# Corrélations dans les graphes d'information hétérogène : prédiction et modélisation de liens à partir de méta-chemins

Hông-Lan Botterman


Thèse de doctorat de Sorbonne Université
Spécialité informatique
École Doctorale Informatique, Télécommunication et Électronique (Paris)

Présentée par
Hông-Lan Botterman

Pour l'obtention du grade de
Docteur de Sorbonne Université

# Corrélations dans les graphes d'information hétérogène

## Prédiction et modélisation des liens à partir de méta-chemins

Soutenue le 19 novembre 2020 devant le jury composé de :

Raphaël Fournier-S'niehotta, maître de conférences, CNAM ..... examinateur
Robin Lamarche-Perrin, chargé de recherche, CNRS ..... directeur
Matthieu Latapy, directeur de recherche, CNRS ..... examinateur
Clémence Magnien, directrice de recherche, CNRS ..... directrice
Gilles Trédan, chargé de recherche, HDR, CNRS ..... rapporteur
Julien Velcin, professeur, Université Lumière Lyon 2 ..... rapporteur

# Corrélations dans les graphes d'information hétérogène

Prédiction et modélisation des liens à partir de méta-chemins

**Résumé :** De nombreuses entités possiblement de natures différentes sont reliées par des liens (physiques ou virtuels) pouvant également être de natures différentes. De telles données peuvent être représentées par un graphe d'information hétérogène (*heterogeneous information network*, HIN). De plus, il existe souvent des corrélations entre entités ou événements de la vie réelle. Une fois ces derniers représentés par des abstractions appropriées telles que les HIN, les corrélations peuvent dès lors se retrouver dans ces graphes particuliers. Motivé par ces considérations, cette thèse s'intéresse aux effets des possibles corrélations entre les liens d'un HIN sur sa structure. Ce présent travail tente de répondre à des questions telles que : y a-t-il effectivement des corrélations entre les liens de différents types ? Si oui, est-il possible de les quantifier ? Que signifient-elles ? Comment les interpréter ? Est-ce que ces corrélations peuvent servir à prédire l'apparition de liens ? À modéliser des dynamiques de co-évolution ? Les exemples étudiés peuvent être divisés en deux catégories. Premièrement, l'utilisation des corrélations pour la prédiction du poids des liens est étudiée. Il est montré que les corrélations entre les liens, et plus particulièrement entre les chemins, peuvent être utilisées pour récupérer et prédire le poids d'autres liens, d'un type spécifié. Deuxièmement, une dynamique de poids de liens est considérée. Il est montré que la co-évolution de liens peut servir, par exemple, à définir un modèle d'attention entre individus et sujets. Les résultats préliminaires sont en accord avec d'autres présents dans la littérature, principalement relatifs aux modèles de dynamiques d'opinions. Globalement, ce travail illustre l'importance des corrélations entre les liens d'un HIN. En outre, il soutient le fait général que différents types de nœuds et liens abondent dans la nature et qu'il peut être important et instructif de prendre en compte cette diversité afin de comprendre l'organisation et le fonctionnement d'un système.

# Correlations in heterogeneous information networks

Prediction and modelling of links from metapaths

**Abstract:** Many entities, possibly of different natures, are linked by physical or virtual links, that may also be of different natures. Such data can be represented by a heterogeneous information network (HIN). In addition, there are often correlations between real-life entities or events. Once represented by suitable abstractions (such as HIN), these correlations can therefore be found in the HIN. Motivated by these considerations, this thesis investigates the effects of possible correlations between the links of an HIN on its structure. This present work aims at answering questions such as: are there indeed correlations between different types of links? If so, is it possible to quantify them? What do they mean? How can they be interpreted? Can these correlations be used to predict the occurrence of links? To model co-evolution dynamics? The examples studied can be divided into two categories. First, the use of correlations for the prediction of the links' weight is studied. It is shown that correlations between links, and more specifically between paths, can be used to recover and, to some extent, predict the weight of other links of a specified type. Second, a link weight dynamics is considered. It is shown that link co-evolution can be used, for example, to define a model of attention between individuals and subjects. The preliminary results are in agreement with others in the literature, mainly related to models of opinion dynamics. Overall, this work illustrates the importance of correlations between the links of an HIN. In addition, it supports the general fact that different types of nodes and links abound in nature and that it could be important and instructive to take this diversity into account in order to understand the organization and functioning of a system.

# Remerciements

Tout d'abord, j'aimerais remercier mes deux encadrants, Robin Lamarche-Perrin et Clémence Magnien. Robin, merci pour ton encadrement tout au long de ce travail. Clémence, merci pour tes relectures et remarques toujours pertinentes.

Merci à Gilles Trédan et Julien Velcin d'avoir accepté de rapporter ce travail ; à Raphaël Fournier-S'niehotta et Matthieu Latapy de participer au jury ; à Raphaël Fournier-S'niehotta et Rushed Kanawati de faire partie de mon comité de suivi.

Je remercie également toutes celles et ceux que j'ai croisés au cours de ces trois années et qui ont contribué, de quelque façon que ce soit, à la réalisation de ce travail.

Enfin, sur un plan plus personnel, j'aimerais remercier ma famille et mes amis pour tout ce qu'ils sont ! Savoir que vous avez de telles personnes à vos côtés chaque fois que vous en avez besoin vous donne la force d'affronter bien des situations. Un merci tout particulier à mes parents pour avoir fait de moi qui je suis et m'avoir donné, parmi tant d'autres choses, la chance d'apprendre.

Ce projet a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union Européenne dans le cadre de la convention de subvention No 732942 (ODYCCEUS).

# Table des matières

**Résumé** . . . iii

**Remerciements** . . . v

**Introduction** . . . 1

**1. Généralités sur les graphes d'information hétérogène** . . . 5
1.1. Choix des graphes d'information hétérogène . . . 5
1.2. Définitions . . . 7
1.2.1. Graphes d'information hétérogène . . . 7
1.2.2. Hypergraphes d'information hétérogène . . . 10
1.2.3. Projection d'hypergraphes . . . 13
1.3. Marches aléatoires contraintes . . . 15
1.3.1. Marches aléatoires contraintes dans les HIN . . . 16
1.3.2. Marches aléatoires contraintes dans les hypergraphes . . . 17
1.4. Résumé . . . 20

**2. Positionnement par rapport au problème de prédiction de liens** . . . 21
2.1. Prédiction et récupération de liens et de leur poids . . . 22
2.1.1. Prédiction et récupération . . . 22
2.1.2. Existence et pondération . . . 22
2.1.3. Évaluation de la prédiction . . . 23
2.2. Difficultés, motivations et positionnement . . . 23
2.3. Méthodes de prédiction . . . 26
2.3.1. Prédiction de l'existence des liens dans les HIN . . . 26
2.3.2. Prédiction du poids des liens dans des graphes multiplexes . . . 29
2.3.3. Prédiction du poids des liens dans les graphes d'information homogène . . . 30
2.3.4. Ajout d'information d'ordre supérieur . . . 34
2.4. Résumé . . . 35

**3. Prédiction de liens manquants par marches aléatoires contraintes** . . . 37
3.1. Approche . . . 38
3.1.1. Modèle de régression linéaire . . . 38
3.1.2. Sélection de modèles et variables explicatives . . . 39
3.1.3. Tâches descriptives et prédictives . . . 40
3.1.4. Discussion des choix . . . 41
3.2. Premier exemple d'application : expliquer et prédire un type de liens . . . 43
3.2.1. Cas d'étude : la coupe du monde 2014 sur Twitter . . . 43
3.2.2. Première régression et interprétation de la solution obtenue . . . 44
3.2.3. Pouvoir prédictif de la solution obtenue . . . 46
3.2.4. Équilibre entre qualité et complexité de la solution . . . 47
3.3. Aller plus loin : interprétation et structure des méta-chemins . . . 50
3.3.1. Cas d'étude : co-publication d'articles scientifiques sur DBLP . . . 50
3.3.2. Orientation des liens au sein des méta-chemins . . . 53
3.3.3. Partitionnement thématique des méta-chemins . . . 54
3.3.4. Décomposition des méta-chemins en sous-graphes . . . 55
3.3.5. Contrôler la qualité des solutions . . . 57
3.4. Analyse empirique d'un exemple complexe . . . 59
3.4.1. Cas d'étude : échange d'information entre individus et médias sur Twitter . . . 59
3.4.2. Régressions linéaires au cours du temps . . . 63

3.4.3. Mesures additionnelles pour expliquer le résultat de la régression . . . . . 65
3.4.4. Informations exogènes pour contextualiser et interpréter le résultat de la régression . . . . . 69
3.5. Intérêt des hypergraphes pour la prédiction de liens . . . . . 71
3.5.1. Représentation hypergraphe et objectif . . . . . 71
3.5.2. Résultats . . . . . 74
3.5.3. Bilan . . . . . 75
3.6. Résumé et perspectives . . . . . 75

**4. Modèle d'évolution du poids des liens . . . . . 81**
4.1. Modéliser l'influence sociale : un problème difficile . . . . . 82
4.1.1. Complexité de l'influence sociale . . . . . 82
4.1.2. Une physique des opinions . . . . . 82
4.2. Évolution des liens d'un HIN . . . . . 84
4.2.1. Expression générale du modèle . . . . . 85
4.2.2. Définition de $\mathbf{T}^{\epsilon}_{\mathcal{P}}(t)$ : marches aléatoires contraintes et limitées . . . . . 86
4.3. Modèle d'attention d'individus . . . . . 90
4.3.1. Deux distributions d'attention initiales . . . . . 91
4.3.2. Trois distributions d'attention initiales . . . . . 94
4.3.3. Influence d'un média . . . . . 101
4.3.4. Discussion d'une extension du modèle aux hypergraphes . . . . . 105
4.4. Résumé et perspectives . . . . . 109

**Conclusion . . . . . 111**

**Références . . . . . 117**

# Introduction

De nombreux systèmes réels consistent en une multitude d'interactions entre différentes entités. C'est le cas par exemple des activités sociales où les individus communiquent ensemble, des réactions chimiques où les produits réagissent entre eux, des espèces en compétition dans un écosystème, etc. Toutes ces situations peuvent être abstraites par des graphes d'information [154]. Dans sa forme la plus simple, un graphe d'information est un ensemble de nœuds potentiellement reliés entre eux par des liens.

Cependant, nombres de systèmes font intervenir des entités de natures différentes interagissant entre elles de plusieurs façons. Prenons par exemple un système discographique. Les entités peuvent représenter des interprètes, des compositeurs, des paroliers, des titres, des styles musicaux, etc. Les relations entre ces entités sont dès lors différentes : un interprète *interprète* un titre tandis qu'un compositeur le *compose* ; la relation entre compositeur-parolier est aussi différente qu'entre interprète-parolier. Un autre exemple souvent utilisé sont les relations purement sociales. Deux personnes peuvent être e.g. amis, apparentés, collègues : les individus peuvent donc être connectés par différents types de relations. On peut aussi les différencier par leur genre, leur âge, etc., ce qui permet de distinguer des groupes d'individus.

Afin de prendre en compte ces informations, le formalisme des *graphes d'information hétérogène* (*heterogeneous information network*, HIN) a été proposé [155]. Grossièrement, un HIN est un graphe dont les nœuds et liens sont typés, étiquetés. Ces graphes permettent ainsi de tenir compte (d'une partie) de la *sémantique* des données à analyser. Du grec *sêmantikos* – signifié – le terme sémantique est utilisé pour parler de la signification des nœuds et liens et, par extension, de la signification des chemins, i.e. des séquences de liens, dans un HIN.

Le concept de *méta-chemin*, i.e. une séquence de types de nœuds et de types de liens, s'est révélé extrêmement utile pour mettre en avant la sémantique dans l'analyse de HIN. Il fournit une sémantique naturelle à un chemin dans un HIN. Dans l'exemple discographique, un interprète $i$ peut être connecté à un genre $g$ par le méta-chemin "interprète-titre-genre" mais aussi par le méta-chemin "interprète-compositeur-titre-genre". Dans le premier cas, $i$ interprète effectivement un titre qui appartient au genre $g$ tandis que dans le second, $i$ est connecté à $g$ parce qu'une personne ayant

composé pour lui a également composé un titre du genre $g$, mais cela ne signifie pas que $i$ chante des titres du genre $g$. Grâce aux méta-chemins, on peut donc systématiquement spécifier la façon dont les nœuds sont connectés dans un HIN. Il s'agit d'un concept central de ce travail.

Une grande partie des évènements quotidiens ne se produisent pas par hasard. Ils résultent bien souvent d'une suite d'actions, de décisions, de faits passés. Dans ces cas, une relation évidente de cause-effet est identifiable. Cependant, ce n'est pas toujours aussi évident et il est parfois impossible d'identifier la cause de l'effet. Soit parce que le temps est ignoré, il n'y a donc pas de notion de cause ni d'effet, soit parce que les évènements sont tous des conséquences d'une cause commune (inconnue), ou encore parce que les évènements sont accidentellement liés. Quelle que soit la raison, nous parlons alors de *corrélations*. Du latin *correlatio*, lui-même de *cum* – avec – et *relatio* – relation, une corrélation est une mesure statistique (souvent exprimée par un nombre) décrivant la "force" d'une relation entre deux ou plusieurs variables. Le concept de corrélation statistique a été introduit par Galton [56] et largement repris par la suite, avec les succès qu'on lui connaît. Le terme corrélation est plus faible que celui de causalité – ainsi que le soutenait Pearson (et bien d'autres), la causalité est une corrélation, mais l'inverse n'est pas vrai [124] – mais est souvent la seule chose que l'on puisse observer lorsqu'on se base uniquement sur des outils statistiques.

Il est tentant de penser que les corrélations dans les systèmes réels se traduisent par des corrélations entre les liens des graphes les représentant. Afin d'investiguer l'idée qu'il y a des corrélations entre les liens d'un HIN, nous nous concentrons sur deux problèmes particuliers : la prédiction du poids des liens et la modélisation de l'évolution du poids des liens.

Pour le premier point, nous nous demandons s'il est possible de récupérer et prédire le poids de certains liens, tout en connaissant le poids d'autres liens ainsi que les corrélations possibles des liens dans le HIN. Cela est motivé par des observations dans la vie de tous les jours, que nous illustrons par l'exemple simplifié suivant. Imaginons qu'une personne n'écoute que de la *tropical house*, ainsi que tous ses amis. Par contre, ses collègues écoutent toute sorte de musique, sans aucune préférence. Supposons que cela s'applique pour de nombreuses personnes. Dans le graphe modélisant cette situation (nœuds : individus et musiques, liens : amitié, relation de travail et écoute), il y a une corrélation entre ce qu'un individu écoute et ce que ses amis écoutent. En revanche, il n'y en aura pas entre ce qu'il écoute et ce que ses collègues écoutent. Insistons bien sur le fait qu'il s'agit uniquement de corrélations : il est impossible de dire, sur base de ces données, si on écoute les mêmes musiques que ses amis, ou si on est ami avec les gens qui écoutent le même style de musique que soi.

