0 \in V_{\Pi_0}$ *un nœud de la couche de départ de ce nœud. On appelle* ***α-diversité du nœud*** u_0 ***suivant le méta-chemin*** Π *le réel suivant :* $D_\Pi^\alpha(u_0) = D_\alpha(x_\Pi(u_0))$

Pour ne pas alourdir encore plus la notation nous ne notons pas le graphe sur notre mesure. En pratique nous ne faisons pas varier le graphe sauf dans le chapitre 4 et à ce moment là nous le préciserons.

Reprenons l'exemple précédant, en regardant $x_{1,2,3}(u_1^1)$ nous voyons qu'à partir du nœud u_1^1 et en se déplaçant au hasard on a 4 fois plus de chance d'arriver sur le nœud u_3^1 que sur le nœud u_3^2. Si nous partions d'un graphe triparti utilisateurs/objets/catégories, nous considérerions que l'utilisateur u_1^1 consomme 4 fois plus d'objets de la catégorie u_3^1 que ceux de la catégorie u_3^2.

Il suffit ensuite d'appliquer un indice de α-diversité pour lui attribuer une note en terme de diversité. Par exemple pour la distribution $x_{1,2,3}(u_1^1)$ de la figure 2.8 on a :

- $D^0_{1,2,3}(u_1^1) = 4$ (le nombre de nœuds atteint)
- $D^1_{1,2,3}(u_1^1) = 2^{-\left(\frac{4}{10}log(\frac{4}{10})+\frac{1}{10}log(\frac{1}{10})+\frac{2}{10}log(\frac{2}{10})+\frac{3}{10}log(\frac{3}{10})\right)} \simeq 3.6$
- $D^2_{1,2,3}(u_1^1) = \left((\frac{4}{10})^2 + (\frac{1}{10})^2 + (\frac{2}{10})^2 + (\frac{3}{10})^2\right)^{-1} \simeq 3.3$
- $D^\infty_{1,2,3}(u_1^1) = \frac{4}{10}^{-1} = 2.5$

2.5 EXTENSIONS DE LA DIVERSITÉ SELON UN MÉTA-CHEMIN.

Rappelons le cadre : un méta chemin donne une distribution à laquelle on applique une mesure de diversité. Nous allons étendre légèrement notre notion. En effet nous allons nous intéresser au comportement global d'une couche, ce qui nous permettra de définir une diversité qu'on appellera collective. Nous pourrons aussi définir une diversité moyenne. Enfin nous utiliserons ce comportement global pour définir une diversité individuelle relative.

2.5.1 *La diversité collective.*

Nous allons d'abord définir la diversité globale d'une couche, que l'on appellera la **diversité collective**. Comme précédemment, cette diversité collective sera définit selon un méta chemin fixé. Pour cela, il nous suffit de redéfinir la distribution suivant un méta chemin sauf qu'on ne part pas d'un nœud particulier mais plutôt de tous les nœuds d'une couche, ou, pour rester avec une vision probabiliste, d'un nœud choisi aléatoirement :

Définition 12 (Probabilité inconditionnelle suivant un méta-chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
Soit $N \in \mathbb{N}^*$ un entier strictement positif. On fixe $\Pi = (i_0, \cdots, i_N)$ un méta chemin et $X_0 \in \boldsymbol{V}_{i_0}$, la variable aléatoire représentant le nœud de départ de la marche aléatoire. La probabilité inconditionnelle suivant un méta-chemin Π est notée pour tout $v_N \in \boldsymbol{V}_{i_N}$: $p_\Pi(v_N)$, et peut être calculée par l'équation :

$$p_\Pi(v_N) = \sum_{v_0 \in \boldsymbol{V}_{i_0}} p_\Pi(v_N \mid v_0)\ \Pr(X_0 = v_0)$$

Ici nous avons donné une définition générale et X_0 peut définir n'importe quelle variable aléatoire. Cependant dans la suite on utilisera un cas particulier simple : une répartition uniforme. Dans ce cas on a une expression plus simple :

$$p_\Pi(v_N) = \frac{\sum_{v_0 \in \boldsymbol{V}_{i_0}} p_\Pi(v_N \mid v_0)}{\mid V_0 \mid}.$$

Si la distribution de départ n'est pas uniforme, on fera référence à la distribution, comme la distribution collective **biaisée**.

On a donc les deux définitions qui suivent, avec les mêmes notations

Définition 13 (Distribution collective suivant un méta chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
La distribution collective suivant un méta chemin Π est définie par $c_\Pi = (p_\Pi(v))_{v \in V_N}$

Définition 14 (α-diversité collective suivant un méta chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
*Soit $\alpha \in \overline{\mathbb{R}}^+$ un réel positif, potentiellement infini. La **diversité collective d'ordre α suivant un méta chemin** Π est définie par $C^\alpha_\Pi = D_\alpha(c_\Pi)$*

La figure 2.10 montre un calcul d'une distribution collective $2, 3$.

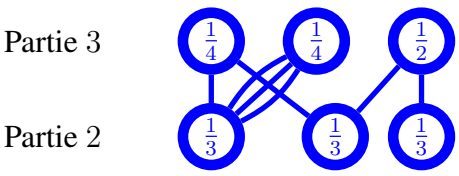


FIGURE 2.10 – Distribution collective de la partie 2

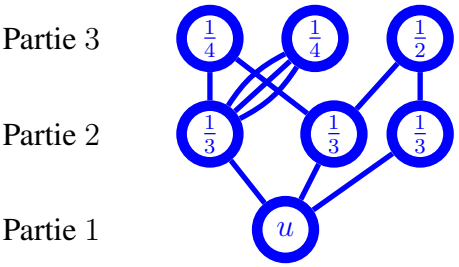


FIGURE 2.11 – Distribution individuelle du nœud u

Une autre façon de voir cette diversité est de rajouter une couche artificielle avant la couche initiale. Cette couche contiendrait un seul nœud relié uniformément à tous les nœuds de départ. La figure 2.11, montre comment faire ceci en pratique en rajoutant le nœud u. Notons que pour la différencier on appelle la distribution (respectivement la diversité) qui ne part d'un seul nœud la distribution (respectivement la diversité) **individuelle**.

Ce lien entre diversité collective et individuelle est un lien fort et structurel. En effet on peut aisément aussi définir la diversité individuelle de u dans la figure 2.11 comme la diversité collective de la figure 2.10. Ce cas est un cas particulier, car tous les liens partant de u sont pondérés identiquement. Mais on pourrait le faire dans le cas général grâce à la diversité collective biaisée par les pondérations des liens initiaux.

2.5.2 *La diversité moyenne.*

Une autre façon de regarder la diversité d'une couche est de prendre la moyenne des scores de diversité des nœuds qui la composent.

Formellement :

Définition 15 (Diversité moyenne suivant un méta chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
Soit $\alpha \in \overline{\mathbb{R}}^+$ un réel positif, potentiellement infini. La diversité moyenne d'ordre α suivant un méta chemin Π de la couche V_0 est définie par

$$m^{\alpha}_{\Pi}(V_0) = \frac{\sum\limits_{u_0 \in V_0} D^{\alpha}_{\Pi}(u_0)}{\mid V_0 \mid}$$

En pratique pour coller aux mieux aux indicateurs des différents domaines ce n'est pas exactement cette diversité que nous allons calculer. Prenons l'indice d'Herfindhal par exemple, pour avoir la diversité d'ordre 2 nous prenons l'inverse de cet indice. Si on fait la moyenne sur la diversité d'ordre 2 nous prendrons donc l'inverse d'une moyenne de l'inverse de ces indicateurs.

Nous préférerons en général, prendre la moyenne des indices de Herfindhal, puis en prendre la moyenne. On notera cette moyenne M pour la différencier de la précédente. Formellement, avec les mêmes notations que précédemment :

Définition 16 (α-diversité moyenne suivant un méta chemin)**.** *Soit $n > 2$ un entier et $G = (V, E)$ un graphe n-parti.*
On fixe Π un méta chemin partant de la couche V_0 et on note pour chaque nœud $v \in \boldsymbol{V}_0$ de la couche initiale, $p(v)$ la distribution partant de v suivant le méta chemin Π

- $M^0_\Pi(V_0) = \frac{\sum_{u \in V_0} R(p(u))}{|V_0|}$
- $M^1_\Pi(V_0) = 2^{\frac{\sum_{u \in V_0} H(p(u))}{|V_0|}}$
- $M^2_\Pi(V_0) = \left(\frac{\sum_{u \in V_0} HHI(p(u))}{|V_0|} \right)^{-1}$
- $M^\infty_\Pi(V_0) = \left(\frac{\sum_{u \in V_0} BPI(p(u))}{|V_0|} \right)^{-1}$

Notons que nous avons créé ces moyennes pour correspondre aux différents domaines. C'est pour cela que nous ne donnons pas de formule générale. Pour la Richness, on remarque que $M^0 = m^0$ ce qui n'est pas le cas pour les autres.

2.5.3 *Différence entre la diversité moyenne et la diversité collective*

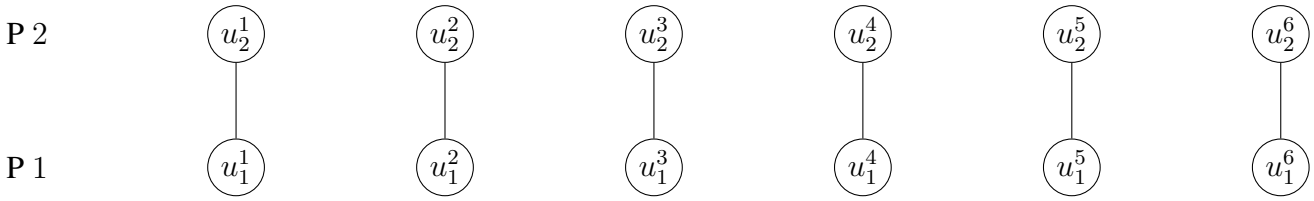


FIGURE 2.12 – Exemple de graphe biparti

Il y a des ressemblances entre la diversité moyenne et la diversité collective d'une couche. Cependant il faut comprendre une différence fondamentale : La diversité moyenne est la moyenne des diversités, tandis que la diversité collective est la diversité du comportement moyen. En effet on peut considérer que la distribution collective représente la distribution du comportement moyen puisque c'est une moyenne sur les nœuds de départ.

Regardons par exemple la figure 2.12. La diversité individuelle de chaque nœud est 1 (quelque soit l'indicateur choisit). Ce qui donne une diversité moyenne de 1. En revanche la distribution collective est totalement uniforme sur tous les nœuds d'arrivés, la diversité collective est donc 6 (ici aussi, quel que soit l'indicateur). S'il y avait n nœud d'arrivé, la diversité collective serait donc n.

2.5.4 *La diversité relative.*

Jusqu'ici nous avons tacitement considéré que la "meilleure" diversité était associée à la distribution uniforme. Ainsi, dans notre exemple musical, ceci revient à faire comme si nous avions considéré que nous vivions dans un monde neutre dans le quel le rock était aussi présent que le rap et que le classique. Dit autrement, c'est considérer qu'une consommation ne peut être parfaitement diverse que si la *playlist* associée est composée de façon équivalente d'un peu de tous les genres.

Maintenant imaginons que le monde produise 90% de rock et 10% de rap. Quelle serait la *playlist* la plus diversifiée ? Celle qui comporterait 50% 50% ou celle qui collerait au mieux avec ce que l'on produit ? Nous n'allons évidement pas répondre à cette question, mais nous allons introduire un indicateur nous permettant de définir une diversité relative à une distribution.

Il existe dans la théorie de l'information une généralisation de nos indicateurs qui permet de fournir une vision relative ([78]) notamment la divergence de Kullback-Leibler [47] qui est une version relative de l'entropie de Shannon. C'est ce que nous allons définir en premier (nous l'appellerons à cette étape α-divergence relative). Cependant on verra ensuite que ce ne sont pas exactement ces indicateurs qui vont nous intéresser en pratique, mais une petite variante de ceux ci (que l'on appellera α-diversité relative).

Tout comme l'entropie de Shannon n'est pas la 1-diversité (le second est le logarithme du premier), ce qu'on appellera true divergence d'ordre 1 n'est pas exactement la divergence de Kullback Leibler [2].

Définition 17 (α-divergence relative)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
La divergence relative d'ordre α est une fonction $D_\alpha : \Delta^* \times \Delta^* \to \mathbb{R}^+$, telle que, pour tout couple de distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$, $q = \langle q_1, \ldots, q_k \rangle \in \Delta^k$, avec $p_i = 0$ quand $q_i = 0$, et $\alpha \in \mathbb{R}^+$, définie ainsi :

$$D^*_\alpha(p \parallel q) = \left(\sum_{\substack{i=1 \\ q_i \neq 0}}^{k} p_i^\alpha q_i^{1-\alpha} \right)^{\frac{1}{\alpha - 1}} \qquad \textit{si } \alpha \neq 1,$$

Comme précédemment les valeurs limites de α définissent aussi des indicateurs ([100]) :

$$D^*_0(p \parallel q) = \left| \{ i \in \{1, \ldots, k\} : p_i \neq 0 \textit{ et } q_i \neq 0 \} \right|,$$

$$D^*_1(p \parallel q) = \left(\prod_{\substack{i=1 \\ q_i \neq 0}}^{k} \left(\frac{p_i}{q_i} \right)^{p_i} \right)^{-1} \quad \textit{avec } {p_i}^{p_i} 1 \textit{ si } p_i = 0 \quad \textit{et} \quad D^*_\infty = (p \parallel q) \left(\max_{\substack{i \leq k \\ q_i \neq 0}} \frac{p_i}{q_i} \right)^{-1}.$$

2. La divergence de Kullback Leibler dkl est définie sur deux distributions de taille k, p et q comme suit : $dkl(p, q) = \sum_{i<k} p_i * log(\frac{p_i}{q_i})$v

Regardons comment cette α-divergence se comporte quand q est la distribution uniforme.

Si $u = \langle 1/k, \ldots, 1/k \rangle$ est la distribution uniforme, alors pour tout $p \in \Delta^{k-1}$ on a $D^*_\alpha(p \,\|\, u) = k/D_\alpha(p)$. Pour avoir un indice de diversité relative comparable à nos diversités classiques nous avons choisi de définir la α-diversité relative ainsi.

Définition 18 (α-diversité relative)**.** *Soit $k \in \mathbb{N}^*$ un entier strictement positif.*
La diversité relative d'ordre α est une fonction $D_\alpha : \Delta^* \times \Delta^* \to \mathbb{R}^+$, telle que, pour pour tout couple de distribution $p = \langle p_1, ..., p_k \rangle \in \Delta^k$, $q = \langle q_1, \ldots, q_k \rangle \in \Delta^k$, avec $p_i = 0$ quand $q_i = 0$, et $\alpha \in \overline{\mathbb{R}}^+$. Définie ainsi :

$$D_\alpha(p \parallel q) = \frac{k}{D^*_\alpha(p \parallel q)}$$

Ainsi, avec $\langle f \rangle$ la distribution uniforme, nous retrouvons les définitions que nous avions posées plus tôt pour la α-diversité, $D_\alpha(p \parallel f) = D_\alpha(p)$.

Maintenant regardons d'autres propriétés classiques. Pour un entier positif $k \in \mathbb{N}^*$ et $p, q \in \Delta^k$ deux distributions, on a

$$D^*_\alpha(p \parallel q) \geq D^*_\alpha(p \parallel p) = 1,$$

Donc

$$D_\alpha(p \parallel q) \leq D_\alpha(p \parallel p) = k,$$

Reprenons notre exemple : supposons que la distribution globale sur une plateforme de musique soit 90% de rock et 10% de rap. Une personne écoutant 50% des deux genre aurait une diversité relative d'ordre 2 se calculant comme suit :

$$D_2(\langle 0.5, 0.5 \rangle \parallel \langle 0.9, 0.1 \rangle) = \frac{2}{\frac{0.5^2}{0.9} + \frac{0.5^2}{0.1}} \approx 0.72.$$

Alors qu'une personne écoutant 80% de rock et 20% de rap aurait une diversité relative d'ordre 2 se calculant comme suit :

$$D_2(\langle 0.8, 0.2 \rangle \parallel \langle 0.9, 0.1 \rangle) = \frac{2}{\frac{0.8^2}{0.9} + \frac{0.2^2}{0.1}} \approx 1.80.$$

Ainsi pour maximiser ce score une *playlist* doit se rapprocher de la distribution de genre globale. Notons que, contrairement au précédent, ce score n'est pas minoré par 1. Cette diversité relative est utile notamment dans la section 5.1.

2.6 CONCLUSION

Cette partie nous a permis d'introduire les objets et les notations nécessaires pour la suite de l'étude. Nous avons commencé par définir les **graphes n-parti**. L'importance de ce formalisme est sa généralité. Nous pourrons grâce à ce formalisme étudier la structure d'un système

avec des objets de plusieurs natures ayant des relations entre eux. Dans les deux parties suivantes nous avons regardé comment la diversité est mesurée dans différents domaines, pour en extraire une famille d'indicateurs ayant des propriétés communes : la α**-diversité**. Une fois les indicateurs de diversité choisis, nous avons défini une façon de trouver des distributions importantes auxquels les appliquer : la marche aléatoire. Nous avons donc pu définir l'objet fondamental de cette étude la α**-diversité selon un méta chemin**. La dernière partie nous donne des petites variations de cet objet qui peuvent aussi nous intéresser par la suite : **diversité collective, diversité moyenne** et **diversité relative**.

À noter que ce chapitre est en lien avec la soumission d'un article collectif [3] ayant pour objet la publication du formalisme complet lié à cette approche. L'une des différences majeures entre l'article soumis et la présente description du formalisme est le contexte des graphes n-partis que nous avons choisi d'utiliser, ce qui permet de se focaliser sur les nœuds des systèmes étudiés. Plus précisément ce sont les nœuds qui sont dans des parties alors que dans le formalisme de l'article (*"Heterogeneous Information Networks"*) ce sont les liens.

3. Article en phase de révision pour le journal Théorical Computer Science (version disponible ici[76]).

Chapitre

3

Analyse d'un graphe triparti décrivant des écoutes musicales.

Maintenant que nous avons vu la partie théorique de cette étude, nous analysons ce que notre méthode donne sur deux exemples issus de la consommation musicale. Ceci nous permet d'abord de montrer comment mettre des données sous forme de graphe triparti et ce que cette structure signifie. De plus il permet de montrer comment nos indicateurs peuvent être utilisés et le sens que nous pouvons leur donner. Enfin il nous permet de tester le comportement de nos indicateurs pour pouvoir les améliorer.

Nous allons d'abord présenter nos deux jeux de données (section 3.1), en montrant comment nous les avons pré-traités pour en faire des graphes triparti, puis en mettant en évidence leurs principales caractéristiques. Ensuite nous allons nous pencher sur un des indicateurs (diversité de Herfindahl : $\alpha = 2$) et l'utiliser pour analyser à la fois la diversité de l'audience d'un genre musical (section 3.2) puis la diversité exprimant le comportement des utilisateurs (section 3.3). Enfin nous compléterons notre analyse en regardant les autres indicateurs (section 3.4).

3.1 Jeux de données

Nous avons choisi deux jeux de données qui viennent du même domaine : l'écoute musicale en ligne, mais qui proviennent de deux plateformes différentes (Lastfm et Amazon). Nous les décrirons plus précisément dans la suite, mais nous pouvons déjà énoncer les différences fondamentales qu'il y a entre ces deux jeux de données. La première est une réelle donnée de consommation, nous allons regarder chaque fois qu'un utilisateur écoute une chanson. La deuxième est extraite des commentaires (*reviews*) : nous regardons quand un utilisateur commente une chanson, ce qui signifie qu'un utilisateur ne va commenter une chanson qu'une seule fois.

Nous allons tout d'abord, dans deux sections différentes, présenter les deux jeux de données et expliquer comment nous les avons pré-traitées pour obtenir un graphe triparti. Puis nous

allons regarder les caractéristiques de ces deux jeux de données.

3.1.1 *MSD / Lastfm : présentation et prétraitement*

Le premier jeu de données que nous avons utilisé provient du projet *Million Song Dataset* (*MSD*) [12]. Ce projet nous donne accès gratuitement à un ensemble de méta données liées à l'activité des utilisateurs sur une plateforme de chanson en ligne. Fournie par *The Echo Nest* (maintenant possédée par *Spotify*), nous pouvons tout d'abord avoir accès à un profil de goût des utilisateurs (jeu de donnée *user taste profile*) [1]. Il contient une liste de triplés (utilisateur, chanson, nombre d'écoute) qui représente le nombre de fois qu'un utilisateur écoute une chanson fixée. Ce jeu de donnée contient approximativement 48 millions de triplés. Ce qui comprend plus d'1 million d'utilisateurs différents et à peu près $300\ 000$ chansons. Grace à cela nous avons pu construire un premier graphe biparti qui constituera les couches 1 (utilisateurs) et 2 (chansons) de notre graphe triparti.

C'est souvent ce type de graphe biparti que nous avons considéré : des utilisateurs reliés à des objets. Ici la pondération des liens correspond simplement au nombre de fois que l'utilisateur écoute cette chanson.

Pour pouvoir avoir une notion de diversité comme on l'a définie précédemment il nous faut y rajouter une notion de catégorisation. Nous pourrions catégoriser les deux entités utilisateur ou chanson, ici nous avons accès à une catégorisation des chansons, ce qui va nous permettre de construire notre troisième couche.

Nous avons exploité le jeu de données *last.fm* [2] duquel nous avons extrait des *tags* (ou étiquettes) associés à des chansons : Pour chaque chanson, le jeu de données nous fournit une liste de tags qui définissent les catégories auxquelles appartient cette chanson. Ces tags ont été associés par les utilisateurs qui ont eux-même étiqueté les catégories musicales des titres proposés par la plateforme. De plus le jeu de données nous donne pour chaque élément de la liste un nombre entier (entre 1 et 100) représentant la force du tag (nombre sur lequel nous n'avons pas beaucoup de précision). Il contient à peu près $500\ 000$ chansons et approximativement le même nombre de tags. Nous pouvons donc constituer un deuxième biparti entre les chansons (couche 2) et les tags (couche 3). Nous avons en fait constitué deux bipartis différents en utilisant la force comme pondération des liens ou non, nous montrerons seulement les résultats sur le deuxième mais nous avons fait les mêmes calculs sur les deux [3].

Comme la catégorisation est faite directement par les utilisateurs, cela donne un caractère endogène à cette catégorisation, ce qui est fondamental ici. En effet nous basons fortement nos outils d'évaluation de la diversité sur cette catégorisation. Ce sera donc une approche différente de regarder un jeu de données dans laquelle la catégorisation est faite par les utilisateurs, par la plateforme, ou par un scientifique extérieur.

1. disponible ici `https://labrosa.ee.columbia.edu/millionsong/tasteprofile`
2. disponible ici `https://labrosa.ee.columbia.edu/millionsong/lastfm`
3. Les résultats sont sensiblement identiques.

Étant donné que ces deux jeux de données ont été fabriqués séparément, nous allons décrire comment nous avons pré-traité ces données pour en faire un triparti plus utilisable. Pour que nos résultats soient les plus expressifs possible, il y a notamment deux choses à travailler. Tout d'abord, nous avons supprimé les tags qui apportaient peu d'information. C'est pourquoi nous avons conservé les tags les plus populaires (nous avons conservé les 1000 tags les plus utilisés sur les plateformes). Ensuite, nous avons éliminé ce qu'on appelle les *impasses*, ici ce sont les chansons qui n'ont jamais été écoutées ou alors qui n'ont pas de tags.

Pour être plus précis, regardons ce qu'expriment les tags. L'utilisation de ces tags est très diverse, en effet on a des tags comme *"rock"* ou *"metal"* qui décrivent clairement le genre musical auquel appartient la chanson, à coté de cela d'autres comme *"webfound"* ou *"polyglotism"* sont plus problématiques. Nous allons voir dans la suite qu'il est quand même intéressant pour notre étude de ne pas conserver seulement les tags qui concernent le genre musical. Nous avons cependant débruité l'ensemble des tags en gardant ceux qui sont les plus populaires c'est-à-dire ceux qui sont associés au plus grand nombre de chansons (ceux qui ont le plus grand degré, ici $d_{3,2}$). Il nous fallait aussi s'assurer que chaque chanson était associée à un identifiant unique (le titre pouvait varier) [4]

Enfin comme annoncé nous avons supprimé les chansons qui n'avaient pas été écoutées et les chansons qui n'avaient pas été taguées. Il en résulte un graphe triparti contenant $1\ 019\ 190$ utilisateurs (couche 1), $234\ 379$ chansons (couche 2) et $1\ 000$ tags (couche 3).

3.1.2 *Amazon : présentation et prétraitement*

Le deuxième jeu de données de cette étude correspond à un enregistrement de tous les commentaires faits entre Mai 1996 et Juillet 2014 sur la plateforme d'achat en ligne *Amazon* [36, 58]. Nous avions accès à beaucoup de catégories d'objets vendus. Pour mieux correspondre au jeu de données précédent nous nous sommes contenté d'étudier les catégories *CDs & Vinyl* et *Digital Music*.

Comme précédemment nous avons exploité deux jeux de données indépendants pour pouvoir constituer un graphe triparti.

Le premier est un jeu de données de notation correspondant aux commentaires (*ratings dataset* [5]). Il contient une liste de quadruplés (utilisateur, objet, note, horodatage) qui signifie qu'un utilisateur a posté une commentaire sur un objet à l'heure horodatée et avec la note donnée. Ici la note ne nous intéresse pas, de plus, pour cette étude nous ne nous intéressons pas à l'aspect dynamique du graphe (ce sera fait dans la section 5.1). Ainsi on peut créer un premier graphe biparti utilisateurs (couche 1) vers un objet, qui sera en fait une chanson, (couche 2). Nous obtenons à peu près 500000 utilisateurs différents et 450000 chansons différentes. Il y a donc deux différences fondamentales avec le biparti précédent : premièrement il y a une différence de nature des liens, ce ne sont pas exactement des liens de consommations entre les utilisateurs

4. pour plus de détails voir ici : `https://labrosa.ee.columbia.edu/millionsong/blog/12-2-12-fixing-matching-errors`.

5. voir `http://jmcauley.ucsd.edu/data/amazon/`.

et les chansons mais des liens de commentaire. Ce qui implique la deuxième différence : il n'y a pas de pondération sur ces liens, un article ne peut être commenté qu'une seule fois par un utilisateur.

Le deuxième jeu de données (*metadata dataset*[5]) nous donne des détails sur tous les produits proposés sur *Amazon* contenant le prix du produit, les autres produits aussi vendus etc. Comme pour *MSD*, ce qui nous intéresse ici c'est la catégorisation de ces produits. Nous avons ici des listes de tags définissant chaque produit. Par exemple nous avons la liste suivante associée à un objet "[*CD& Vinyl, Children's Music, Stories*]". Nous pouvons voir qu'il y a une progression dans la liste qui part d'une catégorie globale et va vers une notion plus précise. Certains produits sont aussi associés à plusieurs listes. Nous avons d'abord choisi d'omettre *CD& Vinyl* et *Digital Music*, puisque c'est la racine de notre sélection (donc commune à tous les objets), puis de considérer chaque autre élément de la liste comme un tag indépendant. Par exemple pour la liste précédente nous avons relié le nœud représentant l'objet aux deux tags *Children's Music* et *Digital Music* avec deux liens distincts non pondérés. Ainsi nous avons un deuxième graphe biparti relayant les objets musicaux (couche 2) aux tags (couche 3). Ici aussi nous avons en réalité fait trois bipartis différents (ceux qui contiennent juste le premier tag de la liste, celui que l'on vient de présenter, et ceux qui contiennent juste le dernier tag de la liste), nous ne présentons que le deuxième. Notons ici aussi une différence fondamentale entre les deux jeux de données. La catégorisation n'est pas faite par les utilisateurs mais par la plateforme, elle est quand même endogène au système mais centralisée.

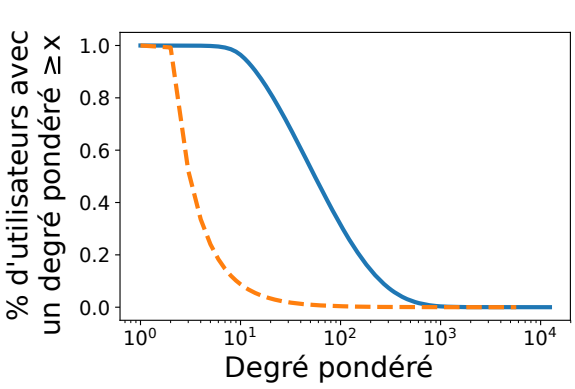
Comme dans le cas précédant nous avons opéré un prétraitement en supprimant les petites catégories (nous avons conservé les 250 les plus utilisées) et les impasses (chansons non commentées ou sans catégories).

Il en résulte un graphe triparti contenant 465 248 utilisateurs (nœuds de la couche 1), 445 514 chansons (nœuds de la couche 2) et 250 tags (nœuds de la couche 3). Comparé au jeu de données *MSD*, le nombre de chansons est similaire, mais le nombre de tags et d'utilisateurs est sensiblement inférieur.

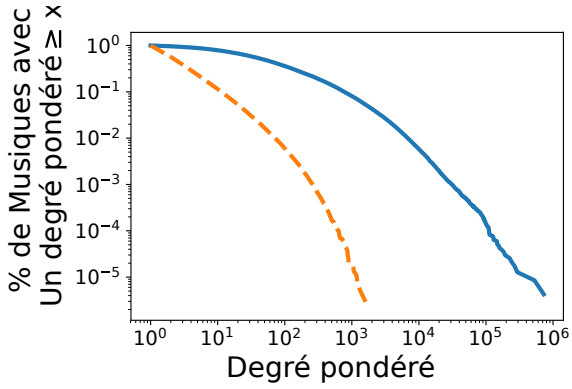
Pour résumer les choses et voir plus clairement les différences et les similarités nous faisons deux tableaux récapitulatifs des deux différents graphes bipartis.

jeu de donnée	nombre de nœuds de la couche 1	nombre de nœuds de la couche 2	nature des liens	nature de la pondération
MSD	1 019 190	234 379	écoute	nombre d'écoutes
Amazon	465 248	445 514	commentaire	pas de pondération

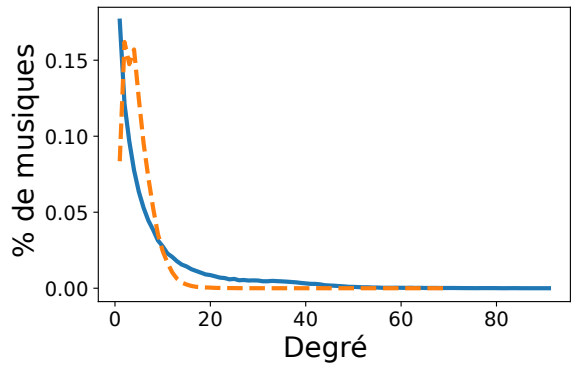
jeu de donnée	nombre de nœuds de la couche 2	nombre de nœuds de la couche 3	nature des liens
MSD	234 379	1000	endogène : faite par les utilisateurs
Amazon	445 514	250	endogène : faite par la plateforme



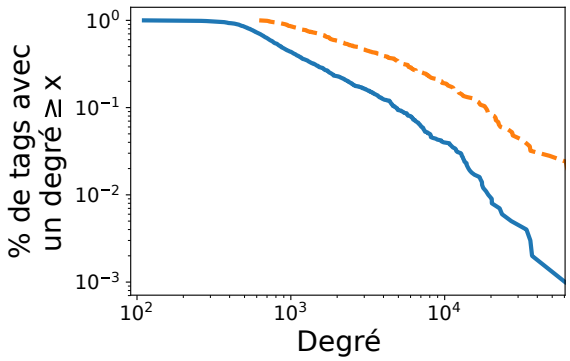
(a) Utilisateurs (degré $1, 2$)



(b) chansons par rapports aux utilisateurs (degré $2, 1$)



(c) chansons par rapports aux tags (degré $2, 3$)



(d) Tags par rapports aux chansons (degré $3, 2$) ($\top$)

FIGURE 3.1 – Les différentes distributions de degrés sur le graphe triparti Million Song Dataset (courbes bleues pleines) et Amazon Dataset (courbes oranges pointillées).

Il y a donc une grande similarité de nature et d'ordre de grandeur de ces données, même si des différences importantes sont à noter.

3.1.3 *Caractéristiques des deux graphes tripartis*

Maintenant que nous avons fixé nos graphes nous pouvons étudier des premières caractéristiques. Nous pouvons regarder les répartitions des 4 degrés différents ($d_{1,2}$, $d_{2,1}$, $d_{2,3}$ et $d_{3,2}$) que nous exposons dans la figure 3.1.

La figure 3.1a présente la distribution du degré pondéré des utilisateurs ($d_{1,2}$) sous forme de cumulative inverse [6]. Nous avons représenté *MSD* en lignes bleues et pleines et *Amazon* en lignes oranges et pointillées. Puisque nous pouvons avoir des degrés de valeurs très différents

6. C'est à dire qu'un point (x, y) sur cette courbe désigne le fait qu'il y a $y\%$ d'utilisateurs qui ont écouté au moins x titres musicaux (avec répétition possible des titres).

nous avons représenté l'axe des abscisses en échelle logarithmique. En effet les degrés maximaux sont $12\,387$ pour *MSD*, $5\,706$ pour *Amazon*, alors que l'ordre de grandeur reste restreint pour les utilisateurs : selon les données de *MSD* un utilisateur a 105 écoutes en moyenne sur 37 chansons différentes. C'est un peu différent sur *Amazon* où un utilisateur écrit (seulement) 5 commentaires. Notons ici que la différence de nature des liens de la couche 1 vers la couche 2 entre les jeux de données est importante ici (le degré pondéré d'un utilisateur sur *Amazon* est en fait son degré puisqu'il ne commentera pas 2 fois le même article). Ces valeurs sont assez représentatives du comportement d'un utilisateur aléatoire (proche des valeurs médianes) ce qui montre que ces comportements sont assez homogènes (malgré quelques utilisateurs avec une consommation intensive).

À l'inverse la figure 3.1b montre que la popularité des chansons est très hétérogène. Elle présente la distribution des degrés pondérés des chansons vers les utilisateurs ($d_{2,1}$). Ici aussi nous avons choisi de la présenter sous forme cumulative inverse, avec une échelle logarithmique en abscisse mais aussi en ordonnée. L'hétérogénéité est visible sur les deux jeux de données. En effet il y a des chansons très populaires (écoutées plus de $100\,000$ fois sur *MSD* et commentées plus de $1\,000$ fois sur *Amazon*). Tandis que la majorité des chansons a un degré pondéré beaucoup plus faible : (91% des chansons sont écoutées moins de $1\,000$ fois sur *MSD*, 64% sont commentées moins de 100 fois sur *Amazon*). Nous pourrions aussi regarder le degré non pondéré.Ceci n'a de sens que pour MSD et montrerait que, là encore, les liens sont répartis de façon hétérogène (78% des chansons sont écoutées par 100 utilisateurs différents et 88% sont commentés par 10 utilisateurs différents).