Pour le second point, i.e. la modélisation de l'évolution du poids des liens, il est tentant de se dire que la plupart des systèmes n'évoluent pas de manière aléatoire mais selon une "loi", prédéfinie

ou non. Dès lors, il serait possible que le poids des liens évolue en fonction du poids d'autres liens dans les HIN, les corrélations des liens dictant l'évolution du système. Reprenons à nouveau l'exemple des musiques mais cette fois-ci, supposons qu'au départ un individu n'écoute pas tout à fait exactement les mêmes musiques que ses amis. Cependant, de par les liens sociaux qui les unissent, ces individus ont tendance à se "rapprocher" en terme d'écoute musicale et vont donc adapter leurs écoutes en fonction de ce que leurs amis respectifs écoutent. Parallèlement, ils ont tendance à se rapprocher socialement des individus qui écoutent les mêmes musiques qu'eux, peu importe s'ils sont initialement amis ou non. Dans ce cas, on voit que les liens sociaux vont influencer les liens de consommation musicale qui eux-mêmes influencent les liens sociaux. On peut donc parler d'une certaine manière de co-évolution des liens en fonction de leurs corrélations.

Dans ce travail, nous nous concentrons sur ces deux exemples de manifestations des corrélations et d'interdépendance dans les HIN. Y a-t-il effectivement des corrélations entre les liens de différents types ? Si oui, est-il possible de les quantifier ? Que signifient-elles ? Comment les interpréter ? Est-ce que ces corrélations peuvent servir à prédire l'apparition de liens ? À modéliser des dynamiques de co-évolution ? Ce travail tente de répondre à ce type de questions.

## Structure du manuscrit

Le chapitre 1 présente les concepts utilisés dans les chapitres suivants du travail. Nous motivons l'utilisation du formalisme des HIN, présentons quelques définitions relatives à ces graphes particuliers ainsi qu'aux hypegraphes d'information hétérogène, généralisation intuitive des HIN. Finalement, nous rappelons le concept de marches aléatoires contraintes par un méta-chemin, concept utilisé tout au long du manuscrit.

Le chapitre 2 discute des méthodes existantes dans la littérature pour résoudre le problème générique de la prédiction de liens. En particulier, nous faisons la distinction entre la prédiction de l'existence d'un lien et la prédiction de son poids. Pour le premier problème, la littérature présentée est relative aux HIN tandis que pour le second problème, nous rapportons des méthodes relatives aux graphes d'information homogène.

Le chapitre 3 constitue la première contribution[1] de ce travail. Il décrit l'approche proposée pour récupérer et prédire le poids des liens dans un HIN : en particulier, une régression basée sur des distributions de probabilités, obtenues par marches aléatoires contraintes par des méta-chemins. Différents cas d'études viennent ensuite illustrer le potentiel de cette approche, mais également

[1] Une partie de ce chapitre a été publiée : Botterman, H.-L. & Lamarche-Perrin, R. Combining path-constrained random walks to recover link weights in heterogeneous information networks. in *International Workshop on Complex Networks* (2019), 97–109.

pointer ses limites.

Le chapitre 4 considère les HINs évoluant au cours du temps et constitue la seconde contribution du travail. Un modèle d'évolution du poids des liens est proposé, également basé sur des méta-chemins. Un exemple d'application est ensuite proposé, à savoir un modèle d'attention : la façon dont l'attention des individus à propos de certains sujets évolue dans le temps. Quelques analyses préliminaires sont effectuées afin de tester la pertinence de l'approche imaginée.

Enfin, le dernier chapitre résume les travaux présentés dans ce manuscrit et discute des orientations possibles pour d'éventuelles futures recherches.

# 1. Généralités sur les graphes d'information hétérogène

Dans ce chapitre, nous introduisons des concepts et définitions de la littérature relatifs aux graphes d'information jugés pertinents pour la compréhension et la cohérence du travail. Ce chapitre n'est pas une présentation exhaustive de tous les objets considérés dans ce travail ; le lecteur intéressé sera renvoyé à la littérature pertinente pour plus de détails.

Ce chapitre est organisé comme suit. Nous motivons le choix des graphes d'information hétérogène en le confrontant avec d'autres notions similaires de la littérature dans la section 1.1. Ensuite, nous présentons plus formellement dans la section 1.2 les définitions relatives aux graphes d'information hétérogène et nous étendons ces définitions aux cas des hypergraphes. Finalement, dans la section 1.3, nous nous intéressons à un processus sur ces graphes, à savoir les marches aléatoires contraintes.

## 1.1. Choix des graphes d'information hétérogène

Dans sa forme la plus élémentaire, un *graphe* ramène un système à une structure abstraite composée d'entités simplifiées, souvent appelées *nœuds* et supposées conserver certaines propriétés des composantes originales du système, ainsi que les connexions par paire entre elles, souvent appelées *liens*. Cette représentation, permettant une vue – certes incomplète – du système, permet d'en révéler quelques caractéristiques. Par exemple, le comportement du système peut être influencé par la topologie du graphe, i.e. la façon dont les entités sont connectées [11, 118].

L'étude des graphes possède une longue et riche histoire, que ce soit e.g. en mathématiques, physique, informatique, économie, écologie ou sociologie. Il est évident que nombre de systèmes réels sont extrêmement complexes. Pour rappel, un *système complexe* est souvent défini comme un système dont on ne peut comprendre ou prédire les propriétés sur la seule base de la connaissance complète de ses constituants. Dès lors, de simples graphes ne peuvent enregistrer toute leur complexité.

Bien que n'importe quel graphe ne soit qu'une modélisation et ne peut donc empêcher une perte

d'information, il a été récemment montré qu'il était possible d'enrichir ces simples graphes en définissant de nouvelles structures, tout en les gardant manipulables. En particulier, le formalisme des graphes d'information hétérogène (HIN) a été introduit, permettant une analyse formelle de systèmes interconnectés et interdépendants [71, 148, 153, 154]. Défini grossièrement, un HIN est un graphe possédant plusieurs types de nœuds et plusieurs types de liens. Une définition plus rigoureuse est donnée dans la section suivante. Avec l'explosion de l'application de la théorie des graphes à l'analyse de toutes sortes de données, de nombreuses façons de modéliser ces dernières sous forme de graphes ont été développées. Nous discutons ici brièvement les concepts et terminologies qui nous semblent similaires à celui des HIN et motivons notre choix pour ce dernier formalisme. Cependant, malgré de nombreux efforts pour accorder tous ces concepts [5, 92], aucune terminologie standard ne semble exister. Cela vient en partie du fait que des mêmes formalismes sont utilisés avec des idées différentes en tête. Nous espérons que cette section permet de se situer, au moins grossièrement, dans cette jungle de termes. Notons que dans ce chapitre, nous nous concentrons uniquement sur les graphes et leurs extensions statiques, i.e. aucune dynamique de nœuds ou de liens n'est envisagée.

Comme déjà présenté, le concept le plus simple est celui de graphe d'information *homogène*. Cela signifie que tous les nœuds (d'une part) et liens (d'autre part) sont de *même type*. Ces graphes, largement étudiés, servent souvent de base pour adapter les techniques d'analyse aux HIN.

Il existe beaucoup de travaux s'intéressant aux graphes composés de plusieurs types de liens. Dans la littérature, ils sont connus sous plusieurs termes : graphes *multi-relationnels* [134, 160, 174], *composites* [182], *multidimensionnels* [16], *à liens colorés* ou encore, graphes *multiplexes* [12]. Tous ces termes désignent donc un graphe composé d'un ensemble de nœuds liés les uns aux autres par différents types de relations. Le terme multiplex permet de particulièrement bien visualiser la chose : le graphe est divisé en plusieurs couches et chaque couche représente un type de liens. Lorsque tous les (resp. une partie des) nœuds sont présents (i.e. actifs) dans chaque couche, le terme graphe à *nœuds alignés* (resp. *nœuds partiellement alignés*) est aussi employé [10]. Un exemple type est un réseau social où les nœuds représentent les individus et les liens représentent différents types de relations entre eux, e.g. liens familiaux, liens professionnels, liens amicaux. Un autre exemple souvent invoqué est le réseau de transport aérien où les différentes compagnies aériennes, i.e. les différents types de liens, relient les aéroports, i.e. les nœuds [26, 166]. Finalement, lorsque les couches n'ont aucun nœud commun, le graphe est *non aligné* [62].

À l'inverse, il existe des modélisations prenant en compte différents types de nœuds connectés par un seul type de liens. Les nœuds peuvent se distinguer les uns des autres par leur fonction, leurs propriétés structurelles [85] ou encore leur nature (e.g. les graphes bipartis). Ils sont souvent

désignés par les termes graphes *à nœuds colorés*, graphes *interconnectés* ou *graphes de graphes*. Des nœuds différents impliquent souvent des liens de nature/fonction différente dans la réalité, mais cela est ignoré dans la modélisation. Ces types de graphes permettent, entre autres, de représenter différents sous-systèmes tels que des systèmes d'infrastructure couplés. Un autre exemple serait un réseau social où chaque nœud est coloré suivant une caractéristique, e.g. tranche d'âge, groupe ethnique, genre [7, 117].

Tous ces concepts peuvent être vus comme des cas particuliers de graphes dits *multicouches* [5, 19, 92]. Dans sa forme la plus générale, un nœud dans une couche peut être connecté à n'importe quel autre nœud de n'importe quelle autre couche. Une couche permet de caractériser les nœuds et/ou liens qui lui appartiennent. La frontière avec ce que nous avons appelé HIN est donc extrêmement ténue. D'ailleurs, il semble qu'il s'agisse plus d'une différence d'objectif de recherche et de communauté scientifique que d'une différence d'objet formel étudié et manipulé. Lorsqu'on s'intéresse au graphe de façon plus abstraite ou plus théorique, il semble que le terme multicouche soit préféré. À l'inverse, lorsque la sémantique des entités en jeu est mise en avant, le terme HIN semble plus usité (cela étant aidé par le formalisme détaillé dans la section suivante). Bien sûr, cela n'est pas une généralité. On notera aussi qu'un HIN est parfois appelé un graphe *multi-modes* [163].

Dans ce travail, nous utiliserons toujours le terme HIN. Ce choix est motivé par la littérature (principalement référencée dans le chapitre suivant) qui nous semble la plus proche de notre problématique et des applications que nous visons : récupérer le poids des liens en ne se limitant pas à une analyse structurelle mais en essayant d'y apporter des informations sémantiques. Ainsi que mentionné, les applications concernent des réseaux sociaux (e.g. données bibliographiques DBLP et données Twitter), réseaux dans lesquels différents agents (e.g. individus, articles, médias, hashtags) interagissent entre eux à l'aide de différents canaux. Le concept de HIN nous semble donc tout à fait approprié.

## 1.2. Définitions

Dans cette section, nous présentons les différentes définitions relatives aux HIN utilisées tout au long du travail. La figure 1.1 illustre toutes les notions introduites dans cette section.

### 1.2.1. Graphes d'information hétérogène

Dans sa forme la plus simple, un graphe se définit comme une paire d'ensembles $G = (V, E)$ où $V$ est un ensemble de nœuds et $E$ un ensemble de liens reliant les nœuds deux à deux. Cependant, dans de nombreux cas, on étend cette définition afin d'englober plus d'informations telles que la

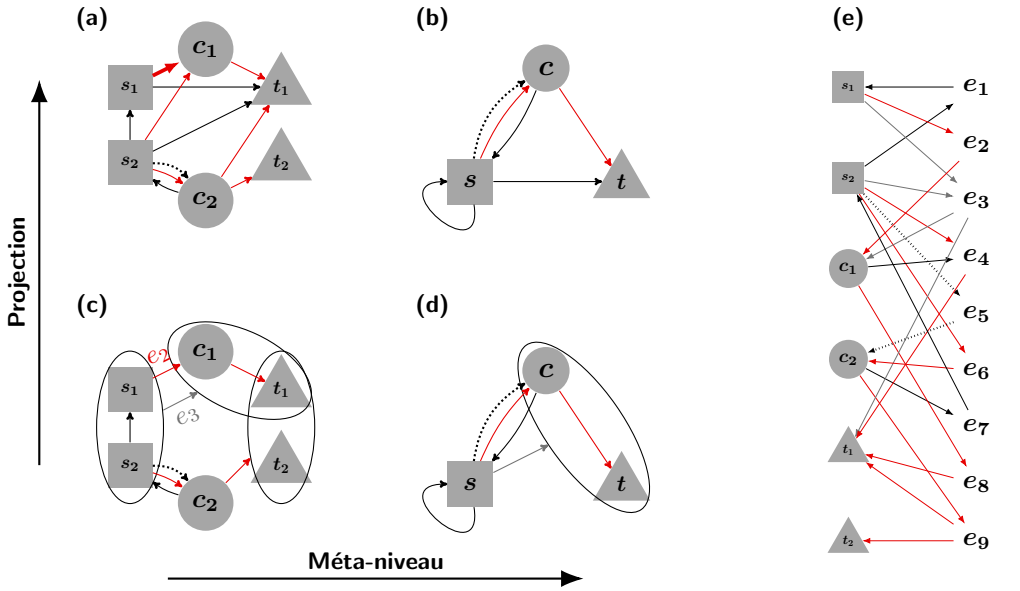



**Figure 1.1.:** Illustration des définitions. (a) Graphe d'information hétérogène $G$ composé de trois types de nœuds (différenciés par leur forme: ■, ● et ▲) et plusieurs types de liens (non différenciés par leur couleur; la couleur servant simplement à mettre en évidence un exemple de méta-chemin et chemins) et (b) son schéma associé $S_G$. Plusieurs liens existant entre deux nœuds sont de types différents: $\rightarrow$ et $\dashrightarrow$ entre $s_2$ et $c_2$. Un exemple de méta-chemin dans le schéma (b) est ■$\rightarrow$ ● $\rightarrow$▲ (rouge), ce qui correspond à un ensemble de chemins dans le HIN (a). (c) Hypergraphe d'information hétérogène $\mathcal{H}$, composé des mêmes (types de) nœuds qu'en (a) et (d) son hyper-schéma associé $S_{\mathcal{H}}$. Un hyperarc peut comprendre plusieurs types de nœuds dans sa tête. En outre, il s'avère que le HIN en (a) est la projection de l'hypergraphe en (c). Lorsque la projection de deux hyperarcs génère deux liens de mêmes types entre deux mêmes nœuds, ce lien est unique et pondéré : lien $(s_1, c_1)$ plus épais, résultat de la projection des hyperarcs $e_2 = (\{s_1\}, \{c_1\})$ et $e_3 = (\{s_1, s_2\}, \{c_1, t_1\})$ qui sont de types différents. La tête de $e_3$ comportant deux types de nœuds différents, deux types de liens sont créés lors de la projection, parfois identiques à d'autres types de liens (figure 1.3 pour plus de détails). (e) Représentation de l'hypergraphe en (c) sous forme de graphe biparti $\mathcal{B}$. Chaque nœud et hyperarc de $\mathcal{H}$ induit un nœud dans $\mathcal{B}$ et la direction des liens est donnée par la matrice d'adjacence signée $\tilde{B}$ (éq. (1.4), figure 1.2 pour plus de détails).

direction d'un lien, son poids, ou permettre plusieurs liens entre deux nœuds et donc relâcher la notion d'ensemble de liens.

**Définition 1.1** (Multigraphe pondéré et orienté)**.** Un *multigraphe pondéré et dirigé* est un tuple $G = (V, E, w, \mu_s, \mu_t)$ avec $V$ un ensemble de liens, $E$ un multi-ensemble de liens, $w : E \to \mathbb{R}^+$ une fonction poids, $\mu_s : E \to V$ (resp. $\mu_t : E \to V$) la fonction qui associe à chaque lien un nœud source (resp. cible).