Regardons maintenant l'autre graphe biparti celui qui relie les chansons aux tags. La Figure 3.1c représente la répartition de degré des chansons par rapport aux tags ($d_{2,3}$). Une grande majorité des chansons ont un tout petit degré, c'est pour cela que nous n'avons pas ici représenté la distribution cumulative mais directement la répartition. En effet 72% des chansons sont liées à moins de 10 tags sur *MSD* (93% sur *Amazon*) sachant qu'il y en a 1000 possibles (250 sur *Amazon*). C'est une caractéristique attendue car les tags sont faits pour définir la chanson, il y a donc un nombre de possibilités restreint de tags à choisir. Rappelons aussi qu'on a supprimé les tags ayant trop peu de chansons, ce qui évidemment change cette répartition. Nous pouvons voir ici directement une différence entre les deux jeux de données. Alors que la courbe représentant *MSD* est complètement décroissante, la courbe représentant *Amazon* ressemble plus à une gaussienne (avec une trainée à droite). Même si nos deux systèmes de caractérisation sont endogènes (ce qui est en partie responsable de cette non uniformité), nous pouvons visualiser la différence ici entre une catégorisation organisée par une seule entité sur *Amazon* (plus uniforme : presque toute les chansons ont le même nombre de tags) et une catégorisation complètement décentralisée (fortement décroissante).

Enfin la figure 3.1d représente la répartition (cumulative inverse) de la distribution des degrés des tags vers les chansons ($d_{3,2}$). Cette courbe (en échelle log-log) montre ici aussi une distribution hétérogène des degrés chez les tags. Comme pour le degré ($d_{2,1}$) il y a des tags très populaires, mais la majorité concerne peu de chanson. Rappelons que l'on a tronqué selon ce degré, c'est pour cela que les deux courbes se retrouvent coupées en dessous de 1000 pour *MSD* et 250 pour *Amazon*.

Globalement, ces deux jeux de données ont des propriétés assez communes sur des systèmes similaires. Par exemple la popularité des chansons est très hétérogène et le comportement moyen des usagers est régulier.

3.2 Analyse de la 2-diversité de l'audience des tags.

Nous avons donc deux jeux de données, qui sont assez conséquents (les deux graphes contiennent un nombre de nœuds qui est de l'ordre du million). Même si la nature des nœuds est proche, il y a dans la nature des liens des différences fondamentales.

Nous allons pouvoir regarder comment nos indicateurs de diversité s'expriment sur ces deux jeux de données. Dans cette section nous nous intéressons à la diversité des tags. Plus précisément la diversité des tags vers les utilisateurs (cela correspond, dans notre formalisme à $D^{\alpha}_{3,2,1}$)). Intuitivement on regarde donc à quel point un tag est relié a des utilisateurs différents. Ainsi, un tag qui sera relié systématiquement aux mêmes utilisateurs aura une note de diversité faible, contrairement à un tag qui serait relié un ensemble plus grand d'utilisateurs et de façon mieux répartie. Cette intuition nous amène à nommer cette diversité : **diversité de l'audiance** (d'un tag). Nos quatre indicateurs ont des vertus différentes, cependant pour commencer avec plus de clarté, nous avons décidé d'en choisir un. Nous allons d'abord nous focaliser sur la 2-diversité (la diversité Herfindhal). Les diversités de variété et de Berger n'évaluant chacun qu'une partie de la diversité, la diversité de Herfindhal nous semble être un bon compromis. La diversité de Shannon étant similaire [7], notre choix entre les deux est arbitraire. Nous nous attarderons dans la suite à comparer ces quatre indicateurs.

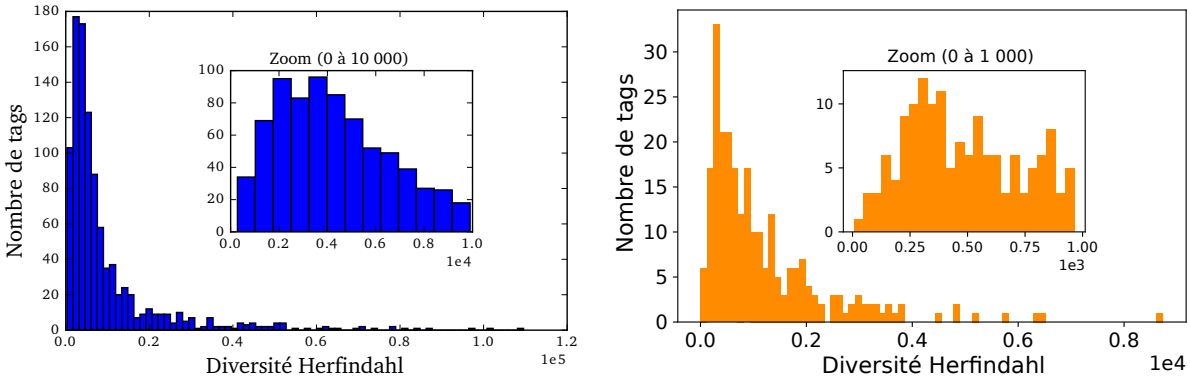
3.2.1 *Une perspective globale*

Nous allons d'abord regarder globalement comment est répartie la diversité de l'audience. La figure 3.2 montre la distribution de la diversité de l'audience pour tous les tags. Cette distribution est très hétérogène, avec une valeur moyenne de 9 699 sur *MSD* (la valeur médiane étant 5 111) et 1 197 sur *Amazon* (la valeur médiane étant 806). Elle exhibe cependant des tags avec une diversité particulièrement haute (supérieure à 100 000 sur *MSD* et proche de 10 000 sur Amazon). Ce qui montre que les tags peuvent avoir un public très différent,pouvant aller d'un public très large à un public très restreint.

Nous proposons aussi de nous concentrer sur les tags n'ayant pas une diversité trop grosse (voir les encarts de la figure3.2). Ces encarts représentent les tags avec une diversité inférieure à 10 000 sur *MSD* (ce qui représente 75% des tags) et 1 000 sur *Amazon* (ce qui représente 59% des tags). [8] Contrairement aux distributions globales, celles-ci sont plus homogènes. Celle de *MSD* est bien répartie et centrée sur une valeur proche des 3 000. Amazon est aussi homogène mais comporte plusieurs pics (300, 500 ...).

7. Les résultats que nos deux indicateurs nous donnent sont proches.

8. Notons qu'avec le zoom, la taille des fenêtres (*bins*) de l'encart n'est pas la même que pour la plus grande échelle.



(a) Million Song Dataset

(b) Amazon Dataset

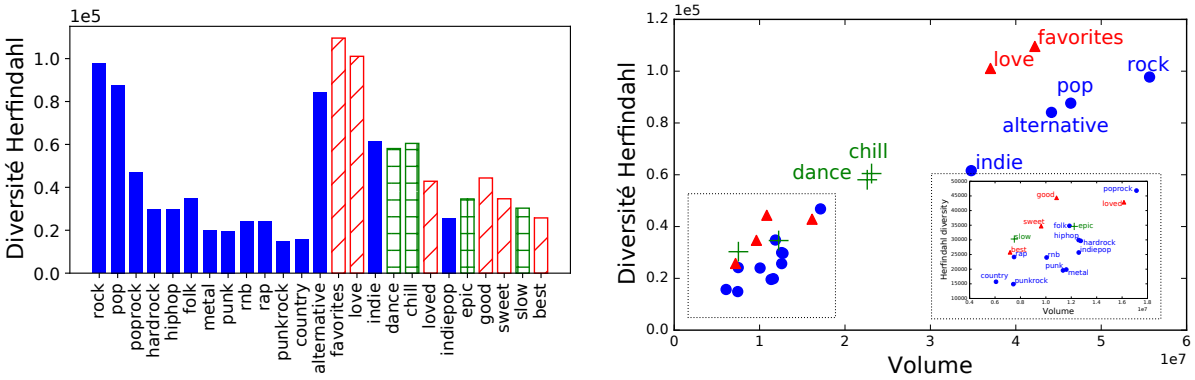
FIGURE 3.2 – Distribution de la diversité de l'audience des tags : *MSD* (à gauche) et *Amazon* (à droite).

Cette diversité suggère de regarder plus profondément le comportement de ce score en fonction du tag. Particulièrement nous allons nous intéresser à ce qui fait que notre indicateur est discriminant pour certains tags.

Pour ce faire, nous allons nous focaliser sur deux petits ensembles de tags et les étudier pour les deux jeux de données séparément. En effet il serait compliqué d'analyser manuellement le millier de tag que nous avons.

3.2.2 *Une concentration sur 25 tags de* MSD

Avant de sélectionner un petit groupe de tags, intéressons nous tout d'abord à la multiplicité des noms différents. Même parmi les 1 000 tags les plus populaires de *MSD*, nous avons une grande variété de noms différents. Ceci est probablement dû au fait que ce sont les utilisateurs qui associent aux chansons des tags. Par exemple il y a des tags clairement utilisés pour décrire le style musical de la chanson choisie (comme *"rock"*,*"metal"* ou *"country"*). Nous allons appeler ces tags *tags de style* et nous les représenterons par du bleu (des lignes pleines, ou des cercles, en fonction des figures). Contrairement à cela, d'autres tags sont moins liés au contenu de la chanson mais plus à un sentiment ou à une émotion ressentie lorsqu'on l'écoute (*"awesome"* ou *"best"*). Nous pouvons trouver aussi la période dans laquelle la chanson a été créée (*"1986"* ou *"70s"*). Il existe même des tags qui décrivent l'endroit ou le moment où la chanson peut être écoutée (*"tosleep"* ou *"shower"*). Pour tous ces types de tags nous utiliserons le terme de *tags génériques* pour les décrire, nous les représentons par des triangles rouge (ou des lignes pointillées rouges). Ces deux catégories ne sont pas forcément antinomiques, en effet il y a des tags qui pourraient appartenir aux deux catégories, notamment à cause de la polysémie des mots. Par exemple *"chill"* définit à la fois une émotion et un style. Nous nous



(a) diversité de l'audience des tags

(b) Herfindahl *vs.* volume

FIGURE 3.3 – Analyse de la diversité d'une sélection de 25 tags sur le MSD

référerons à ces tags en utilisant l'expression *tags intermédiaires*. Nous utiliserons, pour les représenter des croix vertes.

Cette distinction est importante dans le contexte de la diversité. En effet les tags génériques spécifient moins le type de chanson qui peut être écoutée. Nous pouvons donc nous attendre à ce que ces tags soient plus transversaux, c'est-à-dire qu'ils touchent des utilisateurs moins spécifiques. À l'inverse, les tags de style spécifient le genre de chanson, ils auront donc plus tendance à être adaptés à un public particulier. Nous pouvons donc nous attendre à ce que leur public soit moins diversifié. Nous allons voir si nos indicateurs se comportent différemment selon le type de tag.

Notre sélection des 25 tags s'est faite en gardant cette problématique. Ainsi parmi les 50 tags les plus populaires nous avons sélectionné 15 tags de style, 6 tags génériques et 4 tags intermédiaires.

Nous avons la liste de ces tags dans la figure 3.3, avec la diversité Herfindahl d'audience de chacun de ces tags.

De façon surprenante nous observons que les tags les plus diversifiés font partie des deux catégories extrêmes : *"rock"*, *"pop"* et *"alternative"* appartiennent aux tags de style alors que *"favorites"* et *"love"* sont des tags génériques. De même les tags de notre sélection ayant une faible diversité appartiennent aux trois catégories : *"country"* et *"metal"* sont des tags de style ; *"best"* est un tag générique et *"slow"* est un tag intermédiaire.

Ces premières observations soulèvent des interrogations. Il est intéressant de noter par exemple de voir que *"favorites"* et *"best"*, qui devraient définir le même type de chansons puisqu'ils ont un sens similaire, ont des scores de diversité très différents. Ceci nous amène à étudier plus en détail ce que capturent nos indicateurs. Or la démarche classique, lorsque l'on veut évaluer une mesure associé à un nœud dans un graphe, consiste à relativiser cet indicateur vis-à-vis

du degré du nœud. Ici il n'y a qu'un degré possible pour ces nœuds, le degré $d_{3,2}$ [9]. Sur ces données, ce degré représenterait le nombre de fois que le tag est associé à une chanson. Sachant que la diversité que l'on étudie est en fonction des utilisateurs, il nous paraît intéressant de comparer ce score à à un élément lié aux utilisateurs. C'est pourquoi nous avons choisi de rapporter la diversité d'un tag au nombre de fois qu'une chanson qui lui est liée est écoutée. Formellement il s'agit du degré pondéré $d_{3,1}$ sur le graphe projeté par rapport à la couche 2. Nous appelons ce degré : le **volume d'écoute** (que nous abrégerons par volume s'il n'y a pas d'ambiguïté).

La figure 3.3b nous montre la diversité exprimée précédemment en fonction du volume. Nous pouvons clairement voir une corrélation entre le volume et notre score de diversité : plus le volume est grand, plus le score de diversité est élevé. La corrélation est compréhensible car plus un tag est écouté, plus il a de chance de toucher des utilisateurs, donc plus il a de chance de toucher des utilisateurs différents.

Notons que le volume est un majorant de la diversité, c'est le score qu'aurait le tag s'il était écouté une unique fois par chaque utilisateur.

De plus, nous pouvons voir qu'à volume fixé, nous avons une séparation entre les types de tags : par exemple *favorites* et *love* ont un score plus élevé que *pop* et *alternative* alors qu'ils ont un volume sensiblement équivalent. De même, *good* est bien au dessus de *punk* et *métal*.

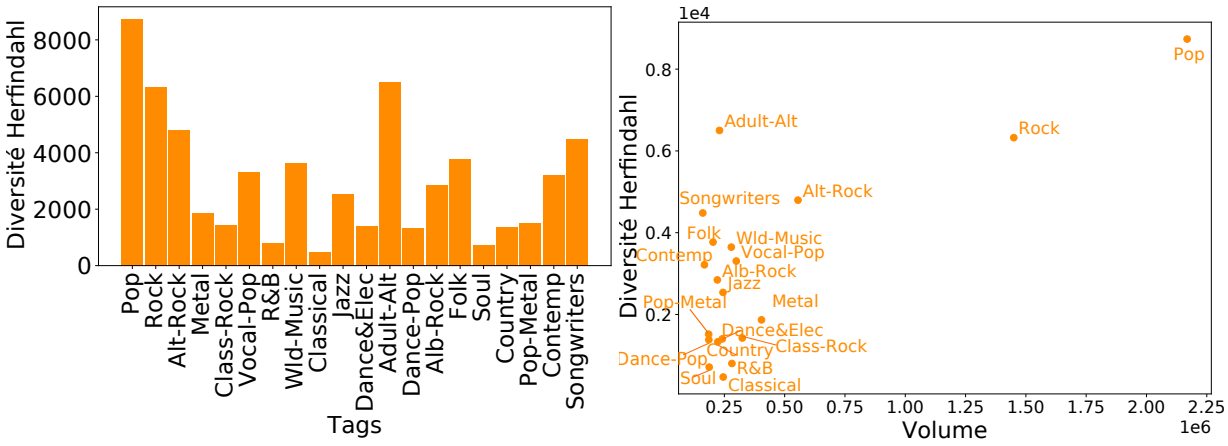
Nous pouvons d'ores et déjà nous poser une question qui reviendra dans cette partie : peut-on trouver une façon de comparer des tags ayant des volumes différents ? Comment comparer, par exemple, la diversité du rock et du rap, sans avoir besoin de considérer le volume pour interpréter la comparaison ? Attention, il nous paraît important, cependant, de ne pas supprimer cette dimension, le nombre d'écoutes est une information importante quand on cherche à analyser ou mesurer la diversité. Cependant nos indicateurs ont l'air d'être trop influencés par cette dimension et donc de ne pas assez mettre en valeur la totalité des dimensions de la diversité.

Regardons maintenant une autre sélection sur *Amazon*.

3.2.3 *Une concentration sur 20 tags d'*Amazon

Comparé au *MSD* les tags utilisés sur *Amazon* sont définis par la plateforme elle-même et non par les utilisateurs. Cet aspect centralisé introduit une manière plus standardisée pour décrire le contenu musical. Ainsi les tags génériques comme *"best"* sont proscrits. Ceci ne veut pas dire que ces tags utilisés sont d'un niveau de précision équivalent, rappelons nous qu'ils proviennent d'une description sous forme de liste qui tend à préciser le contenu musical au fur et à mesure de la progression dans la liste (voir la section 3.1.2). Par symétrie avec le jeu de données *MSD*, nous avons sélectionné 20 tags sur *Amazon*, nous avons pris les 20 tags les plus utilisés. Nous en avons la liste sur la figure 3.4 avec la diversité Herfindhal de chacun de ces tags. Comme précédemment il nous a paru intéressant de mettre en relief cette valeur

9. notons ici que sans pondération le degré est égal au degré pondéré



(a) Diversité de l'audience des tags **(b)** Herfindahl *vs.* volume

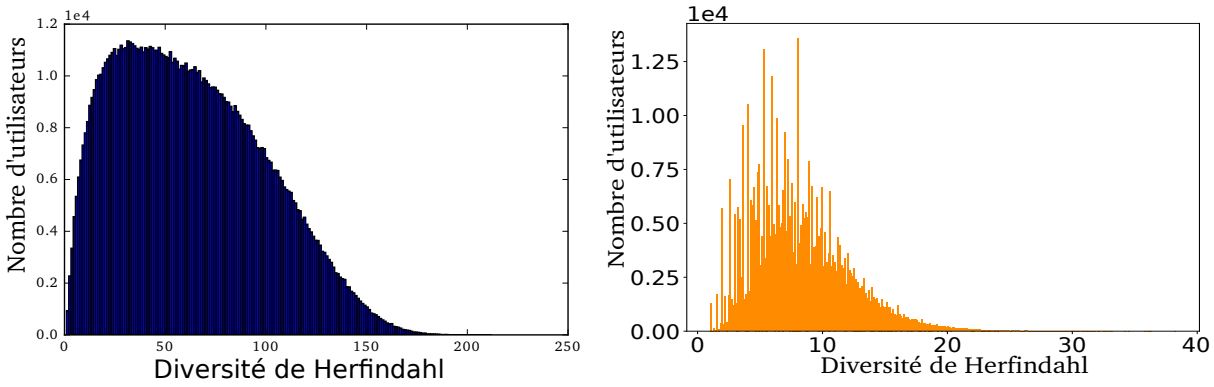
FIGURE 3.4 – Analyse de la diversité d'audience de la selection de 25 tags sur Amazon.

de diversité par rapport au volume. Ici la plupart des tags sélectionnés ont un volume similaire, ce qui nous permet de les comparer. Cependant *Rock* et *Pop* ont des volumes beaucoup plus grands et on voit encore une certaine corrélation car leur diversité est plus grande. Nous allons apporter une solution à ce problème en normalisant notre score par un score moyen dans la section 4.1.2. Cependant cela nécessite d'introduire la notion de modèle qui est le sujet du chapitre 4, pour le moment nous allons rester sur l'utilisation directe des indicateurs du chapitre 2 sur ces deux jeux de données.

3.3 ANALYSE DE LA 2-DIVERSITÉ DE L'ÉCOUTE DES UTILISATEURS

Après avoir analysé la diversité de l'audience des tags ($D^2_{3,2,1}$)) nous allons regarder la diversité du comportement des utilisateurs, c'est à dire formellement en analysant la 2-diversité à l'aide du méta-chemin inverse ($D^2_{1,2,3}$). Nous regardons ainsi comment un utilisateur diversifie son écoute en fonction des tags auxquels les titres qu'il écoute sont associés. Nous l'appellerons **diversité d'écoute**.

Sur la figure 3.5 nous pouvons voir la répartition de la diversité d'écoute des utilisateurs sur *MSD* et *Amazon*. Nous voyons, dans ces deux distributions, une claire homogénéité centrée autour de valeur moyennes (la moyenne et la médiane sont respectivement 63 et 59 sur *MSD*, 8 et 7 sur *Amazon*). Nous pouvons cependant observer que certains utilisateurs ont une diversité d'écoute particulièrement forte. Ceci correspond bien à la distribution des volumes des utilisateurs (voir section 3.1.3). En regardant manuellement ces utilisateurs, nous avons pu confirmer que les utilisateurs avec une forte diversité étaient ceux qui avaient un fort volume (nous les avions identifiés sur la figure 3.1a).



(a) Million Song Dataset

(b) Amazon Dataset

FIGURE 3.5 – Distribution de la diversité d'attention.

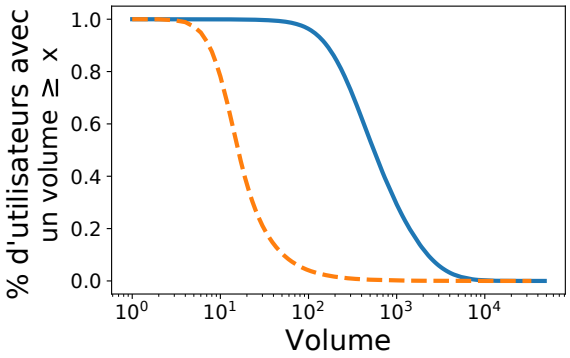


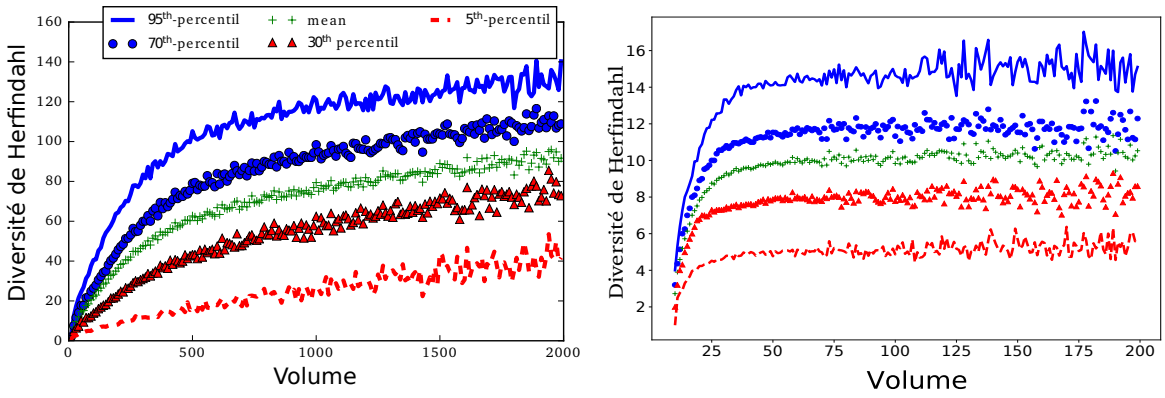
FIGURE 3.6 – Distribution du volume des utilisateurs pour *MSD* (courbe bleu pleine) and *Amazon* (courbe orange pointillées).

Dans la section précédente, après avoir analysé la distribution de diversité nous nous sommes focalisé sur des tags précis. Ici nous n'avons malheureusement pas d'information sur les utilisateurs car ils sont anonymisés. [10]. C'est pourquoi nous ne pouvons pas vraiment faire l'analyse de 25 utilisateurs particuliers.

Par contre, nous pouvons regarder comme précédemment le rapport entre la diversité et le volume. Ici le volume d'un utilisateur u correspond au nombre de tags associés à un titre écouté par u, multiplié par le nombre d'écoutes de ce titre.

Commençons par regarder la distribution du volume des utilisateurs, ce qui est fait dans la figure 3.6. En bleu (et en ligne pleine) nous avons la distribution de volume pour *MSD* et en orange (et pointillé) nous l'avons pour *Amazon*.

10. Un utilisateur est seulement représenté par un identifiant.



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 3.7 – Evolution de la diversité d'attention des utilisateurs en fonction du volume.

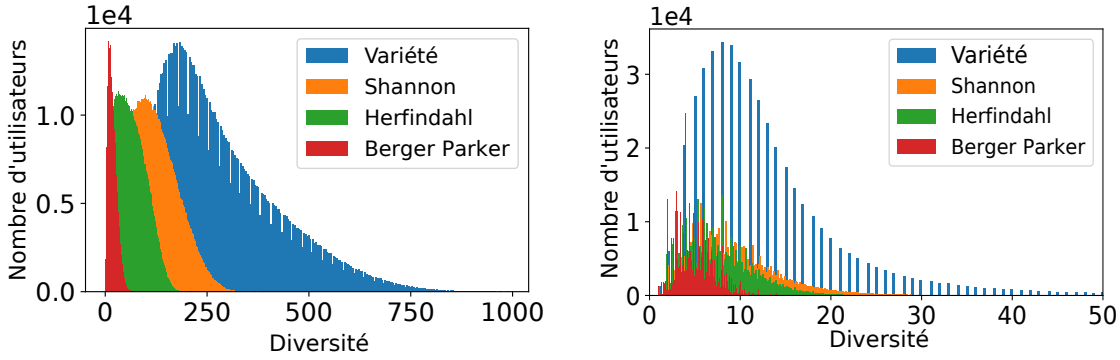
Même si l'axe des abscisses est en échelle logarithmique nous pouvons voir que l'ordre de grandeur est homogène. La grande majorité des utilisateurs (87%) ont un volume entre 10 et $2\,000$ sur *MSD*. Nous pouvons faire la même remarque sur *Amazon* où la majorité des utilisateurs (96%) ont un volume inférieur à 100.

Ceci nous donne assez d'éléments pour étudier comment la diversité d'écoute évolue en fonction du volume. La figure 3.7 nous montre une telle évolution pour les volumes en dessous de $2\,000$ pour *MSD* et en dessous de 200 pour *Amazon* [11]. Pour chaque volume la figure nous montre la moyenne de la diversité d'écoute observée ainsi que différents percentiles ($5^{\text{ième}}$, $30^{\text{ième}}$, $70^{\text{ième}}$ et $95^{\text{ième}}$).

Pour les deux jeux de données nous pouvons voir une influence du volume sur la diversité qui semble évoluer en deux phases. On distingue tout d'abord une première phase décrivant une forte progression (en dessous de 500 pour *MSD*, et 40 pour *Amazon*). Elle est particulièrement forte pour la partie haute de la population (moyenne et au dessus). Dans un second temps, on peut observer une progression certes croissante mais beaucoup moins prononcée, voire même une stagnation pour les plus gros volumes. Ceci met en évidence un phénomène de saturation de la diversité d'écoute des utilisateurs : quand le volume d'écoute atteint un certain seuil (approximativement 500 sur *MSD* et 40 sur *Amazon* selon nos observations), le comportement moyen des utilisateurs semble commencer à avoir atteint une diversité d'écoute qu'il devient difficile de dépasser. Au delà de ces seuils, les utilisateurs écoutent (ou commentent) un type de chanson qui correspondait déjà à ce qui avait été écouté (ou commenté).

Cette redondance expliquerait aussi pourquoi la diversité moyenne (63 pour *MSD*, 8 pour *Amazon*) est loin du maximum théorique (respectivement $1\,000$ et 250). Bien que le volume tend à élargir la perspective musicale d'un utilisateur, son goût vers une quantité limitée de contenus différents limite sa diversité. À ce stade, nous ne pouvons apporter de conclusions formelles et générales à ces remarques mais le chapitre 4 viendra explorer plus en détail et de

11. Pour les grandes valeurs nous aurions eu trop de variabilité, dû au petit nombre d'utilisateurs par volume



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 3.8 – Distribution des quatre indicateurs de diversité pour la diversité d'attention des utilisateurs.

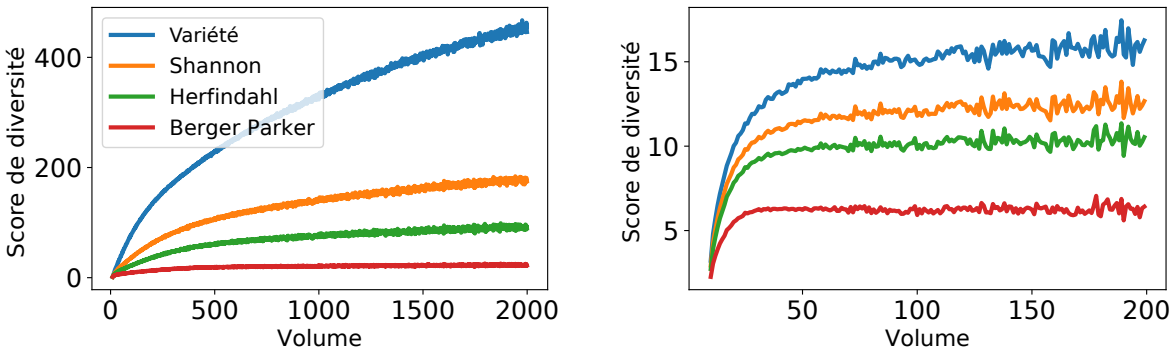
manière plus formelle la relation qu'entretiennent volume et diversité. Le chapitre 4 explore de façon plus construite la relation entre le volume et la diversité.

3.4 ANALYSE DES AUTRES INDICATEURS.

Par souci de simplicité, nous nous étions focalisé sur la diversité Herfindahl. Maintenant, par souci d'exhaustivité, nous allons regarder comment les autres indicateurs se comportent. Rappelons qu'intuitivement l'ordre de l'indicateur exprime un certain équilibre entre deux dimensions de la diversité : avec un α nul on ne prend en compte que le nombre de catégories atteintes, et plus on augmente l'ordre plus l'équilibre entre les abondances proportionnelles est important.

Pour regarder ce que donne cette différence en pratique nous avons tracé les figures 3.8 qui montrent les distributions des quatre indicateurs de diversité d'écoute des utilisateurs sur les deux jeux de données. Nous pouvons voir que plus l'ordre de diversité est grand plus le score calculé de diversité va être faible. De plus, plus cet ordre est grand plus la distribution est resserrée. Enfin, ces distributions sont, toutes les quatre, plutôt homogènes.

Le fait que le score de diversité soit plus faible quand l'ordre augmente, peut se voir aussi sur la courbe suivante. Comme précédemment nous avons regardé sur la figure 3.9 l'évolution de ces scores en fonction du volume (et en supprimant les très gros volume dont la diversité est trop fluctuante étant donné le peu d'utilisateurs associés). Nous pouvons observer le même phénomène de saturation que nous avons remarqué en section 3.3. Cependant les valeurs de saturations sont différentes (ce qui est cohérent avec la remarque précédente qui montre que les scores décroissent avec l'ordre). Par contre le seuil de changement de phase observé est le même (autour de 500 sur *MSD*, 40 sur *Amazon*) ce qui confirme notre analyse sur cet effet de saturation. En analysant comment l'effet de saturation est perçu par les différents indicateurs,



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 3.9 – Évolution des moyennes des quatres indicateurs de diversité en fonction du volume.

nous pouvons pressentir le fait que nos indicateurs analysent les mêmes effets mais avec un ordre de grandeur différent. Notons cependant que si ces valeurs changent en pratique en fonction de l'ordre, l'échelle des valeurs théoriques extrêmes restent identiques (1 et le nombre de nœuds de la couche d'arrivée du méta-chemin).

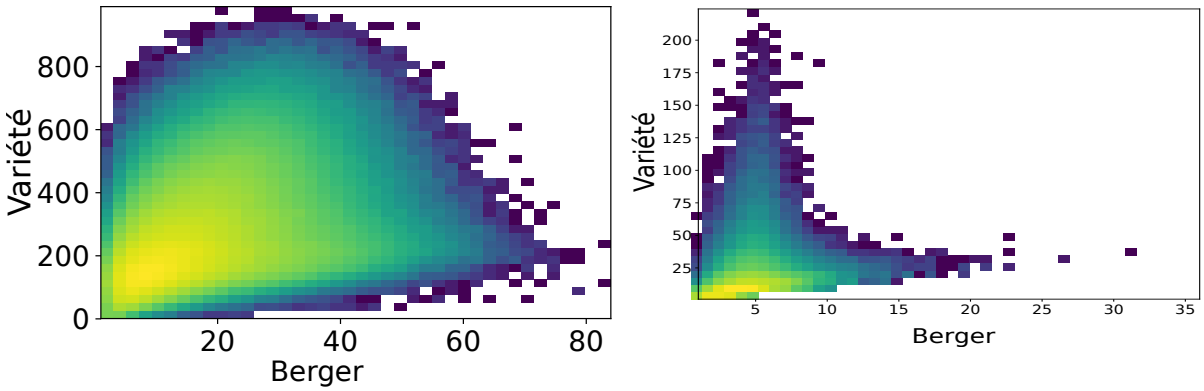
Si cette tendance est vérifiée à un niveau global du graphe triparti, il n'en va pas nécessairement de même au niveau individuel est il est possible que, pris séparément, les différentes diversités d'un nœud aient des rapports différents. C'est pourquoi nous nous tournons maintenant sur une comparaison point à point des différents indicateurs. Nous avons comparé la diversité de Berger et la diversité de variété [12] appliqué à la diversité d'écoute de tous les utilisateurs (Figure 3.10) et la diversité d'audience de tous les tags(Figure 3.11). S'il y a une tendance naturelle montrant une corrélation entre les deux diversité, visible en particulier pour les tags dans *MSD* (figure 3.11a) et *Amazon* (figure 3.11b) et, dans une moindre mesure, par les utilisateurs de *MSD* (figure 3.10a), nous pouvons repérer certains utilisateurs et tags qui ont un comportement de diversité très contrasté.

Par exemple *Album-Rock* et *Rap & HipHop* ont une forte diversité de variété sur *Amazon* mais une faible valeur de diversité de Berger. Cela indique que, bien que les deux atteignent un nombre élevé d'utilisateurs (d'où une forte Richness), un utilisateur très actif écrit régulièrement sur une chanson de ces catégories (d'où une faible diversité Berger).

Inversement, le tag *R & B* touche un nombre relativement faible d'utilisateurs dans *MSD*, mais ce sont des auditeurs plus réguliers de *R& B*. La distribution est alors plus proche d'une distribution uniforme, d'où une valeur Berger relativement élevée.

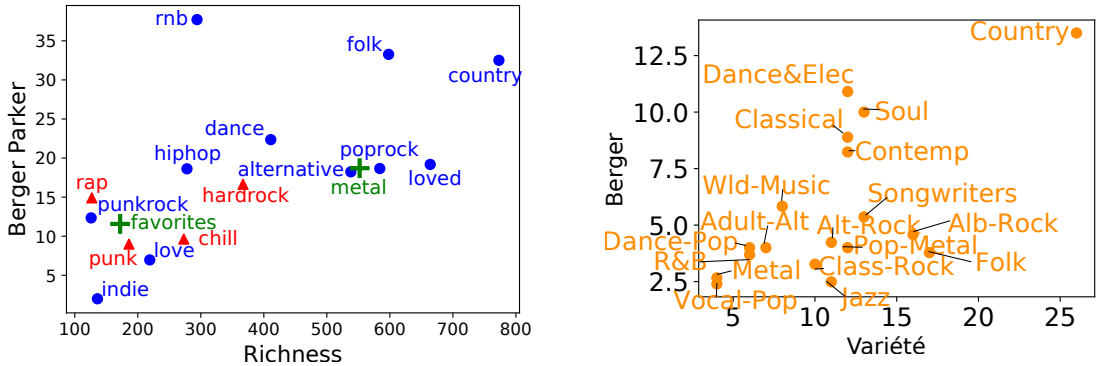
Ces exemples montrent que l'analyse de la diversité à différents ordres fournit une image plus complète de la diversité dans un ensemble de données.