Notons que dans ce travail, nous ne considérons pas les graphes signés, i.e. les poids des liens sont tous de même signe, positif.

Lorsqu'on désire abstraire un système réel par un graphe, les entités du système en question ont parfois une sémantique particulière qu'il est important de garder lors du passage à l'abstraction. Pour ce faire, la notion de *graphe d'information hétérogène* (*Heterogeneous Information Network*,

HIN) a été introduite.

**Définition 1.2** (Graphe d'information hétérogène (HIN))**.** Un *HIN* $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$ est un multigraphe pondéré et dirigé $G$ avec $\mathcal{V}$ l'ensemble de types de nœuds et $\mathcal{E}$ l'ensemble des types de liens. La fonction $\phi : V \to \mathcal{V}$ attribue un type à chaque nœud tandis que $\psi : E \to \mathcal{E}$ associe un type à chaque lien. En outre, la condition suivante doit être satisfaite : si deux liens sont de même type, alors ils partagent les mêmes types de nœuds sources et cibles, i.e. $\forall e_1, e_2 \in E, [\psi(e_1) = \psi(e_2))] \Rightarrow [\phi(\mu_s(e_1)) = \phi(\mu_s(e_2)) \wedge \phi(\mu_t(e_1)) = \phi(\mu_t(e_2))]$.

Lorsque $|\mathcal{V}| = 1 = |\mathcal{E}|$, le graphe est dit *d'information homogène*. Le concept de multigraphe est nécessaire car il y a divers types de liens. Plus précisément, une paire de nœuds ne peut être connectée par plusieurs liens que si ces derniers sont de types différents. Si tel n'est pas le cas, nous choisissons de combiner ces liens de même type en un lien pondéré.

L'ensemble image $\mathcal{V} = \{V_1, ..., V_{|\mathcal{V}|}\}$ (resp. $\mathcal{E} = \{E_1, ..., E_{|\mathcal{E}|}\}$) est l'ensemble de tous les types de nœuds (resp. liens). Par abus de notation, et lorsqu'il n'y a aucune ambiguïté, nous notons $\forall i = 1, ..., |\mathcal{V}|$, $V_i$ comme étant l'ensemble des nœuds de type $V_i$, i.e. $V_i = \{v \in V \,|\, \phi(v) = V_i\}$. De façon analogue, $\forall i = 1, ..., |\mathcal{E}|$, $E_i$ est l'ensemble des liens de type $E_i$, i.e. $E_i = \{e \in E \,|\, \psi(e) = E_i\}$.

La figure 1.1(a) représente un HIN. Ainsi qu'on peut le voir, il n'est pas toujours aisé de distinguer ou démêler les différentes entités présentes. Dans ce cas, il peut être bénéfique de considérer une représentation de plus haut niveau. Pour ce faire, on considère le schéma d'un HIN, i.e. sa description macroscopique. En termes simples, cela correspond au graphe défini sur les types de nœuds et liens (figure 1.1(b)).

**Définition 1.3** (Schéma d'un graphe)**.** Étant donné $H$ un HIN, son *schéma* $S_H = (\mathcal{V}, \mathcal{E}, \nu_s, \nu_t)$ est un graphe dirigé avec $\nu_s : \mathcal{E} \to \mathcal{V} : E^* \mapsto \phi(\mu_s(e))$ la fonction qui attribue un nœud source à un lien et $\nu_t : \mathcal{E} \to \mathcal{V} : E^* \mapsto \phi(\mu_t(e))$ la fonction qui attribue un nœud cible à un lien, avec $e \in E^*$[1].

Deux entités dans un HIN peuvent être liées via différents chemins ayant différentes sémantiques. Pour rappel, un *chemin* $P$ de longueur $n \in \mathbb{N}$ entre $v_0$ et $v_n$ dans un HIN est une séquence de nœuds $v_0, ..., v_n \in V$ reliés par des liens $e_1, ..., e_n \in E$ comme suit : $P = v_0 \xrightarrow{e_1} v_1 \cdots \xrightarrow{e_n} v_n$. Afin de faciliter leur lecture et compréhension, on utilise le concept de méta-chemin [148].

**Définition 1.4** (Méta-chemin)**.** Un *méta-chemin* $\mathcal{P}$ de longueur $n \in \mathbb{N}$ est chemin de longueur $n$ dans le schéma d'un graphe $S_H$. Il s'agit donc d'une séquence de types de nœuds $V_0, ..., V_n \in \mathcal{V}$ reliés par des types de liens $E_1, ..., E_n \in \mathcal{E}$ (avec de possibles répétitions) comme suit: $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots \xrightarrow{E_n} V_n$, avec $\forall\, i = 1, ..., n$, $V_{i-1} = \nu_s(E_i)$ et $V_i = \nu_t(E_i)$.

[1] On peut effectivement prendre n'importe quel élément $e \in E^*$ puisque $\{e \in E | \psi(e) = E^*\}$ est la classe d'équivalence de chacun de ses éléments, avec la relation d'équivalence "a le même type que". Par définition d'un HIN, il suffit de prendre un seul membre de la classe pour connaître les types de nœuds que $E^*$ connecte.

**Définition 1.5** (Méta-chemin inverse)**.** Le *méta-chemin inverse* de $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots \xrightarrow{E_n} V_n$ est un méta-chemin $\mathcal{P}^{-1} = V_n \xrightarrow{E_n^{-1}} \cdots V_1 \xrightarrow{E_1^{-1}} V_0$, avec $E_i^{-1}$ l'inverse[1] de $E_i$ : soit $V_{i-1}$ (resp. $V_i$) le type de nœuds source (resp. cible) de $E_i$, alors, $V_{i-1}$ (resp. $V_i$) est le type de nœuds cible (resp. source) de $E_i^{-1}$ et la sémantique de $E_i^{-1}$ est l'inverse déduite de celle de $E_i$.

**Définition 1.6** (Méta-chemin tronqué)**.** Le méta-chemin $\mathcal{P}^{i,j} = V_i \xrightarrow{E_{i+1}} V_{i+1} \cdots \xrightarrow{E_j} V_j$ est appelée *méta-chemin tronqué* de $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots \xrightarrow{E_n} V_n$ lorsque $\mathcal{P}^{i,j} \subseteq \mathcal{P}$, i.e. $\mathcal{P}^{i,j}$ est une sous-séquence de $\mathcal{P}$.

Étant donnés un chemin $P = v_0 \xrightarrow{e_1} v_1 \cdots \xrightarrow{e_n} v_n$ et un méta-chemin $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots \xrightarrow{E_n} V_n$, on dit que $P$ *satisfait* $\mathcal{P}$ et on note $P \in \mathcal{P}$ si et seulement si $\forall i \in \{1, ..., n\}$, $e_i \in E_i$. Dès lors, un méta-chemin est un ensemble de chemins.

Comme le montre la figure 1.1(b), un exemple de méta-chemin est ■→ ● →▲, en rouge dans le schéma du HIN. Les chemins rouges dans le HIN (figure 1.1(a)) satisfont le méta-chemin puisque tous les segments respectent les conditions énoncées ci-dessus. Comme mentionné, utiliser des méta-chemins est fort pratique pour donner une sémantique à la relation entre deux nœuds. Dans l'exemple de la figure 1.1(a), il existe de nombreux chemins entre $s_2$ et $t_1$ satisfaisant différents méta-chemins et possédant donc des sémantiques différentes. Plus précisément, il existe quatre méta-chemins permettant de relier $s_2$ à $t_1$: ■ → ▲, ■ → ■ → ▲, ■→ ● →▲ et finalement ■ → ■ → ● → ▲. Bien qu'un peu abstraites, les sémantiques associées à ces quatre méta-chemins sont fondamentalement différentes : tantôt directes tantôt alambiquées, certaines avec répétitions, etc. Nous verrons au chapitre 3, à l'aide d'exemples plus concrets, que cette différenciation permet, dans une certaine mesure, d'apporter certaines informations sur l'organisation d'un graphe.

### 1.2.2. Hypergraphes d'information hétérogène

Dans de nombreux réseaux réels, les interactions impliquent plus de deux entités simultanément [14, 109]. Les collaborations scientifiques en sont un parfait exemple, les auteurs étant les nœuds et les articles étant les liens pouvant relier plus de deux auteurs [123], mais plus généralement toute interaction sociale regroupant un nombre quelconque d'invidus [67, 83]. En biologie, les interactions entre protéines sont particulièrement complexes et nécessitent parfois une compréhension/interprétation différente (et complémentaire de celle obtenue avec des graphes) des processus biologiques, e.g. dans la représentation de relations logiques dans les réseaux de signalisation et de régulation [93]. Dans tous ces systèmes, un simple graphe d'information, homogène comme hétérogène, n'est donc pas l'objet le plus approprié pour capturer son organisation.

[1] L'inverse n'est pas forcément défini : soit $E_i$ tel que $E_i^{-1}$ ne soit pas défini dans le HIN. On suppose dès lors que $E_i^{-1} = \{(y, x) \text{ tq } (x, y) \in E_i\}$.

Une façon de remédier à ce problème est de considérer des *hypergraphes*. Il s'agit d'une généralisation d'un graphe en ce sens qu'une relation fait intervenir un nombre quelconque d'entités[1], regroupées par ce qu'on appelle hyperarcs. Les hypergraphes sont un concept bien connu et étudié depuis longtemps [15, 55] mais ont récemment gagné un regain d'intérêt grâce à l'étude intensive des graphes, e.g. [46, 60]. Cependant, à notre connaissance, peu de recherches formelles ont été menées concernant les hypergraphes d'information hétérogène : dans [23], les auteurs parlent d'hypergraphes *unifiés* dans lesquels existent plusieurs types de nœuds et liens mais aucune autre définition n'est donnée. Les définitions présentées ici reprennent des définitions de base, bien connues dans la littérature pour les hypergraphes d'information *homogène*, et sont parfois un peu modifiées afin de prendre en compte soit le caractère dirigé et pondéré de l'hypergraphe, soit la nature hétérogène des entités présentes.

**Définition 1.7** (Hyperarc)**.** Un *hyperarc* $e$ est un couple d'ensembles modélisant la relation entre un nombre quelconque d'entités. Formellement, $e = (T(e), H(e))$ où $T(e)$ (resp. $H(e)$), appelé la *queue* (resp. *tête*) de l'hyperarc, est un ensemble de nœuds.

Cette façon de décomposer un hyperarc est comparable aux fonctions $\mu_s$ et $\mu_t$ définies pour un graphe (section 1.2.1). Comme souvent dans la littérature, on impose $T(e) \cap H(e) = \emptyset$. Lorsque $|T(e)| = |H(e)| = 1$, $\forall e$, on se ramène à un simple graphe : chaque hyperarc se ramène à un lien dirigé ne reliant que deux nœuds à la fois. En pratique, les hypergraphes n'ont donc d'intérêt que lorsque $|T(e)|.|H(e)| > 1$.

**Définition 1.8** (Hypergraphe pondéré et orienté)**.** Un *hypergraphe pondéré et dirigé* est un tuple $\mathcal{H} = (V, E, \tilde{\omega})$ avec $V$ un ensemble de nœuds, $E$ un multi-ensemble d'hyperarcs (hyperliens orientés), une fonction $\tilde{\omega} : E \to \mathbb{R}^+$ associant un poids positif à chaque hyperarc.

**Définition 1.9** (Hypergraphe d'information hétérogène pondéré et orienté)**.** Un *hypergraphe d'information hétérogène pondéré et dirigé* est un hypergraphe pondéré et orienté $\mathcal{H}_0$ muni d'une fonction $\phi : V \to \mathcal{V}$ associant un type à chaque nœud et d'une fonction $\psi : E \to \mathcal{E}$ associant un type à chaque hyperarc, i.e. $\mathcal{H} = (\mathcal{H}_0, \mathcal{V}, , \mathcal{E}, \phi, \psi)$, figure 1.1(c).

Remarquons que la tête et la queue de l'hyperarc peuvent contenir des nœuds de n'importe quel type ; cela sera motivé dans de prochaines applications.

À nouveau, par abus de notation et lorsqu'il n'y a aucune ambiguïté possible, on note $V_i = \{v \in V | \phi(v) = V_i\}$ et $E_i = \{e \in E | \psi(e) = E_i\}$.

[1] Un hypergraphe impliquant $n$ entités ne requiert pas d'avoir toutes les interactions d'ordre inférieur à $n$, e.g. une unique interaction à trois corps ne requiert pas l'existence des trois interactions associées au triangle, à l'inverse des *complexes simpliciaux*.

**Définition 1.10** (Poids d'un nœud)**.** Soit $v \in V$ un nœud appartenant à un hyperarc $e \in \mathcal{E}$. Le *poids d'un nœud dans un hyperarc* désigne l'importance qu'a ce nœud dans cet hyperarc et se note $\lambda_e(v) \in \mathbb{R}^+$. Un nœud possède donc autant de poids qu'il y a d'hyperarcs dans l'hypergraphe. Dès lors, le *poids d'un nœud* s'exprime comme un vecteur à valeurs réelles positives $\boldsymbol{\lambda}(v) = [\lambda_{e_1}(v), ..., \lambda_{e_{|E|}}(v)]$, avec $\lambda_e(v) \in \mathbb{R}^+$ le poids de $v$ dans l'hyperarc $e$.

On définit deux matrices d'*incidence* $H^+$ et $H^-$ comme suit : $\forall\, v \in V, \forall\, e \in E$, $H^+(v,e) = 1$ si $v \in T(e)$, 0 sinon et de façon similaire, $H^-(v,e) = 1$ si $v \in H(e)$, 0 sinon. Bien qu'un peu contre-intuitives, nous gardons ces notations $H^+$ et $H^-$ puisqu'elles sont utilisées dans la littérature, e.g. [44]. Ces matrices permettent de voir un hypergraphe orienté comme un graphe biparti orienté $\mathcal{B} = (V_\mathcal{B}, E_\mathcal{B})$. Par définition, $V_\mathcal{B}$ est l'union de deux ensembles de nœuds : le premier est composé des nœuds de l'hypergraphe $V$ tandis que le second représente les hyperarcs $E$. De cette façon, le nœud $v \in V$ (resp. l'hyperlien $e \in E$) a son homologue $v_B \in V_\mathcal{B}$ (resp. le nœud $e_B \in V_\mathcal{B}$), ainsi qu'illustré aux figures 1.1(c), 1.1(e) et 1.2. Par abus de notation, on a donc $V_\mathcal{B} = V \cup E$. La matrice d'*adjacence* $B \in \mathbb{R}^{|V|\times|E|}$ du biparti, construite à partir des matrices d'incidence $H^+$ et $H^-$, est telle que $B(v,e) = \delta_{1,H^-(v,e)} - \delta_{1,H^+(v,e)}$, avec $\delta_{ij} = 1$ si $i = j$, 0 sinon, le delta de Kronecker. Notons que cette matrice $B$ est en fait l'unique matrice d'incidence définie dans [55]. Cependant, nous verrons dans la section 1.3 l'intérêt de définir deux matrices d'incidence. Les poids des hyperliens et nœuds de $\mathcal{H}$ peuvent également être pris en compte, définissant une matrice d'adjacence pondérée $\tilde{B}$.