12. Afin d'alléger la lecture de cette partie nous n'avons présenté dans le manuscrit que la comparaison entre ces deux ordre extremes de diversités mais une étude des autres indicateurs confirment les analyses obtenues ici.



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 3.10 – Diversité de Berger comparée à la diversité de variété (attention des utilisateurs).



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 3.11 – Diversité de Berger comparée à la diversité de variété (audience des tags).

3.5 CONCLUSION ET PERSPECTIVES

Nos indicateurs nous permettent de pouvoir analyser individuellement et globalement la diversité d'écoute des différents utilisateurs et, de façon complètement symétrique, la diversité d'audience des tags. Pour les utiliser il nous faut faire deux choses. D'abord, mettre nos données sous la forme de graphe triparti. Il y a des choix à faire : quels liens prendre en compte ? Quelle pondération ? Quels nœuds faut-il éliminer (avec le problème des *impasses*). Selon ces choix, et selon ce chemin choisi, il nous faut ensuite définir le sens que nous avons donné à cet indicateur (ici audience ou écoute).

Fondamentalement, nous avons aussi mis en exergue l'importance du volume pour nos indicateurs. C'est une dimension importante de la diversité, mais il est nécessaire de regarder en quoi elle influence nos indicateurs, ce qui est l'objet du prochain chapitre.

À noter que ce chapitre est en partie publié dans des actes d'une conférence internationale avec comité de lecture, sous la référence [72].

Chapitre

4

Normalisation de la diversité par des modèles de génération aléatoire.

Si le précédent chapitre a permis de mettre en évidence l'intérêt de l'approche proposée dans cette thèse pour l'analyse de la diversité de l'écoute musicale en ligne, une question est restée en suspens lorsque nous avons confronté la valeur de nos indicateurs aux volumes associés aux nœuds. La section 3.2 en particulier a montré que le volume avait une forte influence sur nos indicateurs. Pour pouvoir comparer deux éléments ayant des volumes très différents, il nous faut une manière de normaliser nos indicateurs par rapport au volume. En analysant l'effet de saturation que nous avons mis en évidence, une simple normalisation par le volume (ou même son logarithme) est insuffisante car le volume semble avoir une influence plus complexe qu'une influence linéaire ou logarithmique.

D'autre part, si l'analyse (comme celle proposée dans le précédent chapitre) permet de mettre en évidence des phénomènes intéressants, pouvoir expliquer pourquoi ces phénomènes sont présents nous semble également très important. Plus précisément, si le volume semble, pour l'instant, être un ingrédient fondamental, il y a peut être d'autres éléments qui peuvent "expliquer" la diversité.

Cette double question motive l'utilisation de modèles. Ainsi, dans la philosophie de l'utilisation de modèles de référence, nous allons tenter de calculer une diversité moyenne, c'est-à-dire la diversité qu'auraient les nœuds de notre graphe si une partie du graphe avait été générée aléatoirement. Ce score moyen nous permettra de normaliser le score de diversité proposé dans le chapitre 2, et ainsi de comparer des scores de nœuds ayant des volumes différents, sans que ceux-ci ne faussent l'interprétation. De plus, en mettant en valeur les différentes façons de générer les graphes aléatoirement et les différents paramètres que nous prenons en compte, nous pourrons compléter l'analyse et apporter des explications sur l'observation de la (non) diversité dans nos jeux de données.

Cette démarche rejoint aussi la distinction entre les deux méta-chemins que nous avons faite dans le chapitre précédent. En effet, la diversité d'audience des tags étant facile à analyser

individuellement (sur notre sélection de 25 tags) à l'aide du méta-chemin $3 \rightarrow 2 \rightarrow 1$, la question intéressante sera quelle normalisation adopter pour pouvoir comparer par exemple le *"rock"* et le *"rap"* qui ont des volumes très différents. Parallèlement à cela la diversité d'écoute des utilisateurs est intéressante car nous avons une masse importante de données (de l'ordre du million). Ainsi nous utiliserons la courbe d'évolution de cette diversité par rapport au volume pour la comparer à celle attendue par nos modèles aléatoires. Ce qui nous semble également intéressant dans ce chapitre est le triple regard que nous avons sur la situation. Premièrement nous avons analysé les deux jeux de données, ce qui nous a en retour permis de mieux cerner ce que capturaient nos indicateurs et la manière dont ils réagissaient aux différentes structures relationnelles. Deuxièmement, et c'est ce que nous allons présenter dans ce chapitre, nous allons proposer des scores calculés sur des graphes générés aléatoirement mais issus de modèles, respectant certaines propriétés, observées dans les jeux de données (toute la difficulté ici étant de déterminer lesquelles et comment les préserver).Enfin, nous allons compléter cette dernière approche par un travail analytique qui va permettre de déterminer précisément la valeur de diversité à partir des paramètres des modèles, ce qui renforcera l'intérêt de l'utilisation de modèles de génération aléatoire de graphes. C'est en faisant des aller-retour entre ces trois visions : analytique, simulée et réelle que nous avons pu avancer.

Tout d'abord nous allons montrer comment normaliser nos scores d'$\alpha - diversit$ à l'aide de modèles de génération de graphes aléatoire (section 4.1). Ensuite, nous prendrons du recul et proposerons un cadre formel pour dériver analytiquement cette normalisation dans la section 4.2. Nous partirons d'un cas simple et contraint du modèle, avant d'élargir les paramètres possibles. Ensuite, dans la section 4.3, nous comparerons différents scores de diversité issus de différents modèles et montrerons comment cette démarche permet d'expliquer le phénomène de saturation mis en évidence dans le chapitre précédent. Enfin,dans la section 4.4 nous allons prendre du recul sur les données pour regarder ce qui se passe quand on génère des graphes à partir de distributions de degré et non de graphes réels.

4.1 NORMALISATION DE LA DIVERSITÉ À L'AIDE DU CONFIGURATION MODEL.

En science des réseaux, une façon courante de prendre en compte l'effet d'une propriété sur un score est de comparer ce score à ce qui est sa valeur attendue pour des réseaux aléatoires ayant des propriétés similaires. C'est généralement effectué à l'aide de modèles qui génèrent des réseaux aléatoires. C'est exactement ce que nous allons faire ici. Dans cette section nous nous servirons de la diversité d'audience des tags pour l'illustrer. La question est donc de normaliser nos mesures et de relativiser l'importance du volume dans nos indicateurs, ce qui nous motive à regarder des points précis : nos deux sélections de tags (section 3.2.3 et section 3.2.2). Nous allons d'abord présenter comment générer un graphe à partir d'un autre, puis comment l'utiliser en l'appliquant sur nos jeux de données.

4.1.1 *Générer un graphe aléatoirement à partir d'un autre*

Dans notre contexte, nous avons utilisé une variante du *configuration model* [65] qui assigne aléatoirement des arêtes en conservant une seule propriété : le degré de chaque nœud. Nous parlons de variante car nous l'appliquons à des graphes bipartis, ce qui nous impose deux séquences de degrés (pondérés) fixés. Dans les faits, la solution algorithmique est plus simple car il n'y a pas, dans notre cas, de liens internes aux couches. Pour notre exemple, nous l'avons utilisé pour mélanger la partie inférieure du graphe triparti , c'est à dire les liens entre la couche 1 et 2, afin de réaffecter au hasard les liens entre les utilisateurs et les chansons.

Plus précisément, nous avons généré des graphes tripartis ayant le même nombre de nœuds et de liens mais tels que les liens entre les couches 2 et 1 sont répartis uniformément (avec leur poids) entre les nœuds en fonction de leur degré observé. Les autres liens entre la couche 3 et 2 restent inchangés.

Cela signifie que, par rapport au graphe original, les musiques du graphe généré ont exactement les mêmes tags et sont écoutées le même nombre de fois. Cependant, les utilisateurs qui les écoutent sont eux aléatoires. De plus, chaque utilisateur écoute le même nombre de titres différents et le nombre d'écoute par titre est conservé. Ainsi, chaque utilisateur a le même nombre d'écoute mais les musiques qu'il écoute sont sélectionnées aléatoirement de manière uniforme. Attention ici nous parlons d'uniformité, mais c'est une uniformité biaisée car comme nous respectons les degrés pondérés, un utilisateur a plus de chance d'écouter une musique très écoutée qu'une musique rarement écoutée. L'uniformité est donc à comprendre comme une uniformité vis à vis des liens proposés par la couche 2. Symétriquement une musique a plus de chance d'être écoutée par un utilisateur qui écoute beaucoup de musique qu'un utilisateur qui en a écouté peu.

Nous utiliserons les notations suivantes

> **Notation 2** (Graphe généré aléatoirement)**.** *Soit $\mathbb{T}$ un graphe triparti et $k, i \in \mathbb{N}$ des entiers.*
>
> — *On notera $\mathrm{Rand}(\mathbb{T})$ un graphe aléatoirement généré à partir de $\mathbb{T}$.*
> — *S'il y a plusieurs (k) graphes générés, on notera $\mathrm{Rand}_i(\mathbb{T})$ le $i^{ème}$ graphe généré à partir de $\mathbb{T}$, Dans ce cas on notera $\mathrm{R}_{\mathbb{T}}$, la suite des graphes générés : $\mathrm{R}_{\mathbb{T}} = (\mathrm{Rand}(\mathbb{T})_i)_{i<k}$.*
> — *Soit v un nœud de $\mathbb{T}$, on notera $c_i(v)$ la copie de v dans $\mathrm{Rand}_i(\mathbb{T})$.*

4.1.2 *Diversité moyenne, diversité normalisée*

Une fois que nous avons expliqué comment générer des graphes aléatoirement, nous allons les utiliser pour calculer une **diversité moyenne**. Ensuite nous diviserons nos indicateurs observés par cette moyenne pour calculer une diversité normalisée.

Ici il nous faut tout d'abord étendre la notion de diversité moyenne que l'on a définie précédemment (définition 16). En effet ce que l'on calculait était, pour un graphe fixé, la moyenne

des scores de diversité sur un ensemble de nœuds. Ici nous ne voulons pas regarder un ensemble de nœuds, mais fixer un nœud et faire varier les graphes.

Définition 19 (α-diversité moyenne par génération aléatoire)**.** *Soit $\mathbb{T}$ un graphe n-parti, $k \in \mathbb{N}$ un entier. On fixe Π un méta chemin partant de la couche V_0 et $v_0 \in \boldsymbol{V}_0$ nœud de la couche initiale. De plus on suppose que l'on a généré k copies de $\mathbb{T}$, dont la suite est notée $\mathrm{R}_{\mathbb{T}}$.*

Pour chaque v nœud on note $p(v)$ la distribution partant de v suivant le méta chemin Π.

- $M^0_\Pi(\mathrm{R}_\mathbb{T}, v) = \frac{\sum_{i<k} R(p(c_i(v)))}{k}$
- $M^1_\Pi(\mathrm{R}_\mathbb{T}, v) = 2^{\frac{\sum_{i<k} H(p(c_i(v)))}{k}}$
- $M^2_\Pi(\mathrm{R}_\mathbb{T}, v) = \left(\frac{\sum_{i<k} HHI(p(c_i(v))}{k}\right)^{-1}$
- $M^\infty(\mathrm{R}_\mathbb{T}, v) = \left(\frac{\sum_{i<k} HHI(p(c_i(v))}{k}\right)^{-1}$

Nous allons utiliser ces scores calculés sur des graphes générés aléatoirement pour normaliser les mesures précédentes. Formellement :

Définition 20 (α-diversité normalisée par génération aléatoire suivant un méta chemin)**.** *Soit $\alpha \in \overline{\mathbb{R}}^+$ un réel positif, potentiellement infini et v un nœud d'un graphe n-parti $\mathbb{T}$. On fixe $\mathrm{R}_\mathbb{T}$ une suite de graphes générée à partir de $\mathbb{T}$. La diversité normalisée par génération aléatoire d'ordre α suivant un méta chemin Π de la couche V_0 est définie par*

$$N^\alpha_\Pi(\mathrm{R}_\mathbb{T}, v) = \frac{D^\alpha_\Pi(v)}{M^\alpha_\Pi(\mathrm{R}_\mathbb{T}, v)}$$

Dans la suite nous allons regarder en particulier la 2-diversité normalisée.

Le résultat de la figure 4.1 reprend la sélection de tags faite dans la section 3.2.2 sur le jeu de donnée *MSD*, et trace leur 2-diversité normalisée en fonction du volume. Comme espéré, la comparaison avec le modèle aléatoire compense l'impact du volume sur la façon dont la diversité est calculée. À première vue, aucune corrélation particulière ne peut maintenant être observée entre le volume et la diversité herfindahl normalisée. Cela n'empêche pas les tags à volume élevé d'avoir toujours une diversité relativement élevée, car ils ont de meilleures chances d'atteindre un public plus large. Nous voyons notamment que la diversité augmente légèrement avec le volume.

Par ailleurs, la diversité normalisée semble rétablir l'équilibre entre des tags très proches sur le plan sémantique. Les tags *love* et *loved* par exemple, obtiennent maintenant une diversité similaire (environ $0,29$) bien qu'ils aient un volume différent (*love* est presque 4 fois plus utilisé que *loved*). Des observations similaires peuvent être tirées pour les tags *indie* et *indiepop*.

De plus, cette courbe montre que la diversité herfindahl normalisée peut faire la distinction entre les tags génériques (situés sur la partie supérieure de la courbe) et les tags de style

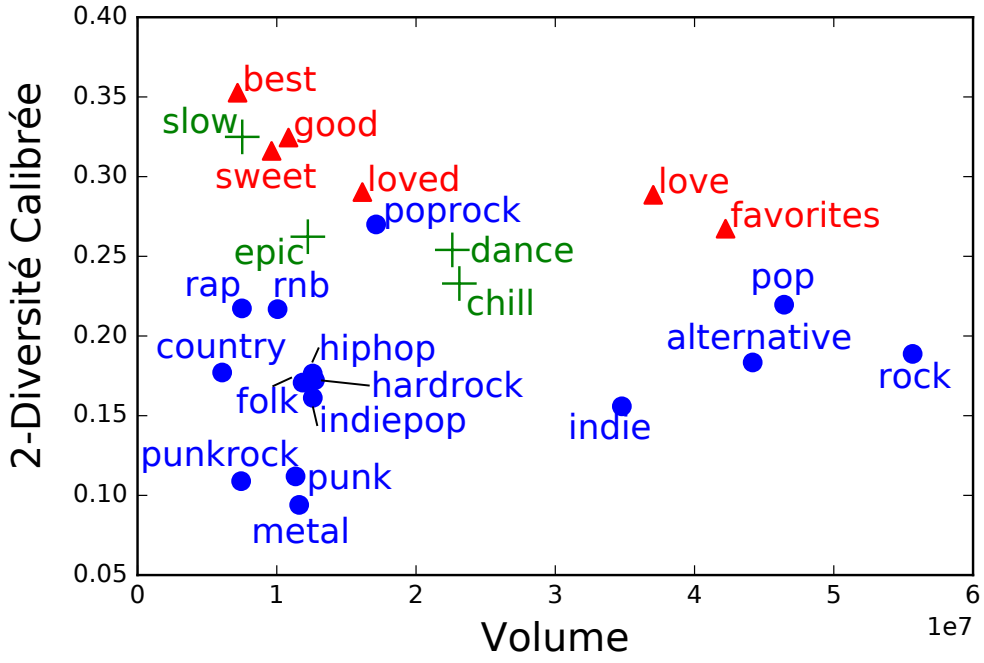


FIGURE 4.1 – 2-Diversité d'audience des tags normalisée(*MSD*)

(principalement dans la partie inférieure) : les tags ont tous une grande diversité (entre $0,23$ et $0,35$) tandis que les tags de style ont tendance à avoir une valeur inférieure (de $0,09$ à $0,22$).

La seule exception est *poprock* qui a un score relativement élevé pour une tag de style $(0,27)$ et malgré son faible volume (17 millions d'écoutes). Il est difficile de tirer une conclusion avec cette valeur seulement, cependant les tags *rock* et *pop* pris indépendamment ayant également une bonne diversité, nous pouvons émettre l'hypothèse que *poprock* profite de leur diversité respective et atteint les deux publics donc un public plus large et diversifié.

Enfin, le fait que la diversité normalisée soit désormais indépendante du volume permet des comparaisons de musiques avec des volumes similaires. Par exemple, on peut le constater, bien qu'ils aient tous un volume similaire (environ 10 millions d'écoutes), *rap* et *rnb* touchent un public beaucoup plus diversifié que *metal* et *punk*, qui semblent former des sous-groupes plutôt denses et cohérents d'auditeurs.

Des observation similaires peuvent être faites à partir de la sélection de 25 tags dans le jeu de données *MSD* (cf section 3.2.3, nous avons refait la même courbe cette fois sur notre sélection d'*Amazon*. Nous voyons clairement que, de façon semblable à ce que nous avons observé sur le cas *MSD*, bien que *Pop* et *Rock* aient un volume élevé, leur diversité normalisée est inférieure à celle des autres tags, tels que *Adult-alternative* ou *Singer-Songwriters* pour citer les plus hauts scores. Ce dernier en particulier pourrait évidemment représenter un tag générique puisqu'il est plus susceptible de représenter une qualité du chanteur (qui a également composé la musique) qu'une propriété du contenu musical. En tant que tel, il aurait tendance à catégo-

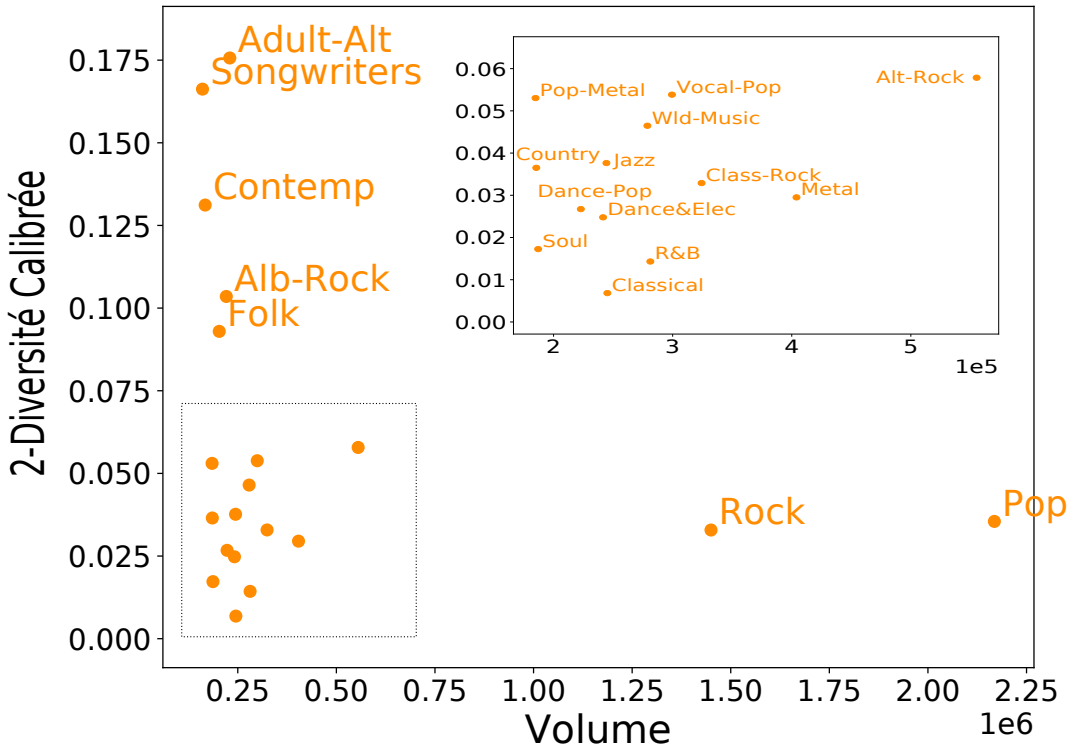


FIGURE 4.2 – 2-Diversité d'audience des tags calibrée (*Amazon*)

riser des musiques appartenant à des types de musique distincts, ce qui pourrait expliquer ce score.

Encore une fois, bien qu'il soit difficile de tirer des conclusions générales sur un nombre restreint de tags (cela nécessiterait des observations plus fines sur les utilisateurs afin de mieux justifier la diversité), les deux résultats (Figures 4.1 et 4.2) renforcent notre affirmation selon laquelle la diversité normalisée fournit un bon moyen d'analyser la diversité des tags dans ce contexte. On voit en particulier que nous avons diminué l'influence du volume qui influençait trop nos indicateurs initiaux.

Notons que nous avons choisi une une façon particulière de générer des graphes (nous ne changeons que l'ensemble des liens entre la couche 2 et la couche 1 en préservant les degrés, et nous ne touchons pas aux liens entre la couche 3 et la couche 2). Ceci correspond à une question précise que nous nous sommes posée sur les données : quel serait la diversité d'un tag si les utilisateurs écoutaient des musiques de façon aléatoire, tout en gardant la même intensité d'écoute chez les utilisateurs, et la même intensité d'audience chez les musiques. Cette dernière partie est importante à noter, nous avons gardé des grandes distinctions entre l'écoute des utilisateurs : il y a des utilisateurs *très actifs*(qui écoutent beaucoup de musiques) et des utilisateurs *peu actifs* qui en écoutent peu. De même il y a des musiques plus *populaires* (écoutées) que d'autres.

Ainsi, le choix du modèle de génération aléatoire est fortement lié au contexte et nous en

étudierons d'autres (développés dans la section 4.3). Notons que ce qui a guidé notre choix ici a été de conserver le volume des nœuds dans les graphes générés par le modèles, afin de pouvoir comparer équitablement les valeurs de diversité avant et après la génération aléatoire.

4.2 PRÉVOIR ANALYTIQUEMENT LA DIVERSITÉ MOYENNE

Nous n'avons pas parlé d'implémentation ni de temps de calcul. Sachant que nous devons nous adapter à des données potentiellement grosses, nous pouvons anticiper que, même avec des algorithmes efficaces, générer des centaines, des milliers, voire des millions de graphes aléatoires, et calculer à chaque fois nos indicateurs, peut s'avérer coûteux en temps et en espace de calcul. C'est une des raisons qui nous amène à chercher analytiquement des résultats permettant de déterminer, sans simulation, la diversité attendue.

4.2.1 *Une formalisation générale du problème.*

Nous allons prendre du recul vis à vis de notre situation pour fixer un cadre formel beaucoup plus abstrait mais qui nous servira de base pour regarder plusieurs cas pratiques liés à différents choix à faire lors de la génération aléatoire. Pour cela, souvenons-nous que, si on simplifie le contexte et oublie un moment la structure relationnelle sur laquelle nous travaillons, le problème peut se résumer en l'analyse de distributions de probabilités sur un ensemble d'entités fini (ceux de la couche d'arrivée). Ainsi notre problème peut se voir comme l'affectation de valeurs (celles issues des nœuds de la couche de départ) sur ces entités. Formellement, soit les éléments suivants :

- Soit $n \in \mathbb{N}$ un entier : représentant le nombre d'entités (nombre de sommets de la couche d'arrivée dans notre contexte).
- Soit $m \in \mathbb{N}$ un entier : représentant le nombre de valeurs à distribuer.
- Fixons $(v_i)_{i<m} \in (\mathbb{R}^+)^m$ une suite de réels : représentant la suite des valeurs (les distributions issues des marches aléatoires dans notre contexte).
- Pour tout $i < m$, fixons X_i une variable aléatoire à valeur dans $[\![0, n-1]\!]$: représentant l'entité à laquelle nous allons affecter la valeur i.
- Pour tout $k < n$ nous nous intéresserons à $Y_k = \sum_{i<m} (\mathbb{1}_{<X_i=k>} v_i)$: ceci représente la somme des valeurs affectées à l'entitié k. Mis dans notre contexte l'ensemble des Y_k représente l'abondance proportionnelle.
- Pour pouvoir représenter nos indicateurs de façon générale nous fixons 3 différentes fonctions f définies de $\mathbb{R}^+$ dans $\mathbb{R}$:
  - $f_0 : x \mapsto \mathbb{1}_{x>0}$: utile pour la richness
  - $f_1 : x \mapsto xlog(x)$: utile pour la diversité Shannon
  - $f_2 : x \mapsto x^2$: utile pour la diversité Herfindahl

— Nous pouvons ainsi définir, pour une fonction f donnée, la diversité $D_f(Y) = \sum_{k<n} f(Y_k)$ [1]

Grâce à ce cadre, la question que nous nous posons est de regarder l'espérance de D_f que nous noterons $E(D_f)$.

4.2.2 *Un premier cas simple.*

Théorème 5 (Cas de base)**.** *Sous l'hypothèse des contraintes suivantes :*

— *Quel que soit* $i < m$, $v_i = 1/m$.

— *Quel que soit* $k < n$, $P(X_i = k) = 1/n$ *(par souci de simplification notons* $a = 1/n$*).*

— *Les* X_i *sont des variables indépendantes*

On a : $E(D_f) = n \sum_{i \leq m} f(i/m)\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$

Démonstration. Par définition : $E(D_f) = E(D_f(Y)) = (\sum_{k<n} f(Y_k))$
Par linéarité de l'espérance : $E(D_f) = \sum_k E(f(Y_k))$.
Par l'équation de transfert : $E(D_f) = \sum_{k<n} \sum_{s \in S} f(s)P(Y_k = s)$ (où S est l'ensemble des valeurs possibles des Y_k).
Or, S est l'ensemble des valeurs de Y_k possibles c'est-à-dire $\{i/m, i \in [\![0, m]\!]\}$ [2].
Fixons maintenant $i \in [\![0, m]\!]$, $k \in [\![0, n-1]\!]$ et calculons $P(Y_k = i/m)$.
Rappelons que $Y_k = \sum_{i<m} (\mathbb{1}_{<X_i=k>} v_i)$.
De plus, pour $j < n$, $P(X_j = k) = a$.
Comme les variables X_j sont indépendantes : $P(Y_k) = \binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$.
Ainsi : $E(D_f) = \sum_{k<n} \sum_{i<m} f(i/m)\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$.
Comme la somme ne dépend pas de k on a bien :

$$E(D_f) = n \sum_{i \leq m} f(i/m)\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$$

.

□

Regardons maintenant comment, en faisant varier f (donc les indicateurs de diversité) nous pouvons aboutir à des résultats plus précis.

Dans le cas ou $f = f_0$ on a le résultat suivant :

1. Attention ce système nous permet de généraliser et de poser des calculs simplement, il a pourtant deux défauts : il ne prend pas en compte la diversité Berger, et ce ne sont pas exactement nos indicateurs de α-diversité qui sont définis.

2. Ce sont des sommes composées de i fois le nombre $1/m$, avec $i \in [\![0, m]\!]$.

Théorème 6 (Cas de base pour $f = f_0$)**.** *Sous l'hypothèse des contraintes suivantes :*

- *Quel que soit* $i < m$, $v_i = 1/m$.
- *Quel que soit* $k < n$, $P(X_i = k) = 1/n$ *(par souci de simplification notons* $a = 1/n$*).*
- *Les* X_i *sont des variables indépendantes*

On a : $E(D_{f_0}) = n(1 - (1 - 1/n)^m)$

Démonstration. Par le théorème 5 on a : $E(D_{f_0}) = n \sum_{i \leq m} f_0(i/m)\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$.
En remplaçant f_0 par sa définition on a : $E(D_f) = n \sum_{i \leq m} \mathbb{1}_{i/m>0}\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$
On sait que $i/m = 0$ ssi $i = 0$.
Donc $E(D_f) = n\left(\left(\sum_{i \in [\![1,m]\!]} \binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}\right) - \binom{m}{0}\alpha^0(1-\alpha)^{(m-0)}\right)$
Or $\sum_i \binom{m}{i}\alpha^i(1-\alpha)^{(m-i)} = (a + (1-a))^m = 1^m = 1$.
Ce qui nous donne : $E(D_f) = n(1 - \binom{m}{0}\alpha^0(1-\alpha)^{(m-0)})$
En simplifiant :

$$E(D_{f_0}) = n(1 - (1 - 1/n)^m)$$

□

Avant de prouver notre prochain théorème nous avons besoin du lemme suivant :

Lemme 4 (binôme et carré)**.** $\sum_{i \leq m} i^2\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)} = \frac{m(n+m-1)}{n^2}$

Démonstration. Posons g la fonction définie de $\mathbb{R}$ dans $\mathbb{R}$ telle que : $g : x \mapsto \sum_{i \leq m} x^i\binom{m}{i}(1-\alpha)^{(m-i)}$.
g est une fonction polynomiale donc dérivable et :
$\forall x,\ g'(x) = \sum_{1 \leq i \leq m} ix^{i-1}\binom{m}{i}(1-\alpha)^{(m-i)}$
donc $\forall x,\ xg'(x) = \sum_{i \leq m} ix^i\binom{m}{i}(1-\alpha)^{(m-i)}$.
La fonction $x \mapsto xg'(x)$ est aussi dérivable et :
$\forall x,\ x(xg'(x))' = \sum_{i \leq m} i^2x^i\binom{m}{i}(1-\alpha)^{(m-i)}$.
En appliquant en a, $(x(xg'(x))')(a) = \sum_{i \leq m} i^2\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$
Regardons maintenant la fonction g d'une autre manière :
$\forall x,\ g(x) = (x + 1 - a)^m$
donc $g'(x) = m(x + 1 - a)^{m-1}$
on multiplie par x : $xg'(x) = xm(x + 1 - a)^{m-1}$
et $(xg'(x))' = m(x + 1 - a)^{m-1} + xm(m-1)(x + 1 - a)^{m-2}$ [3]
ainsi $x(xg'(x))' = xm(x + 1 - a)^{m-1} + x^2m(m-1)(x + 1 - a)^{m-2}$.

3. on a besoin de $m > 2$, ce qui est le cas en pratique

En appliquant en a, $(x(xg'(x))')(a) = am(a+1-a)^{m-1} + a^2m(m-1)(a+1-a)^{m-2}$.
On simplifie $(x(xg'(x))')(a) = am + a^2m(m-1)$.
On remplace a par $1/n$, et on factorise par m pour avoir :
$(x(xg'(x))')(a) = \frac{m(n+m-1)}{n^2}$. En réunissant nos deux égalités nous avons :

$$\sum_{i \leq m} i^2 \binom{m}{i} \alpha^i (1-\alpha)^{(m-i)} = \frac{m(n+m-1)}{n^2}$$

□

Dans le cas où $f = f_2$ on a l'équation suivante :

Théorème 7 (Cas de base pour $f = f_2$)**.** *Sous l'hypothèse des contraintes suivantes :*

- *Quel que soit $i < m$, $v_i = 1/m$.*
- *Quel que soit $k < n$, $P(X_i = k) = 1/n$ (par soucis de simplification notons $a = 1/n$).*
- *Les X_i sont des variables indépendantes*

On a : $E(D_{f_2}) = \frac{n+m-1}{m*n}$

Démonstration. Par le théorème 5 on a : $E(D_{f_2}) = n \sum_{i \leq m} f_2(i/m)\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$.
Par définition de f_2 : $E(D_{f_2}) = n \sum_{i \leq m} (i/m)^2\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$. On peut sortir m de la somme : f_2 : $E(D_{f_2}) = \frac{n}{m^2} \sum_{i \leq m} i^2\binom{m}{i}\alpha^i(1-\alpha)^{(m-i)}$. On reconnaît notre somme connue, grâce au lemme précédent on a $E(D_{f_2}) = \frac{n}{m^2} \times \frac{m(n+m-1)}{n^2}$
En simplifiant :

$$E(D_{f_2}) = \frac{n+m-1}{m*n}$$

□

$f = x \mapsto x^2$.

4.2.3 *Introduire une différence entre les entités.*

Nous allons maintenant complexifier ce cas basique pour voir comment notre équation peut évoluer. La première chose que nous allons supprimer est l'uniformité entre les entités. Précisément, nous allons considérer que la probabilité qu'une valeur soit affectée à une entité dépend de l'entité . Ceci nous permettra, dans la suite, de pouvoir adapter nos équations aux simulations qui préservent le degré. En effet, considérer que le degré des entités est fixée c'est en quelque sorte fixer une probabilité d'être affecté à une entité. Formellement on a le théorème suivant

Théorème 8 (Cas de base avec dépendance de l'entité)**.** *Sous l'hypothèse des contraintes suivantes :*

- *Quel que soit* $i < m$, $v_i = 1/m$.
- *Quel que soit* $k < n$, $P(X_i = k) = a_k$ *(où* $a_k \in \mathbb{R}$ *et* $\sum_{k<n} a_k = 1$ *).*
- *Les* X_i *sont des variables indépendantes*

On a :

f_0 : $E(D_{f_0}) = \sum_{k<n} (1 - (1 - a_k)^m)$

f_2 : $E(D_{f_2}) = \sum_{k<n} \frac{a_k(1+a_k(m-1))}{m}$

Démonstration. Nous avons exactement les mêmes preuves que précédemment. □

4.2.4 *Introduire des valeurs différentes*

Maintenant que nous avons regardé ce qui se passait avec des valeurs toutes identiques, nous allons considérer le cas plus général où les valeurs peuvent varier, mais sont connues. Si on replace cela dans le contexte de la α-diversité ceci nous permettra de faire varier les degrés des nœuds de départs ainsi que la pondération des liens. Dans ce cas, le cas de f_0 n'est pas très intéressant puisqu'il ne s'intéresse pas aux valeurs. Nous allons donc nous focaliser sur f_2.[4]

Théorème 9 (Des valeurs distinctes)**.** *Sous l'hypothèse des contraintes suivantes :*

- *Quel que soit* $i < m$, $v_i \in \mathbb{R}$. *(où* $\sum_{i<m} v_i = 1$*).*
- *Quel que soit* $k < n$, $P(X_i = k) = a_k$ *(où* $a_k \in \mathbb{R}$ *et* $\sum_{k<n} a_k = 1$ *).*
- *Les* X_i *sont des variables indépendantes*

On a : $E(D_{f_2}) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k(1 - \sum_{i<m} {v_i}^2) \right)$

Démonstration. Soit $k < n$, reprenons la définition de Y_k : $Y_k = \sum_{i<m} v_i \mathbb{1}_{X_i=k}$.
Donc $\sum_{k<n} Y_k^2 = \sum_{k<n} \sum_{i<m} \sum_{j<n} v_i v_j \mathbb{1}_{X_i=k} \mathbb{1}_{X_j=k}$.
Ainsi par linéarité de l'espérance $E(\sum_{k<n} Y_k^2) = \sum_{k<n} \sum_{i<m} \sum_{j<n} v_i v_j P(X_i = k, X_j = k)$.
Or, pour $k < n$, $P(X_i = k, X_j = k) = a_k$ si $i = j$ et a_k^2 sinon.
Donc $E(\sum_{k<n} Y_k^2) = \sum_{k<n} (\sum_{i<m} a_k {v_i}^2 + \sum_{i<m} \sum_{j<m \wedge j \neq i} v_i v_j a_k^2)$

4. Nous avons aussi essayé d'étudier f_1 mais son étude était plus compliquée.

En factorisant par rapport à a_k on a $E(\sum_{k<n} Y_k^2) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k \sum_{i<m} (v_i \sum_{j<m \wedge j \neq i} v_j) \right)$

Ainsi

$$E(D_{f_2}) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k (1 - \sum_{i<m} {v_i}^2) \right)$$

.

□

4.2.5 *Influence de la redondance*

Pour finir nous allons nous intéresser à la supposition d'indépendance des X_i. Cette supposition signifie qu'il n'y a en quelque sorte pas de mémoire dans la façon d'associer des valeurs, c'est-à-dire que le fait d'associer une valeur à une entité n'influence pas le choix de l'entité suivante. Nous allons regarder ce qui se passe si l'on contredit cette supposition. Nous utilisons le terme redondance car on va considérer qu'il est plus probable d'associer une valeur à une entité si on en a déjà associé une. Pour donner une intuition dans le contexte de la diversité dans l'écoute musicale vue auparavant, cela reviendrait à prendre en compte la probabilité d'écouter un titre appartenant à une catégorie déjà considérée par l'utilisateur.

Théorème 10 (Cas avec redondance.)**.** *Sous l'hypothèse des contraintes suivantes :*

- *Quel que soit* $i < m$, $v_i \in \mathbb{R}$. *(où* $\sum_{i<m} v_i = 1$*).*
- *Quel que soit* $k < n$, $P(X_i = k) = a_k$ *(où* $a_k \in \mathbb{R}$ *et* $\sum_{k<n} a_k = 1$ *).*
- *Quel que soit* $k < n$, $P(X_i = k \wedge X_j = k) = a_k \beta_k$ *(où* $\beta_k \in \mathbb{R}$*)*

On a : $E(D_{f_2}) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + \beta_k (1 - \sum_{i<m} {v_i}^2) \right)$

Démonstration. Comme pour la preuve du théorème 9 on a :
$E(\sum_{k<n} Y_k^2) = \sum_{k<n} \sum_{i<m} \sum_{j<n} v_i v_j P(X_i = k, X_j = k)$.
Or, pour $k < n$, $P(X_i = k, X_j = k) = a_k$ si $i = j$ et $a_k \beta_k$ sinon.
Donc $E(\sum_{k<n} Y_k^2) = \sum_{k<n} (\sum_{i<m} a_k {v_i}^2 + \sum_{i<m} \sum_{j<m \wedge j \neq i} v_i v_j a_k \beta_k)$.

En factorisant par rapport à a_k on a $E(\sum_{k<n} Y_k^2) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + \beta_k \sum_{i<m} (v_i \sum_{j<m \wedge j \neq i} v_j) \right)$.

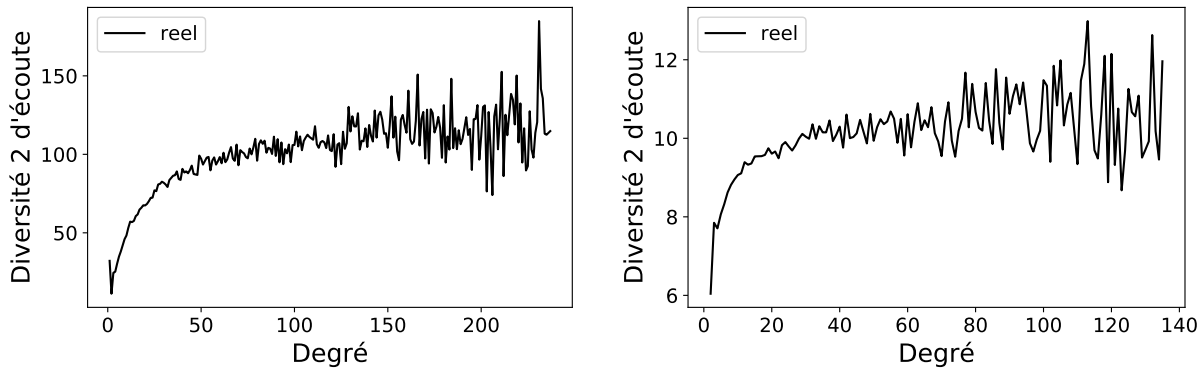
Ainsi

$$E(D_{f_2}) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + \beta_k (1 - \sum_{i<m} {v_i}^2) \right).$$

□

4.3 ÉTUDE DE LA SATURATION : PERTINENCE DES MODÈLES PROPOSÉS.

Nous allons nous appuyer sur ces différents résultats pour étudier l'impact des différentes hypothèses sur la diversité et mettre cet impact en perspective avec la diversité observée sur nos jeux de données. Pour cela, nous allons reprendre la courbe présentant l'évolution de la diversité d'écoute des utilisateurs (figure 4.3) et la comparer à celle issues des différents modèles issus des théorèmes 6 à 10. Cette confrontation entre la diversité observée dans les jeux de données et celles issues de modèles théorique va ainsi permettre d'identifier plus précisément quels sont les éléments incorporés dans les modèles qui sont le mieux à même de reproduire une diversité réelle. Cette démarche permet ainsi de fournir des éléments explicatif à la (non) diversité observée et analysée dans le chapitre 3. Notons que, par souci d'homogénéité, nous avons choisi d'analyser l'évolution de la diversité en fonction du degré (ce qui correspond dans notre cas au volume d'écoute d'un utilisateur) en prenant le cas de la diversité d'ordre 2. L'un des points d'attention dans ce qui suit est la question du phénomène de saturation mis en évidence section 3.3. C'est cette évolution qui va constituer l'élément clef pour évaluer à quel point la diversité induite par les modèles est réaliste ou non. Notons que pour la 2-diversité nous allons regarder $\frac{1}{E(D_{f_2})}$.

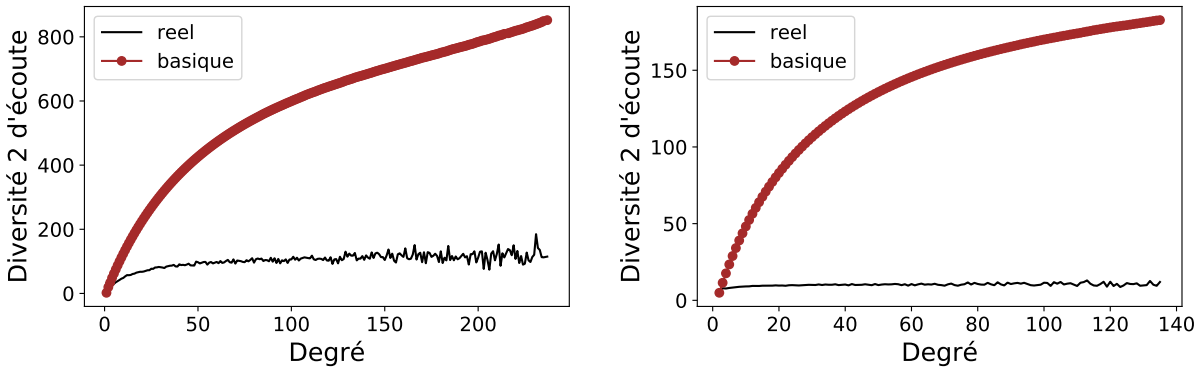


(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.3 – Évolution de la 2-diversité d'écoute des utilisateurs.

4.3.1 *Écouter les tags de manière uniforme.*

Commençons par un modèle très naïf et supposons que les utilisateurs sélectionnent des styles de musique de manière uniforme. La diversité attendue dans cette situation est décrite par l'équation donnée par le théorème 7 dans laquelle m correspond alors au degré de l'utilisateur et n au nombre de tags. $\frac{1}{E(D_{f_2})} = \frac{m*n}{n+m-1}$. En comparant l'évolution de la diversité attendue par le modèle à celle observée en pratique (figure 4.4), on peut remarquer que le phénomène



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.4 – Étude du phénomène de saturation à partir du modèle de base (baseline)

de saturation est déjà visible (même s'il reste très léger) dans ce cas très idéalisé. En revanche, le seuil de saturation est nettement plus élevée (elle tend vers le nombre de tags) que celle observée, qui stagne autour de 150 (resp. 30) dans le cas de *MSD* (resp. *Amazon*). Notons que nous pouvons remarquer ça directement sur l'équation à n fixé : $\frac{1}{E(D_{f_2})} = \frac{m*n}{n+m-1} \xrightarrow[m\to\infty]{} n$. Il est intéressant de voir que dans ce cas totalement idéalisé il y a déjà ce phénomène de saturation. Cependant sa limite est exactement le maximum possible : le nombre de tags.

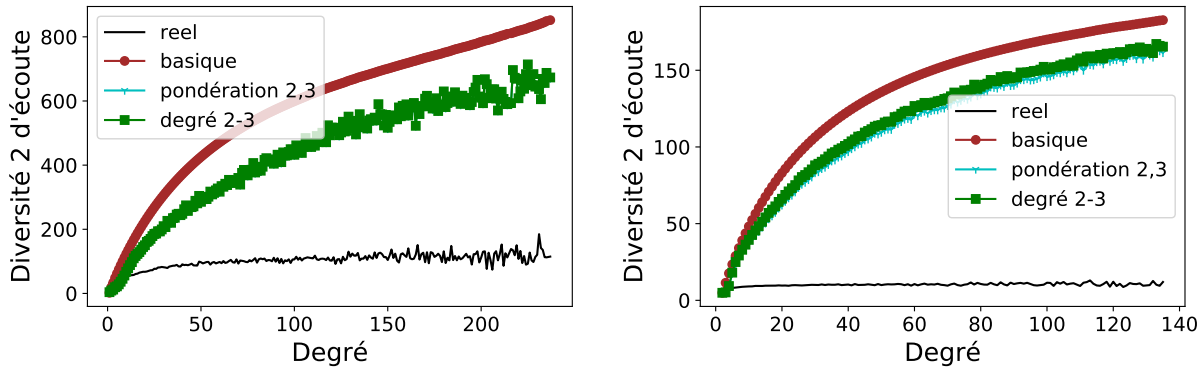
En partant de ce modèle de référence, nous allons maintenant intégrer dans le modèle différents éléments afin de se rapprocher le plus possible d'une situation réaliste, en suivant la progression faite pour établir les théorèmes 6 à 9. Ceci va nous permettre de jauger de l'influence de chacune des hypothèses sur la diversité observée. Ainsi, nous allons successivement étudier l'influence de la différentiation des entités (sous section 4.3.2), de la différentiation des valeurs (sous section 4.3.3)et de la redondance (sous section 4.3.4) au regard de nos jeux de données.

4.3.2 *Conserver un biparti, générer un autre.*

Supposons tout d'abord que la structure relationnelle entre utilisateur et titre soit conservée mais que les tags associés aux titres soient choisis aléatoirement. Il existe deux manières d'effectuer l'appariement entre titre et musique : Soit nous conservons le degré pondéré de chaque musique, soit nous conservons la pondération de chaque lien partant de cette musique [5]. Nous appellerons la première génération **degré** $2,3$ et la deuxième **pondération** $2,3$.

Dans les deux cas nous avons un ensemble de valeurs qui évolue, nous sommes donc dans le cas du théorème 9 (sans variation de l'entité) qui établit que la diversité attendue est $\frac{1}{E(D_{f_2})} =$

5. Pour cette étude nous avons utilisé la pondération des liens $2-3$ sur *MSD* liée à la *force* des tags



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.5 – Étude du phénomène de saturation à partir des modèles de conservation du degré et de la pondération

$\left(\sum_{i<m} {v_i}^2 + a(1 - \sum_{i<m} {v_i}^2)\right)^{-1}$. Il nous faut déterminer dans les deux cas quelle est la suite (v_i).

Il nous faut introduire quelques notations : Pour u un utilisateur de degré $d_u = d^*_{1,2}(u)$, considérons la suite $(s_k)_{k<d_u}$ des musiques composant son voisinage $(E_{1,2}(u))$. Pour $k < d_u$, on peut noter $d_k = d^*_{2,3}(s_k)$, et de la même manière on a $(t^j_k)_{j<d_k}$ l'ensemble des tags composant son voisinage $(E_{2,3}(u))$. Regardons maintenant la pondération des liens concernés. Pour $k < d_u$ et $j < d_k$ notons $x_k = \frac{w_{1,2}(u,s_k)}{d_{1,2}(u)}$ et $y^j_k = \frac{w_{2,3}(s_k,t^j_k)}{d_{1,2}(s_k)}$.

Pour la génération *degré* $2,3$, pour chaque musique s_k (où $k < d_u$), on crée d_k liens, tous de pondération $1/d_k$. Chaque lien sera pris en compte par la marche aléatoire une fois ayant pour valeur $x_k * d_k$. Ainsi $(v_i) = ((x_k * d_k)_{j<d_k})_{k<d_u}$.

Pour la génération *pondération* $2,3$, pour chaque musique s_k (où $k < d_u$), on a d_k liens dont la $j^{\text{ème}}$ pondération (pour $j < d_k$) est égale à y^j_k . Chaque lien sera pris en compte par la marche aléatoire une fois ayant pour valeur $x_k * y^j_k$. Ainsi $(v_i) = ((x_k * y^j_k)_{j<d_k})_{k<d_u}$.

Nous avons tracé l'évolution des diversités associées à ces deux variantes du modèle sur la figure 4.5. [6] Nous voyons aussi que nous avons encore un phénomène de saturation mais que la limite semble plus faible que celle de l'équation simple. Enfin nous pouvons voir que les deux nouvelles courbes sont très proches, ce qui montre [7] que ces deux façons de simuler sont assez proches. Cependant on reste très loin de la limite observée.

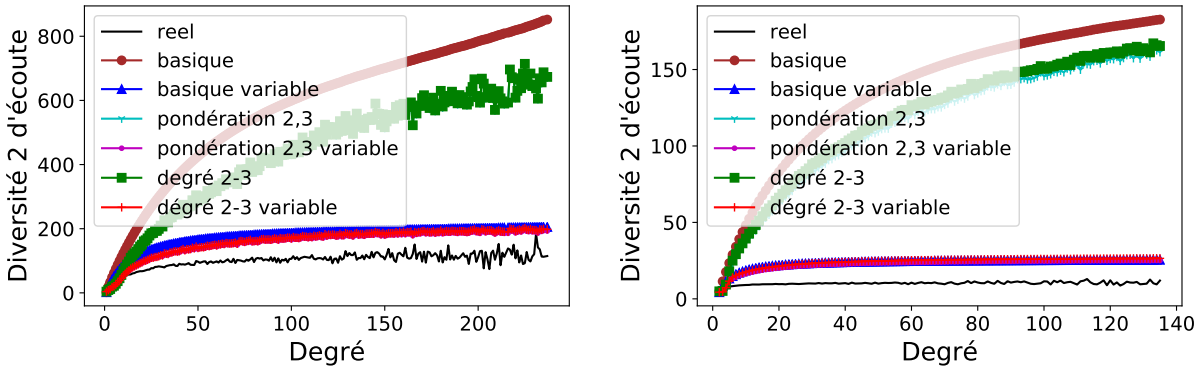
Pour nous intéresser plus à cette limite nous pouvons regarder le comportement de nos équation à l'infini. Supposons que $\sum_{i<m} v_i^2 \to 0$ (ce que l'on peut prendre comme première approxi-

6. Par symétrie avec les autres sections, nous avons tracé cette évolution par rapport au degré, même si l'équation ne dépend pas exactement du degré

7. sur *MSD*, puisque sur *Amazon* il n'y a pas de pondération $2-3$

mation) on a $\frac{1}{E(D_{f_2})} = \left(\sum_{i<m} {v_i}^2 + a(1 - \sum_{i<m} {v_i}^2)\right)^{-1} \xrightarrow[\sum_{i<m} v_i^2 \to 0]{} \frac{1}{a} = n$. La limite de $\sum_{i<m} v_i^2$ étant légèrement plus haute que 0 nous avons une saturation un peu plus basse que n.

4.3.3 *Prendre en compte le degré des tags.*



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.6 – Étude du phénomène de saturation à partir de modèles qui prennent en compte ou non le degré des tags.

Dans la perspective de proposer un modèle encore plus réaliste, nous allons maintenant prendre en compte le degré des tags. Nous avons vu que les tags ont des volumes d'écoute très différents. Par exemple le *rock* a un volume 6 fois plus élevé que celui du *metal*. Ce qui signifie qu'il y a 6 fois plus de chance qu'un titre écouté soit du *rock* plutôt que du *metal*. Ce phénomène est d'autant plus important que l'on se souvient de la grande hétérogénéité des degrés (et volumes) des tags (voir figure 3.1). C'est ce que nous allons intégrer ici. Dans notre formalisme cela revient à donner des probabilités plus grandes à certaines entités. Nous avons fait cela à partir du théorème 8, puis dans le théorème 9. Nous allons donc regarder comment se comportent nos trois premières équations quand on intègre ce phénomène (nous l'avons nommé **variable**). Nous avons donc les deux équations suivantes : $\frac{1}{E(D_{f_2})} = (\sum_{k<n} \frac{a_k(1+a_k(m-1))}{m})^{-1}$ et $\frac{1}{E(D_{f_2})} = \left(\sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k(1 - \sum_{i<m} {v_i}^2)\right)\right)^{-1}$. Dans les deux cas il faut déterminer la probabilité de tomber sur une entité. Fixons un tag t, notons d son degré et v son volume, notons aussi s_d la somme des degrés de tous les tags et s_v la somme des volumes de tous les tags. Pour l'équation de base, on imagine que les liens sont mis directement entre les utilisateurs et les tags. La probabilité (a) d'être associé à une entité représentée par le tag t est donc $a = v/s_v$. Pour les deux équations précédentes, on associe les nœuds de la couche 2 avec ceux de la couche 3 donc la probabilité d'associer à une entité la valeur $a = d/s_d$. Prendre en compte le degré ou le volume des tags est une bonne manière de prendre en compte ce que

nous pouvons appeler la **popularité** des tags. La courbe 4.6 représente donc l'évolution de ces six équations en fonction du degré. Nous voyons qu'il y a une nette différence entre les trois courbes prenant en compte la popularité des tags et les trois autres, les modèles prenant en compte la popularité. En effet les équations prenant en compte la popularité des tags sont beaucoup plus proches de la diversité observée que les autres. Ce qui montre que la popularité des tags influence plus fortement nos indicateurs de diversité que les autres éléments considérés dans les modèles jusqu'à présent. Notons que la popularité des tags est déjà une conséquence de la manière dont les utilisateurs écoutent leurs musiques. Mais il est intéressant de séparer les phénomènes pour une meilleure compréhension.

Regardons, comme précédemment la limite de nos équations :

— l'équation basique :

$$\frac{1}{E(D_{f_2})} = (\sum_{k<n} \frac{a_k(1+a_k(m-1))}{m})^{-1} \overrightarrow{\sum_{i<m} v_i^2 \to 0} \left(\sum_{k<n} a_k^2\right)^{-1}.$$

— Les équations pour la couche $2-3$:

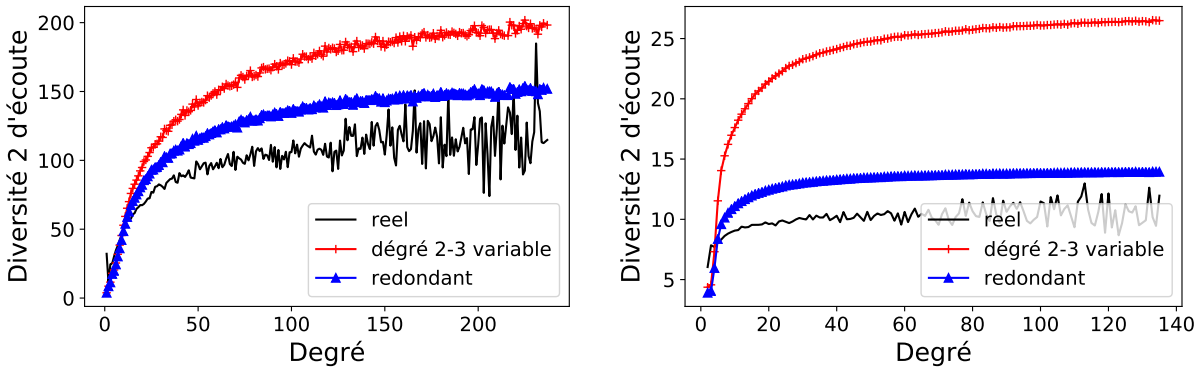
$$\frac{1}{E(D_{f_2})} = \left(\sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k(1-\sum_{i<m} {v_i}^2)\right)\right)^{-1} \overrightarrow{\sum_{i<m} v_i^2 \to 0} \left(\sum_{k<n} a_k^2\right)^{-1}.$$

Notons que cette limite est la 2-diversité des degrés (ou volumes) des tags. Sur nos jeux de données elle vaut approximativement 27 sur *Amazon* et 215 sur *MSD*. Comme précédemment nos limites sont un peu plus faibles puisque $\sum_{i<m} v_i^2$ tend vers une limite plus faible que 1. Puisque nos trois nouvelles courbes sont très proches, nous allons simplement en garder une pour la suite. Prenons la pondération $2-3$ variable (avec dépendance du degré).

Notons aussi que dans le cas de la génération qui conserve le degré nous avons un modèle très semblable au configuration model (sur les parties $2-3$). Il y a cependant une différence, certes les degrés entre la couche 2 et la couche 3 sont fixés (donc exactement les mêmes que sur le graphe de départ), mais les degrés de la couche 3 ne le sont pas. Ils restent proches du degré du graphe départ et pour un tag fixé, son degré moyen sur tous les graphes générés est égal à son degré sur le graphe de départ. Nous avons aussi vérifié que la courbe représentant le configuration model était très proche de la courbe de l'équation $2-3$ degré variable.

4.3.4 *Introduire de la redondance.*

Maintenant que la diversité attendue dans nos modèles est proche de celle observée en pratique,nous allons ajouter un dernier phénomène. Comme dans la sous section 4.2.5, nous allons considérer que si une valeur est associée une entité, une autre valeur aura plus de chance de l'être elle aussi à cette entité. Pour notre exemple, nous allons considérer qu'une personne qui



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.7 – Étude du phénomène de saturation à partir du modèle contenant de la redondance

a écouté du rock a plus de chance de réécouter du rock qu'autre chose. Nous avons vu grâce au théorème 10 que nous avons l'équation suivante :

$$\frac{1}{E(D_{f_2})} = \left(\sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + \beta_k(1 - \sum_{i<m} {v_i}^2)\right)\right)^{-1}.$$

Pour chaque indice k représentant un tag t, la valeur de β_k est la probabilité de réécoute de ce tag. Il y a plusieurs façons de considérer cette probabilité, nous avons choisi de la calculer globalement : Nous regardons le nombre n_0 de chemins issus du méta-chemin $3 \to 2 \to 1 \to 2 \to 3$ qui partent de t, parmi ceux là il y en a n_1 qui reviennent vers t (sans compter le retour par le même chemin). Ainsi pour ce tag $\beta_k = \frac{n_1}{n_0}$. Sur nos données, nous ajoutons une nouvelle équation à notre étude de la saturation en regardant la figure 4.7. Nous voyons que notre nouvelle courbe est encore plus proche de l'équation observée. Nous regardons encore une fois la limite de cette équation

$$\frac{1}{E(D_{f_2})} = \left(\sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + \beta_k(1 - \sum_{i<m} {v_i}^2)\right)\right)^{-1} \xrightarrow[\sum_{i<m} v_i^2 \to 0]{} \left(\sum_{k<n} a_k \beta_k\right)^{-1}.$$

Sur nos deux jeux de données, cette somme est égale à 164 sur *MSD* et 14 sur *Amazon*. Comme précédemment ces limites sont un peu plus hautes que les limites des courbes puisque $\sum_{i<m} v_i^2$ ne tend pas exactement vers 0.

4.3.5 *Influence du degré des utilisateurs dans la diversité.*

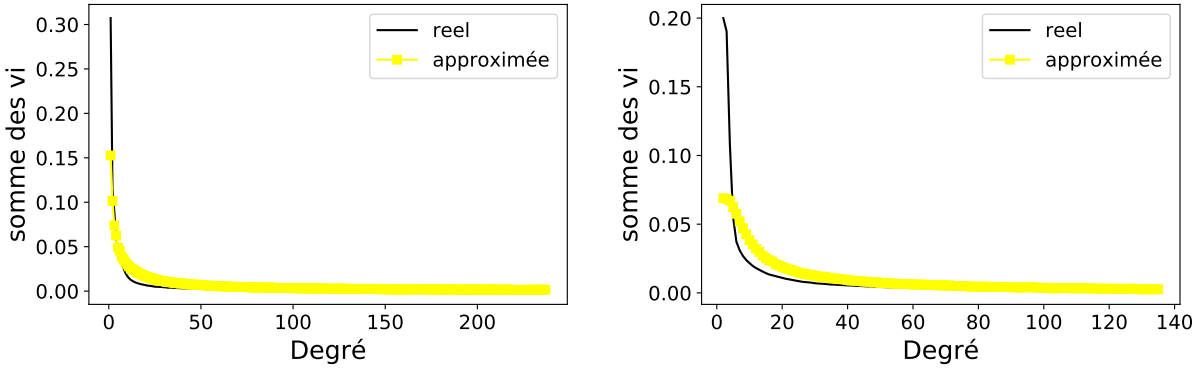
Intéressons nous maintenant au terme $\sum_{i<m} v_i^2$. C'est le seul terme qui dépend encore du nœud de départ. Comme il ressemble à un terme que l'on aurait pu obtenir avec nos indicateur il

rentre aussi dans notre formalisme. Nous pouvons donc calculer son espérance. Notons ce terme D pour ne pas le confondre avec l'équation précédente.

Il y a plusieurs choix à faire par rapport à ceux que l'on a fait précédemment, que nous n'allons pas détailler. Mais nous avons une équation nous donnant une approximation. Prenons le cadre du théorème 9 :

$$E(D) = \sum_{k<n} a_k \left(\sum_{i<m} {v_i}^2 + a_k(1 - \sum_{i<m} {v_i}^2) \right).$$

Ici notre équation n'est pas exactement la bonne puisqu'on n'a pas un ensemble de valeur fixes. Cependant on peut quand même arriver à approximer le $E(D)$ avec les degrés moyens des différentes couches et le degré de l'utilisateur de départ. Nous avons tracé cette approximation



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.8 – Comparaison entre la somme des v_i^2 et son approximation.

en jaune (et en point) par rapport à D : la somme observée des v_i^2 sur la figure 4.8. Nous voyons que notre courbe colle bien avec le D observé. Ce qui nous motive à remplacer D par son espérance dans l'équation de redondance, nous appelons cette nouvelle courbe **redondant approximé** et la traçons dans la figure 4.9. Il n'y a pas une très grosse différence entre les deux courbes.

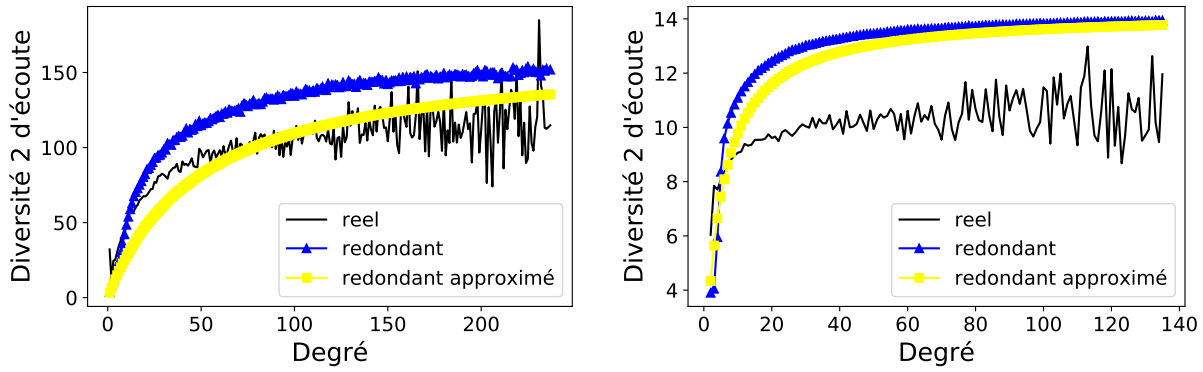
4.3.6 *Expliquer la diversité.*

Grace au travail de cette section, nous avons pu identifier différents paramètres de nos équations qui influence fortement les score de diversité d'écoute des utilisateurs. Ces paramètres peuvent être vus comme des ingrédients qui expliquent la (non) diversité des utilisateurs. Nous en avons identifié trois :

— les a_k : exprimant la popularité des tags,

— le degré de l'utilisateur (visible dans D),

— les β_k exprimant la redondance d'écoute des utilisateurs en fonction du tag.



(a) *Million Song Dataset* **(b)** *Amazon Dataset*

FIGURE 4.9 – Étude du phénomène de saturation à partir de modèle redondant approximé ou non

4.4 GÉNÉRATION DE GRAPHES ALÉATOIRES À PARTIR DE LOIS DE DISTRIBUTION DONNÉES.

Pour terminer ce chapitre dédié aux modèles, nous allons approfondir l'analyse des indicateurs proposés en nous éloigant un temps des données empiriques pour considérer des modèles reproduisant des propriétés fixées (et non des structures relationnelles fixées comme auparavant). Plus précisément, nous allons nous interroger sur le poids de propriétés structurelles des graphes dans la mesure de la diversité. Pour ce faire, nous allons considérer plusieurs propriétés liées à la distribution des degrés (uniformes, hétérogènes, homogènes, ...) et analyser la diversité de graphes respectant ces propriétés.

4.4.1 *Un formalisme et un algorithme pour construire des graphes aléatoires.*

Dans la science des réseaux, et particulièrement dans notre thèse, la distribution des degrés des nœuds est une propriété qui s'avère fondamentale. En effet connaître la distribution de degré des nœuds d'un graphe nous donne beaucoup d'informations sur sa structure et nous avons vu, tout au long de notre étude, à quel point le comportement de nos indicateurs est fortement dépendant de cette propriété. Nous avons donc décidé de générer des graphes à partir de distributions de degrés pré-déterminées. Plus précisément, et étant donné quatre distributions de degré ($d_{1,2}, d_{2,1}, d_{2,3}$ et $d_{3,2}$), il s'agit d'étudier la diversité issue d'un graphe triparti choisi aléatoirement uniformément parmi l'ensemble des graphes tripartis respectant ces distributions.

Afin de limiter le champ d'analyse nous allons nous concentrer sur l'une des diversités que nous avons déjà considérée précédemment, à savoir la 2-diversité mesurer à partir du méta

chemin $1 \to 2 \to 3$ et nous allons étudier la moyenne de cette diversité sur l'ensemble des graphes générés.

4.4.1.1 *Construire un graphe à partir de distributions de degrés.*

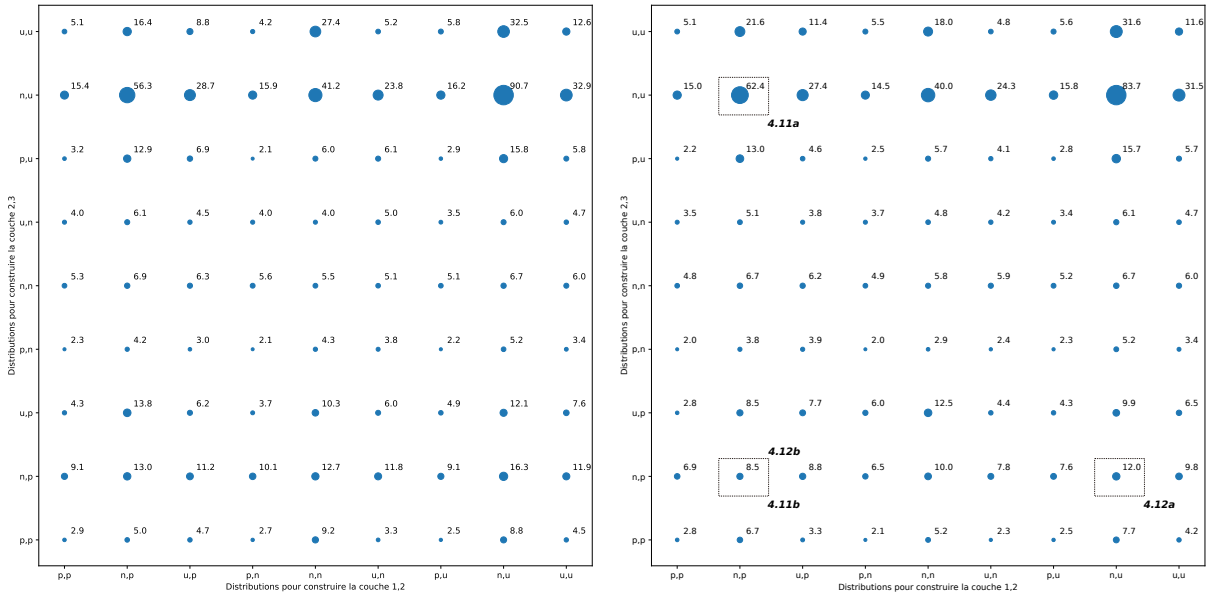
Pour créer ces graphes nous construisons séparément et de la même manière deux graphes bipartis (B_1 celui entre les couche 1 et 2 et B_2 celui entre les couches 2 et 3) qui formeront notre graphe triparti. Prenons deux distributions de degrés x (utilisées pour fixer les degrés $d_{1,2}$) et y (utilisée pour fixer les degré $d_{2,1}$). Nous n'allons pas adopter le même comportement avec x et y. En effet nous allons fixer pour chaque nœud de la couche 1 son degré correspondant dans la distribution x, et puis pour chaque lien partant de ce nœud, nous allons fixer grâce à y une probabilité de se lier à chaque nœud de la couche 2. Ceci est une façon détournée d'imposer les degrés de la couche 2. En effet plus la probabilité de tomber sur un nœud est grande plus le degré de ce nœud sera grand. Mais il sera grand en espérance et non fixé à une valeur donnée. Nous choisissons ce paradigme pour plusieurs raisons. Premièrement il nous permet, de donner une chance aux nœuds de la couche 2 de ne pas être atteints (d'avoir un degré nul). Ceci nous permet d'être au plus proche des graphes observés, il arrive que des musiques ne soient pas écoutées par exemple. Cela nous demandera de traiter ces nœuds (impasses) à part, pour les supprimer de notre jeu de données, ce qui nous oblige à les pré-traiter comme sur des données classiques. De manière similaire cette façon de faire nous permet aussi de ne pas borner le degré des nœuds, ce qui empêche des effets de seuil. Deuxièmement il nous permet d'avoir plus de graphes possibles pour des distributions fixées en se donnant une liberté par rapport à la distribution y. Avoir plus de graphes possibles est une manière de stabiliser les moyennes que nous allons faire par la suite. Troisièmement, cette manière de faire permet de se rapprocher de la façon dont nos graphes analysés se construisent. En effet nous pouvons ici considérer que ce sont les nœuds de la couche 1 qui dirigent la façon dont se créent les liens. En pratique, nous avons par exemple les utilisateurs qui choisissent les musiques qu'ils écoutent. Par ailleurs, comme nous le verrons dans les exemples du chapitre suivant, ce sont les ONG qui décident des pays dans lesquels ils s'implantent, des émissions qui invitent des personnalités, des reviewers qui sélectionnent un objet etc.