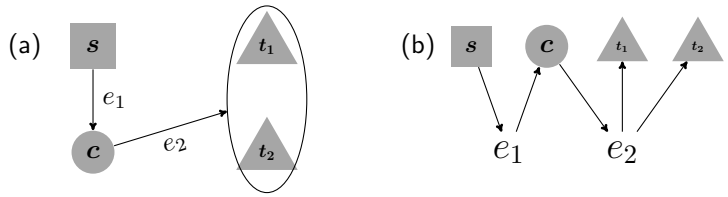



**Figure 1.2.:** Hypergraphe (a) et graphe biparti associé (b). L'orientation des (hyper)liens est conservée dans le biparti. Les signes des poids de la matrice d'adjacence $B$ "indiquent" cette direction : un signe négatif (resp. positif) signifie que le lien part de $V$ (resp. $E$) vers $E$ (resp. $V$).

Afin d'éviter toute confusion par la suite lorsqu'on définira une marche aléatoire sur un hypergraphe (section 1.3), nous rappelons les définitions des degrés. Soit un hyperarc $e \in E$. Son degré sortant est $\delta^+(e) = \sum_{v\in V} \lambda_e(v) H^+(v,e)$ tandis que son degré entrant est $\delta^-(e) = \sum_{v\in V} \lambda_e(v) H^-(v,e)$. Ainsi, son degré vaut $\delta(e) = \delta^+(e) + \delta^-(e)$. Dans la suite, nous supposerons toujours que $\tilde{\omega}(e)$ égale le nombre de fois que l'hyperarc typé $e$ apparaît dans l'hypergraphe. Afin de prendre en compte d'autres informations propres à l'hyperarc, nous définissons un nouveau poids $\omega(e)$ égal au nombre de nœuds pondérés dans $e$ multiplié par $\tilde{\omega}(e)$, le tout à une certaine puissance, i.e.

$\omega(e) = \tilde{\omega}(e)\,[\delta(e)]^{\gamma}$, $\gamma \in \mathbb{R}$. Dès lors, des valeurs négatives de $\gamma$ privilégieront les hyperarcs composés de peu de nœuds tandis que l'inverse se passe pour des $\gamma$ positifs. Soit un nœud $v \in V$. Son degré sortant est $d^{+}(v) = \sum_{e \in E} \omega(e) H^{+}(v, e)$ tandis que son degré entrant est $d^{-}(v) = \sum_{e \in E} \omega(e) H^{-}(v, e)$. Ainsi son degré vaut $d(v) = d^{+}(v) + d^{-}(v)$.

Pour illustration, reprenons l'exemple de la figure 1.2(a) et définissons les poids des nœuds tels que $\lambda_{e_1}(s) = \lambda_{e_1}(c) = \lambda_{e_2}(t_1) = 1$ et $\lambda_{e_2}(c) = \lambda_{e_2}(t_2) = 2$. De cette façon, les degrés des hyperarcs sont $\delta(e_1) = 1 + 1 = 2$ et $\delta(e_2) = 2 + (1 + 2) = 5$. Chaque hyperarc n'apparaît qu'une fois: $\tilde{\omega}(e_1) = \tilde{\omega}(e_2) = 1$, mais $\omega(e_1) = 1 \times 2^{\gamma}$ et $\omega(e_2) = 1 \times 5^{\gamma}$. Enfin, le degré du nœud, e.g., $c$ égale $d(c) = 5^{\gamma} + 2^{\gamma}$.

Soient deux nœuds $v_0, v_n \in V$. Un hyper-chemin de longueur $n \in \mathbb{N}$ de $v_0$ à $v_n$ dans $\mathcal{H}$ est une séquence $P = v_0 \xrightarrow{e_1} v_1 \cdots \xrightarrow{e_n} v_n$ telle que $v_0 \in T(e_1)$, $v_n \in H(e_n)$ et $\forall\, i = 1, ..., n-1, v_i \in T(e_{i-1}) \cap H(e_i)$.

À partir des fonctions $\phi$ et $\psi$, nous définissons un hyper-schéma $\mathcal{S}_{\mathcal{H}} = (\mathcal{V}, \mathcal{E})$. Il s'agit d'une méta-structure associée à un hypergraphe $\mathcal{H}$ (déf. 1.9). Formellement,

**Définition 1.11** (Hyper-schéma)**.** Soit un hypergraphe $\mathcal{H}$. Un *hyper-schéma* est un hypergraphe dirigé défini sur les types de nœuds $\mathcal{V}$ et d'hyperarcs $\mathcal{E}$. Chaque hyperarc $E^* \in \mathcal{E}$ s'écrit $(T(E^*), H(E^*))$ (figure 1.1(d)).

À nouveau, la décomposition des hyperarcs est comparable aux fonctions $\nu_s$ et $\nu_t$ définies pour un schéma (section 1.2.1).

Grâce à ce nouvel objet, on peut facilement définir un méta-chemin dans les hypergraphes; il s'agit tout simplement d'un chemin dans l'hyper-schéma.

**Définition 1.12** (Hyper-méta-chemin)**.** Soit un hypergraphe $\mathcal{H}$. Un *hyper-méta-chemin* $\mathcal{P}$ de longueur $n \in \mathbb{N}$ est une séquence de types de nœuds et d'hyperarcs $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 ... V_{n-1} \xrightarrow{E_n} V_n$ tel que $\forall i = 1, ..., n$, $\exists\, e \in E_i$ tel que $\exists\, u_1 \in T(e)$ avec $\phi(u_1) = V_{i-1}$ et $\exists\, u_2 \in H(e)$ avec $\phi(u_2) = V_i$.

### 1.2.3. Projection d'hypergraphes

De nombreuses questions se posent lorsque différentes représentations d'un même système sont envisagées : dans quelle mesure deux représentations différentes se chevauchent-elles ? Est-il possible d'établir une correspondance canonique entre les deux représentations ? Ici, afin de faire le lien entre hypergraphe et graphe, nous définissons la projection : la *projection* d'un hypergraphe $\mathcal{H}$ se résume à un graphe $G = (V, E_G)$ (figures 1.1(a), 1.1(c), 1.3(a) et 1.3(c). L'ensemble des nœuds reste le même et donc la fonction $\phi$ également. En revanche, cela change pour les liens. On définit la projection de chaque hyperarc comme $\pi(e) = \{(u, v) \subseteq T(e) \times H(e)\} \subseteq E_G$, d'où

$E_G = \cup_e \pi(e)$. Chaque élément de cet ensemble est un lien dont le type est donné par l'hyperarc dont il est issu, moyennant quelques ajustements. En particulier, puisque les nœuds appartenant à la queue ainsi qu'à la tête de l'hyperarc peuvent être de n'importe quel type, la projection $\pi(e)$ peut générer, e.g. deux liens $e_1 = (u_1, v_1)$ et $e_2 = (u_2, v_2)$, avec $\phi(u_1) \neq \phi(u_2)$ et/ou $\phi(v_1) \neq \phi(v_2)$. Par définition d'un HIN, $e_1$ et $e_2$ ne peuvent donc être de même type. Dans ce cas, le type de l'hyperarc $\psi(e)$ donne lieu à deux types de liens. De façon générale, $\psi(e)$ donne lieu à $|\phi[T(e)]| \times |\phi[H(e)]|$ types de liens, avec $\phi[T(e)]$ (resp. $\phi[H(e)]$) la notation abusive désignant l'ensemble des types de nœuds présents dans $T(e)$ (resp. $H(e)$). En outre, il n'est pas interdit que deux hyperarcs de types différents génèrent des liens de même type, ainsi qu'illustré à la figure 1.3(b)). Dans ce cas, tous les liens de $u$ à $v$ de type $E^*$ et de poids[1] $\tilde{w}_e(u, v)$ sont rassemblés en un seul et on définit un nouveau poids $w$ qui est la somme de tous les poids $\tilde{w}$. En notant abusivement $\psi[\pi(e)]$ l'ensemble des types de liens obtenus en projetant $e$, le poids du lien $(u, v) \in E_G$ de type $E^*$ s'exprime comme $w(u, v) = \sum_{e \text{ tq } E^* \in \psi[\pi(e)]} \tilde{w}_e(u, v)$. La projection est donc un graphe $G = (V, E_G)$ dont les poids des liens sont donnés par $w$. Concernant le poids d'un nœud $u$, celui-ci est défini comme la somme des poids du nœuds $u$, $\boldsymbol{\lambda}(u)$, dans l'hypergraphe $\mathcal{H}$, i.e. $\lambda_G(u) = \sum_{e \in E} \lambda_e(u)$ (figure 1.3(c)).

D'autres choix des poids des liens et des nœuds dans $\mathcal{H}$ et sa projection sont bien entendus possibles et ces choix ne sont pas anodins. Par exemple, dans les processus de marche aléatoire, ces poids peuvent véritablement biaiser la marche et impacter le résultat final (section 1.3). Cela sera discuté plus en détails dans un chapitre ultérieur.

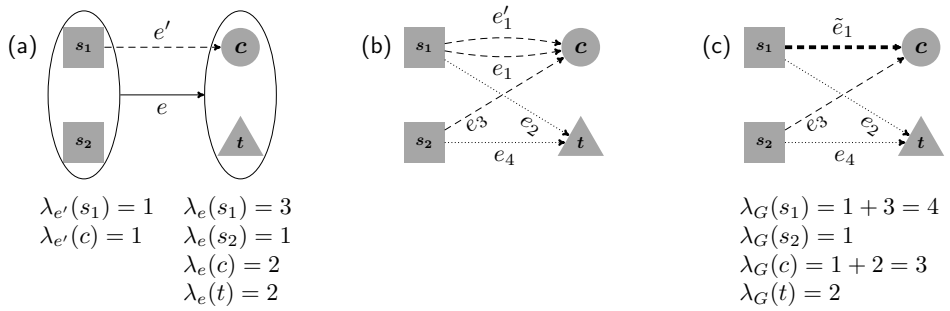



**Figure 1.3.:** Projection d'un hypergraphe. (a) Hypergraphe composé de 3 types de nœuds et 2 types d'hyperarcs. L'hyperarc $e = (\{s_1, s_2\}, \{c, t\})$ comprend 3 types de nœuds, d'où $|\phi[T(e)]| \times |\phi[H(e)]| = 2$. (b) La projection $\pi(e) = \{e_1, e_2, e_3, e_4\}$ est donc composée de 2 types de liens, ⇢ et ⇢, en fonction des types de nœuds source et cible dans du graphe projeté. Le type ⇢ (de $e_1$ et $e_3$) est le même que celui de $e'_1 = \pi(e')$. (c) En conséquence, $e_1$ et $e'_1$ sont deux liens reliant $s_1$ à $c$ de même type. Leur poids est donné par le poids de l'hyperarc dont ils sont issus, i.e. $\tilde{\omega}(e)$ et $\tilde{\omega}(e')$ respectivement. Par définition d'un HIN, ils ne peuvent former qu'un seul lien, $\tilde{e}_1$, de type ⇢ et de poids $w = \tilde{\omega}(e) + \tilde{\omega}(e')$. La projection de l'hypergraphe en (a) est donc le graphe représenté en (c). La projection est donc une fonction surjective.

[1] Chaque lien $(u, v)$ de type $\psi(e)$ possède un poids $\tilde{w}$ qui lui vient directement de l'hyperarc dont il est issu, i.e. $\tilde{w}_e(u, v) = \tilde{\omega}(e)$. Pour rappel, celui-ci représente le nombre d'occurrences de $e$ dans les données.

## 1.3. Marches aléatoires contraintes

L'étude des marches aléatoires possède une longue tradition dans la théorie des graphes [120]. Bien qu'étant un outil relativement simple, les marches aléatoires se sont révélées utiles afin de e.g. calculer des mesures de centralité [21, 119], trouver des communautés [128, 136], mais aussi pour comprendre comment la structure d'un graphe, e.g. la distribution de degrés, pouvait impacter des processus de diffusion [84].

Une marche aléatoire sur un graphe est un processus stochastique dans lequel un marcheur se déplace sur une structure sous-jacente statique représentée par un graphe [118]. Dans sa forme la plus simple, à chaque pas de temps, un marcheur saute d'un nœud vers un nœud voisin avec une probabilité uniforme. De nombreuses autres versions ont été proposées afin de e.g. cibler des nœuds possédant certaines propriétés [50], explorer des probabilités de transition non linéaires [149] ou encore considérer l'aspect temporel des liens [152]. Dans les HIN, le concept de marches aléatoires contraintes par un méta-chemin (*Path-Constrained Random Walk*, PCRW) est particulièrement apprécié, le méta-chemin donnant une sémantique à la marche [148, 153]. *Grosso modo*, l'idée d'une PCRW est qu'un marcheur ne puisse sauter d'un nœud sur un autre dans le HIN qu'en empruntant certains liens dont le type est spécifié par le méta-chemin. Cette notion est la base de la méthode de prédiction des poids de liens proposée au chapitre 3.

Soit $X_i \in V_i$ une variable aléatoire représentant la position d'un marcheur aléatoire dans l'ensemble $V_i$.

**Définition 1.13** (Chaîne de Markov)**.** Une *chaîne de Markov* $\mathcal{M}$ à temps discret est une séquence de variables aléatoires $X_0, X_1, ... \subseteq \Omega$, avec $\Omega$ l'espace d'états, ayant la propriété de Markov : $\mathbb{P}(X_i = v_i \,|\, X_{i-1} = v_{i-1}, ..., X_0 = v_0) = \mathbb{P}(X_i = v_i \,|\, X_{i-1} = v_{i-1})$, à condition que ces probabilités conditionnelles soient bien définies, i.e. $\mathbb{P}(X_0 = v_0, \cdots, X_{i-1} = v_{i-1}) > 0$. Lorsque cette probabilité est indépendante de $i$, la chaîne de Markov est dite *homogène*.

Dans ce travail, nous associons une marche aléatoire, commençant en $X_0$ à une chaîne de Markov $\mathcal{M}$ homogène et finie, i.e. l'espace d'états est fini (ici $|\Omega| = n$). Cela permet de définir

**Définition 1.14** (Matrice de transition)**.** Une *matrice de transition* est une matrice $\mathbf{T} \in \mathbb{R}^{n \times n}$ capturant les probabilités de mouvements d'un marcheur errant aléatoirement sur le graphe : $\mathbf{T}_{u,v} := \mathbb{P}(X_i = v \,|\, X_{i-1} = u)$. Cette matrice est stochastique sur les lignes : tous les éléments sont non négatifs et la somme des éléments d'une même ligne égale 1.

Dans le reste de cette section, nous présentons les PCRW dans les HIN et étendons aux hypergraphes d'information hétérogène.

### 1.3.1. Marches aléatoires contraintes dans les HIN

Soient un HIN $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$ et un méta-chemin $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots V_{n-1} \xrightarrow{E_n} V_n$. Il peut y avoir des répétitions dans cette séquence de types de nœuds et liens.

**Définition 1.15** (Probabilité de transition contrainte par un type de lien)**.** Étant donnés le type de liens $E_i \equiv \mathcal{P}^{i,i+1} \in \mathcal{E}$ et un couple de nœuds $(v_{i-1}, v_i) \in V_{i-1} \times V_i$, la probabilité de transition pour un marcheur de passer de $v_{i-1}$ à $v_i$ suivant $E_i$ se note $\mathbb{P}_{E_i}(v_i|v_{i-1})$ et est définie par

$$\mathbb{P}_{E_i}(v_i|v_{i-1}) = \frac{w_{E_i}(v_{i-1}, v_i)}{\sum_k w_{E_i}(v_{i-1}, v_k)} \tag{1.1}$$

où $w_{E_i}(v_j, v_k)$ est le poids du liens de type $E_i$ entre les nœuds $v_j$ et $v_k$, i.e. $w_{E_i}(v_j, v_k) = w(e)$ où $e$ est l'unique lien dans $E_i$ tels que $\mu_s(e) = v_j$ et $\mu_t(e) = v_k$. Si un tel lien n'existe pas, le poids est mis à zéro.