Ainsi, formellement, pour ce graphe biparti nous fixons deux distributions x, y représentant respectivement la distribution de degré $d_{1,2}$ et la distribution de probabilité qui associe à chaque nœud de la couche 2 la probabilité qu'un lien partant de la couche 1 arrive sur ce nœud. Le principe décrit ci-dessus permet donc de générer un graphe biparti (B_1) respectant les distribution $d_{1,2}$ et $d_{2,1}$ et, de la même manière, un second graphe biparti (B_2) respectant les distributions $d_{2,3}$ et $d_{3,2}$. En fusionnant les deux graphes bipartis obenus, nous obtenons un graphe triparti respectant les quatre distributions. Pour résumer nous avons un algorithme qui, à partir de 4 distributions, nous construit un graphe choisi uniformément parmi les graphes qui respectent ces distributions.

4.4.1.2 *Caractériser les distributions : 3 familles possibles.*

Maintenant que le principe est posé, il nous reste à déterminer quelles distributions nous allons considérer pour les quatres distributions de degrés. En particulier, il nous faut choisir entre considérer une distribution homogène (tous les individus ont à peu près les mêmes proportions) ou une distribution hétérogène (il y a une grande diversité dans les proportions). En ce qui concerne les distributions homogènes, elles seront déterminées dans la suite par une loi normale (gaussienne) à deux paramètres, sa moyenne μ et sa variance σ. En ce qui concerne les distributions hétérogènes, elles seront elles définies par une loi de puissance (de paramètre a). Enfin, et dans le but de compléter ces deux cas extrêmes par un troisième type de distribution, nous allons considérer également par la suite une loi uniforme, simplement définie par ses bornes (a et b). Avec ces trois familles nous avons une base intéressante pour faire varier les propriétés des graphes aléatoires que nous allons analyser.

4.4.2 *Évaluer le score de diversité selon 4 types de distributions de degré.*



(a) (1000, 10000, 100000) **(b)** (100000, 10000, 1000)

FIGURE 4.10 – Moyenne de la diversité-2 (1,2,3) selon les distributions utilisées

Nous pouvons ainsi regarder comment la 2-diversité moyenne $1, 2, 3$ se comporte selon les 4 distributions que nous fixons. Fixons d'abord des paramètres arbitraires pour les 3 familles :

— $a = -1.2$ pour nos lois de puissance,

— $\mu = 10$ et $\sigma = 1$ pour nos lois normales,

— $a = 1$ et $b = 10$ pour nos lois uniformes.

Ainsi nous avons trois lois possibles pour chacune des 4 distributions ce qui nous donne $3^4 = 81$ possibilités. Il nous reste à fixer la taille des couches c'est-à-dire un triplet de trois nombres entiers n_1, n_2, n_3 définissant respectivement le nombre de nœuds des couches 1, 2 et 3 . Nous avons tracé ces 81 points sur la figure 4.10 pour deux triplets de tailles $(1000, 10000, 100000)$ (figure 4.10a) et $(100000, 10000, 1000)$ (figure 4.10b). En abscisse nous avons les noms des familles des deux distributions permettant de construire le biparti B_1 et en ordonnée ceux qui permettent de construire le biparti B_2. *"p"* représente la loi de puissance, *"n"* représente la loi normale et *"u"* représente la loi uniforme. Les points encadrés seront analysés dans la section suivante. La taille des points (donnée explicitement par le nombre étiqueté pour chaque point) est la 2-diversité moyenne suivant le chemin $1 \to 2 \to 3$.

Nous remarquons tout d'abord qu'il a très peu de différence selon l'ordre du triplet n_1, n_2, n_3.Nous avons analysés les six permutations possible (non montré ici) et les résultats sont similaires. Même en divisant par 10 ou 100 la taille de toutes les couches nous n'avons pas vu de changement dans les résultats (seulement dans les ordres de grandeurs).

Ensuite nous pouvons voir que les lignes sont plus constantes que les colonnes. Ce qui signifie que l'influence du biparti B_2 sur notre score de diversité est plus forte que celle du biparti B_1. Ceci confirme ce que nous avons pu remarquer sur nos jeux de donnés : la structure de la couche d'arrivée est plus influente que celle de la couche de départ.

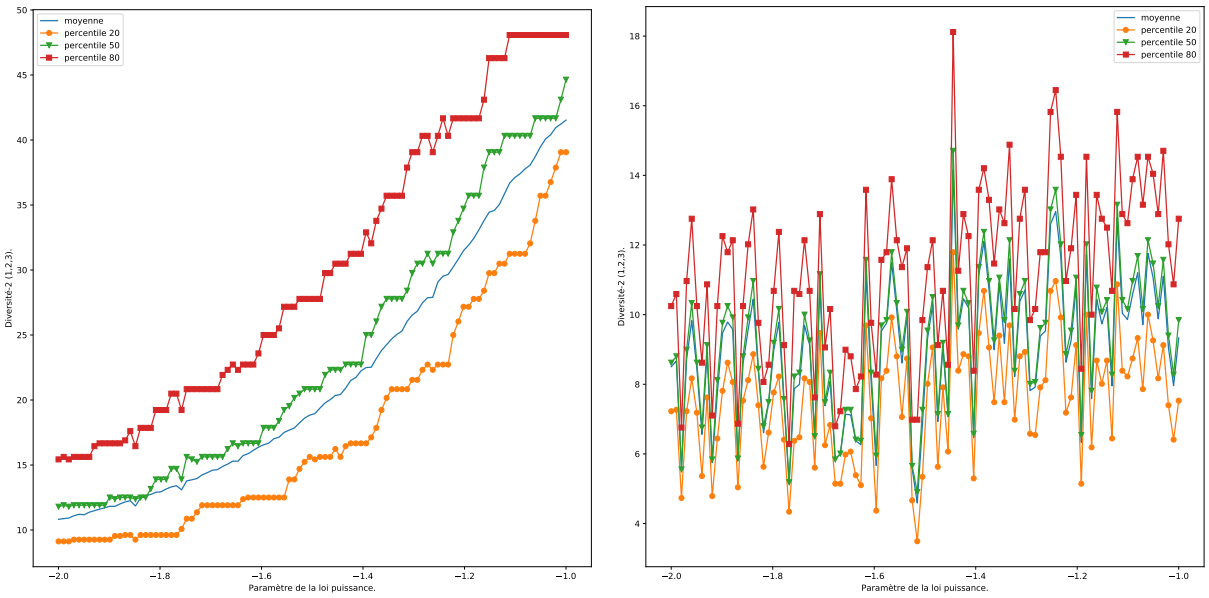
Regardons maintenant les différences entre les familles. Sans surprise plus nos distributions de degré sont homogènes, plus nos scores de diversité sont élevés. Notons que les lois uniformes et normales jouent un rôle symétrique en terme d'homogénéité selon qu'elles soient première ou deuxième du couple. En effet pour que la distribution de degré soit très homogène il faut qu'elle suive une loi normale. Mais pour que la probabilité de liens amène à une distribution de degré homogène il faut que la probabilité de liens suive une loi uniforme, puisqu'il faut que chaque lien ait la même probabilité d'arriver sur chacun des nœuds. Ainsi les lignes et colonnes étiquetées n, u sont celles qui génèrent des graphes de plus forte diversité moyenne. À l'inverse les lignes et colonnes étiquetées p, p ou p, n correspondent à des graphes de faible diversité moyenne.

4.4.3 *Faire évoluer un paramètre avec des familles fixées.*

Analysons maintenant certains de ces points en particulier en faisant varier un des paramètres. Regardons quels points pourraient nous intéresser. Nous avons vu dans les sections précédentes que le degré et le volume des nœuds la couche de départ de notre chemin influençaient fortement nos indicateurs utilisant ce chemin. Nous pouvons dans notre cas le contrôler en fixant les degrés constants $1, 2$, et $2, 3$. Ainsi chaque nœud de la couche 1 aura le même degré $(1, 2)$ et le même volume $(1, 2, 3)$. Pour notre étude nous allons fixer le degré $d_{1,2}$ à 10 et le degré $d_{2,3}$ à 5. Ainsi nous n'étudierons que des nœuds de départ (couche 1) qui auront un degré de 10 et un volume de 50.

Notons que fixer ces degrés constant à une valeur c revient dans notre formalise à fixer deux distributions normales avec une moyenne égale à c et une variance très faible. Nous avons donc

déjà fixé deux distributions (la première et la troisième). Il nous reste à fixer les deux autres, dont l'une des deux variera selon un paramètre. Une bonne façon de faire varier l'homogénéité d'une distribution est de faire varier le paramètre a de la loi de puissance. Nous ferons varier ce paramètre entre -2 (très hétérogène) et -1 (plus homogène). Pour la dernière distribution nous allons fixer aussi deux cas possibles : un cas parfaitement homogène avec une distribution uniforme et un cas hétérogène avec une loi de puissance (de paramètre $a = -1.2$). Nous avons donc pu tracer ces deux cas pour deux cas différents : quand le paramètre varie pour construire le biparti B_1 (figure 4.11) et quand il varie pour construire le biparti B_2 (figure 4.12). Nous avons encadré dans la figure 4.10b les points qui correspondaient à nos 4 figures. La figure 4.11b et la figure 4.12b correspondent au même point puisqu'il y a deux lois de puissance, nous faisons varier le paramètre du biparti B_1 pour la première et du biparti B_2 pour la seconde. Nous avons fait cela sur la figure 4.10b puisque nous avons pris les mêmes tailles de couches (100000, 10000, 1000), en se basant sur les rapports entre les ordres de grandeurs sur *MSD* ou *Amazon*. Sur chacune de ces figures nous avons tracé la diversité moyenne des nœuds en bleu (trait plein) et en plus, pour regarder la façon dont ces scores sont répartis entre les individus, nous avons tracé différents percentiles (le percentile 20 en ronds orange, la médiane en triangles verts et le percentile 20 en carrés rouges).

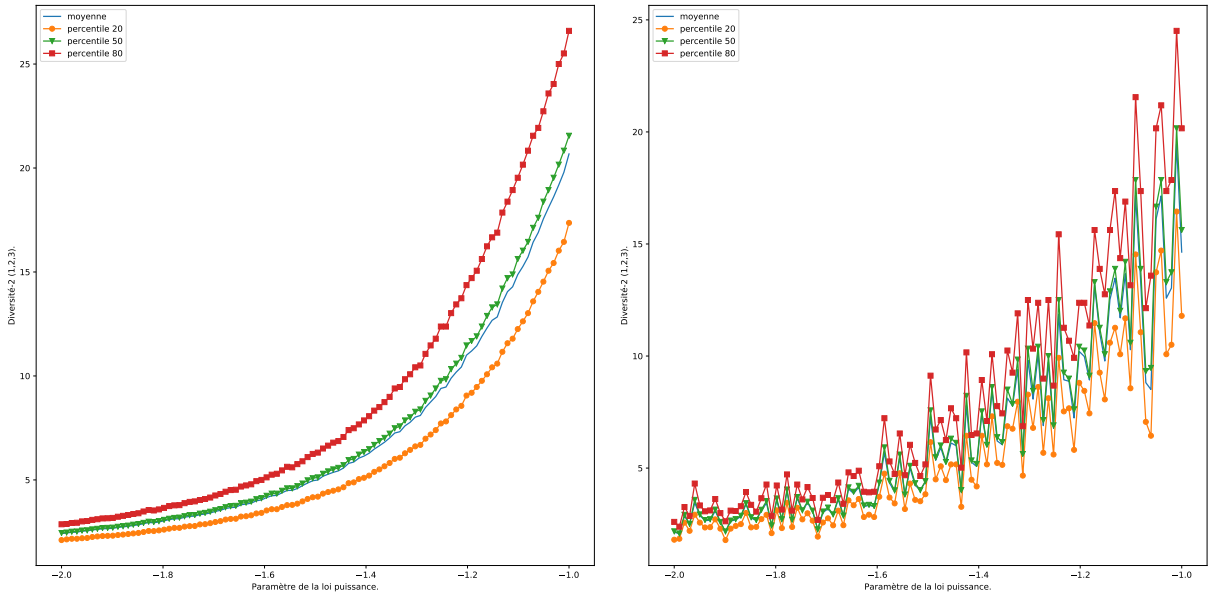


(a) Avec une affiliation $3, 2$ uniforme

(b) Avec une affiliation $3, 2$ en loi de puissance

FIGURE 4.11 – Différents percentiles de la diversité-2 (1,2,3) en fonction de la façon de construire la couche $1, 2$

Pour y voir plus clair, prenons le vocabulaire de l'exemple de la base de données *MSD* : des utilisateurs écoutent des musiques qui sont taguées. Nous fixons le nombre de musiques écoutées par les utilisateurs à 10 et le nombre de tags qui caractérisent une musique à 5. Et nous étudions comment l'hétérogénéité dans la façon dont les musiques et les tags sont sélectionnés influence la diversité d'écoute des tags.



(a) Avec une affiliation $2,1$ uniforme

(b) Avec une affiliation $2,1$ en loi de puissance

FIGURE 4.12 – Différents percentiles de la diversité-2 (1,2,3) en fonction de la façon de construire la couche $2,3$

La figure 4.11a nous montre qu'un utilisateur peut augmenter sa diversité d'écoute en homogénéisant sa façon de sélectionner sa musique : c'est-à-dire en mieux répartissant ses choix de musiques. Ceci est le cas quand la façon dont les musiques sont taguées est uniforme. En effet la figure 4.11b nous montre que lorsque les musiques sont taguées de façon hétérogène, le fait de choisir un panel de musiques plus hétérogène n'améliore que peu la diversité d'écoute. Ceci est visible puisque d'abord l'échelle est plus faible, ensuite la courbe est beaucoup plus variable et enfin l'effet d'augmentation de la diversité par rapport au paramètre a vu dans la figure 4.11a n'est que très peu visible ici. La figure 4.12a nous montre qu'on peut augmenter la diversité d'écoute des utilisateurs en homogénéisant la façon d'associer des tags aux musiques. Cette évolution est encore plus lisse que sur la figure 4.11a, les percentiles étant aussi plus centrés sur la moyenne. Cette évolution est rendue plus difficile si la façon dont les utilisateurs sélectionnent leur musique est hétérogène, comme on le voit sur la figure 4.12b.

Pour résumer, nous avons regardé à quel point il était possible d'améliorer le score de diversité d'un individu en améliorant l'homogénéité globale du système ou alors l'homogénéité de cet individu. De plus nous avons encore constaté que nos score de diversités étaient plus fortement influencés par la structure du deuxième biparti plutôt que par celle du premier.

4.5 CONCLUSION

Cette section consacrée aux modèles a été l'occasion de montrer L'intérêt de la modélisation pour trois perspectives. Tout d'abord, elle permet de normaliser les indices bruts de diversités (section 4.1) et donc de permettre une meilleure interprétation de leur valeur relative. Ensuite, nous avons vu que les résultats analytiques issus des différents modèles proposés permettaient de fournir des indications sur les observations faites sur les jeux de données empiriques (section 4.2 et section 4.3). Enfin, en dépassant le contexte purement empirique et en considérer des lois de probabilités caractéristiques en lien avec la génération aléatoire de graphe, il a été possible d'approfondir la compréhension des propriétés structurelles qui influencent le plus la diversité dans les graphes tripartis (section 4.4). Nous avons ainsi pu montrer empiriquement que trois phénomènes en particulier permettent de reproduire la (non) diveristé observée sur le comportement des utilisateurs : la popularité hétérogène des tags, la redondance dans l'écoute des titres musicaux et le degré des utilisateurs. Finalement avec trois phénomènes (popularité des tags, redondance d'écoute, degré des utilisateurs) nous arrivons à coller au comportement moyen des utilisateurs. Enfin, dans le but de sortir du cadre purement empirique, nous avons proposé l'étude de graphes générés suivants des lois de distribution de degrés en considérant trois lois caractéristique (uniforme, homogène, hétérogène) et avons ainsi précisé l'impact de ces caractéristique sur les diversités observées.

À noter que ce chapitre est en partie publié dans le journal Information Processing & Management (version disponible ici [73]).

Chapitre

5

Au delà de l'analyse statique des graphes tripartis.

Les trois précédents chapitres nous ont permis de détailler le cœur de notre étude : ce que notre méthode est capable d'apporter lorsque nous avons des données sous forme de graphe triparti statique. Dans ce chapitre nous allons dépasser cette vision pour montrer ce que nos méthodes peuvent apporter dans un cadre plus général ou dans des cadres particuliers. Cette partie nous permet aussi de regarder d'autres jeux de données, où les types d'entités sont différentes et de voir comment peuvent s'interpréter nos indicateurs dans des cas concrets.

Tout d'abord nous allons regarder comment la dimension temporelle ainsi que la diversité relative s'analysent sur un jeu de données concernant les personnalités invitées dans les médias et à la radio (section 5.1). Puis nous regarderons sur un jeu de données représentant l'implantation des ONG dans le monde, ce que nos indicateurs peuvent apporter quand le nombre de couches est plus grand que 3 (section 5.2).

5.1 Diversité temporelle et relative.

Nous allons étudier la diversité des personnes invitées à la radio et à la télévision. Ce contexte va nous permettre d'illustrer comment prendre en compte l'aspect temporel dans l'utilisation de nos indicateurs. En effet, ici, en plus de savoir quelle personne a été invitée sur quelle chaîne, nous savons aussi à quel moment a eu lieu l'émission. Ce qui nous permet d'analyser comment évoluent les abondances proportionnelles au cours du temps, et bien entendu la diversité.

Nous allons d'abord présenter les données dans la section 5.1.1 et montrer comment nous définissons un graphe 4-parti à partir de celles-ci. Puis nous analyserons dans les sections 5.1.2 et 5.1.3 la diversité relativisée. Ensuite nous étudierons dans la section 5.1.4 ce que nous apporte la dimension temporelle. Enfin dans la section 5.1.5 nous montrerons en quoi la dimension temporelle apporte de nouveaux points de vue sur la diversité relativisée. Notons que

sur cette section nous abandonnons la diversité-2 pour se focaliser sur la 1-diversité. Ceci est motivé par le fait que la diversité relative est à l'origine liée à la divergence qui est elle-même fortement liée à l'entropie de Shannon [1].

5.1.1 *Présentation des données.*

Le jeu de données considéré dans cette section est constitué de $3\ 660\ 024$ n-upplets décrivant chacun la présence d'un.e invité.e dans une émission de télévision ou de radio. Ce jeu de données est collecté depuis 2012 par l'Institut National de l'Audiovisuel [2]. Chaque ligne décrit l'invitation d'un invité à la télévision ou à la radio. Ces n-upplest contiennent 34 informations précisant en particulier la date, l'heure, l'invité, l'émission, la chaîne, ainsi que des descriptions des invités, de l'émission, de l'animateur... Nous n'avons gardé que 5 de ces informations : la date, la chaîne, l'émission, le nom de l'invité et ce qui est appelé les *qualités* de l'invité. Ces qualités décrivent pourquoi il est invité dans une émission de télévision ce qui en général est résumé par son métier (journaliste, violoncelliste, homme politique etc). Ainsi nous avons pu construire un graphe 4-parti composé de 471 qualités (couche 1), $92\ 765$ invités (couche 2), 6408 émissions (couche 3) et 118 chaînes (couche 4).

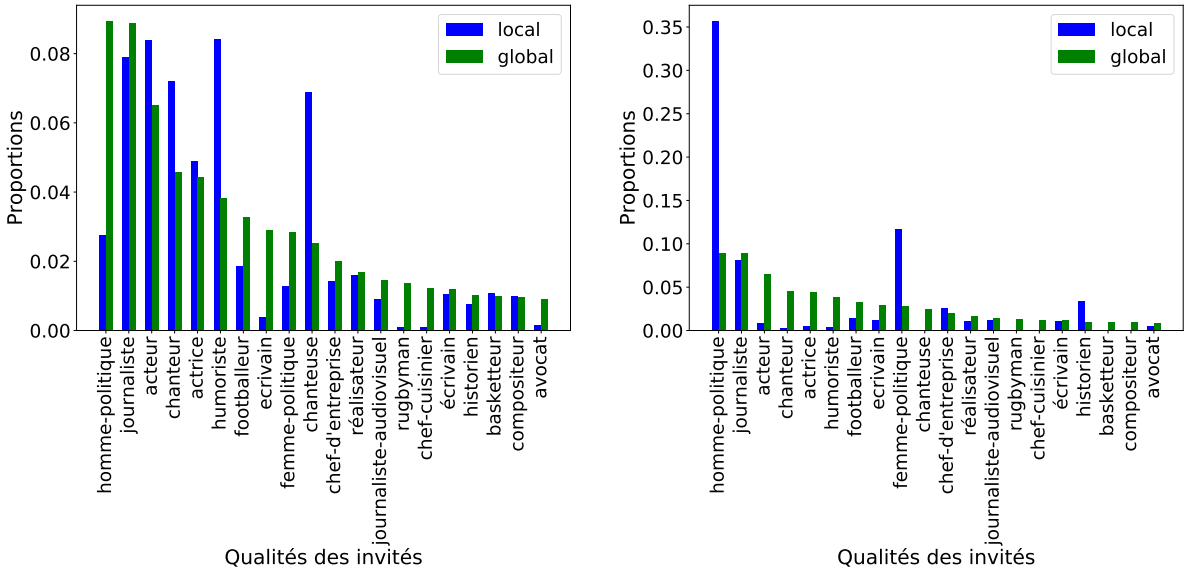
5.1.2 *Utilisation de la diversité relative.*

Dans le chapitre précédent, nous avons regardé comment normaliser la diversité pour pouvoir comparer la diversité de nœuds ayant des degrés différents. Nous avons alors mis en valeur l'impact du degré sur la diversité. Ici nous allons mettre en valeur l'impact de notre catégorisation sur la diversité. Plus précisément : nous avons, jusque là, considéré qu'avoir une diversité parfaite signifiait être connecté à tous les nœuds d'arrivée de la même manière. Dans notre exemple de consommation musicale, avoir une diversité parfaite revenait à consommer exactement autant de *rock*, *rap*, *techno* etc. Or que se passe-t-il s'il y a beaucoup plus de *rock* que de *rap* proposées à la consommation ? N'y a-t-il pas une consommation que l'on pourrait décrire comme plus diversifiée mais qui n'est pas totalement uniforme ? C'est dans cette optique que nous allons étudier ce jeu de données. Nous allons nous intéresser pour le manuscrit à la diversité $4 \to 3 \to 2 \to 1$, c'est-à-dire la diversité d'une chaîne en fonction de la diversité de *qualité* de ses invités. Pour cela intéressons-nous aux abondances proportionnelles obtenues en suivant ce chemin. Les figures 5.1a et 5.1b nous montrent respectivement celles que nous avons obtenues pour *france 4* et *public sénat* [3]. En plus de chaque proportion en "bleu" nous avons représenté à sa droite (et en vert) la proportion qu'on obtient en partant de la totalité des chaînes (c'est à dire l'abondance proportionnelle collective). Nous pouvons voir par exemple

1. Nous avons vu que la 1-diversité et la 2-diversité sont fortement corrélées, tous les calculs et les courbes auraient pu être fait sur cet indicateur et sur les autres.

2. L'article suivant utilise ce jeu de données (`https://www.inaglobal.fr/television/article/invites-des-talk-shows-et-emissions-de-divertissement-tous-les-memes-9796`)

3. Nous n'avons affiché que les 15 plus grandes proportions (touchées globalement) par souci de lisibilité



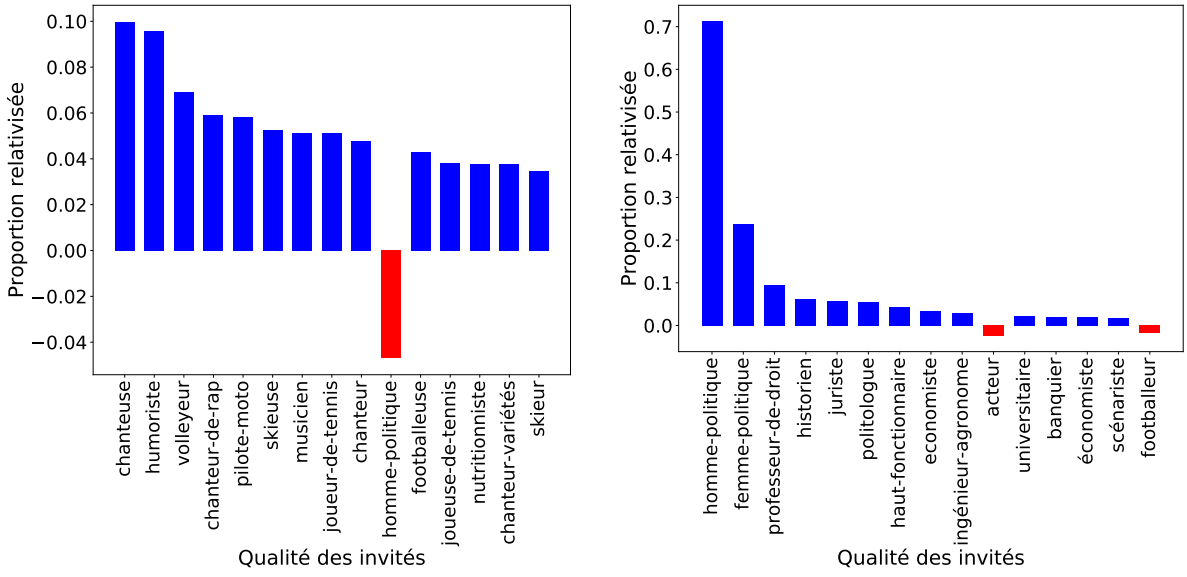
(a) Sur france 4

(b) Sur public sénat

FIGURE 5.1 – Distribution d'abondances proportionnelles pour deux chaînes de télévision

que si, dans le cas des deux chaînes considérées, la grande majorité des invités sont des journalistes, cette impression est relativisée par le fait que la proportion de journalistes est en réalité plus faible que celle des journalistes invités dans l'ensemble des émissions. Nous pouvons voir aussi que la proportion d'hommes politiques est forte chez les deux, mais beaucoup plus faible que la globalité pour *france 4* et beaucoup plus forte que la globalité pour *public sénat*. Notre but est donc de regarder comment se comporte une distribution de proportion abondante d'une chaîne par rapport à la distribution globale. Notons qu'ici la distribution globale est celle que l'on utilise pour la diversité collective. Pour cela nous allons utiliser la diversité relative (définition 18). Cette diversité vient de la divergence de Kullback-Leibler. Pour comprendre comment se comporte notre mesure de diversité, regardons son impact sur les abondances proportionnelles. Pour une chaîne notons $< p >$ son abondance proportionnelle et $< q >$ celle de l'ensemble des chaînes. Chaque qualité d'invité i est donc associée à deux proportions : p_i à partir de la chaîne fixée et q_i à partir de tout le monde. La 1-diversité relativisée de notre chaîne est donnée : $-2^{\sum_i p_i * log(p_i/q_i)}$.

Nous pouvons donc regarder ce qu'il se passe si nous regardons la distribution des $< p_i * log(p_i/q_i) >$. Nous avons regardé ces deux distributions pour les deux mêmes chaînes *france 4* (figure 5.2a) et *public sénat* (figure 5.2b). Ainsi lorsque $p_i > q_i$, c'est-à-dire que la proportion de la tendance i est plus grande que la proportion globale de la tendance i, la proportion $p_i * log(p_i/q_i)$ est positive, à l'inverse, lorsque $p_i < q_i$, cette proportion est négative. Ainsi une proportion positive (en bleu) indique une sur-représentation de la tendance sur cette chaîne, et une proportion négative (en rouge) indique une sous-représentation. Ceci nous permet de mieux visualiser les différences entre chaque chaîne et la globalité. Les *hommes politiques*



(a) Sur france 4

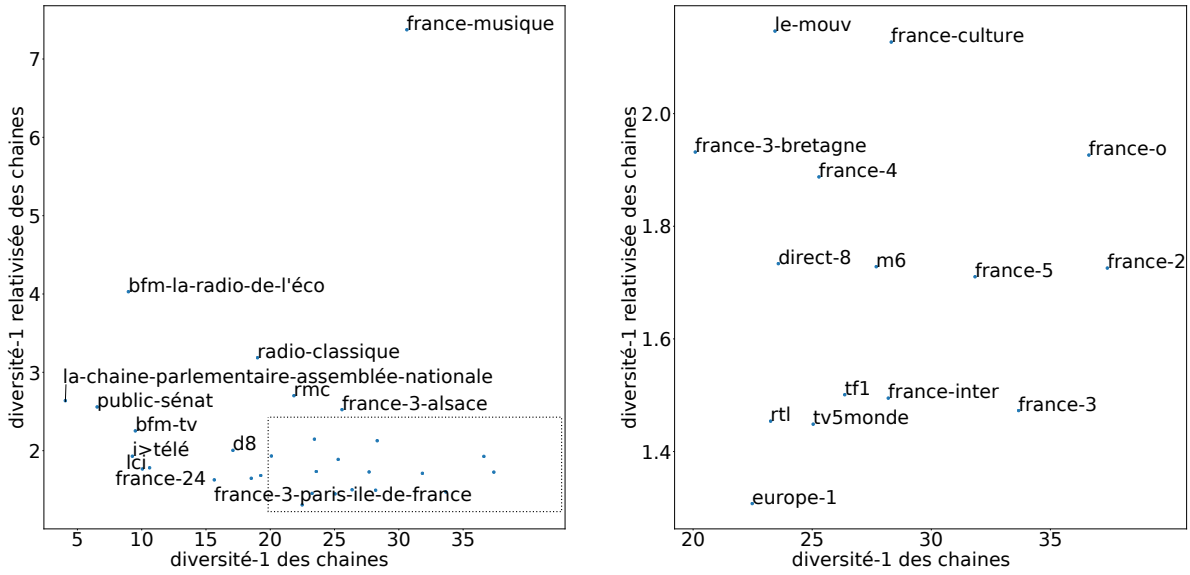
(b) Sur public sénat

FIGURE 5.2 – Distribution relativisée pour deux chaînes de télévision

sont sous-représentés sur *france 4* et sur-représentés sur *public sénat* et on a le phénomène inverse pour les *acteurs*. Notons que l'échelle sur la courbe représentant *public sénat* est plus grande car les hommes et les femmes politiques y sont sur représentés.

Nous avons analysé ce que donne la 1-diversité relativisée dans ce contexte. La figure 5.3 montre la 1-diversité relativisée par rapport à la 1-diversité. Comme il y avait une forte concentration de points dans le carré hachuré (1-diversité relativisée inférieure à 2.5 et 1-diversité supérieure à 20) nous avons fait un zoom sur cette partie (figure 5.3b). Nous pouvons remarquer que les chaînes qui se démarquent par une grande diversité relativisée sont celles qui invitent des personnalités de qualités différentes de celles invitées par la globalité des chaînes. Par exemple la chaîne de radio *bfm la radio de l'éco* obtient un bon score de diversité relativisée, ce que nous pouvons expliquer (en regardant les distributions relativisées) par le fait qu'elle invite beaucoup plus de *chefs d'entreprises* que la globalité. De même, comme nous l'avons vu précédemment *france 4* invite un milieu plus artistique que le reste (*chanteuse*, *chanteur*, *humoriste*) etc.

Notre diversité relative nous permet donc de voir en quoi une chaîne se distingue des autres en terme de qualité d'invité. Encore une fois ici, nos indicateurs (relativisés ou non) sont une bonne façon d'avoir une vision d'ensemble de phénomènes, que les distributions des abondantes proportionnelles permettent de détailler.



(a) Ensemble de la courbe.

(b) Zoom sur le rectangle en pointillé.

FIGURE 5.3 – 1-diversité relativisée par rapport à la 1-diversité des chaînes.

5.1.3 *Le cas particulier des invités politiques*

Les résultats de diversité relative sont fortement liés à la façon de catégoriser les entités (ici le lien entre un invité et sa qualité). En effet les choix faits pour la création de la base de donnée, influencent fortement nos indicateurs. Ici, par exemple, le choix a été de genrer les qualités (on distingue *chanteur* de *chanteuse*), mais aussi d'avoir une certaine précision dans la catégorisation (par exemple on distingue *pianiste* de *violoncelliste*). Si ces catégories étaient regroupées, nous aurions des comportements d'indicateurs différents. Dans la suite, nous allons donner un axe politique à nos données, car nous allons nous intéresser seulement aux hommes et femmes politiques, et nous allons remplacer la couche qualité, par une couche **tendance politique**.