**Définition 1.16** (Probabilité de transition contrainte par un méta-chemin)**.** Étant donné $(v_0, v_n) \in V_0 \times V_n$, la probabilité d'atteindre $v_n$ à partir de $v_0$ suivant le méta-chemin $\mathcal{P}$, dénotée par $\mathbb{P}_{\mathcal{P}}(v_n|v_0)$, est simplement définie par la marche aléatoire de $v_0$ à $v_n$ le long des chemins satisfaisant $\mathcal{P}$ et s'exprime comme

$$\mathbb{P}_{\mathcal{P}}(v_n|v_0) = \sum_{v \in V_{n-1}} \mathbb{P}_{E_n}(v_n|v) \times \mathbb{P}_{\mathcal{P}^{0,n-1}}(v|v_0) \tag{1.2}$$

avec $\mathbb{P}_{\mathcal{P}^{0,1}}(v_1|v_0) = \mathbb{P}_{E_1}(v_1|v_0)$ la base de la récurrence.

La distribution de probabilités ainsi obtenue dans $V_n$, notée $\mathbb{P}_{\mathcal{P}}(:|v_0)$, a une importance toute particulière dans la méthode proposée au chapitre 2.

**Remarque 1.1** (Nœud puits)**.** Il peut arriver qu'il n'existe aucun lien de type $E_{ij}$ entre $v_i \in V_i$ et $v_j \in V_j$. Par conséquent, la probabilité de transition n'est pas définie. Pour contrer ce petit désagrément, on munit chaque ensemble $V_k$ d'un *nœud puits* $h_k$ sur lequel pointent tous les nœuds déconnectés. En outre, tous les nœuds puits sont connectés entre eux et aucun lien ne peut partir d'un nœud puits vers un autre nœud "normal", i.e. qui n'est pas nœud puits. Formellement, $\forall\, V_k \in \mathcal{V}$, $V_k^h := V_k \cup \{h_k\}$. $\forall\, E_{ij} \in \mathcal{E}$, si $w_{E_{ij}}(v_i, v_j) = 0$, $\forall\, v_j \in V_j$ alors $w_{E_{ij}}(v_i, h_j) = 1$, sinon $w_{E_{ij}}(v_i, h_j) = 0$. De plus, $\forall\, E_{ij} \in \mathcal{E}$, $w_{E_{ij}}(h_i, h_j) = 1$ et $\forall\, v_j \in V_j$, $w_{E_{ij}}(h_i, v_j) = 0$. De cette façon, les probabilités de transition sont bien définies.

### 1.3.2. Marches aléatoires contraintes dans les hypergraphes

À l'instar des graphes, plusieurs propositions de marches aléatoires dans les hypergraphes existent avec des applications en, e.g. classement, partionnement, traitement d'images [33, 44, 183]. Cependant, à notre connaissance, le fait de les contraindre par des méta-chemins n'a pas encore été étudié dans la littérature. Pour ce faire, nous procédons en quatre étapes : *a)* nous rappelons tout d'abord les probabilités de transition dans un hypergraphe ; *b)* nous les contraignons par un type de liens ; *c)* nous les contraignons par un méta-chemin de longueur 1 et finalement ; *d)* nous les contraignons par un méta-chemin de longueur quelconque. Soit donc un hypergraphe d'information hétérogène $\mathcal{H}$.

Nous définissons tout d'abord la probabilité de passer d'un nœud $u$ à un autre $v$ sans aucune contrainte. Cette procédure s'effectue en deux temps : *i)* démarrant d'un nœud $u$, le marcheur choisit un hyperarc $e$ tel que $H(e) \supset e$ avec une probabilité proportionnelle à son poids $\omega(e)$ ; *ii)* ensuite, il choisit un nœud $v \in T(e)$ avec une probabilité proportionnelle à son poids $\lambda_e(v)$ (figure 1.4(a)). Cela est toujours possible puisque, par convention, $|H(e)| \neq \emptyset \neq |T(e)|$.

**Définition 1.17** (Probabilité de transition)**.** La probabilité de transition de $u$ à $v$ dans un hypergraphe $\mathcal{H}$ s'exprime comme

$$\mathbb{P}(v|u) = \sum_{e \in E} \frac{\omega(e) h^+(u,e)}{d^+(u)} \frac{h^-(v,e)\lambda_e(v)}{\delta^-(e)}, \tag{1.3}$$

On voit donc l'intérêt de définir les deux matrices d'incidence $H^+$ et $H^-$. Cette probabilité est bien définie. En effet, $\forall u, v$, $\mathbb{P}(v|u) \geq 0$ et

$$\begin{aligned}\sum_{v \in V} \mathbb{P}(v|u) &= \sum_{v \in V} \sum_{e \in E} \frac{\omega(e) h^+(u,e)}{d^+(u)} \frac{h^-(v,e)\lambda_e(v)}{\delta^-(e)} \\ &= \sum_{v \in V} \sum_{e \in E} \frac{\omega(e) h^+(u,e)}{\sum_{e \in E} \omega(e) h^+(u,e)} \frac{h^-(v,e)\lambda_e(v)}{\sum_{v \in V} \lambda_e(v) h^-(v,e)} = 1.\end{aligned}$$

Lorsque les poids des nœuds sont indépendants des hyperarcs, i.e., $\lambda_e(v) = \lambda(v)$, $\forall v \in e, \forall e \in E$, il existe une expression des poids $w$ dans le graphe projeté de l'hypergraphe pour laquelle une marche aléatoire dans l'hypergraphe est équivalente à une marche aléatoire dans ce graphe projeté. Cependant, cela n'est pas le cas lorsque les poids des nœuds dépendent des hyperarcs [33].

Remarquons que dans le cas d'un graphe simple, $\sum_{e \in E} \omega(e) h^+(u,e) h^-(v,e)$ se réduit au poids du lien dirigé entre $u$ et $v$, i.e. $\omega(\{\{u\},\{v\}\})$ où $(\{u\},\{v\}) \equiv (u,v)$ est le seul lien entre $u$ et $v$ et $\delta^-(e) = 1$. Dès lors, l'éq. (1.3) se réduit à $\mathbb{P}(v|u) = \frac{w(u,v)}{d^+(u)}$, et on retrouve bien la formule pour les graphes simples.

**Remarque 1.2** (Équivalence avec des graphes bipartis)**.** Pour faire le parallèle avec les graphes bipartis, cela revient à sélectionner un nœud à chacune des étapes (figure 1.4(b)). Le choix de ces nœuds n'est pas uniforme mais dépend des poids des nœuds et liens. En particulier, la matrice $\tilde{B}$ se définit par

$$\tilde{B}(v_B, e_B) = \lambda_e(v)\delta_{1,B(v,e)} - \omega(e)\delta_{-1,B(v,e)}. \tag{1.4}$$

Cela signifie que le poids du lien entre $v_B$ et $e_B$ dans le biparti $\mathcal{B}$ est l'opposé du poids de $e$ dans $\mathcal{H}$ lorsque $v$ appartient à la queue de $e$ tandis que ce poids est le degré de $u$ dans l'hyperlien $e$ dans $\mathcal{H}$ lorsque $u$ appartient à la tête de $e$. Dans ce biparti, l'éq. (1.3) devient

$$\mathbb{P}_{biparti}(v|u) = \sum_{e\in E} \frac{\tilde{B}(u,e)\mathbb{1}_{\{B(u,e)<0\}}}{\sum_{e'\in E}\tilde{B}(u,e')\mathbb{1}_{\{B(u,e')<0\}}} \frac{\tilde{B}(v,e)\mathbb{1}_{\{B(v,e)>0\}}}{\sum_{v'\in V}\tilde{B}(v',e)\mathbb{1}_{\{B(v',e)>0\}}}, \tag{1.5}$$

avec la fonction indicatrice $\mathbb{1}_{\{x\in A\}} = 1$ si $x \in A$, 0 sinon. Cette probabilité est bien définie: elle est non négative et $\sum_{v\in V}\mathbb{P}_{biparti}(v|u) = 1$. Cela permet de voir le même processus sous deux angles différents. Dans la suite, nous nous concentrons sur la formulation dans l'hypergraphe, mais les adaptations sont tout aussi directes pour le point de vue biparti.

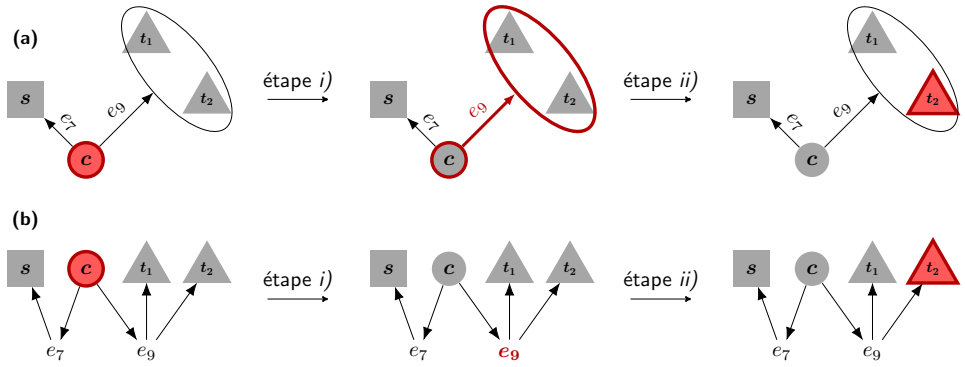



**Figure 1.4.:** Illustration du calcul des probabilités de transition (éqs (1.3) et (1.5)) dans les (a) hypergraphes et (b) graphes bipartis équivalents (éléments sélectionnés en rouge). Le marcheur, initialement en $c$ choisit *i)* $e_9$ avec probabilité $\omega(e_9)/[\omega(e_7)+\omega(e_9)]$ (resp. $\tilde{B}(c,e_9)/[\tilde{B}(c,e_7)+\tilde{B}(c,e_9)]$) dans l'hypergraphe (resp. le biparti) et ensuite; *ii)* $t_2$ avec probabilité $\lambda_{e_9}(t_2)/[\lambda_{e_9}(t_1)+\lambda_{e_9}(t_2)]$ (resp. $\tilde{B}(t_1,e_9)/[\tilde{B}(t_1,e_9)+\tilde{B}(t_2,e_9)]$) dans l'hypergraphe (resp. biparti).

À présent que nous avons défini la probabilité de transition sans contrainte $\mathbb{P}(v|u)$ (éq. (1.3)) dans un hypergraphe, on aimerait exploiter le fait que nous avons des informations supplémentaires sur les nœuds et hyperarcs, i.e. les fonctions $\psi$ et $\phi$. À l'instar des HIN, nous contraignons les probabilités de transition par des types d'hyperarcs, voire des hyper-méta-chemins, ce qui permet de donner une sémantique aux probabilités.

Soit un type d'hyperarcs $E^* \in \mathcal{E}$. Le degré sortant d'un nœud contraint par $E^*$ se définit comme $d^+_{E^*}(u) := \sum_{e\in E^*}\omega(e)H^+(v,e)$.

**Définition 1.18** (Probabilité de transition contrainte par un type d'hyperarcs)**.** La probabilité de transition contrainte par le type d'hyperarcs $E^*$, notée $\mathbb{P}_{E^*}(v|u)$, se définit de manière naturelle en ne considérant que les hyperarcs vérifiant cette condition, i.e. $e \in E^*$,

$$\mathbb{P}_{E^*}(v|u) = \sum_{e \in E^*} \frac{\omega(e)h^+(u,e)}{d^+_{E^*}(u)} \frac{h^-(v,e)\lambda_e(v)}{\delta^-(e)}. \tag{1.6}$$

À nouveau, il est aisé de vérifier que $\sum_v \mathbb{P}_{E^*}(v|u) = 1$.

Étant donné que les hyperarcs sont hétérogènes, il peut y avoir plusieurs types de nœuds dans la tête de l'hyperarc. Malheureusement, la probabilité définie par l'éq. (1.6) ne contraint en rien sur le type du nœud $v$. Dès lors, nous considérons un méta-chemin de longueur 1, à savoir $\mathcal{P} = V_s \xrightarrow{E^*} V_c$ qui, contrairement aux simples HINs, n'équivaut pas à un type de liens.

Cela nous permet de définir la probabilité de transition suivant un méta-chemin de longueur 1. Concrètement, la première étape de la transition est la même qu'avant. La différence réside dans le choix du nœud une fois que le marcheur a choisi son hyperarc (i.e. l'étape *ii)*). Cette fois, le marcheur ne peut se diriger que vers un nœud $v$ ayant le type désiré, i.e. $v \in V_c$ (figure 1.5). Pour ce faire, on définit $\delta^-_{V_c}(e) := \sum_{v \in V_c} \lambda_e(v)h^-(v,e)$, le degré entrant de l'hyperarc $e$ respectant $\mathcal{P}$ et plus particulièrement $V_c$.

**Définition 1.19** (Probabilités de transition contrainte par un méta-chemin de longueur 1)**.** Soient $u \in V_s$ et $v \in V_c$ et $\mathcal{P} = V_s \xrightarrow{E^*} V_c$, la probabilité de transition contrainte par $\mathcal{P}$ dans un hypergraphe est donnée par

$$\mathbb{P}_{\mathcal{P}}(v|u) = \sum_{e \in E^*} \frac{\omega(e)h^+(u,e)}{d^+_{E^*}(u)} \frac{h^-(v,e)\lambda_e(v)}{\delta^-_{V_c}(e)}, \tag{1.7}$$

Comme dit, la différence se trouve au niveau de l'étape *ii)* : nous divisons donc par $\delta^-_{V_c}(e)$, où $\delta^-_{V_c}(e) \leq \delta^-(e)$, ce qui accroît donc bien les probabilités dans le sens $\mathbb{P}_{\mathcal{P}}(v|u) \geq \mathbb{P}_{E^*}(v|u)$. Cette probabilité (éq. (1.7)) est bien définie et $\sum_{v \in V_c} \mathbb{P}_{\mathcal{P}}(v|u) = 1$.

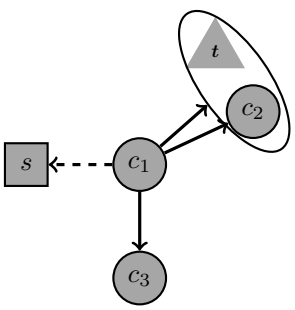



**Figure 1.5:** Hypergraphe dirigé composé de trois types de nœuds et de deux types d'hyperarcs. Soit le méta-chemin de longueur 1 $\mathcal{P}$ : ◯ → ◯. Dans une modélisation graphe appliquée jusqu'à présent, contraindre la probabilité par le méta-chemin $\mathcal{P}$ ou par le type de lien → revient au même: $\mathbb{P}^G_{\mathcal{P}}(c_2|c_1) = \mathbb{P}^G_{\rightarrow}(c_2|c_1) = 2/3$. Dans la modélisation hypergraphe avec $\omega(e) = \delta^-(e)$ et $\lambda_e(v) = 1, \forall e, v$, si on contraint simplement par le type de liens →, $\mathbb{P}^{\mathcal{H}}_{\rightarrow}(c_2|c_1) = 1/2$. Ce qui est différent de contraindre la probabilité par le méta-chemin $\mathcal{P}$, i.e. $\mathbb{P}^{\mathcal{H}}_{\mathcal{P}}(c_2|c_1) = 3/4$.

Nous pouvons enfin généraliser l'éq. (1.7) à un méta-chemin de longueur quelconque.

**Définition 1.20** (Probabilités de transition contrainte par un méta-chemin)**.** Soient deux nœuds

$v_0 \in V_0, v_n \in V_n$ et un méta-chemin $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 ... V_{n-1} \xrightarrow{E_n} V_n$. La probabilité d'atteindre $v_n$, démarrant de $v_0$, contrainte par $\mathcal{P}$, notée $\mathbb{P}_{\mathcal{P}}(v_n|v_0)$, se définit récursivement comme suit

$$\mathbb{P}_{\mathcal{P}}(v_n|v_0) = \sum_{v \in V_{n-1}} \mathbb{P}_{\mathcal{P}^{n-1,n}}(v_n|v) \times \mathbb{P}_{\mathcal{P}^{0,n-1}}(v|v_0) \tag{1.8}$$

avec $\mathbb{P}_{\mathcal{P}^{n-1,n}}(v_n|v)$ défini par l'éq. (1.7).

Les probabilités résultant des expressions (1.2) et (1.8) sont, en principe, différentes. Il sera donc intéressant de les comparer afin d'observer comment la modélisation d'un système peut impacter les analyses qu'on en fait.

## 1.4. Résumé

Dans ce chapitre, nous avons motivé le choix des graphes et hypergraphes d'information hétérogène en les comparant avec d'autres concepts de la littérature, parfois ne différant que par la terminologie utilisée. Ce choix est motivé par l'application que nous aimerions mener, à savoir, récupérer le poids des liens dans des réseaux sociaux dans lesquels différents agents interagissent à travers différents canaux. Des définitions formelles des objets qui seront traités tout au long de ce travail ont été présentées, avec un accent sur les notions d'hypergraphes. Le lien entre graphes et hypergraphes a été aussi discuté au moyen de projections. Finalement, les marches aléatoires contraintes par des méta-chemins ont été présentées. Ces dernières sont la base de la méthode de récupération du poids des liens (chapitre 3) ainsi que d'un modèle d'évolution des liens (chapitre 4).

# 2. Positionnement par rapport au problème de prédiction de liens

La prédiction de liens dans un graphe consiste soit à récupérer les liens non observés, i.e. prédiction de liens *structurels*, soit à prédire l'apparition d'interactions futures, i.e. prédiction de liens *temporels*, à partir d'informations passées et présentes observées dans le graphe. Dans le premier cas, les liens existent déjà mais sont manquants ou imparfaitement observés dans les données. Cela peut être dû à un échantillonnage, à la volonté du nœud/agent de ne pas donner accès à toutes ses données (e.g. les applications sociales en ligne), à des méthodes de collectes de données défectueuses, etc.

La prédiction de liens a été et est un sujet de recherche actif de la théorie des graphes, aussi bien théorique que pratique. Par conséquent, de nombreux algorithmes ont été développés, principalement pour résoudre le problème de la prédiction de l'existence des liens. Néanmoins, prédire l'existence d'un lien n'est pas toujours suffisant. Par exemple, dans un réseau social, le fait de savoir que deux personnes sont liées ne dit rien sur la fréquence de leurs interactions ni sur l'intensité de leur amitié. Par conséquent, récupérer le poids réel du lien peut apporter des informations utiles. Par exemple, dans les systèmes de recommandation, le poids peut représenter la côte qu'un utilisateur donnerait à un article [37, 101, 169]. Actuellement, le problème de la prédiction du poids des liens est principalement abordé dans le cas des graphes d'information homogène.

Ce chapitre est composé de trois sections. Nous présentons le problème générique de la prédiction de liens dans la section 2.1. Ce problème est divisé en deux sous-problèmes : la prédiction de l'existence des liens et la prédiction du poids des liens. Nous expliquons ensuite dans la section 2.2 les difficultés générales de tels problèmes et discutons des choix que nous faisons pour résoudre le problème de la prédiction du poids des liens dans les HIN. Afin de situer nos choix par rapport à l'état de l'art, une partie de la littérature jugée pertinente est présentée dans la section 2.3. Enfin, nous fournissons un bref résumé dans la section 2.4.

## 2.1. Prédiction et récupération de liens et de leur poids

Dans cette section, nous précisons les termes de prédiction et de récupération. Nous présentons ensuite les problèmes de l'existence et de la pondération des liens.

Nous considérons dans cette section un HIN $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$.

### 2.1.1. Prédiction et récupération

Le terme *prédiction* fait souvent référence à une notion de temps. En effet, le problème de la prédiction de l'existence de liens peut se définir comme suit : étant donné un instantané du graphe au temps $t$ (totalement observable), la tâche est de prédire l'ensemble des liens qui se seront formés[1] à l'instant $t' > t$ [104]. Le problème de la *récupération* est fort similaire, à la différence que nous ne nous projetons pas dans le temps. En particulier, la tâche est de récupérer les liens déjà présents dans le système à l'instant $t$ mais qui, pour une raison quelconque, sont manquants dans les données d'observation. Une autre façon intuitive de distinguer ces deux problèmes est que le premier cherche à comprendre les causes menant à la formation de liens tandis que le second cherche plutôt des corrélations entre des données du graphe pour comprendre la présence ou non de liens. Néanmoins, ces deux problèmes sont étroitement liés [34, 91] : ils sont formellement très proches et il arrive quelques fois que le terme prédiction soit utilisé bien qu'aucune notion de temps ne soit en jeu. Dans la suite du travail, nous nous concentrons sur le problème de récupération : nous ne prédirons jamais ni l'existence, ni le poids de liens futurs. Cependant, nous utiliserons aussi bien le terme récupération que prédiction, tout en ayant en tête que le temps est ignoré.

### 2.1.2. Existence et pondération

Ainsi que mentionné, nous ne nous attardons que sur la prédiction de liens manquants. Soient $u, v \in V$. La tâche de prédiction de l'*existence* des liens revient à déterminer si le lien $(u, v)$ existe, i.e. si $(u, v) \in E$. La tâche de prédiction du *poids* des liens revient quant à elle a déterminer le poids du couple de nœuds $(u, v)$, i.e. $w(u, v)$.

Il existe deux façons de résoudre le second problème. La première façon consiste à d'abord prédire l'existence des liens et ensuite prédire le poids des liens prédits. Les poids prédits sont donc strictement positifs. La seconde façon est de prédire simultanément l'existence et le poids des liens: les poids prédits sont donc positifs et un lien non existant est vu comme un lien de poids nul.

Dans la suite, nous utilisons l'expression *prédiction de liens* pour parler du problème générique et lorsque la situation l'exige, nous précisons s'il s'agit de la prédiction de l'existence ou de la

[1] La prédiction de la disparition de liens, bien que possible, n'est pas discutée ici. La tâche qui nous intéresse est simplement de prédire les futurs liens.

pondération de ces liens.

#### Prédiction dans les HIN

Dans un HIN, le problème de la prédiction des liens se fait par rapport à un type de liens. Cela signifie que ce ne sont pas uniquement des liens qui doivent être prédits, mais des *liens typés*. La plupart du temps, tous les liens à prédire sont de même type $E^*$: déterminer si $(u, v) \in E$ avec $\psi[(u, v)] = E^*$ ou déterminer le poids $w_{E^*}(u, v)$. Notons tout de même que certains travaux se sont penchés sur la prédiction de plusieurs types de liens simultanément [25].

### 2.1.3. Évaluation de la prédiction

L'évaluation de la qualité de la prédiction de l'existence des liens dépend du type de sortie de l'algorithme utilisé. Typiquement, l'algorithme attribue un score de vraisemblance $s_{uv}$ à chaque couple $(u, v) \in U \setminus E$, avec $U$ l'univers des liens typés possibles. Il renvoie au final soit un classement de liens potentiels – ceux ayant les scores les plus élevés étant classés en premier – soit le score lui-même. Lorsque ce sont des classements, les $k$ premiers sont estimés comme prédits. Lorsque ce sont les scores qui sont retournés, un lien est effectivement prédit si le score associé est supérieur à un seuil fixé. De nombreuses métriques peuvent ensuite être appliquées afin d'évaluer la performance de la méthode, e.g. précision, rappel, aire sous la courbe (AUC).

Pour la prédiction des poids, deux mesures sont principalement utilisées : l'erreur quadratique moyenne RMSE et le cœfficient de corrélation de Pearson $R^2$, calculés entre la matrice des poids prédits et la matrice des poids observés. Dans ce travail, nous nous concentrons uniquement sur le $R^2$ (et $R^2$ ajusté, expliqué dans suite).

## 2.2. Difficultés, motivations et positionnement

Il existe des difficultés relatives aux problèmes de la prédiction de liens. Ces difficultés sont d'autant plus nombreuses lorsque les problèmes sont attaqués de manière supervisée. Nous évoquons ici quelques-unes d'entre elles.

Lorsque la prédiction de l'existence des liens est vue comme une tâche de classification binaire, le problème du *déséquilibre des classes*, i.e. une disproportion entre le nombre des étiquettes positives et négatives, est presque toujours rencontré. Cela est dû au fait que, quel que soit le graphe considéré (orienté, multigraphe, etc.), parmi tous les liens possibles, seuls quelque-uns existent réellement: les graphes réels sont fort clairsemés. Ce problème est d'autant plus prononcé que le graphe est grand. Il y a donc un risque d'obtenir un très faible taux d'erreur en prédisant toujours

négativement. Il existe plusieurs façons d'y remédier (sans pour autant garantir l'efficacité), e.g. sous- ou sur-échantillonner. Dans le premier cas, cela peut conduire à une perte d'information non négligeable tandis que dans le second, le temps de calcul peut devenir très important. Une autre façon est de pénaliser le modèle d'apprentissage. Dans le problème de prédiction du poids de liens, cela se traduit par le fait que le modèle doit être capable de prédire pour la plupart des couples un poids nul, tout en étant capable de prédire un poids positif pour les liens existant réellement.

Un autre problème est le *choix des prédicteurs*. Tout d'abord, il faut être capable de choisir des prédicteurs informatifs, qui apportent effectivement de l'information pertinente pour prédire les poids. Malheureusement, il est bien souvent ardu de savoir *a priori* si le prédicteur est pertinent ou non. Ensuite, il est toujours appréciable que les prédicteurs ne soient pas trop gourmands en ressources ou temps de calcul, principalement lorsque de grands graphes sont traités. Des solutions existent pour sélectionner les prédicteurs. Par exemple, des méthodes basées sur la variance imposent aux prédicteurs d'avoir une variance supérieure à un certain seuil ou encore des méthodes basées sur des tests statistiques univariés, e.g. test du $\chi^2$.

Un problème lié est celui du *choix du modèle*. En se concentrant uniquement sur la prédiction du poids, ce problème est vu comme un problème de régression consistant donc à prédire une variable quantitative réelle. De nombreux modèles de régressions existent, linéaires ou non. Le choix dépend évidemment de l'application considérée et du but recherché. L'idéal est toujours d'obtenir un modèle le plus simple *et* le plus explicatif/prédictif possible. Cependant, ces deux souhaits sont souvent conflictuels. Dès lors, vaut-il mieux obtenir le modèle maximisant les *mesures de validation* choisies (e.g. cœfficient de détermination, RMSE) quel que soit le nombre de prédicteurs ou avoir un modèle comprenant au plus $p$ prédicteurs (pour une interprétation plus aisée et une complexité moindre) au prix de moins bons scores de validation ? Dans le dernier cas, lors d'une régression linéaire, une régularisation peut être envisagée, i.e. une contrainte sur la norme des cœfficients.

La motivation principale de ce travail est de présenter une méthode pour prédire le poids des liens structurels dans les HIN. Une autre question fondamentale déjà évoquée doit être posée : le problème de la prédiction de l'existence des liens peut-il être considéré comme un problème de prédiction du poids des liens ? Dit autrement, la prédiction du poids doit-elle être menée en une ou deux phases: prédire l'existence des liens manquants et ensuite le poids des liens prédits, ou alors simultanément prédire l'existence et le poids des liens manquants ? Dans le second cas, un lien non existant a un poids nul.

Afin de répondre à cette question, il semble important de soulever un autre point : celui de la *signification* du poids. Nous avons déjà mentionné le fait que nous ne traitons que les poids positifs (section 1.2.1). Cela peut être vu comme e.g. le volume d'écoulement à travers les liens. Dès lors,

bien que structurellement faux, prédire un poids extrêmement faible ou nul (i.e. absence de lien) revient *presque* au même: un écoulement extrêmement faible aura *presque* la même importance que s'il n'existait pas. Cela est aussi à prendre de manière *relative* : imaginons un nœud $v$ relié à $m = m_1 + m_2$ autres dont $m_1$ avec un poids très élevés comparés aux $m_2$ autres et relâchons une substance sur ce nœud $v$. Cette substance va principalement se diffuser au travers des $m_1$ liens ; que les $m_2$ liens existent effectivement ou non n'a, dans certains cas, pas beaucoup d'importance.

Dans ce travail, nous faisons le choix de ne considérer qu'une seule tâche, i.e. simultanément prédire l'existence et le poids des liens manquants. Ce choix est motivé par le fait que ce problème est vu comme un problème de régression. De plus, dans les applications que nous proposons au chapitre 3, le poids d'un lien peut être interprété comme le volume d'écoulement ou la capacité d'un lien à diffuser une information. Bien sûr, cela est discutable, en particulier lorsqu'on pense aux travaux menés sur l'importance des liens faibles, principalement en sciences sociales [66].

Le concept de méta-chemin s'est avéré extrêmement utile pour calculer des similarités entre les nœuds et prédire des liens dans les HIN, nous y reviendrons plus en détails dans la section 2.3.1. Nous faisons donc des méta-chemins la base de notre méthode de prédiction du poids des liens. À l'instar de [181], nous faisons l'hypothèse que la similarité entre un couple de nœuds est corrélée au poids du lien entre ces nœuds. Cette hypothèse est aussi faite en regard de la signification donnée aux liens, comme discuté précédemment.

À la différence des graphes d'information homogène, les nœuds dans un HIN sont similaires *relativement à un type de liens* (section 2.3.1 pour plus de détails), d'où l'intérêt des méta-chemins. Nous aimerions donc que la méthode proposée permette d'agréger plusieurs valeurs de similarités, relatives à différents méta-chemins, afin d'inférer le poids des liens. Cette agrégation doit être facilement *interprétable* : nous devons être capables de quantifier, de manière simple, l'importance de chaque similarité dont la sémantique est donnée par le méta-chemin associé.

Le choix d'une méthode basée sur la similarité, comparé aux méthodes probabilistes ayant recueilli beaucoup de succès (section 2.3.3 et e.g. [4, 34, 76]), est aussi motivé par le fait que ces dernières peuvent être impossibles à calculer pour de grands graphes. C'est aussi, en partie, pour cette raison que de nouvelles mesures de similarité sont très fréquemment proposées.