5.1.3.1 *Présentation de la classification politique.*

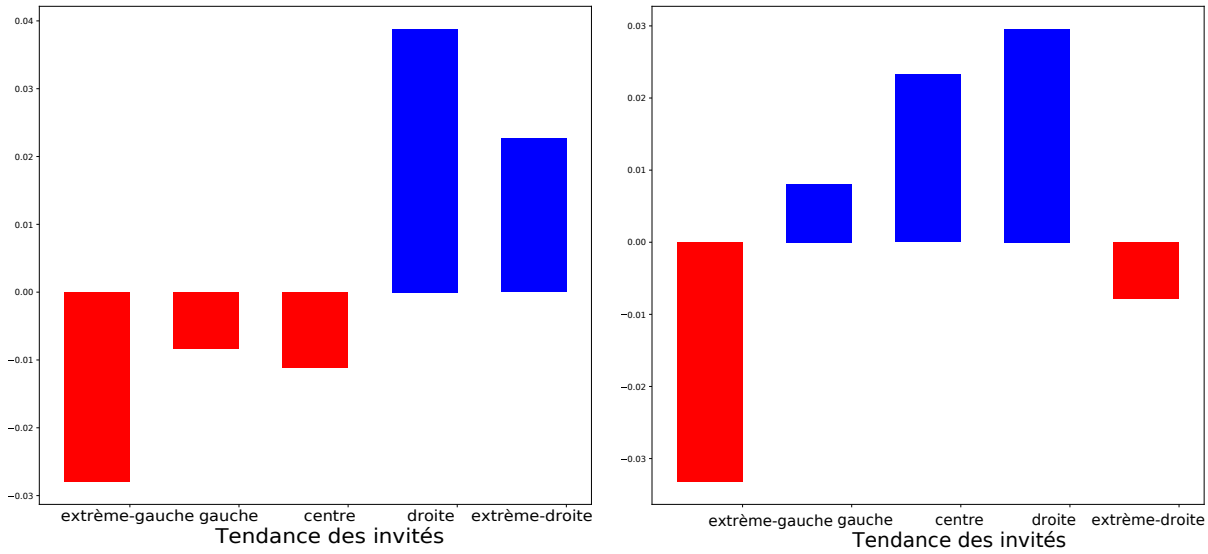
Nous voulons nous focaliser sur les hommes et femmes politiques pour regarder la couleur politique des invités. Pour cela nous avons sélectionné seulement les invités étant qualifiés comme homme ou femme politique et qui avaient été invités plus de 100 fois pendant notre période (c'est-à-dire que leur degré $2,3$ est supérieur à 100). Ensuite nous avons récupéré leur parti politique sur leur page Wikipédia. Sachant que certains politiques ont plusieurs partis, ou changent de parti pendant la période, nous avons sélectionné le dernier parti noté sur sa page. Puis nous avons assigné à chaque parti politique une tendance : L'extrême gauche rassemble : 'FI', 'Ensemble !', 'NPA', 'PRG', 'PCF' et'LO' ; la gauche : 'PS' et 'EELV' ; Le centre : 'MoDem', 'LREM', 'UDI', 'LRC - Cap21', 'UDF' et 'Résistons' La droite : 'LR' et 'UMP' ;

Enfin l'extrême droite est composée du 'FN' et de 'RN'. Ceci nous donne un nouveau graphe 4-parti avec 5 tendances (couche 1), 877 invités (couche 2), 3950 émissions (couche 3) et 101 chaînes (couche 4).

Il est clair que notre façon d'associer à un invité une tendance politique pourrait être travaillée avec une meilleure connaissance politique. Il est important de noter que cette catégorisation influence fortement nos indicateurs de diversité.

5.1.3.2 *Retour sur la diversité relative.*

Maintenant que nous avons présenté notre façon de construire une nouvelle couche 1, nous pouvons revenir sur notre travail de relativisation de la diversité. Nous avons regardé les distributions relativisées de i-télé (figure 5.4a) et lci (figure 5.4b). Ceci nous permet rapidement de voir des différences de profils, les invités d'i-télé ayant tendance à être plus de droite, tandis que les invités de lci ont tendance à être plus dans le centre de l'espace politique.

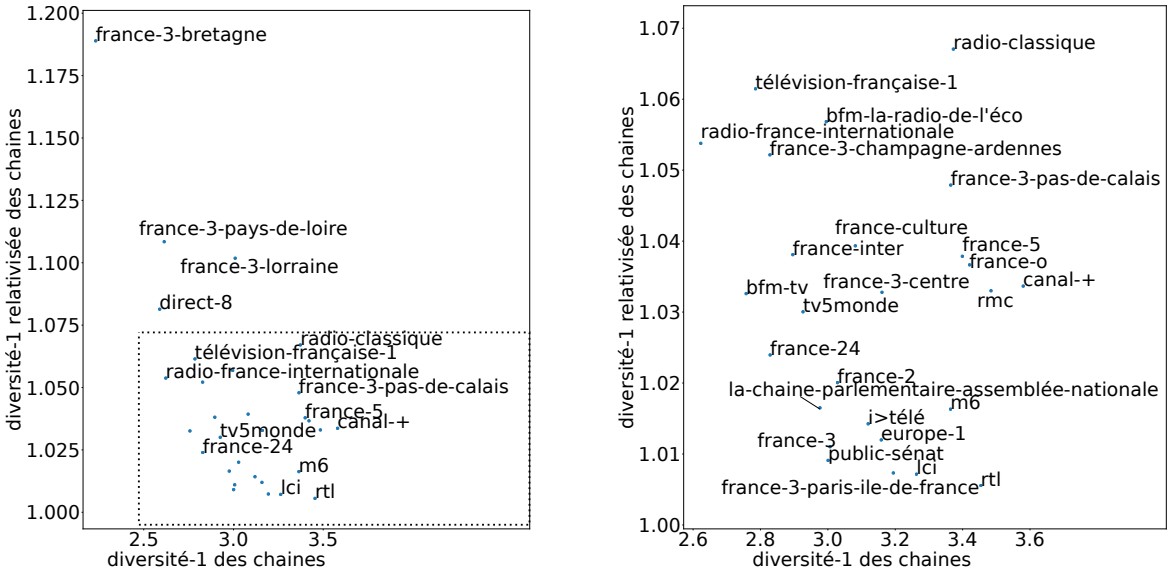


(a) Sur itélé.

(b) Sur lci.

FIGURE 5.4 – Distribution relativisée de deux chaînes de télévision.

Nous pouvons ainsi afficher pour les grandes chaînes leur diversité 1-relativisée par rapport à leur 1-diversité classique (figure 5.5a). Ici aussi nous avons extrait la partie inférieure de cette figure (diversité relativisée inférieure à 1.075) pour les tracer sur la figure 5.5b. Ainsi on observe que *France 3 Bretagne* a une 1-diversité plus faible mais qu'elle s'éloigne beaucoup de la façon d'inviter globale, alors que *rtl* a une diversité assez forte, mais très proche du comportement global des chaînes. Notons, grâce à cet exemple, que deux scores de diversité relativisée peuvent être espérés selon que l'on veuille avoir une politique d'invitation proche ou éloignée de celle de la totalité des chaînes.



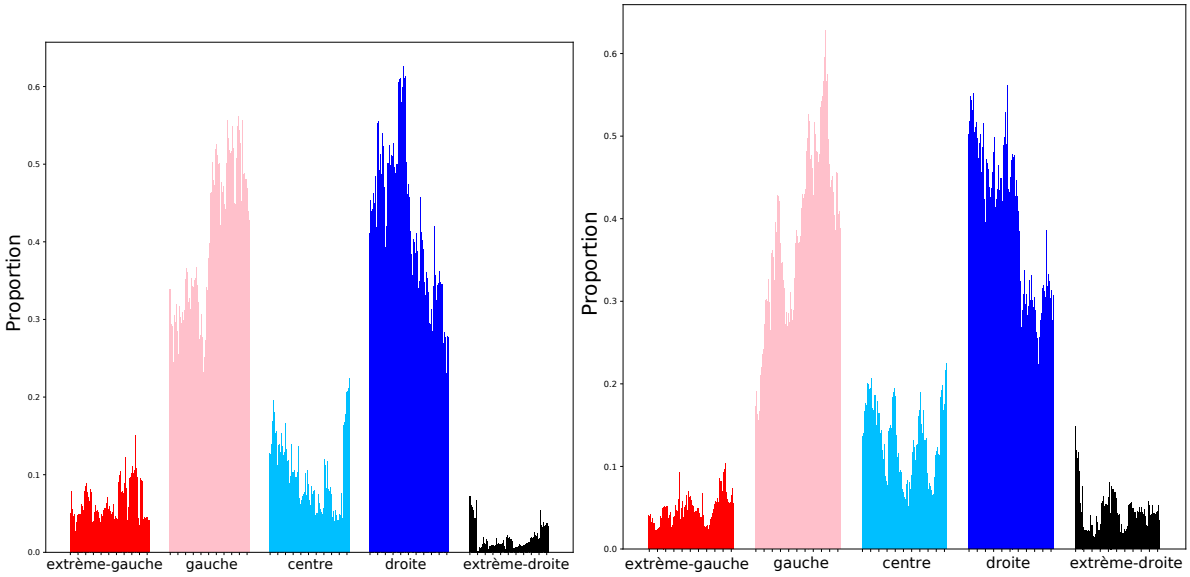
(a) Ensemble de la courbe

(b) Zoom sur le rectangle en pointillé

FIGURE 5.5 – 1-diversité relativisée par rapport à la 1-diversité des chaînes

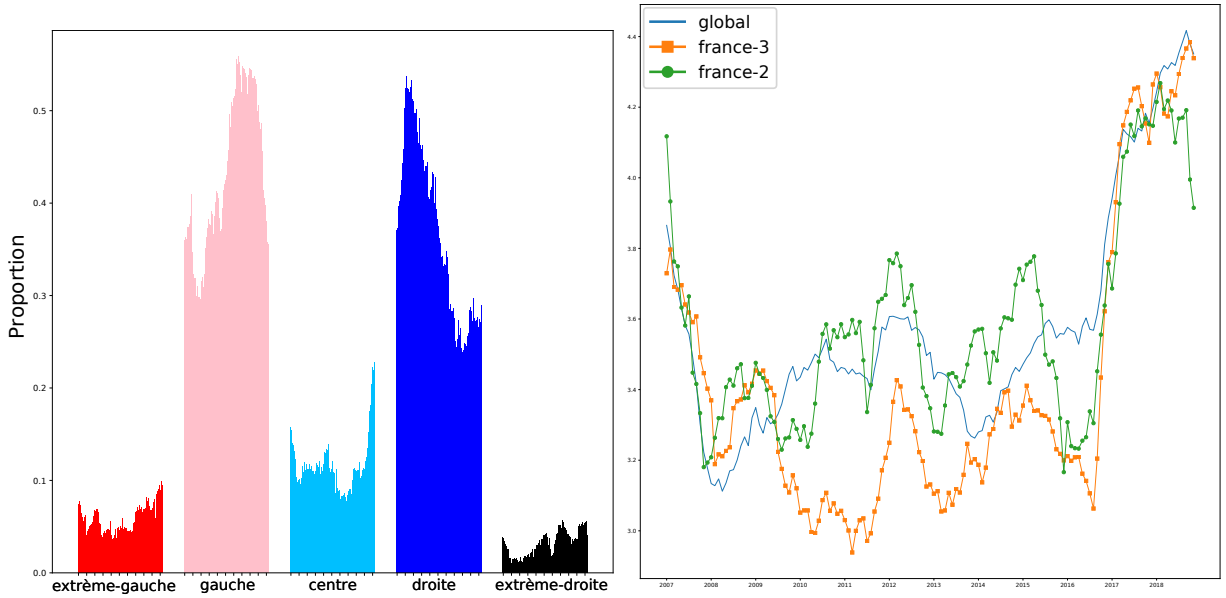
5.1.4 *Dimension temporelle et diversité.*

Cette diversité relative est un indicateur qui va prendre encore plus de sens quand nous allons ajouter une dimension temporelle à nos données. En effet nous allons utiliser une information que nous n'avons pas encore traitée : la date à laquelle une chaîne invite un invité. Concrètement nous avons simplement découpé notre graphe en mois et nous n'avons plus un graphe mais 143 graphes dont chacun représente un mois : du mois de janvier 2007 au mois de novembre 2018. Il y a donc 12 années mais nous n'avons pas l'information pour le dernier mois de 2018. Ceci crée des difficultés, notamment parce que certains nœuds des graphes sont présents seulement sur certains graphes car ils n'ont pas d'activité durant un mois. Cependant, pour notre étude, nous allons regarder les chaînes et particulièrement les grandes chaînes (celles qui invitent le plus de personnes), ce qui élude ce problème car elles sont toutes présentes dans tous les graphes. Ainsi nous avons regardé l'évolution des tendances des invités sur notre période de 12 ans pour deux chaines *france 2* (figure 5.6a) et *france 3* (figure 5.6b). Nous avons sur ces figures pour chaque tendance (de gauche à droite : extrême gauche en rouge, gauche en rose, centre en bleu clair, droite en bleu foncé et extrême droite en noir) l'évolution de leur abondance proportionnelle mois après mois. Nous pouvons y voir par exemple que les invités à tendance de droite (appartenant à des partis de droite) laissent place au fil du temps à des invités à tendance de gauche qui sont ensuite remplacés par des individus à tendance centriste. Nous pouvons aussi regarder globalement l'évolution des tendances (figure 5.7a) des invités sur toutes les chaînes. Nous retrouvons l'évolution des tendances qui correspond aux alternances de gouvernement : il y a des pics d'invitations à droite, puis à gauche et enfin au centre. Ceci suit clairement l'alternance des gouvernements.



(a) Sur france 2. **(b)** Sur france 3.

FIGURE 5.6 – Distribution de l'évolution des tendances sur deux chaînes.



(a) Distribution de l'évolution des tendances sur la globalité des chaînes

(b) Évolution de la diversité des tendances sur la globalité des chaines

FIGURE 5.7 – Évolution de la distribution et de la diversité des tendances

Nous pouvons maintenant nous intéresser à l'évolution des diversités qui correspondent à ces tendances. Nous voyons cependant une grande variabilité dans nos distributions, ce qui ferait une grande variabilité de notre diversité. Pour résoudre un peu ce problème nous avons décidé

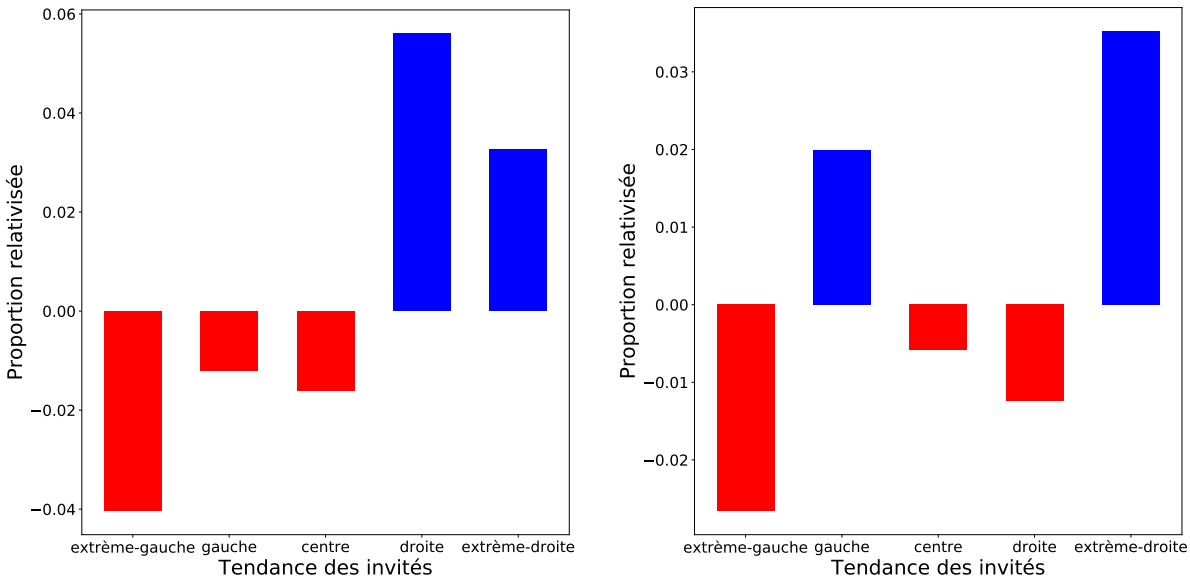
que nous allons passer d'un découpage par mois à un découpage par année glissant sur les mois. Précisément le premier graphe représentera la période allant de janvier 2007 à décembre 2007 compris, puis le deuxième commencera à février 2007 et terminera en janvier 2008 etc. Grâce à cette méthode nous avons pu tracer (figure 5.7b) l'évolution par rapport au temps de la diversité globale (en bleu), ainsi que celle de *france 2* (en vert et points ronds), puis *france 3* (en orange et points carrés). Nous pouvons tout d'abord voir une corrélation entre les deux courbes, laissant à penser que la diversité est plus une affaire de conjecture temporelle que de décision des chaînes. En regardant l'évolution globale de la diversité nous pouvons voir effectivement que des pics de diversité correspondent au moment des élections. Ces pics sont dus au phénomène d'alternance que nous venons d'exprimer. Mais aussi à la diversité plus forte avant les élections qui augmente la part des "petites" tendances (extrême droite, centre, extrême gauche). De plus on voit une très forte diversité fin 2018 qui s'explique aussi par une réapparition à la télévision du centre et de l'extrême gauche. Ceci est potentiellement en partie impliqué par notre façon de catégoriser les invités. Ces deux pics d'invitations sont par exemple dus à la forte présence de deux hommes politiques : *Mélenchon* et *Macron*. La question est ont-ils réellement apporté de la diversité politique en amenant à la télévision de nouvelles tendances (*extrême gauche* et *centre*) ou est-ce que c'est simplement notre façon de les catégoriser ainsi qui crée cette nouvelle diversité ?

5.1.5 *Dimension temporelle et diversité relative*

Maintenant que nous avons regardé la diversité temporelle d'une chaîne nous pouvons nous poser les mêmes questions qu'au début de cette section : comment se situe la distribution d'une chaîne à un moment précis par rapport à la globalité. Ici pour l'exemple nous avons pris le comportement de *france 2* en 2013. Nous pouvons le comparer au comportement de *france 2* tant sur toute la période $(2007 - 2018)$, c'est ce que nous avons fait dans la figure 5.8a. Mais nous pouvons le comparer aussi avec le comportement de toutes les chaînes en 2013, ce qui nous donne la figure 5.8b. Enfin nous pouvons mêler les deux méthodes et regarder le comportement de *france 2* en 2013 relativement à toutes les chaînes sur toute la période, ce que nous avons fait dans la figure 5.9a. En comparant les deux premières figures on peut voir que les deux analyses nous offrent des points de vues différents et complémentaires. Tout d'abord, *france 2* a invité plus de personnalités à tendance de gauche en 2013 et moins de personnalités à tendance de droite comparativement au profil de ses invités sur l'ensemble de la période,. Par contre la deuxième figure nous montre que *france 2* a invité plus d'invités à tendance de droite que les autres chaînes à ce moment là et moins d'invités à tendance de gauche.

La figure 5.9b nous montre l'évolution des trois diversités-1 relatives de *france 2* correspondantes à trois éléments de comparaisons : En orange (et points ronds) nous la comparons avec le comportement de la chaine sur l'ensemble de la période (comme la figure 5.8a). En vert (et points carrés) l'élément de comparaison est fixé sur une période, mais sur la globalité des chaînes comme la figure 5.8b). Enfin en bleu l'élément de comparaison est le graphe complet (comme pour la figure figure 5.9a). Nous pouvons tout d'abord remarquer que les trois courbes sont fortement corrélées, car les trois évolutions se suivent. Ce qui signifie que chaque

comportement se distingue en même temps des autres chaînes et des autres moments. De plus on voit que la courbe relativisée par rapport aux autres chaînes est plus basse. Ceci peut s'interpréter en disant que les programmes d'une chaîne sont plus impactés par la conjecture du moment que par une volonté globale de la chaîne. Ainsi la dimension temporelle nous donne plusieurs nouvelles façons de comparer des distributions, et donc de définir des nouvelles diversités relativisées qui ont des sens différents et complémentaires.



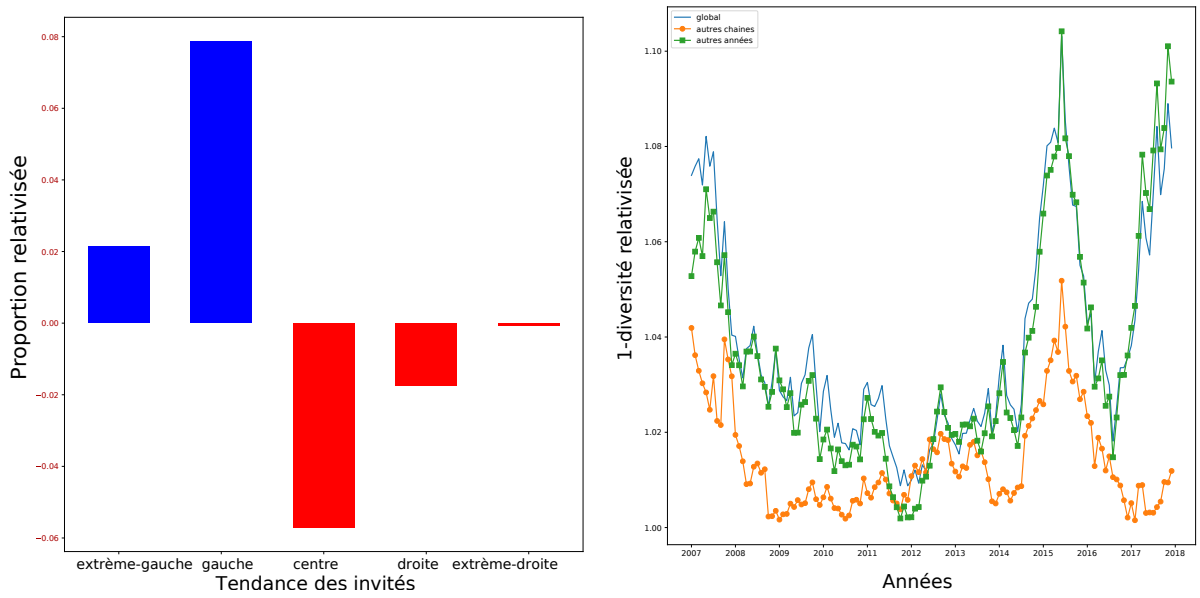
(a) Par rapport à toute la période sur france 2.

(b) Par rapport à toutes les chaînes en 2013.

FIGURE 5.8 – Distribution relativisée sur france 2 en 2013 selon différents comportements globaux.

5.1.6 *Conclusion*

Nous avons pu regarder l'intérêt d'un nouvel indicateur : la diversité relativisée. Cet outil est une nouvelle façon de mettre en valeur la diversité, qui se distingue de la diversité normalisée. En effet la diversité normalisée nous permet de regarder l'influence de propriétés structurelles des graphes (notamment le degré des nœuds) dans les scores de diversité. C'est outil ne fait aucune distinction entre les catégories, mais intègre le principe qu'il y aura des catégories qui seront plus atteintes que d'autres. La diversité relative, elle, se base sur une sorte de comportement *attendu* mesuré à l'aide du comportement collectif. Elle fait donc une différence entre les catégories atteintes. Quand la diversité normalisée prend en compte la différence de degré des nœuds d'arrivée entre les individus c'est une manière d'intégrer une distribution attendue différente de la distribution uniforme. Cependant la différence fondamentale reste : elle ne distingue pas les catégories.

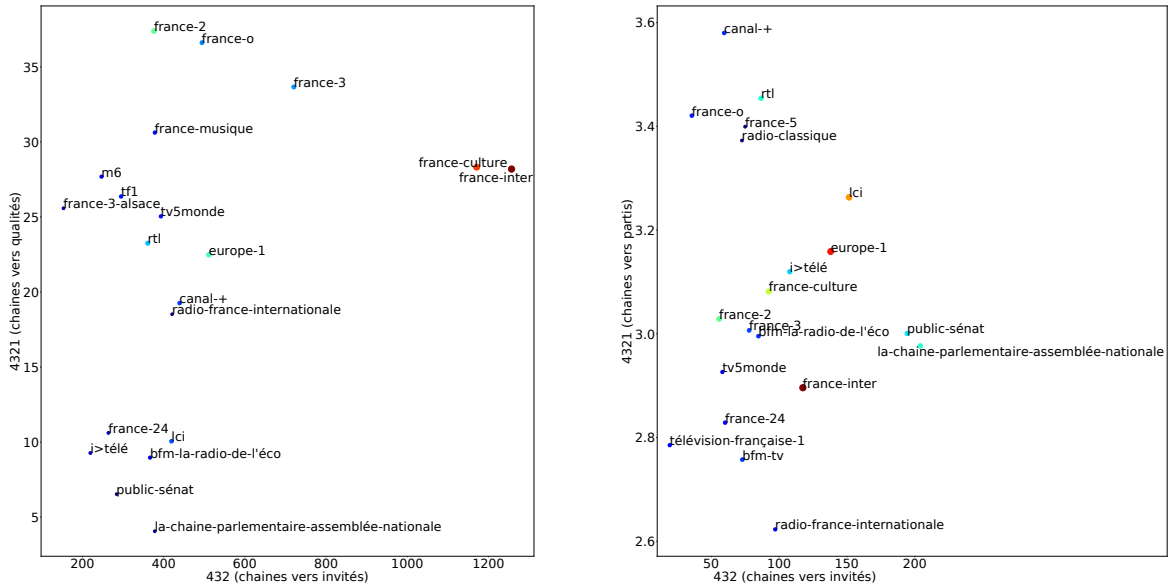


(a) Par rapport à toutes les chaines durant toute la période. **(b)** 1-diversité relativisée sur france 2 en 2013.

FIGURE 5.9 – Distribution et diversité relativisée sur france 2 en 2013.

La diversité relative pose aussi la question de la définition d'un comportement global, question qui prend encore plus de sens lorsque nous ajoutons la dimension temporelle, en offrant de nouveaux points de comparaison. Analyser l'évolution de la diversité dans le temps est une question qui est déjà en soit intéressante et elle nous permet d'identifier des situations particulières et d'en chercher de premières explications. Ceci nous donne une nouvelle problématique sur le traitement des données. En effet la façon de découper le graphe en périodes influence nos indicateurs. Il est différent par exemple de la couper par mois ou par année. Nous pouvons aussi regarder des découpages *glissants* où la période considérée comporte une partie de la période précédente.

Ces indicateurs, nouveaux et anciens, bien qu'ils soient parlant en soit, sont aussi une bonne façon d'identifier les situations et les comportements que nous allons cibler, pour ensuite regarder plus précisément certaines distributions. La diversité relative nous a d'ailleurs apporté une nouvelle façon de visualiser une distribution par rapport à une autre : les **distributions relativisées**. Les deux différentes façons de catégoriser et les problématiques qui y sont liées, nous montrent que notre travail doit se compléter en amont d'une fine connaissance des données, puisque nos indicateurs en sont fortement dépendants. Dans cette optique de meilleure visualisation, nous allons conclure sur ce jeu de données avec une particularité nouvelle (par rapports aux jeux de données précédents) que nous n'avons pas encore exploitée : c'est un graphe qui comporte 4 couches. Nous pouvons donc comparer des diversités liées à des chemins différents. Par exemple sur la figure 5.10, nous regardons la diversité $4 \to 3 \to 2 \to 1$ des chaînes par rapport à la diversité $4 \to 3 \to 2$, avec deux couches 1 différentes, la première représente les qualités pour la figure 5.10a et la deuxième représente les tendances politiques pour la figure 5.10b. *France 2* se situe en haut à gauche de la figure 5.10a, ce qui signifie



(a) Avec les qualités. **(b)** Avec les tendances politiques.

FIGURE 5.10 – Diversité $4 \to 3 \to 2 \to 1$ des chaînes par rapport à la diversité $4 \to 3 \to 2$.

qu'elle n'a pas une grande diversité d'invités, mais qu'ils correspondent à des qualités bien distinctes, au contraire de *france culture* qui a une plus grand diversité d'invités mais qui ont des qualités similaires. De même, sur la figure 5.10b, *canal +* semble avoir peu d'invités différents, comparé aux autres chaînes, mais toutes les tendances politiques sont représentées de manière équilibrée, à la différence de *public sénat* qui a une grande diversité d'invités mais sont plus ciblés en termes de tendance politique. Nous allons regarder ce que ce genre d'étude peut nous apporter dans la section suivante.

5.2 ÉTUDE D'UN GRAPHE À 6 PARTIES.

Si cette section ne présente pas en soi des résultats nouveaux vis-à-vis des chapitres précédents, nous avons choisi de rendre compte de ce travail dans le manuscrit car il nous semble compléter utilement la description de l'approche proposée dans cette thèse. Il permet notamment d'illustrer, sur un exemple plus complexe que les graphes précédents (les couches du graphe ne suivent pas une structure linéaire) comment adapter l'interprétation des scores de diversité dans d'autres contextes que ceux vus précédemment. Ils permettent aussi de rendre compte d'une collaboration pluri-disciplinaire conduite avec un géographe [4].

Nous allons analyser un autre graphe, cette fois-ci avec la particularité d'être un graphe 6 parties. Ce graphe représente l'implantation des ONG dans l'espace qui peut être découpé de différentes manières (pays, continent, région). En plus de cela, chaque projet des ONG est

4. Thomas Rosenthal, doctorant en géographie, UMR 8504 Géographie Cités, Université Panthéon-Sorbonne

associé à une thématique. L'objectif de cette section est donc de regarder comment peuvent s'utiliser nos indicateurs lorsque le graphe comporte plus de 3 couches. De plus ce travail a pu être faits en collaboration avec un géographe. Ce qui nous permet de montrer comment notre travail se complète avec une vision des données (que nous apporte le géographe) et comment il peut servir d'outil dans plusieurs domaines scientifiques. Nous choisissons donc de répondre à une problématique géographique : comment les ONG et les thématiques se comportent par rapport à trois échelles de découpage de l'espace différentes. Ces trois échelles étant trois couches de notre graphe complétées par les thématiques et les ONG, cette problématique est adaptée à l'étude de la diversité sur un graphe à beaucoup de couches.

5.2.1 *Sources*

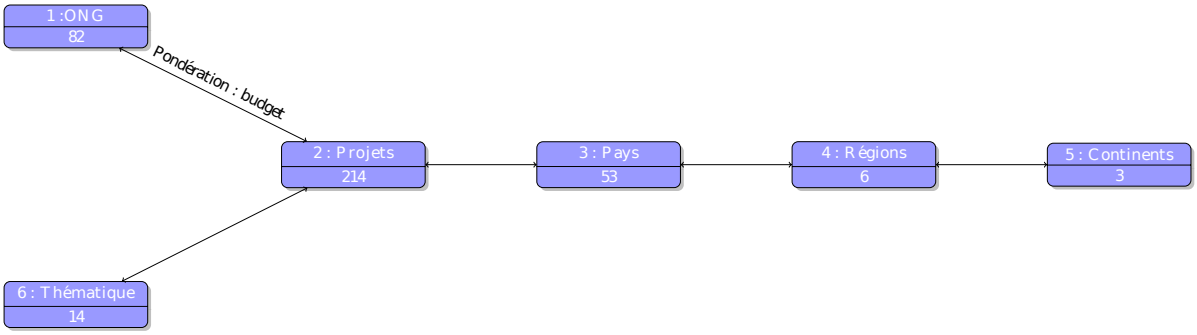


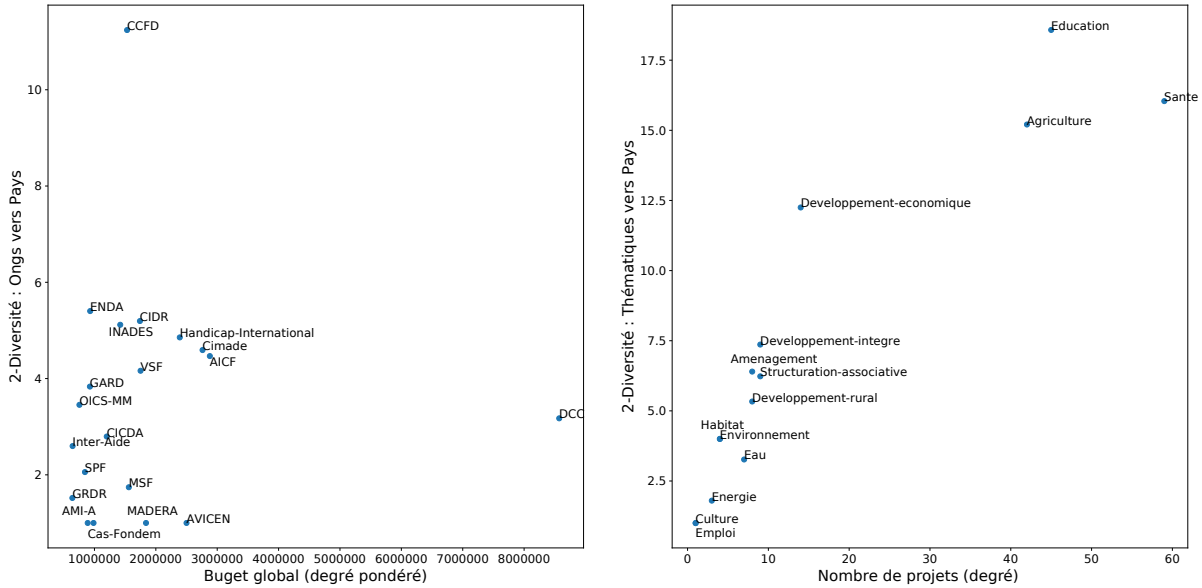
FIGURE 5.11 – Résumé des différentes couches de notre graphe.

Ces données sont tirées de l'ouvrage « L'Autre Coopération ([1]) édité par le Ministère des Affaires Étrangères et le Ministère de la Coopération. C'est le premier véritable ouvrage dans lequel on retrouve des analyses de l'univers de la société civile française en voie de structuration. On y retrouve en annexe un tableau des 259 projets financés par les deux ministères (la Coopération se concentrant sur l'Afrique subsaharienne francophone, et les Affaires Étrangères sur le reste du Monde) durant l'année 1989. C'est une des premières traces de la coopération entre l'État et la société civile française dans le cadre de l'aide internationale (aide humanitaire et aide au développement confondues). Notons que nous avons une représentation seulement des projets qui sont financés par l'Etat Français. Pour chacun de ces projets nous avons une description du projet, le pays dans lequel il est monté, la région (un groupement de pays) ; le nom de l'ONG et enfin le budget alloué par l'état français. Par exemple nous avons : *formation d'un responsable d'éducation populaire, Rwanda, Afrique noire, ccfb,* $62200F$. Nous allons garder toutes ces informations, nous changeons le nom de certaines régions (Afrique noire devient par exemple Afrique subsaharienne), nous associons un continent à chaque région, et enfin nous associons à chaque description une thématique [5] (ici par exemple la thématique sera *éducation*). Nous supprimons [6] aussi quelques projets qui ne sont

5. Le livre fournit une thématique, mais elle a été retraitée manuellement par le géographe
6. ici aussi manuellement avec l'aide du géographe.

pas exactement liés à de l'aide au développement : édition d'un livre, organisation d'un séminaire, préparation d'un événement en France qui regroupe les ONG etc. Nous avons donc 6 couches distinctes : Les ONG(1), les Projets(2), les Pays (3), les Régions (4), les Continents (5) et les Thématiques (6). Les pays sont liés aux régions qui sont eux liés aux continents. Les projets sont eux liés aux thématiques, aux pays, mais aussi aux ONG. Nous rajoutons une pondération aux liens entre les ONG et les projets qui est simplement le budget alloué. Nous avons résumé les liens entre les couches ainsi que leur taille (nombre de nœud dans chaque couche) dans la figure 5.11. Nous pouvons voir que cette structure n'est pas linéaire. En effet, contrairement aux graphes étudiés précédemment, il y a une couche, la couche Projets, qui est liée à 3 autres couches. Cette structure inhabituelle nous permet de voir de nouveaux aspects de notre méthode. Nous voyons aussi à quel point la façon de créer ce graphe est fortement déterminée par une expertise du domaine scientifique associé aux données et à la question de recherche. Dans cette optique, la collaboration avec un géographe est importante au moment de la création du graphe.

5.2.2 *La diversité spatiale des ONG et des thématiques*



(a) ONG

(b) Thématiques.