Évidemment, tous ces choix sont discutables. Nous aurons l'occasion, à travers les différents exemples présentés dans le chapitre 3, de relever une partie de leurs inconvénients mais également, nous l'espérons, de possibles intérêts.

## 2.3. Méthodes de prédiction

Dans cette section, nous rapportons quelques méthodes de résolution des problèmes de prédiction de liens présentés précédemment. Cela permet de situer nos choix par rapport à l'état de l'art.

De manière générale, prédire le poids des liens est nettement moins fréquent dans la littérature que prédire leur existence. Le problème est principalement abordé dans les graphes et, à notre connaissance, très peu a été fait concernant les HIN. Pour ces raisons, nous passons en revue certaines méthodes de prédiction de l'existence des liens dans les HIN tandis que nous abordons principalement les méthodes de prédiction du poids des liens pour les graphes[1]. En outre, puisque nous nous concentrons uniquement sur les graphes statiques, les études mettant l'accent sur la dynamique du graphe ne sont pas rapportées.

### 2.3.1. Prédiction de l'existence des liens dans les HIN

À notre connaissance, il n'existe pas de travaux se concentrant sur la prédiction du poids des liens dans les HIN. Cependant, les deux problèmes de prédiction (existence et pondération) étant relativement proches (section 2.1), nous évoquons une partie des méthodes existantes pour prédire l'existence des liens dans les HIN.

Certaines de ces méthodes se basent sur des mesures de similarité : étant donné une mesure de similarité, un lien est prédit entre deux nœuds s'ils sont similaires [69, 103, 159, 184, 190, 191]. Dans les HIN, les mesures de similarité entre nœuds reposent souvent sur le concept de méta-chemin, permettant de donner une sémantique précise à ces mesures. Étant donné l'importance de ces mesures dans le reste du travail (et dans la littérature), nous parcourons les plus importantes d'entre elles avant de nous plonger dans les méthodes de prédiction de l'existence des liens proprement dites.

#### Mesures de similarité

Nous nous intéressons uniquement aux mesures de similarité basées sur les *chemins* et non aux mesures basées sur des propriétés topologiques locales. Dans les HIN, les mesures de similarité entre deux entités prennent en compte non seulement la structure mais également les méta-chemins entre elles. Parmi ces mesures, PathCount (PC [156]) et Path Constrained Random Walk (PCRW [97], cf. section 1.3) sont deux mesures de base et ont donné naissance à de nombreuses extensions [47, 70, 82, 187].

[1] Pour la prédiction de l'existence des liens dans les graphes, se référer à [105, 107] en général, [72, 104] dans les réseaux sociaux et [151] pour les méthodes spécialement basées sur la similarité, ainsi que les références qui y sont citées.

Les méthodes relatives à PC reposent sur le nombre de chemins entre une paire de nœuds étant donné un méta-chemin. Cependant, PC a le désavantage de favoriser les nœuds fortement connectés puisque le degré des nœuds n'est pas pris en compte. En d'autres mots, $m$ chemins entre deux nœuds $u_1$ et $u_2$ de degré resp. $d_1$ et $d_2$ ont la même importance qu'entre deux nœuds $u'_1$ et $u'_2$ de degré resp. $d'_1 > d_1$ et $d'_2 > d_2$ ; les similarités entre les couples $(u_1, u_2)$ et $(u'_1, u'_2)$ sont donc potentiellement identiques, ce qui n'est pas toujours sensé. Afin de prendre en compte cette remarque, la mesure Normalized PathCount NPC divise le nombre de chemins entre chaque couple de nœuds par le degré des nœuds source et cible [156]. Notons que ce résultat "normalisé" ne l'est que sur l'entièreté du chemin, et non sur chaque étape du chemin: seuls les degrés des nœuds sources et cibles sont pris en compte. PathSim mesure la similarité entre deux objets de même type le long d'un méta-chemin symétrique $\mathcal{P}$, i.e. un méta-chemin égal à son inverse $\mathcal{P} = \mathcal{P}^{-1}$ [157]. D'après cette mesure, pour que deux nœuds soient similaires, ils doivent non seulement être fortement connectés, mais également partager une visibilité comparable. Malheureusement, PathSim ne se calcule que le long de méta-chemins symétriques, ce qui est restrictif puisque nombre de méta-chemins sont asymétriques et la relation entre deux nœuds de types différents peut s'avérer intéressante en pratique. Deux autres mesures basées sur PathSim intègrent plus d'informations telles que le degré et la transitivité [73, 175].

Les méthodes relatives à PCRW sont basées sur des marches aléatoires et donc la probabilité d'atteindre un nœud à partir d'un autre étant donné un méta-chemin. Une adaptation, HeteSim [147], mesure la probabilité de rencontre entre deux marcheurs démarrant des extrémités opposées d'un chemin satisfaisant un méta-chemin donné. Cependant, cette méthode requiert la décomposition des relations atomiques pour les méta-chemins de longueur impaire. Cette décomposition crée artificiellement un nouveau nœud, situé à égale distance des extrémités du méta-chemin. Cela permet aux marcheurs de se rencontrer sur ce nœud, mais s'avère très coûteuse en temps et calculs pour les grands graphes. Pour pallier ce problème, AvgSim [111] calcule la similarité entre deux nœuds recourant à des marches aléatoires contraintes par un méta-chemin et son inverse. Cette méthode est fort appréciable pour les HIN non dirigés puisque dans ces cas, il est tout aussi raisonnable de parcourir un chemin dans une direction que dans l'autre.

Toutes ces mesures se basent sur le *dénombrement* des chemins satisfaisant un méta-chemin. Une autre façon de mesurer la similarité entre deux nœuds est d'utiliser le concept d'*information*. Par exemple, Shakibian *et al.* quantifient l'information fournie par les méta-chemins basée sur la composition de plusieurs matrices de co-occurrence [99, 185] suivant les méta-chemins. Une fois les matrices de co-occurrence extraites, ces dernières sont analysées à l'aide de mesures statistiques telles que l'énergie, l'inertie, l'homogénéité locale, la corrélation et la mesure d'information de la

corrélation afin de calculer la similarité entre les nœuds [143].

#### Prédiction de l'existence des liens

Certaines méthodes généralisent directement des approches *topologiques* des graphes simples. Par exemple, dans [39] est proposé un prédicteur de liens multirelationnel, basé sur une extension à pondération probabiliste de l'indice d'Adamic/Adar [3]. Les poids sont basés sur le dénombrement de triades (i.e. graphlets à trois nœuds) dans le réseau. En outre, les auteurs montrent que les méthodes supervisées fournissent de meilleurs résultats que les non supervisées, particulièrement dans le cas des HIN pour lesquels il est fort difficile de connaître à l'avance quel type d'information est utile.

D'autres méthodes utilisent le concept de *méta-chemins* pour calculer des similarités (voir précédemment) et ensuite se servent d'un algorithme de classification, dont les *features* sont les similarités, afin d'inférer l'existence d'un lien [25, 30, 156, 179]. Sun *et al.* proposent la méthode PathPredict en deux étapes [156]. La première définit des caractéristiques topologiques basées sur un méta-chemin (e.g. PathCount, RandomWalk) qui sont ensuite utilisées dans la seconde phase, consistant en une régression logistique prédisant l'existence ou non d'un lien. Dans [25], Cao *et al.* s'attaquent à la prédiction de plusieurs types de liens simultanément. Ils introduisent une mesure de parenté (*relatedness measure*, RM), basée sur le principe d'homophilie entre différents types de nœuds afin de calculer la probabilité d'existence d'un lien. Testée au moyen d'applications dans le domaine de la bio-informatique, leur méthode (i.e. la mesure RM) fournit, en moyenne, de meilleurs résultats comparés à NPC, PCRW, AvgSim, mais met aussi en avant sa sensibilité aux paramètres de l'algorithme, limitant son application.

Un inconvénient commun à toutes ces méthodes vient du fait qu'elles reposent sur des mesures de similarité dépendant majoritairement du degré de connectivité des paires de nœuds. Par conséquent, ces mesures ne tiennent pas compte d'informations additionnelles fournies par le méta-chemin lui-même. Afin de contourner ce problème, Shakibian *et al.* formulent une méthode non supervisée basée sur la théorie de l'information pour prédire l'existence des liens dans un HIN [142]. L'idée est de définir une entropie de lien basée sur les méta-chemins en vue d'estimer la probabilité de l'existence d'un lien. Plus précisément, la probabilité qu'une paire de nœuds soit connectée est formulée comme une auto-information conditionnelle de l'existence de ce lien sachant un méta-chemin. Dès lors, contrairement aux méthodes basées sur la similarité quantifiant la quantité de connectivité entre un couple de nœuds, leur méthode a pour but d'estimer la quantité d'information passant par les chemins entre deux nœuds. Un avantage de cette méthode est la capacité du modèle à bien prédire les liens, même lorsque le nombre de liens entre les nœuds dans le graphe

devient faible. Néanmoins, dans certains cas, cette méthode fournit de moins bons résultats que les méthodes citées précédemment, en particulier lorsque sont utilisés des méta-chemins moins informatifs. Finalement, bien qu'un nombre quelconque de méta-chemins puissent être utilisés dans cette méthode, les expérimentations menées n'en incluent qu'au plus deux, dont les poids sont déterminés manuellement.

Toujours en se basant sur les méta-chemins, Shakibian *et al.* proposent la méthode MKLP (*Multi-kernel One class Link Predictor*) dans laquelle une machine à vecteurs de support (SVM) à une classe (OC-SVM) est appliquée sur les paires de nœuds positifs, i.e. des paires de nœuds effectivement connectées, pour former le prédicteur de liens [144]. Cette méthode est particulièrement adaptée pour contrer le problème du déséquilibre des classes, quasiment incontournable dans la prédiction de liens. Cependant, comme le soulèvent les auteurs, la question cruciale de MKLP est de savoir comment trouver une fonction de noyau (*graph kernel* : intuitivement, une fonction qui mesure la similarité entre deux graphes) appropriée pour un ensemble donné de données. En outre, la précision des fonctions de noyau définies dépend des méta-chemins employés.

Il existe également des approches *probabilistes*. Yang *et al.* utilisent la propagation de l'influence (*influence propagation* [32]) à travers des relations hétérogènes [174]. Leur méthode, *Multi-Relational Influence Propagation* (MRIP), s'avère utile dans le cadre des graphes épars, comme c'est le cas de nombreux graphes réels et de leur cas d'application : des relations de coauteurs scientifiques. Dans [43], Dong *et al.* proposent un modèle *ranking factor graph* (RFG) afin de prédire et de recommander des liens dans un réseau social. Ensuite, les auteurs font l'hypothèse que, même si les gens créent des liens dans différents graphes (i.e. différents types de liens), les principes sous-jacents sont similaires. Cela leur permet de trouver différents motifs sociaux à travers les HIN, i.e. sous-graphes significativement sur-représentés, et de développer un modèle basé sur le transfert (cf. apprentissage par transfert [121]), qui combine les motifs et d'autres informations structurelles.

### 2.3.2. Prédiction du poids des liens dans des graphes multiplexes

Comme mentionné dans le chapitre 1, les graphes multiplexes peuvent être vus comme des cas particuliers de HIN. Des méthodes de prédiction de liens spécifiques aux graphes multiplexes ont été développées, e.g. [87, 116, 129, 132]. Nous discutons ici d'une méthode de prédiction des poids dans un multiplexe pondéré non-orienté [146]. La couche dans laquelle la prédiction doit être faite est appelée couche cible tandis que les $L-1$ autres couches sont les couches prédictives. L'ensemble de nœuds $V$ est le même pour toutes les couches mais chaque couche possède un sous-ensemble différent de liens. La prédiction de poids est décomposée en deux sous-problèmes : prédiction de l'existence des liens et prédiction du poids des liens.

Pour la prédiction de l'existence des liens, un score global est calculé entre chaque paire de nœuds, $s_{uv}$, afin d'inférer la présence du lien $(u, v)$ dans la couche cible. Ce score global est une moyenne pondérée des scores calculés pour chaque couche. Le score de $(u, v)$ dans la couche $l$ est simplement donné par 1 si le lien $(u, v)$ existe dans cette couche $l$, 0 sinon. Le poids du score dans la couche prédictive $l$ est donné par la probabilité d'avoir un lien dans la couche cible étant donné que ce lien existe dans la couche $l$. La qualité de cette première étape est mesurée au moyen de l'AUC et de la précision.

Ensuite, pour prédire le poids d'un lien $(u, v)$, les auteurs supposent que le poids de ce lien dépend des liens qui relient $u$ et $v$ et leurs voisins respectifs. En particulier, ils cherchent le voisin $a$ (resp. $b$) de $u$ (resp. $v$) pour lequel le score, calculé lors de l'étape précédente, est le plus élevé. Une moyenne pondérée est ensuite déduite des poids entre des voisins sélectionnés, supposés connus, et le nœud correspondant du lien ciblé. Le poids est utilisé pour la normalisation au cas où les scores $s_{au}, s_{uv}$ et $s_{bv}, s_{uv}$ présentent une différence importante. En formule,

$$w_{uv} = \frac{\beta_1 w_{au} + \beta_2 w_{bv}}{2}$$
$$\text{avec } \beta_1 = 1 + \frac{s_{au} + s_{uv}}{s_{au}} \text{ et } \beta_2 = 1 + \frac{s_{bv} + s_{uv}}{s_{bv}},$$

avec $w_{uv}$ et $s_{uv}$ le poids et le score entre $u$ et $v$. La qualité de cette étape est calculée grâce à l'erreur quadratique moyenne normalisée.

## 2.3.3. Prédiction du poids des liens dans les graphes d'information homogène

### Méthodes basées sur la similarité

Les méthodes de *similarité* sont basées sur l'hypothèse que le poids du lien incident à deux nœuds est proportionnel à la similarité entre ceux deux nœuds. Par exemple, dans le cas d'un graphe non dirigé, Zhao *et al.* emploient une régression linéaire de la forme $w_{uv} = c \cdot s_{uv}$, avec $w_{uv}$ le poids du lien entre $u$ et $v$, $s_{uv}$ la similarité entre $u$ et $v$ et $c$ le cœfficient de la régression, déterminé par la résolution d'un problème d'optimisation [181]. Notons que les poids sont tous normalisés pour se trouver dans [0,1], suivant différentes fonctions de normalisation, e.g. logistique, linéaire, exponentielle négative. La similarité est calculée sur base des voisins communs et s'exprime comme $s_{uv} = \sum_{z \in \mathcal{N}(u) \cap \mathcal{N}(v)} I(z)$, où $\mathcal{N}(u)$ est le voisinage de $u$ et $I(z)$ est un indice pouvant prendre différentes valeurs. En notant $k_u$ (resp. $s_u$) le degré (resp. degré pondéré) du nœud $u$, les différents indices $I(z)$ s'expriment comme repris dans la table 2.1.