FIGURE 5.12 – 2-diversité des ONG et des thématiques en fonction des pays touchés.

Nous pouvons comparer deux chemins différents qui arrivent sur la même couche, ici les pays (couches 3), mais qui partent de deux couches différentes : les ONG (chemin $1, 2, 3$) [7]

7. Dans cette partie pour alléger le manuscrit on notera les méta-chemins par un n-uplet, par exemple $(1, 2, 3)$, plutôt que $1 \to 2 \to 3$. De plus nous parlerons de chemins, plutôt que de méta-chemins, puisqu'il n'y a pas d'ambiguïté ici.

et les Thématiques (chemin $6, 2, 3$). Ce sont deux diversités qui se ressemblent puisque les proportions abondantes sont associées aux mêmes entités (ici les pays) mais ces diversités ne concernent pas les mêmes entités (Thématique ou ONG). Voyons comment le sens des analyses est similaire mais dépend de l'entité de départ.

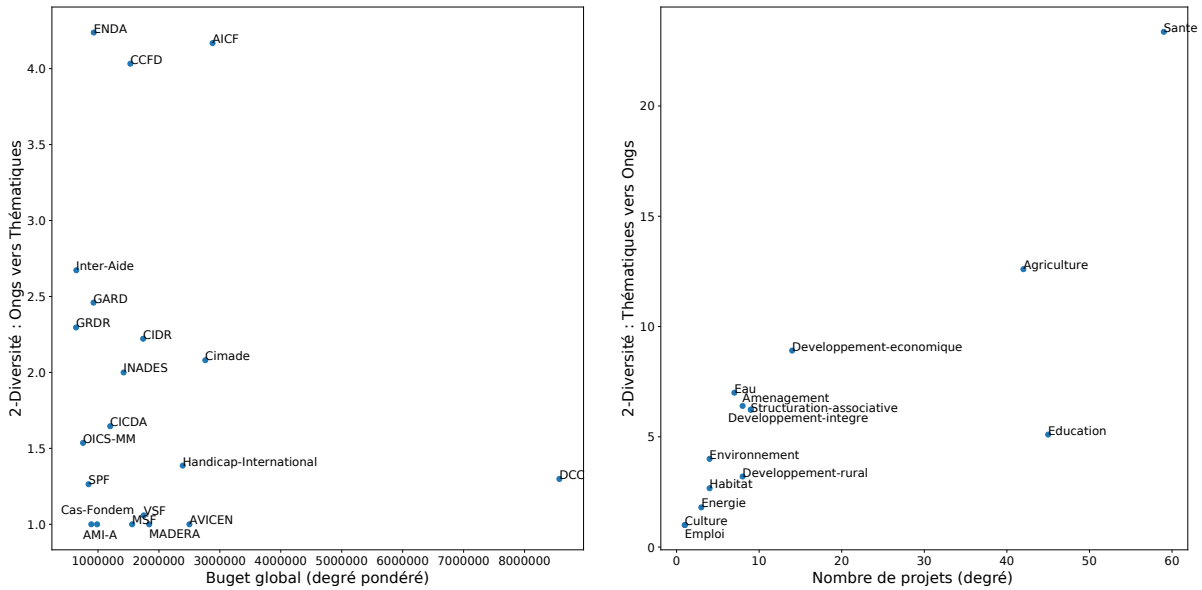
Nous traçons donc, dans la figure 5.12a [8], la 2-diversité des pays visés par les ONG en fonction de leur degré pondéré c'est-à-dire ici de l'ensemble des financements. Nous voyons qu'il n'y a pas de corrélation entre les financements accordés par l'État français et la diversité des pays où les ONG interviennent, comme l'illustre le cas du CCFD, qui touche une diversité 3 fois plus grande de pays que VSF avec autant de financements. De même la DCC, avec des financements 4 fois plus importants que ceux du CCFD ou de VSF, touchent autant de pays. Cependant, la DCC fait figure d'exception dans l'année 1989, ce qui s'explique par le fait que c'est une ONG dont l'action consiste à envoyer des volontaires dans les pays en développement ; ce sont les frais liés aux volontariats qui sont ici couverts par le Ministère de la Coopération et le Ministère des Affaires Étrangères. Cela étant, la courbe nous montre également que, hormis les exceptions que sont le CCFD et la DCC, il y a un pôle central d'ONG qui ont tendance à se concentrer sur un faible nombre de pays.

Comme nous l'avons annoncé nous avons aussi tracé la courbe (figure 5.12b) de la 2-diversité des thématiques selon les pays visés (chemin $6, 2, 3$), en fonction du degré (c'est-à-dire le nombre de projets). Nous voyons une corrélation assez forte entre la diversité thématique par pays et le nombre de projets. Ce qui semble assez évident, la solidarité internationale se décline en un grand nombre de thématiques d'action. Ainsi, si l'on voit des thématiques très influentes, comme l'agriculture, la santé ou l'éducation, c'est que ce sont des enjeux par essence mondiaux. D'autres enjeux comme l'emploi et l'énergie ne sont pas moins mondiaux, cela étant, en 1989, les enjeux humanitaires de premier ordre sont privilégiés par les ONG et par les bailleurs publics dans le déploiement de l'aide venue de France.

Pour étudier la corrélation entre les deux nous pouvons regarder la diversité sur un chemin allant des ONG aux thématiques $(1, 2, 6)$ et sur le chemin inverse : Thématique vers ONG $(6, 2, 1)$. Ces chemins complètent l'analyse de la corrélation entre les deux. Nous avons représenté ces deux diversités par rapport à leur degré pondéré sur la figure 5.13.

Nous remarquons sur la figure 5.13a une forte corrélation entre la diversité des projets (en abscisse) et la diversité des thématiques par ONG (en ordonnée). Plus il y a de projets sur une thématique donnée (par exemple agriculture et santé), et plus il y a d'ONG concernées. Il semblerait que seule la thématique de l'éducation fasse exception : en effet, en regardant de plus près la distribution de l'éducation, nous voyons que 2 ONG se répartissent plus de 60% des financements de l'Etat. La DCC (Délégation Catholique pour la Coopération) et le CCFD (Comité Catholique Français pour le Développement) ont une proportion exceptionnelle de projets financés par l'Etat. Cela est certainement lié à leur historique d'action du catholicisme social (ce sont parmi les plus vieilles structures non gouvernementales conduisant des projets

8. AICF : Action Internationale Contre la Faim. CCFD : Comité Catholique contre la Faim et pour le Développement. DCC : Délégation Catholique pour la Coopération. GRDR : Groupe de Recherche et de Réalisation pour le Développement Rural. INADES : Institut Africain pour le Développement Économique et Social. VSF : Vétérinaires Sans Frontières.



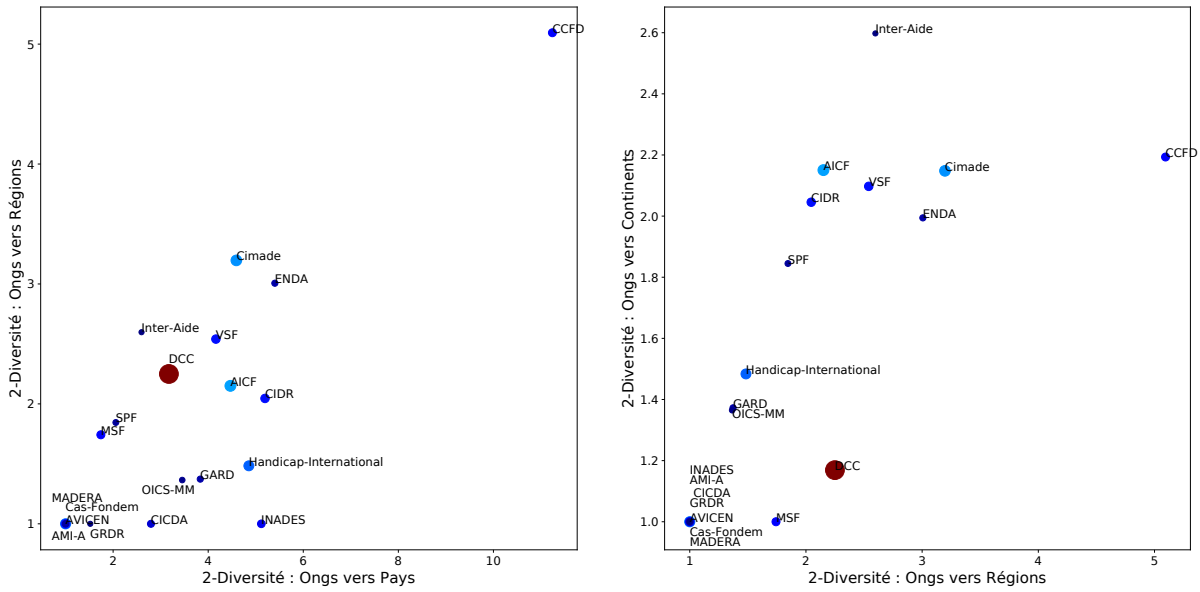
(a) ONG vers les thématiques. **(b)** Thématiques vers les ONG.

FIGURE 5.13 – 2-diversité des ONG vers les thématiques et inversement.

dans le Monde). Cela est sûrement lié aussi à la thématique de l'éducation en elle-même, qui nécessite du savoir faire acquis dans la durée : c'est certainement cela qui explique ce choix de l'Etat de financer ces deux structures en priorité. A l'inverse, on peut voir que des thématiques comme l'eau sont diversifiées en ONG malgré un faible nombre de projets.

Toutefois, on peut voir que les financements ne sont pas particulièrement déterminants dans la diversité d'action des ONG. En effet, la figure 5.13b nous permet de visualiser le poids des financements publics dans la diversité des thématiques des ONG. Nous pouvons voir une exception notable, qu'est la DCC qui reçoit beaucoup de subventions pour mener quasi exclusivement des projets d'éducation. D'autres structures, comme le CCFD et INADES, peuvent recevoir autant de financements pour une diversité de thématiques abordées variant du simple au double.

Jusqu'ici, nous pouvons voir qu'il y a une corrélation entre le nombre de projets et la diversité des thématiques selon les ONG. Nous pouvons voir également qu'il n'existe pas de liens évidents entre les financements et la diversité d'action des ONG. Certaines structures sont, par nature, diversifiées et considèrent le développement comme un enjeu total (comme le CCFD), quand d'autres structures sont plus spécialisées (comme VSF pour l'agriculture ou la DCC pour l'éducation).



(a) Comparaison pays et régions.

(b) Comparaison continents et régions

FIGURE 5.14 – 2-diversité des ONG vers les différents découpages de l'espace

5.2.3 *Trois façons de découper l'espace*

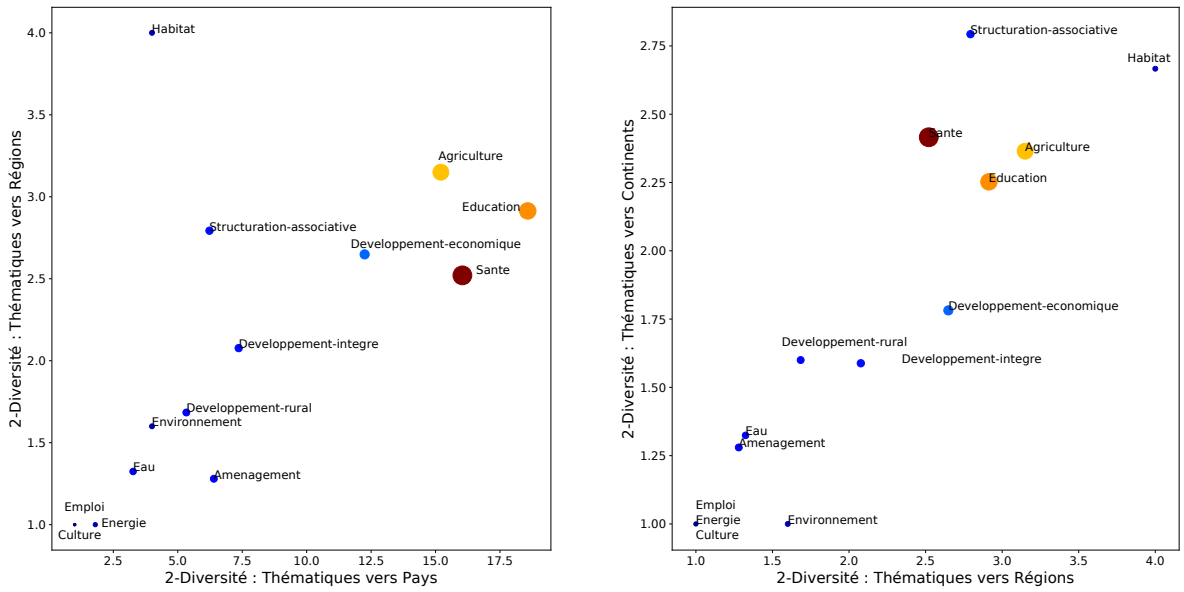
Regardons maintenant l'influence de notre découpage géographique sur nos scores de diversité. Cela signifie que nous allons croiser les scores de diversité sur des chemins qui commencent de la même manière. Par exemple en partant des ONG on peut regarder la diversité des pays touchés $(1, 2, 3)$, la diversité des régions touchées $(1, 2, 3, 4)$ et enfin la diversité des continents touchés $(1, 2, 3, 4, 5)$. Comme pour l'entité de départ nous pouvons regarder ces diversités sur la même figure. Par exemple nous avons comparé celle qui touche aux régions à celle qui touche aux pays (figure 5.14a) et celle qui touche aux régions par rapport à celle qui touche aux continents (figure 5.14b). Ceci nous permettra de saisir les motifs d'action des ONG françaises et de préciser notre vision de la distribution spatiale à différentes échelles des projets menés par les ONG françaises, quand elles sont financées par l'État.

Au premier abord la figure 5.14a nous présente une corrélation entre la diversité de pays et la diversité de régions investis par les ONG, nous pourrions être amenés à penser que c'est une corrélation logique : plus de pays sont touchés, et logiquement plus de régions le sont aussi. Nous pouvons voir toutefois sur la figure 5.14b, que si des ONG comme le CCFD ou Handicap International ont une répartition linéaire (liens logiques entre le nombre de pays, de régions et de continents touchés), bien des structures se détachent de cette linéarité.

En effet, on peut voir que la dynamique d'une structure comme Inter-Aide qui se déploie dans 3 continents malgré un faible nombre de projets, est suivie par une grappe de structures telles que La Cimade, AICF ou VSF. Ces différentes structures sont déjà diversifiées et réparties dans différents continents et les financements étatiques ne font qu'appuyer cette dynamique. De la

même manière, on peut voir qu'un groupe conséquent de structures, telles que le GRDR ou INADES touchent un seul continent (l'Afrique) malgré un nombre de pays touchés différents.

En conclusion, on peut considérer que pour avoir une meilleure approche de la répartition spatiale des ONG, il est nécessaire de travailler de manière *scalaire*, c'est-à-dire en changeant les façons de segmenter l'espace, ce qui nous donne une analyse plus fine des distributions spatiales. En effet, on pourrait penser que, puisque le CCFD a de nombreux projets touchant de nombreuses thématiques, ce serait logiquement la structure la plus diversifiée spatialement. Nous pouvons voir que selon une approche par pays, par régions ou par continents, les analyses dues aux différents chemins diffèrent et se complètent.



(a) Comparaison pays et régions.

(b) Comparaison continents et régions

FIGURE 5.15 – 2-diversité des Thématiques vers les différents découpages de l'espace

Regardons la même chose pour les thématiques dans la figure 5.15. Nous voyons apparaître un processus assez similaire : quand on peut penser que la relation entre les diversités de pays, régions et continents touchés sont assez linéaires, de nombreux contre-exemples viennent à l'encontre de cette tendance. La figure 5.15b, toutefois, nous présente une corrélation linéaire entre la proportion de régions et de continents touchés. La figure 5.15a, quant à elle, nous fait apparaître 3 principaux groupes : un groupe composé de l'agriculture, la santé et l'éducation, qui sont les thématiques les plus riches en projets et en ONG. Logiquement elles sont bien réparties en pays, en régions et en continents. Ensuite, nous voyons apparaître un deuxième groupe, composé des thématiques moins prisées, telles que l'eau, le développement économique et l'habitat : assez logiquement, il y a peu de projets, donc un faible nombre de pays, de régions et de continents touchés, et ce de manière corrélée.

Enfin, un troisième groupe est composé de deux exceptions notoires : les thématiques de l'habitat et de la structuration associative (aide à la construction de sociétés civiles locales) sont

très fortement représentées en régions et en continents pour un faible nombre de pays touchés. Par exemple, en regardant sa distribution associée, la thématique de l'habitat est développée dans 4 pays différents : le Maroc, le Botswana, le Vietnam et le Pérou. Soit dans 4 régions et dans 3 continents différents : ce qui en fait la thématique la plus diverse en termes de continents de notre échantillon.

Par conséquent, grâce aux différents chemins pris en compte, nous voyons bien le rôle de l'échelle dans notre réflexion : plus on la prend en compte, et plus notre information géographique est précise et permet de rendre compte de dynamiques spécifiques.

5.2.4 *Conclusion de ce travail en collaboration avec un géographe*

Cette section nous permet de voir que la multitude de chemins dans un groupe nous permet une richesse d'interprétation que ce soit en regardant des chemins partant de ou arrivant sur la même couche. De plus, il nous permet de voir en quoi notre travail peut être un outil de collaboration avec d'autres domaines scientifiques. La connaissance des données est nécessaire en amont quand il s'agit de construire le graphe et en aval quand il s'agit d'interpréter et d'expliquer nos scores. Comme nos indicateurs sont une façon de donner un score à une distribution, ils nous permettent de les comparer facilement : constituer des groupes, découvrir des corrélations, déterminer les singularités. Une fois que ce travail est fait, il est intéressant d'analyser les distributions, ce qui serait très difficile et long à faire avant ce travail. Ainsi nos indicateurs peuvent parfois être utilisés pour être interprétés directement, ou alors, plus simplement, pour orienter la recherche. La multiplicité des chemins renforce cette difficulté. Par exemple aux 10 chemins que nous avons traités ici, nous avions plus d'un millier de distributions associées. Une fois que nous nous étions entendus sur la construction du graphe et sur les chemins intéressants, nous avons pu utiliser nos méthodes pour générer facilement toutes ces courbes.

5.3 Conclusion

Nous avons analysé deux nouveaux jeux de données qui nous ont permis de dépasser les graphes tripartis classiques. Tout d'abord dans la première section nous avons dépassé le côté statique en regardant comment nos indicateurs peuvent intégrer une dimension temporelle. Nous avons pu aussi regarder dans cette section la diversité relative et voir en quoi elle prenait une nouvelle importance avec la dimension temporelle. Dans la deuxième section nous avons dépassé le nombre de 3 couches et regardé ce qu'impliquait la multitude de chemins possibles. Nous avons vu aussi que des jeux de données de nature très différentes peuvent être adaptés à nos analyses de diversité. À chaque fois, la nature des entités, c'est-à-dire la nature des couches, nous donne de nouvelles interprétations possibles pour nos indicateurs.

Chapitre

6

Conclusion et perspectives.

Nous avons rédigé ce manuscrit avec une certaine progression. En introduction (chapitre 1) nous avons montré les motivations qui nous poussent à analyser la diversité sur des données relationnelles. Le chapitre 2 nous a permis d'introduire le formalisme nécessaire à toute notre étude. En particulier nous avons unifié plusieurs indicateurs de diversité venant de plusieurs domaines scientifiques pour en extraire une famille avec des propriétés mathématiques intéressantes : l'α**-diversité**. De plus nous avons pu combiner ceci avec une marche aléatoire définie sur des graphes n-partis pour définir l'élément central de notre étude l'α**-diversité selon un chemin**. Ensuite dans le chapitre 3 nous avons pu montrer comment ces méthodes permettaient d'analyser deux jeux de données musicales *MSD* et *Amazon*. Nous avons pu voir que nos indicateurs se rapprochaient d'une notion de diversité intuitive. Nous avons pu détecter entre autres deux phénomènes fondamentaux. Premièrement la diversité des tags comme celle des utilisateurs est fortement dépendante de leur volume. Deuxièmement en analysant cette dépendance nous nous sommes rendu compte qu'elle n'était pas linéaire mais qu'il y avait un phénomène de saturation. Pour approfondir cette question nous nous sommes penchés sur l'étude de la diversité sur des graphes aléatoires dans le chapitre 4. Tout d'abord, en analysant la diversité sur des graphes construits à partir d'un graphe réel. Ensuite en prouvant des formules analytiques qui correspondent à ces générations aléatoires. Enfin nous avons étudié la diversité sur des structures aléatoires respectant différentes distributions de degrés fixées. Tout ceci nous a permis de définir l'α**-diversité normalisée** qui nous est adaptée à la notion intuitive de la diversité. De plus ceci nous a permis de mieux comprendre les phénomène qui influençaient les scores de diversités. Le chapitre 5 nous permet d'ouvrir l'horizon de notre étude. En effet nous avons regardé deux nouveaux jeux de données qui nous ont permis pour le premier d'analyser des données temporelles et pour le deuxième d'analyser un graphe ayant 6 parties. Nous avons aussi vu dans cette section l'intérêt de l'α-diversité relative notamment quand la dimension temporelle rentre en jeu. Ainsi nous avons formalisé nos méthodes, traité des exemples, perfectionné ces méthodes grâce aux observations précédentes puis traité de nouveaux exemples différents.

Nous avons exploré une grande partie de l'espace que couvrait notre formalisme et nos mé-

thodes. Il y a beaucoup de débouchés possibles à ceci. Tout d'abord nous pouvons regarder d'autres jeux de données, ce qui amène encore des significations différentes à nos indicateurs. Nous présenterons dans la section 6.1 deux autres jeux de données que nous avons traités mais dont nous n'avons pas parlé car ils ne mettent pas en œuvre une démarche nouvelle. Nous regarderons ensuite certaines débouchés possibles qu'offre nos méthodes et comment les perfectionner dans la section 6.2. Puis nous regarderons les deux principales possibilités qu'offrent notre travail. Tout d'abord, et c'était la motivation première de ce travail, nous pourrions l'utiliser comme une base pour l'analyse de la recommandation. C'est ce que nous expliquons dans la section 6.2.2. Enfin nous nous interrogerons (dans la section 6.3) sur l'approche quantitative de notre travail pour montrer en quoi nous pensons qu'il peut être productif : il se marie bien avec une étude qualitative .

6.1 D'AUTRES JEUX DE DONNÉES TRAITÉS.

Nous n'avons pas présenté tous les jeux de données que nous avons traités durant notre travail. En effet il nous a paru important durant toute notre étude de confronter nos méthodes à des données venant d'horizons différents. Nous faisons ici un retour sur deux autres jeux de données que nous avons traités en utilisant des analyses que nous avons déjà vues dans le reste du manuscrit. Ces deux jeux de données traitent des médias : le premier sur la consommation de médias (lecture d'articles) sur le site *melty* [1]. Le deuxième traite de l'utilisation de ces médias sur *twitter* par des utilisateurs colorés politiquement.

6.1.1 *Données de consommation de médias : Melty*

Comme annoncé, ce jeu de données présente la consommation de médias par les utilisateurs du site *melty*. Ce site est une plateforme d'information qui se caractérise par des articles pour un public particulièrement jeune. Notre analyse de la diversité intéressait les développeurs de ce site pour pouvoir potentiellement améliorer la qualité de leur service.

6.1.1.1 *Traitement des données et construction du graphe triparti*

Dans le cadre de la collaboration ANR AlgoDiv [2] (lien site en note de bas de page), nous avons ainsi pu avoir accès à des données de consommation pendant une semaine (6 jours exactement). Précisément nous avons un fichier dans lequel chaque ligne indique un utilisateur (anonymisé grâce à un identifiant) et le nom de l'article que cet utilisateur a consulté. De plus chaque article est catégorisé dans un dossier (**folder**) qui définit à peu près le sujet de l'article. Dans la majorité des cas le folder désigne une personnalité (exemple : *Jenifer*, *Sarah Michelle Gellar*) un film (exemple : *Le Seigneur des Anneaux* ou *Thor*) une série (exemple : *Downton*

1. www.melty.fr
2. http://algodiv.huma-num.fr/

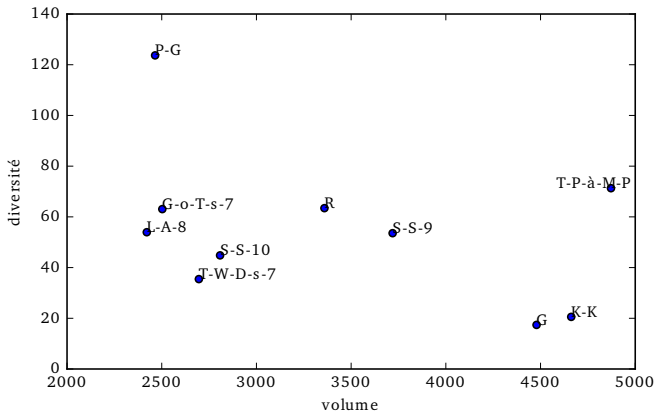
Abbey Saison 6 ou *New Girl saison 6*) et enfin une émission de télévision (exemple : *The Voice Kids* ou *La Nouvelle Star 2016*). Dans de rares cas nous pouvons aussi avoir des sujets plus globaux par exemple *Allemagne*, *esport club*, *TF1* et *OL*.

Notons qu'ici nous avons pu travailler directement avec un développeur du site, ce qui nous a permis d'avoir une meilleure façon de *nettoyer* les données.

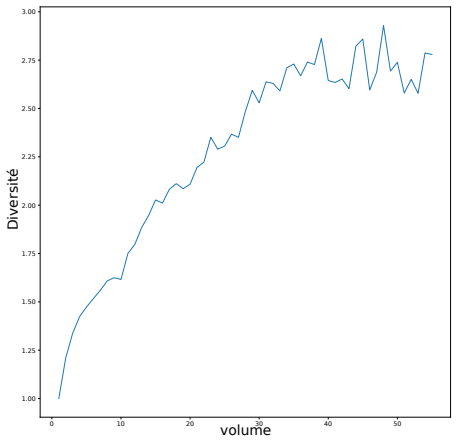
Nous avons donc créé un graphe triparti avec des utilisateurs (couche 1 : $2\ 587\ 433$ nœuds) reliés aux articles qu'ils ont consultés (couche 2 : $306\ 010$ nœuds) qui sont eux mêmes reliés aux folders (couche 3 : 4401 nœuds).

Nous avons par exemple remarqué qu'il y avait un nombre important d'utilisateurs (80%) qui ne consultaient qu'un article : ce qui est dû à la courte durée et une authentification non obligatoire. Au contraire nous avons aussi des utilisateurs qui consultent énormément d'articles (pouvant aller pour certains à plus de 1000 par jour), utilisateurs qui semblent, selon les dire des développeurs, être des robots, qu'il n'avaient pas encore pu supprimer de leurs données. Nous avons donc tronqué notre jeu de données, pour ne garder que des utilisateurs ayant entre 5 et 100 articles : ce qui nous donne 111358 utilisateurs, 102676 articles et 3496 folders.

6.1.1.2 *Application de nos méthodes.*



(a) 2-diversité d'audience $(3, 2, 1)$ des folders.



(b) Évolution de la 2-diversité en fonction du volume.

FIGURE 6.1 – Diversité sur Melty

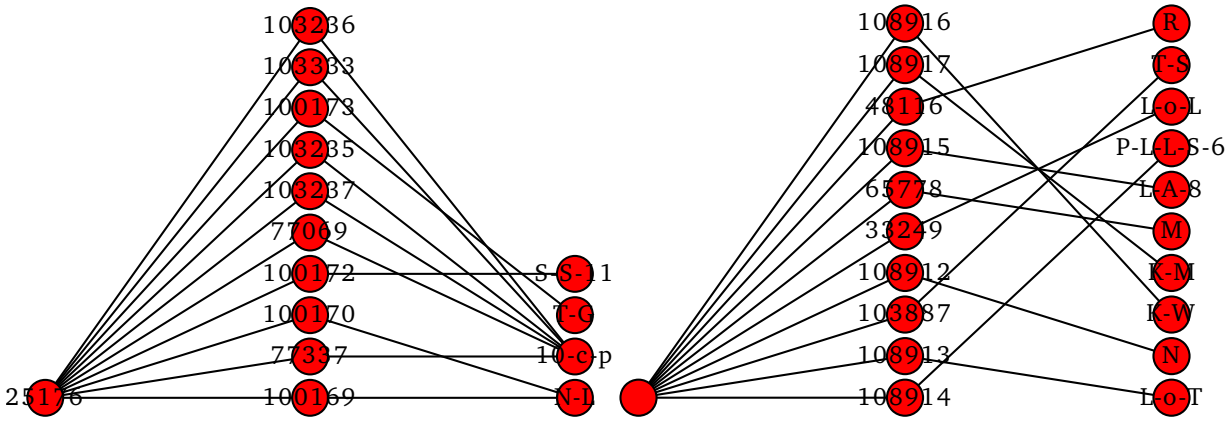
Nous avons réitéré les mêmes études que ce que nous avons fait dans le chapitre 3. La figure 6.1a montre la 2-diversité des 10 folders ayant le plus de consultations (volume $3, 2, 1$). Pour plus de lisibilité nous n'avons noté que les initiales des folders : on peut trouver la liste des abréviations en notes [3].

3. 'G' : 'Gaming', 'R' : 'Rihanna', 'P-G' : 'Pokemon-Go', 'T-P-à-M-P' : 'Touche-Pas-à-Mon-Poste',

Nous voyons qu'à volume fixé *'Touche Pas à Mon Poste'* est beaucoup plus diversifié en terme d'utilisateurs que *'Kim-Kardashian'*. Ce qui signifie que beaucoup d'utilisateurs différents consultent des articles sur *'Touche Pas à Mon Poste'* alors qu'une population très spécifique lit des articles sur *'Kim-Kardashian'*. Cela peut déjà intéresser les créateurs du site, car ils ne vont pas mettre en valeur l'article de la même manière selon que la population ciblée soit un lecteur lambda ou un lecteur spécifique.

Pour comparer nos résultats, nous avons aussi regardé si le phénomène de saturation que nous avons observé précédemment se retrouve aussi ici. Nous avons donc tracé la figure 6.1b montre l'évolution de la diversité-2 des utilisateurs par rapport aux folders consultés $(1, 2, 3)$ en fonction de leur volume. Nous voyons le même phénomène de saturation que précédemment même si notre suppression des utilisateurs très actifs nous empêche de voir la courbe pour les gros volumes. Dans les faits, nous avons quand même regardé la diversité de ces *gros* utilisateurs, certains avaient des diversités assez étonnement proches d'un comportement de consommation complètement aléatoire, ce qui nous a conforté dans l'idée que ces utilisateurs étaient peut-être des robots. Ce qui donne un autre intérêt à nos indicateurs.

Nous pouvons aussi grâce à nos méthodes détecter des comportements spécifiques des utilisateurs. Par exemple, nous avons regardé les graphes induits par deux utilisateurs particuliers dans la figure 6.2. Nous les avons sélectionnés, par leur diversité-2. Le premier profil ayant un score à peu près égal à 3 (figure 6.2a) et le deuxième un score de 10 (figure 6.2b). Ces deux profils ont pourtant le même volume (10). Ce sont deux profils assez différents. Le premier sélectionne des articles ciblés sur seulement 4 folders avec une attention très particulière pour *10-c-p* (*10 couples parfaits*). Le deuxième a une diversité parfaite puisqu'il a consulté 10 articles venant de 10 folders différents. Là aussi on voit l'intérêt pour le site d'identifier rapidement les utilisateurs qui ont par exemple une faible diversité pour voir si le site peut arriver à d'avantage leur apporter des nouveaux sujets.



(a) Utilisateur ayant une 2-diversité égale à 3 **(b)** Utilisateur ayant une 2-diversité égale à 10

FIGURE 6.2 – Graphe induit pour deux utilisateurs particuliers de Melty

'G-o-T-s-7' : 'Game-of-Thrones-saison-7', 'S-S-9' : 'Secret-Story-9', 'L-A-8' : 'Les-Anges-8', 'K-K' : 'Kim-Kardashian', 'S-S-10' : 'Secret-Story-10', 'T-W-D-s-7' : 'The-Walking-Dead-saison-7'

Comme nous avons mené ce court travail en collaboration avec la plateforme, nous commençons à voir ici comment pourraient être utilisées nos méthodes de l'intérieur d'une plateforme pour étudier et améliorer la diversité de lecture des utilisateurs.

6.1.2 *Données d'utilisation des médias sur Twitter : Datapol*

Nous avons participé à une *hackaton* [4] où nous avons analysé pendant quelque jours des données venant de Twitter. Plus précisément nous avons utilisé les données du *Politoscope* [5] qui a analysé des millions de comptes Twitter pendant la campagne présidentielle française.

6.1.2.1 *Traitement des données et construction du graphe triparti*

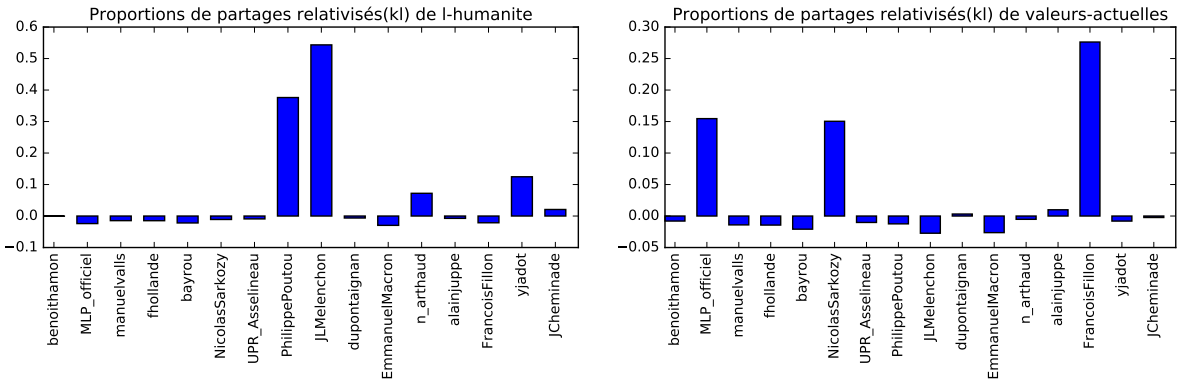
Ces comptes Twitter ont été associés à un candidat (ou potentiel candidat) de la présidentielle grâce à un algorithme de détection de communauté en analysant les retweets entre comptes. Ainsi plus d'un million de comptes ont été associés à un candidat politique. Nous dirons que chaque candidat représente une communauté. Nous avons déjà un premier biparti $(1, 2)$: communauté -compte. De plus pour chacun des tweets nous avons regardé s'il y avait une url citée. Nous avons extrait ainsi la source de chaque url qui peut correspondre à un média (*lemonde.fr* ou *lefigaro.fr* par exemple) mais aussi à une plateforme web (*facebook*, *youtube*). Pour éliminer le bruit nous n'avons pris que les 500 sources les plus citées. Nous créons donc une nouvelle couche (3) qui correspond à ces sources. Et chaque fois qu'un tweet cite une source on place un lien entre le compte qui a tweeté et la source. Ainsi nous avons un graphe triparti comportant trois couches : communautés (couche 1 : 16 nœuds), compte (ou utilisateur) (couche 2 : $1\ 227\ 233$ nœuds) et source (couche 3 : 500 nœuds). Notons que le biparti $2, 3$ est très dense car il contient plus de 18 millions de liens.

6.1.2.2 *Application de nos méthodes.*

Sur ces données nous nous sommes particulièrement intéressés comme à la section 5.1, aux distributions et diversités relativisées puisque les tailles des différentes communautés sont très disparates. Par exemple nous avons regardé les distributions relativisées des sources par rapport aux communautés $(3, 2, 1)$ pour deux différentes sources sur la figure 6.3. Le journal *l'humanité* étant un journal communiste (figure 6.3a) il n'est pas étonnant de le voir relativement fortement cité par les communautés bien ancrés à gauche (*Poutou*, *Mélenchon*, *Jadot*, *Arthaud*). Inversement le journal *valeurs actuelles* étant un journal d'extrême-droite (figure 6.3b) il se retrouve relativement cité par les communautés situées à droite (*Fillon*, *Sarkozy*, *Le Pen*).

4. DATAPOL est organisé par le médialab de Sciences Po en collaboration avec le CEVIPOF, l'Ecole de Journalisme de Sciences Po, la Bibliothèque de Sciences Po, l'Institut des Systèmes Complexes, le Public Data Lab, la société Linkfluence, l'Institut National de l'Audiovisuel, les Décodeurs du journal Le Monde, et la société Matlo, avec le soutien du Google News Lab.

5. Projet Politoscope, CNRS Institut des Systèmes Complexes Paris Ile-de-France (ISC-PIF), `http://politoscope.org`



(a) Pour l'humanité.

(b) Pour valeurs actuelles.

FIGURE 6.3 – Distributions relativisées des sources par rapport aux communautés.

Cette distribution relative nous a permis donc de cataloguer les médias par la façon dont ils sont utilisés par les communautés. Avec l'aide de politologues en présence, nous avons pu fixer les candidats sur un axe politique et en récupérer une note pour chaque média [6]. Mais ceci nous a permis de comparer la 1-diversité relativisée par rapport à cet axe, que nous avons repris de cette collaboration sur la figure 6.4.

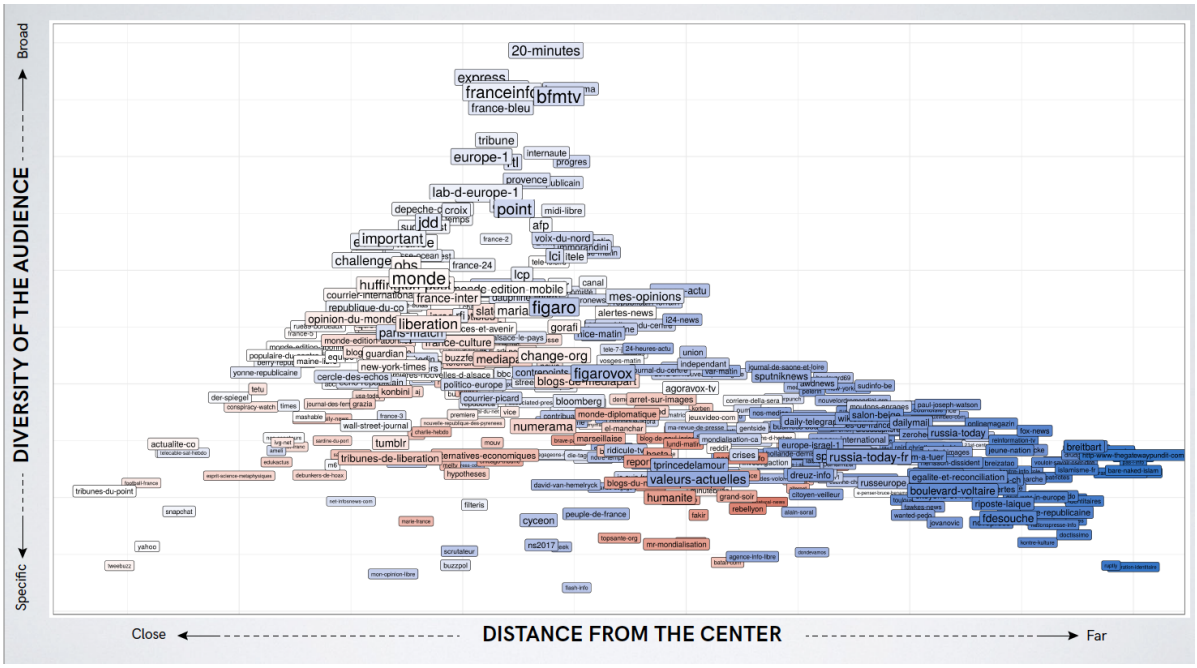


FIGURE 6.4 – Diversité des médias par rapport à leur position politique.

Nous voyons qu'il y a une plus forte diversité sur les médias situés au centre. Notons ici que

6. Ceci a été fait pour la visualisation car nous n'avons pas mené à terme cette méthode.

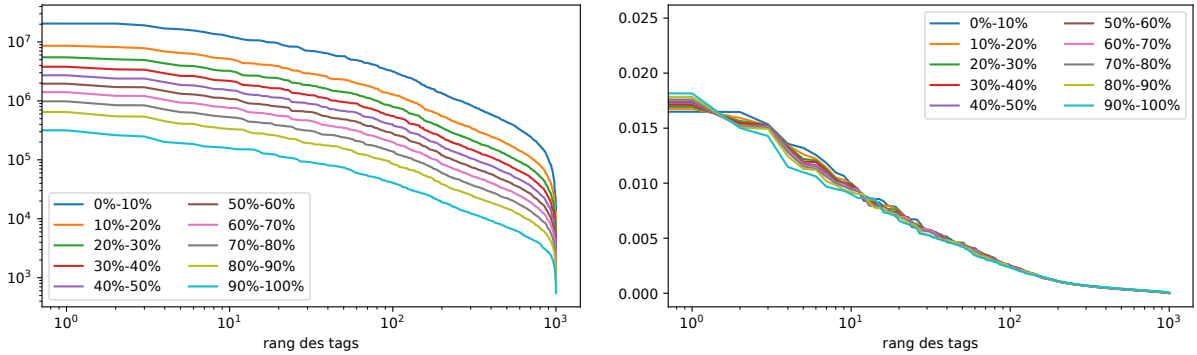
notre façon de catégoriser les médias et la manière de calculer la diversité sont basées sur la même chose : la distribution relative, ce qui peut aussi expliquer en partie cette corrélation.

Ce travail nous montre en quoi les données Twitter peuvent faire assez facilement une bonne base d'analyse pour nos indicateurs. Comme nous l'avons dit précédemment il est cependant nécessaire d'avoir des connaissances sur les données, ici médiatiques et politiques, pour pouvoir créer un triparti pertinent, et en tirer des conclusions utilisables.

6.2 Perfectionnement de nos méthodes.

Regardons ici comment certains pans de nos méthodes peuvent être utilisés et perfectionnés. Nous regarderons d'abord comment nous pouvons changer nos indicateurs pour dépasser ou accompagner l'α-diversité. Enfin nous présenterons deux sujets que nous avons déjà traités mais qui mériteraient chacun une étude plus spécifique et plus approfondie : la diversité temporelle ainsi que l'utilisation de la génération aléatoire pour analyser et comprendre la diversité
.

6.2.1 *Dépasser l'α-diversité.*



(a) Sur la distribution non normalisée.

(b) Sur la distribution normalisée.

Figure 6.5 – Analyse de la loi de zipf

Dans tout le manuscrit nous nous sommes focalisé sur la α-diversité, mais nous pouvons regarder d'autres indicateurs de diversité que l'on pourrait définir sur une distribution. Fixons $\langle p_1, \cdots, p_n \rangle$ une distribution. Rappelons nous que nous nous intéressons à l'équilibre de cette distribution. Une première idée que nous pourrions avoir est d'ordonner cette distribution. En effet ceci nous permettrait d'assimiler cette suite de proportion à une suite de points monotone. Si nous avons assez de points on peut imaginer une courbe monotone. Plus cette courbe serait pentue, plus les proportions seraient différentes, moins la distribution serait équilibrée. Formellement on fixe σ une permutation telle que $\langle p_{\sigma(1)}, \cdots, p_{\sigma(n)} \rangle$ est ordonnée ($\forall i$,

$p_{\sigma(i)} \geq p_{\sigma(i+1)}$). Ceci ressemble à l'analyse de la *loi de zipf* [54] qui concerne à l'origine l'occurrence des mots dans un livre. Ici ce qui nous intéresse ce n'est pas vraiment de regarder si on suit cette loi mais juste de faire la même analyse. Nous avons donc tracé cette courbe dans la figure 6.5 pour deux distributions $\langle p \rangle$ différentes mais qui correspondent toutes les deux à l'écoute des tags par les utilisateurs sur *MSD*. La figure 6.5a est plus adaptée à l'analyse du zipf classique : nous regardons pour un utilisateur *l'occurrence* de chaque tag, c'est à dire le nombre de fois qu'un utilisateur écoute une musique de chaque tag,. Pour l'obtenir il suffit de multiplier chacune des proportions par le volume de l'utilisateur. Pour le dire autrement : c'est le nombre de chemins qui arrivent à chaque tag . La figure 6.5b correspond à la distribution séquentielle $1, 2, 3$ classique. Pour chacune des figures nous n'avons pas pris le comportement d'un utilisateur mais le comportement moyen global des utilisateurs. Nous les avons tout de même séparés en groupes selon leur volume. La courbe la plus haute (en bleu foncé) représente les 10% d'utilisateurs ayant les plus gros volumes, la courbe la plus basse (en bleu clair) représente les utilisateurs ayant le plus petit volume. Nous voyons qu'il y a un comportement moyen car les courbes sont parallèles sur la figure 6.5a, celles qui sont plus hautes le sont simplement parce qu'elles représentent des plus hauts volumes, donc les tags ont un plus grand nombre d'occurrences. La figure 6.5b le confirme puisqu'elle relativise en quelque sorte le nombre d'occurrence pour en faire des proportions. Comme il y a ici un comportement moyen, il serait intéressant de l'utiliser pour le comparer à un comportement individuel. Calculer une distance entre la courbe d'un individu et la courbe moyenne permettrait de regarder à quel point il est diversifié par rapport aux autres.

Dans cet état d'esprit de comparaison avec une courbe parfaite nous avons l'**indice de Gini** [87]. Nous n'avons pas beaucoup regardé cet indice dans notre étude mais il serait intéressant de le calculer et de le comparer à nos indicateurs.

Nous avons dans toute cette étude considéré que toutes les catégories étaient aussi distinctes les unes que les autres. Ce qui n'est pas évident. Si un utilisateur écoute 50% de black métal et 50% de hard métal est-il vraiment aussi diversifié qu'un utilisateur qui écoute 50% de black métal et 50% de musique classique ? Évidemment nous pourrons reprocher ça au système de catégorisation mais nous pouvons proposer ici une autre méthode. Supposons que l'ensemble des catégories soit muni d'une distance d. Pour i et j des catégories le nombre $p_i * p_j d(i,j)$ serait grand si les deux proportions sont grandes et que la distance est grande. De plus le produit $p_i * p_j$ est plus grand si ces deux nombres sont proches que si ils sont éloignés. Pour l'illustrer regardons la moyenne des deux $m = \frac{p_i + p_j}{2}$: $p_j * p_i = (m - (p_j - m)) * (m - (p_i - m) = m^2 - (p_j - m)^2$. Ainsi $m^2 > p_i * pj$. Ainsi ce nombre $p_i * p_j d(i,j)$ est d'autant plus grand qu'il note l'équilibre entre deux parties fortement distantes. Nous pourrions donc considérer ce nombre $D = \sum_i \sum_j p_i * p_j d(i,j)$ qui nous donnerait un bon indice de diversité.

Ceci nous donne une piste que nous n'avons pas non plus étudiée. Tout d'abord parce que nous n'avons pas eu la connaissance des données pour fixer une distance sur une couche. Ensuite il se calcule de manière quadratique sur le nombre de catégories, alors que jusqu'ici nos indicateurs étaient linéaires. Cependant ce serait une bonne extension de nos indicateurs d'α-diversité.

6.2.2 *Diversité et dimension temporelle.*

Nous avons abordé dans la section 5.1 ce que la dimension temporelle peut apporter à l'analyse de la diversité. La première solution que nous avons adoptée est de décomposer notre graphe en plusieurs graphes, chacun associé à une période de temps. Cette méthode revient à fixer une durée T et à prendre des photos du graphe pendant une période de taille T. Dans notre étude nous avons un petit peu complexifié ce système puisque nous avons pris une période d'un an, mais nous l'avons fait démarrer chaque mois et non chaque année. Ainsi nous avons fait *glisser* nos photos. Cependant, comme la moyenne des diversités n'est pas la diversité moyenne, avoir la diversité de tous les mois d'une année ne nous donne pas la diversité de cette année. Il nous faut donc fixer bien la temporalité qui nous intéresse et il est un peu problématique de devoir la fixer en créant les graphes. Pour cela nous pourrions redéfinir notre méthode en intégrant directement la dimension temporelle dans les graphes. Il serait intéressant de redéfinir nos indicateurs dans ce contexte, probablement en utilisant des *link stream* ([52]). Ce formalisme inscrit les nœuds et les graphes dans la durée. Et nous pouvons regarder comment nos marches aléatoires se redéfiniraient sur ce formalisme. Nous pouvons aussi trouver un intermédiaire plus simple en horodatant simplement les liens. Notre algorithme peut simplement calculer la marche aléatoire en vérifiant que la date des liens correspond à la période voulue. Cependant le problème des *impasses* serait encore plus présent et nous ne pouvons pas seulement le supprimer en créant le graphe puisque des impasses pourraient exister pendant une période mais pas durant la période globale, puisqu'il y a moins de lien. De plus, nous avons défini notre algorithme fondamentalement en calculant les distributions puis nos indicateurs en détruisant les distributions (puisqu'il peut être complexe en espace de garder toutes les distributions). Ici il serait intéressant de définir une notion de distribution évolutive pour garder l'information nécessaire pour calculer nos indicateurs.

6.2.3 *Génération aléatoire.*

Nous avons accordé une bonne partie de notre étude et une section à ce sujet. Cependant il offre une multitude de variables ce qui mériterait là aussi de continuer à être analysé. Tout d'abord nous avons regardé un modèle de génération de graphe à partir d'un graphe fixé. Nous avons exploré la conservation du degré de certaines couches, et, grâce à nos formules analytiques, pu analyser une forme de redondance. Notons d'abord que cette dernière formule ne correspond pas à une façon de générer des graphes que nous aurions définie. Il serait intéressant de définir d'autres méthodes qui prennent en compte la redondance. Ici aussi la redondance se décline en plusieurs choses : est-ce les utilisateurs qui réécoutent les même musiques, ou alors qui réécoutent des musiques du même type (tag). Ou encore est-ce les tags qui sont plus écoutés par un type d'utilisateurs... Ensuite nous avons décidé de nous intéresser aux générations aléatoires qui préservent la couche de départ (ou au moins les degrés) puisque nous voulions garder au mieux le volume. Il en est de même pour les formules analytiques que nous avons montrées, elles sont fortement liées à notre sens de génération. Que donneraient-elles dans le cas où c'est la couche de départ qui varie ? De plus nous n'avons montré que des formules s'appliquant à la 2-diversité et la 0-diversité. Qu'en est-il des autres indicateurs ? Enfin la géné-

ration aléatoire de graphe à partir de distributions montre à elle toute seule toute la variabilité du sujet. En effet nous avions fixé seulement 3 lois alors que d'autres distributions existent. Nous n'avons regardé qu'un petit jeu de paramètres pour chaque loi. Et, même si le cadre est assez général, il garde la spécificité précédente en se focalisant seulement sur les degrés.

Il y a donc une grande étude à faire sur le sujet qui est évidemment plus grande que ce que nous avons pu traiter. Cependant la partie que nous avons explorée nous parait cohérente puisqu'elle est basée sur ce qui nous semblait influencer le plus nos indicateurs (par exemple le volume des éléments de départ). Elle nous a même permis de découvrir d'autres choses qui l'influençaient, certaines redondances, et les degrés des individus d'arrivée. Nous n'avons donc exploré qu'une partie du grand champ possible mais qui nous a fortement aidé dans la normalisation et la compréhension de nos résultats.

6.3 ANALYSE DES ALGORITHMES DE RECOMMANDATION.

Nous avons motivé notre travail par de l'analyse de la recommandation mais il était nécessaire de construire nos méthodes et de les perfectionner avant de les appliquer. Regardons ici comment nous pourrions utiliser notre travail pour analyser l'influence de la recommandation. Symétriquement à ce que nous avons fait dans l'introduction, nous allons regarder ceci dans deux cas possibles : quand l'information est sous forme de trace, nous regardons donc ceci de l'extérieur ; ou au contraire quand nous avons accès aux algorithmes de l'intérieur, accès au code, aux données, etc.

6.3.1 *Quand l'information est traitée de l'extérieur.*

Tout d'abord, ce qui est le plus proche de ce que nous avons fait, nous pouvons analyser des traces d'utilisateurs et comparer par exemple différentes plateformes. Si ces plateformes fournissent un système de recommandation nous pouvons analyser globalement son influence. Ensuite nous pouvons aussi simuler différents algorithmes et regarder comment ils influenceraient la diversité des choix des utilisateurs (eux aussi simulés). En supposant par exemple qu'il suivent toujours ce qui est recommandé, ou alors en mettant une probabilité de choix pour chacun. En regardant l'influence de ces différents algorithmes et de leurs paramètres sur la diversité collective ou sur les diversités individuelles nous pourrons juger de la diversité créée par un algorithme. Si nous avons accès à des données de recommandation, par exemple si nous avons des listes de recommandations, nous pouvons créer le nœud qu'elle représente. Par exemple en reprenant l'exemple de *MSD* : une recommandation serait une liste de musiques. Nous pourrions créer un nœud par liste et le relier avec chacune des musiques de la liste. Ainsi nous aurions une couche parallèle à celle des utilisateurs représentant les recommandations que nous pourrions étudier. Une question importante serait de savoir comment représenter l'ordre de la liste de recommandation. Nous pourrions utiliser une pondération particulière ou simplement tronquer la liste quand elle est trop longue, puisqu'on sait qu'elle ne sera majoritairement pas regardée. Si nous avons accès aux deux informations en même

temps, c'est à dire les traces de consommations et les informations sur les recommandations, ceci nous permet de combiner ces questions. Là l'analyse des diversités selon plusieurs chemins rentrerait en compte. Pour regarder toutes ces évolutions, il serait nécessaire aussi de prendre en compte la dimension temporelle comme nous l'avons fait et expliqué.

6.3.2 *Quand nous avons accès à l'intérieur.*

Imaginons encore une fois que nous avons une plateforme comme celle de *MSD*. Nous pourrons donc tester plusieurs algorithmes de recommandation sur des groupes d'utilisateurs et regarder la diversité d'écoute des utilisateurs pour voir l'influence des différents algorithmes. De plus nous pouvons insérer à l'intérieur de la recommandation un poids pour améliorer la diversité. La fonction de poids pourrait évidemment être calquée sur nos indicateurs. Enfin nous pourrions implémenter pour l'utilisateur des indicateurs représentant sa propre diversité en temps réel. Notons qu'intégrer la dimension temporelle aux algorithmes de classement, notamment le faire en temps réel, est déjà un problème en soi [5]. Il y a donc beaucoup de choses que nous pourrions essayer en se basant sur ce travail pour analyser et améliorer la diversité induite par les algorithmes de recommandation.

6.4 Une étude quantitative complémentaire avec des connaissances qualitatives.

Toute cette étude peut être presque intégralement qualifiée de *quantitative*. En effet nous analysons des graphes c'est à dire que les seules informations que nous avons sont les nœuds, les liens et la pondération des liens. Même si la nature des couches nous indique le sens des nœuds et a fortiori des liens, il est important de se rendre compte que ce modèle de représentation supprime énormément d'information. Par exemple quand un utilisateur écoute une musique, l'écoute-t-il en travaillant ? Veut-il la faire écouter à ses amis ? Il est possible que la musique soit lue automatiquement et qu'il ne soit même pas devant son ordinateur. Il serait impossible de traiter des dizaines de millions de liens comme nous l'avons fait en considérant toute cette information (il serait d'ailleurs déjà compliqué de la collecter). C'est donc l'avantage et l'inconvénient de notre étude. Nous nécessitons une forte connaissance qualitative des données. D'abord avant d'appliquer nos méthodes : pour construire le graphe (notamment la caractérisation des éléments, nous avons vu à quel point nos indicateurs y sont sensibles). Pendant : pour savoir lesquels appliquer et pour savoir quelles sont les informations à trouver. Et après : pour interpréter les différents scores. Revenons un petit peu sur la caractérisation des éléments. Nous pourrions aussi mêler notre travail à des façons informatiques utilisées pour catégoriser des éléments[55] [82].

C'est dans cette optique que nous avons défini nos méthodes et codé notre programme. En effet les méthodes sont assez générales pour s'appliquer à un ensemble très divers de données et notre programme est assez interactif pour pouvoir calculer à la demande tout ce qui a été défini ici. Ce programme a été, de plus fait pour être efficace. Les plus gros jeux de données

que nous avons traités étaient composé de 10^8 lignes (chaque ligne représentera un lien) et nous avons pu calculer tous nos indicateurs en à peu près 40 minutes [7] (la première moitié du temps étant utilisée pour importer le graphe en mémoire). Notre travail est donc fondamentalement un travail fait pour la collaboration avec des scientifiques provenant d'autres disciplines scientifiques. Notamment il pourrait être utilisé en sociologie du numérique pour apporter des réponses aux auteurs d'Imaginer la sociologie numérique qui se demandent[10] "comment le développement des écosystèmes numériques transforme-t-il actuellement les manières de savoir sur le social ".

7. Sur un intel core i7 : 16M de mémoire et 8 cœurs.

Bibliographie

[1] *L'Autre Coopération*. Ministère des Affaires Étrangères et le Ministère de la Coopération, 1991.

[2] J Aczél and C Alsina. Synthesizing judgements : A functional equations approach. *Mathematical Modelling*, 9(3-5) :311–320, 1987.

[3] János Aczél and Zoltán Daróczy. On measures of information and their characterizations. *New York*, page 168, 1975.

[4] Charu C Aggarwal and S Yu Philip. Semantic based collaborative filtering, November 26 2002. US Patent 6,487,539.

[5] Marie Al-Ghossein, Pierre-Alexandre Murena, Talel Abdessalem, Anthony Barré, and Antoine Cornuéjols. Adaptive collaborative topic modeling for online recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 338–346, 2018.

[6] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2) :211–36, 2017.

[7] W Ross Ashby. Requisite variety and its implications for the control of complex systems. In *Facets of systems science*, pages 405–417. Springer, 1991.

[8] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. Optimal greedy diversity for recommendation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[9] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239) :1130–1132, 2015.

[10] Nicolas Baya-Laffite and Bilel Benbouzid. Présentation : imaginer la sociologie numérique. *Sociologie et sociétés*, 49(2) :5–32, 2017.

[11] Wolfgang H Berger and Frances L Parker. Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937) :1345–1347, 1970.

[12] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[13] T Martijn Bezemer and Wim H Van Der Putten. Ecology : diversity and stability in plant communities. *Nature*, 446(7135) :E6, 2007.

[14] Julia Cagé, Nicolas Hervé, Marie-Luce Viaud, et al. L'information à tout prix. Technical report, 2017.

[15] Sylvain Castagnos, Armelle Brun, and Anne Boyer. When diversity is needed... but not expected ! 2013.

[16] Pablo Castells, Neil J Hurley, and Saul Vargas. Novelty and diversity in recommender systems. In *Recommender systems handbook*, pages 881–918. Springer, 2015.

[17] Satya R Chakravarty and Wolfgang Eichhorn. An axiomatic characterization of a generalized index of concentration. *Journal of Productivity Analysis*, 2(2) :103–112, 1991.

[18] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

[19] David Chavalarias. Fake news : l'arbre qui cache la forêt après le brésil, à qui le tour ? 2018.

[20] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[21] Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3) :261–273, 2008.

[22] Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119) :348–361, 1920.

[23] Aisling J Daly, Jan M Baetens, and Bernard De Baets. Ecological diversity : measuring the unmeasurable. *Mathematics*, 6(7) :119, 2018.

[24] William H Dutton, Sharon Eisner Gillett, Lee W McKnight, and Malcolm Peltu. Bridging broadband internet divides : reconfiguring access to enhance communicative power. *Journal of Information Technology*, 19(1) :28–38, 2004.

[25] Freeman J Dyson. Statistical theory of the energy levels of complex systems. i. *Journal of Mathematical Physics*, 3(1) :140–156, 1962.

[26] David Encaoua and Alexis Jacquemin. Degree of monopoly, indices of concentration and threat of entry. *International economic review*, pages 87–105, 1980.

[27] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33) :E4512–E4521, 2015.

[28] Paul Geroski et al. The choice between diversity and scale. *Davis, E., Geroski, PA, Kay, J. A., Manning A., Smales, C., Smith, SR and Szymanski, S.(eds.)*, pages 29–45, 1992.

[29] Jack P Gibbs and Walter T Martin. Urbanization, technology, and the division of labor : International patterns. *American sociological review*, pages 667–677, 1962.

[30] S.E. Gillett. In praise of policy diversity, position paper for oii broadband forum. *None*, 2003.

[31] Corrado Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121) :124–126, 1921.

[32] Nicholas J Gotelli and Robert K Colwell. Estimating species richness. *Biological diversity : frontiers in measurement and assessment*, 12 :39–54, 2011.

[33] Gernot Grabher and David Stark. Organizing diversity : evolutionary theory, network analysis and postsocialism. *Regional studies*, 31(5) :533–544, 1997.

[34] Marshall Hall and Nicolaus Tideman. Measures of concentration. *Journal of the american statistical association*, 62(317) :162–168, 1967.

[35] Leslie Hannah and John Anderson Kay. *Concentration in modern industry : Theory, measurement and the UK experience*. Springer, 1977.

[36] Ruining He and Julian McAuley. Ups and downs : Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517. ACM, 2016.

[37] Gene Helfman, Bruce B Collette, Douglas E Facey, and Brian W Bowen. *The diversity of fishes : biology, evolution, and ecology*. John Wiley & Sons, 2009.

[38] Orris C Herfindahl. Concentration in the us steel industry. *Unpublished PhD. Dissertation, Columbia University*, 1950.

[39] Mark O Hill. Diversity and evenness : a unifying notation and its consequences. *Ecology*, 54(2) :427–432, 1973.

[40] Albert O Hirschman. *National power and the structure of foreign trade*. University of California Press, 1945.

[41] Sönke Hoffmann. *Concavity and additivity in diversity measurement : re-discovery of an unknown concept*. Univ., FEMM, 2006.

[42] Stuart H Hurlbert. The nonconcept of species diversity : a critique and alternative parameters. *Ecology*, 52(4) :577–586, 1971.

[43] Amin Javari and Mahdi Jalili. A probabilistic model to resolve diversity–accuracy challenge of recommendation systems. *Knowledge and Information Systems*, 44(3) :609–627, 2015.

[44] Lou Jost. Entropy and diversity. *Oikos*, 113(2) :363–375, 2006.

[45] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage : a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1) :1–42, 2016.

[46] Andrei Nikolaevich Kolmogorov and Guido Castelnuovo. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, 1930.

[47] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79–86, March 1951.

[48] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Quantifying search bias : Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 417–432, 2017.

[49] Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P Gummadi, and Saptarshi Ghosh. Characterizing information diets of social media users. In *ICWSM*, pages 218–227, 2015.

[50] Ni Lao and William W Cohen. Fast query execution for retrieval models based on path-constrained random walks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 881–888. ACM, 2010.

[51] Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1) :53–67, 2010.

[52] Matthieu Latapy, Tiphaine Viard, and Clémence Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8(1) :61, 2018.

[53] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380) :1094–1096, 2018.

[54] Wentian Li. Zipf's law everywhere. *Glottometrics*, 5 :14–21, 2002.

[55] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3) :329–348, 2014.

[56] Robert H MacArthur. Patterns of species diversity. *Biological reviews*, 40(4) :510–533, 1965.

[57] Robert M May. Patterns of species abundance and diversity. *Ecology and evolution of communities*, pages 81–120, 1975.

[58] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52. ACM, 2015.

[59] Kevin Shear McCann. The diversity–stability debate. *Nature*, 405(6783) :228, 2000.

[60] Samuel J McNaughton. Diversity and stability of ecological communities : a comment on the role of empiricism in ecology. *The American Naturalist*, 111(979) :515–525, 1977.

[61] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather : Homophily in social networks. *Annual review of sociology*, 27(1) :415–444, 2001.

[62] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm : an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7) :959–977, 2018.

[63] Mitio Nagumo. Über eine klasse der mittelwerte. In *Japanese journal of mathematics : transactions and abstracts*, volume 7, pages 71–79. The Mathematical Society of Japan, 1930.

[64] Masatoshi Nei. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3) :583–590, 1978.

[65] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2) :167–256, 2003.

[66] Phuong Nguyen, Pier-Paolo Saviotti, Michel Trommetter, and Bernard Bourgeois. Variety and the evolution of refinery processing. *Industrial and corporate change*, 14(3) :469–500, 2005.

[67] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. Measuring online social bubbles. *PeerJ Computer Science*, 1 :e38, 2015.

[68] Safiya Umoja Noble. *Algorithms of oppression : How search engines reinforce racism*. nyu Press, 2018.

[69] Helga Nowotny, Peter Scott, Michael Gibbons, and Peter B Scott. *Re-thinking science : Knowledge and the public in an age of uncertainty*. Cambridge : Polity Press, 200.

[70] Eugene P Odum. *Fundamentals of ecology*. WB Saunders company, 1959.

[71] Eli Pariser. *The filter bubble : What the Internet is hiding from you*. Penguin UK, 2011.

[72] Rémy Poulain and Fabien Tarissan. Quantifying the diversity in users activity : an example study on online music platforms. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 3–10. IEEE, 2018.

[73] Rémy Poulain and Fabien Tarissan. Investigating the lack of diversity in user behavior : The case of musical content on online platforms. *Information processing & management*, 57(2) :102169, 2020.

[74] Heikki P Pursiainen. Consistency in aggregation, quasilinear means and index numbers. *Quasilinear Means and Index Numbers (December 4, 2008)*, 2008.

[75] Lijing Qin and Xiaoyan Zhu. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[76] Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. Measuring diversity in heterogeneous information networks. *arXiv*, pages arXiv–2001, 2020.

[77] C Radhakrishna Rao. Rao's axiomatization of diversity measures. *Wiley StatsRef : Statistics Reference Online*, 2014.

[78] Alfréd Rényi. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Contributions to the Theory of Statistics*, pages 547–561, Berkeley, CA, 1961. University of California Press.

[79] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Contributions to the Theory of Statistics*, pages 547–561. The Regents of the University of California, 1961.

[80] Stephen A Rhoades. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79 :188, 1993.

[81] Jacques Riget and Jakob S Vesterstrøm. A diversity-guided particle swarm optimizer-the arpso. *Dept. Comput. Sci., Univ. of Aarhus, Aarhus, Denmark, Tech. Rep*, 2 :2002, 2002.

[82] Céline Robardet. *Contribution à la classification non supervisée : proposition d'une méthode de bi-partitionnement*. PhD thesis, Lyon 1, 2002.

[83] BRUCE Runnegar, KSW Campbell, and MF Day. Rates and modes of evolution in the mollusca. *Rates of evolution. Allen and Unwin, London*, pages 39–60, 1987.

[84] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.

[85] Eric D Schneider and James J Kay. Life as a manifestation of the second law of thermodynamics. *Mathematical and computer modelling*, 19(6-8) :25–48, 1994.

[86] David O Sears and Jonathan L Freedman. Selective exposure to information : A critical review. *Public Opinion Quarterly*, 31(2) :194–213, 1967.

[87] Amartya Sen, Master Amartya Sen, Sen Amartya, James E Foster, James E Foster, et al. *On economic inequality*. Oxford University Press, 1997.

[88] Claude E Shannon and Warren Weaver. The mathematical theory of communication. 1949. *Urbana, IL : University of Illinois Press*, 1963.

[89] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3) :379–423, 1948.

[90] Elena V Shevchenko, Dmitri V Talapin, Nicholas A Kotov, Stephen O'brien, and Christopher B Murray. Structural diversity in binary nanoparticle superlattices. *Nature*, 439(7072) :55, 2006.

[91] Gerald Silverberg, Giovanni Dosi, and Luigi Orsenigo. Innovation, diversity and diffusion : A self-organisation model. *The Economic Journal*, 98(393) :1032–1054, 1988.

[92] Edward H Simpson. Measurement of diversity. *nature*, 1949.

[93] John Maynard Smith. Trees, bundles or nets ? *Trends in ecology & evolution*, 4(10) :302–304, 1989.

[94] Andrew R Solow and Stephen Polasky. Measuring biological diversity. *Environmental and Ecological Statistics*, 1(2) :95–103, 1994.

[95] Yicheng Song, Nachiketa Sahoo, and Elie Ofek. When and how to diversify—a multicategory utility model for personalized content recommendation. *Management Science*, 65(8) :3737–3757, 2019.

[96] Andrew Stirling. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28 :1–156, 1998.

[97] Andy Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15) :707–719, 2007.

[98] Fabien Tarissan. *Au cœur des réseaux. Des sciences aux citoyens*. 2019.

[99] Rasmus K Ursem. Diversity-guided evolutionary algorithms. In *International Conference on Parallel Problem Solving from Nature*, pages 462–471. Springer, 2002.

[100] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7) :3797–3820, 2014.

[101] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.

[102] Ge Wang and Ming Jiang. Axiomatic characterization of nonlinear homomorphic means. *Journal of Mathematical Analysis and Applications*, 303(1) :350–363, 2005.

[103] Martin L Weitzman. On diversity. *The Quarterly Journal of Economics*, 107(2) :363–405, 1992.

[104] PAUL H Williams, CJ Humphries, and RI Vane-Wright. Measuring biodiversity : taxonomic relatedness for conservation priorities. *Australian systematic botany*, 4(4) :665–679, 1991.

[105] Han-Xin Yang, Zhi-Xi Wu, Changsong Zhou, Tao Zhou, and Bing-Hong Wang. Effects of social diversity on the emergence of global consensus in opinion dynamics. *Physical Review E*, 80(4) :046108, 2009.

[106] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10) :4511–4515, 2010.

[107] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.