En plus de mesures de similarité classiques telles que Adamic/Adar, voisins communs, etc., entre

| | Voisins communs | Adamic Adar | Allocation de ressources |
|---|---|---|---|
| Non-pondéré | 1 | $\frac{1}{\log k_z}$ | $\frac{1}{k_z}$ |
| Pondére | $w_{uz} + w_{vz}$ | $\frac{w_{uz}+w_{vz}}{\log(1+s_z)}$ | $\frac{w_{uz}+w_{vz}}{s_z}$ |
| *Reliable-route* | $w_{uz} \cdot w_{vz}$ | $\frac{w_{uz} \cdot w_{vz}}{\log(1+s_z)}$ | $\frac{w_{uz} \cdot w_{vz}}{s_z}$ |

**Table 2.1.:** Différentes formes de l'indice $I(z)$ pour le calcul de la similarité entre deux nœuds.

les nœuds, Fu *et al.* utilisent des mesures de centralité de liens comme *features* de méthodes supervisées [51]. Afin de calculer les mesures de centralité des liens, ils font usage du graphe adjoint $L(G)$ (*line graph*) du graphe original $G$. Pour rappel, les nœuds de $L(G)$ sont les liens de $G$ et deux nœuds de $L(G)$ sont connectés si les liens correspondants dans $G$ sont adjacents. Dès lors, les mesures de centralité des liens sont les mesures de centralité "usuelles" calculées sur les nœuds du graphe adjoint. Leurs résultats montrent qu'il est bénéfique pour la prédiction du poids des liens de prendre en compte des informations (i.e. mesures de centralité) sur les liens, même si la mesure *Ressource Allocation* calculée directement sur les nœuds est la plus significative, en accord avec les travaux de [186] pour la prédiction de liens manquants. Notons qu'ils utilisent des méthodes supervisées telles que les forêts d'arbres décisionnels (RF) ou les SVM, qui, malgré toutes leurs forces, ont quelques désavantages à être utilisées pour de la régression. Par exemple, les RF ne peuvent pas extrapoler, ce qui peut être préjudiciable si les poids dans les ensembles d'entraînement et de validation ne sont pas dans les mêmes fourchettes de valeurs. Pour les SVM, il faut définir des paramètres supplémentaires (comparé aux tâches de classification) qui ne sont pas évidents. En outre, les méthodes basées sur les SVM ne conviennent pas toujours bien pour l'apprentissage et la prédiction d'échantillons à grande échelle, car elles résolvent le vecteur support au moyen d'un calcul quadratique, ce qui peut consommer beaucoup d'espace et de temps de calcul lorsqu'il s'agit de matrices d'ordre élevé.

#### Méthode basée sur l'ensemble des voisins

Zhu *et al.* soulèvent le problème de l'hypothèse de corrélation entre les poids et les indices de similarité : bien que cela soit avéré dans de nombreuses situations, il ne s'agit pas d'une généralité. Dès lors, en s'appuyant sur l'information structurelle des plus proches voisins d'un nœud, Zhu *et al.* proposent une méthode permettant dans un premier temps de prédire l'existence d'un lien et ensuite d'inférer les poids de ces liens prédits [189]. La méthode repose sur l'hypothèse que la formation de poids de liens est régulée par des regroupements locaux (*local clustering*) dans lesquels les liens ont tendance à avoir des poids similaires. À nouveau, les poids sont normalisés pour appartenir à l'intervalle [0,1]. Contrairement à la méthode présentée dans [181], les poids sont

directement déterminés par une mesure ne nécessitant pas de résoudre un problème d'optimisation (cœfficient $c$ de la régression linéaire). Sur un ensemble de six jeux de données, les auteurs montrent que la qualité de la prédiction des poids est relativement similaire à celle obtenue par la méthode dans [181] : de meilleurs résultats par rapport au $R^2$ mais de moins bons pour le RMSE.

#### Méthodes probabilistes

Une autre approche consiste à envisager des modèles *probabilistes*. Parmi ces derniers, il y a les méthodes à blocs stochastiques (SBM), un modèle connu pour apprendre la structure communautaire des réseaux non pondérés [76]. Nous rappelons brièvement son fonctionnement afin de comprendre son extension aux graphes pondérés.

Initialement, le SBM est un modèle probabiliste d'interactions par paires entre $n$ nœuds. Chaque nœud $i$ appartient à un groupe (ou bloc) latent, parmi $K$ possibles, dénoté par $z_i$. Finalement, la probabilité de chaque lien $A_{ij}$ ($A_{ij} \in \{0,1\}$) est donnée par le paramètre $\theta_{z_i z_j}$ ($\boldsymbol{\theta} \in \mathbb{R}^{K\times K}$). Dit autrement, en supposant que l'existence de chaque $A_{ij}$ soit conditionnellement indépendante étant donnés $\mathbf{z}$ et $\boldsymbol{\theta}$, $A_{ij}$ est une variable aléatoire suivant une distribution de Bernoulli, i.e. $A_{ij} \sim B(1, \theta_{z_i z_j})$. La fonction de vraisemblance du SBM est donnée par $\mathbb{P}(\mathbf{A}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{ij} \theta_{z_i z_j}^{A_{ij}} (1 - \theta_{z_i z_j})^{1-A_{ij}}$. En prenant le log et en réécrivant cela suivant une famille exponentielle, l'objectif du SBM est de trouver les paramètres $\mathbf{z}$ et $\boldsymbol{\theta}$ telle que la probabilité logarithmique d'observer $\mathbf{A}$ soit maximisée

$$\log\left[\mathbb{P}(\mathbf{A}|\mathbf{z}, \boldsymbol{\theta})\right] = \sum_{ij} \left[T(A_{ij}) \cdot \eta(\theta_{z_i z_j})\right],$$

où $T(A_{ij}) = (A_{ij}, 1)$ (resp. $\eta(\boldsymbol{\theta}) = (\log[\boldsymbol{\theta}/(1-\boldsymbol{\theta})], \log[1-\boldsymbol{\theta}])$) est la fonction vectorielle des statistiques exhaustives (resp. des paramètres naturels) d'une variable de Bernoulli.

Afin d'étendre le SBM aux graphes pondérés, noté WSBM (*weighted* SBM), $A_{ij}$ n'est plus tirée d'une loi de Bernoulli mais d'une famille exponentielle, paramétrée par $\theta_{z_i z_j}$ [4]. À titre d'exemple, supposons des poids réels et une distribution normale $\mathcal{N}(\mu, \sigma^2)$. $\theta_{z_i z_j}$ devient le paramètre de distribution des poids entre les groupes $(z_i, z_j)$, i.e. $\theta_{z_i z_j} = (\mu_{z_i z_j}, \sigma^2_{z_i z_j})$.

Notons que WSBM ne prend en compte que l'information contenue dans les poids des liens. Afin de tirer parti de l'existence et des poids, Aicher *et al.* introduisent le modèle bWSBM (*balanced* SBM) [4]. En notant $(T_e, \eta_e)$ (resp. $(T_w, \eta_w)$) la famille des distributions de l'existence (resp. des poids) dans le modèle SBM, la fonction de vraisemblance devient

$$\log\left[\mathbb{P}(\mathbf{A}|\mathbf{z}, \boldsymbol{\theta})\right] = \alpha \sum_{ij\in E} \left[T_e(A_{ij}) \cdot \eta_e(\theta_{z_i z_j})\right] + (1-\alpha) \sum_{ij\in W} \left[T_w(A_{ij}) \cdot \eta_w(\theta_{z_i z_j})\right],$$

avec $\alpha \in [0,1]$ le paramètre régulant la contribution de chaque terme, $E$ (resp. $W$) l'ensemble des interactions observées (resp. des poids des liens), avec $W \subset E$.

Finalement, incorporer le degré des nœuds est aussi faisable en modifiant $(T_e, \eta_e)$ du bWSBM par $T_e(A_{ij}) = (A_{ij}, -k_i k_j)$ et $\eta_e(\boldsymbol{\theta}) = (\log \boldsymbol{\theta}, \boldsymbol{\theta})$, $k_i$ étant le degré de $i$. Cette correction dcSBM (*degree corrected* SBM) est appréciable pour gérer les distributions de degrés à queue lourde, connues pour provoquer de mauvais résultats pour le SBM [90].

#### Méthodes basées sur l'apprentissage profond

D'autres travaux se basent sur l'*apprentissage profond*. L'idée est de construire un algorithme qui apprend efficacement les informations complexes et non observables sur les nœuds, i.e. les vecteurs de nœuds, à partir de relations simples et observables entre les nœuds, i.e. les poids de liens, et qui utilise ces informations pour prédire le poids des liens inconnus. Par exemple, Hou *et al.* utilisent un réseau de neurones complètement connecté, baptisé modèle **R** (pour Relation) [79]. Il est composé *i)* d'une première couche (*mapping layer*) transformant les nœuds en vecteurs, format d'entrée indispensable pour les réseaux de neurones ; *ii)* d'une couche d'entrée activée par une fonction des nœuds ; *iii)* d'un nombre ajustable de couches cachées complètement connectées dont la fonction d'activation (i.e. ce qui permet d'introduire des non-linéarités dans le réseau) est $f(x) = \max\{0, x\}$. L'objectif de ces couches est d'apprendre à extraire des informations abstraites concernant le poids et finalement ; *iv)* d'une couche de sortie dont l'activation est une fonction linéaire $f(x) = kx$, renvoyant le poids entre les deux nœuds passés en entrée à partir des informations obtenues des couches précédentes. Leurs expérimentations ont montré que l'apprentissage profond surpassait les trois méthodes probabilistes présentées précédemment, i.e. WSBM, bSBM et dcSBM.

Un travail actuellement mené par Hou *et al.* est de construire une première approche générique d'apprentissage profond reposant sur le plongement de nœuds pour la prédiction de liens [78]. L'idée est de généraliser leur modèle **R** en utilisant d'autres méthodes de plongement de nœuds tels que word2vect [112], node2vect [68], plongement localement linéaire [137], tout en gardant la phase d'apprentissage des poids des liens du modèle **R**. Cela conduit au modèle **S**. Leur conclusion est que le plongement du modèle **R** surpasse les autres plongements envisagés, confortant l'efficacité de leur premier modèle.

#### Méthodes sur des graphes signés

Les travaux que nous avons cités jusqu'à présent sont menés sur des graphes non signés. Cependant, dans des situations réelles, des liens signés peuvent être utiles. Par exemple, dans un réseau social, lorsque la relation entre deux personnes est haineuse (resp. amicale), il peut être sensé de munir

ce lien d'un poids négatif (resp. positif). Ces graphes sont dits *signés*. Dans ce type de graphes, Kumar *et al.* prédisent le poids des liens manquants à l'aide du produit de deux nouvelles mesures: *fairness* et *goodness* [94]. La *fairness* d'un nœud permet de mesurer le niveau d'équité de ce nœud lorsqu'il évalue les autres nœuds, tandis que la *goodness* du nœud indique intuitivement à quel point ce nœud est apprécié des autres.

Une autre méthode applicable aux graphes signés est proposée dans [95], où le problème générique de prédiction de liens (existence et pondération) est vu sous un angle *algébrique*. En particulier, les auteurs montrent qu'un certain nombre de noyaux de graphes et d'autres algorithmes de prédiction de liens peuvent être interprétés comme la transformation spectrale de la matrice d'adjacence ou laplacienne du graphe. En limitant les méthodes de prédiction de liens aux noyaux de cette forme, un problème de minimisation pouvant être réduit à un problème de régression unidimensionnelle est défini. Ce formalisme permet d'estimer les paramètres des différents noyaux de graphes et de comparer visuellement les noyaux de graphes en utilisant l'ajustement des courbes (*curve fitting*).

### 2.3.4. Ajout d'information d'ordre supérieur

L'objectif de ce travail n'est pas de prédire des hyperliens ou hyperarcs comme dans e.g. [145, 176, 180] mais simplement d'utiliser des informations provenant des relations impliquant plus de deux nœuds à la fois afin de prédire le poids d'un lien, i.e. relation dyadique, dans un HIN, à l'instar de [23, 100]. Li *et al.* modélisent un réseau social par un hypergraphe et utilisent une méthode de classement, inspirée de [183], pour estimer la proximité entre les nœuds et finalement prédire les liens, i.e. des interactions reliant uniquement deux nœuds sont prédites [100]. Dans [23], le problème est un peu différent puisqu'il s'agit de recommander des musiques en essayant de combiner plusieurs types d'informations sur les médias sociaux et signaux acoustiques musicaux. Les auteurs modélisent toutes ces informations par un *hypergraphe unifié*, i.e. un hypergraphe non orienté avec plusieurs types de nœuds et hyperliens. Leur problème de recommandation se base sur le filtrage collaboratif ainsi qu'une méthode de classement, fort inspirée de [100]. Ce travail est intéressant car il prend explicitement en compte les différents types de nœuds et liens dans sa tâche de recommandation.

Dans [59], Gerow *et al.* proposent de retrouver le poids des liens dans un hypergraphe dense et pondéré modélisant un réseau social académique. Selon eux, les liens sont caractérisés par deux quantités : leur masse et leur similarité avec d'autres liens. En utilisant la masse des liens à travers les voisins communs de deux nœuds $u$ et $v$, pondérée par la similarité de ces mêmes liens, Gerow *et al.* proposent des scores entre $u$ et $v$, comparables au poids du lien $(u, v)$. Afin d'évaluer leur méthode, ils ne s'attaquent pas à comparer directement les poids mais ils calculent la corrélation de

Spearman entre le classements des scores obtenus et celui des véritables poids des liens. Finalement, ils pointent le fait que cette méthode est surtout adaptée pour un réseau social déjà très dense, i.e. une partie de la force du modèle est due à la structure des données elle-même.

## 2.4. Résumé

Dans ce chapitre, nous avons défini les problèmes de la prédiction et de la récupération de l'existence et du poids des liens. Nous avons aussi exposé des difficultés générales liés au problème générique de la prédiction de liens et des possibles solutions. Le problème de la prédiction du poids étant fort peu abordé dans les HIN, nous avons principalement rapporté quelques méthodes de prédiction de l'existence des liens dans ces graphes. Ces méthodes ont permis de mettre en avant la richesse des HIN et l'importance de prendre en compte la sémantique contenue dans les nœuds et liens afin d'améliorer la prédiction de l'existence des liens. Nous avons poursuivi avec les méthodes pour la prédiction du poids des liens dans les graphes d'information homogène. Ce bref parcours de la littérature nous a permis de nous rendre compte que, malgré l'importance et les avantages des HIN, peu de choses ont été faites pour prédire le poids des liens dans ces graphes. En essayant de combiner différents ingrédients ayant fait leurs preuves pour prédire le poids des liens dans les graphes d'information homogène et ceux adaptés aux HIN, nous avons rendu clair notre objectif de prédiction de poids dans les HIN et motivé le travail présenté dans le chapitre suivant.

# 3. Prédiction de liens manquants par marches aléatoires contraintes

Soient un HIN $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$ avec $G = (V, E, w, \mu_s, \mu_t)$ un graphe orienté et pondéré et un type de liens cible $E_c \in \mathcal{E}$. L'objectif de la méthode présentée dans ce chapitre est de prédire le poids des liens de type $E_c$, i.e. $w_{E_c}(u, v)$, $\forall (u, v)$. Toutefois, nous ne tentons pas de prédire les poids tels qu'ils sont présents dans le graphe mais plutôt des poids normalisés entre [0,1]. Cette normalisation correspond aux distributions des marches aléatoires contraintes par des méta-chemins. Cela signifie que l'objectif réel est de trouver la distribution de probabilités $\mathbb{P}_{E_c}(: |u)$, $\forall u \in V$ (section 1.3.1). Puisque $\sum_{v \in V} \mathbb{P}_{E_c}(v|u) = 1$, chaque valeur $\mathbb{P}_{E_c}(v|u)$ peut s'interpréter comme l'importance relative que $u$ accorde à $v$. Lorsque le degré pondéré sortant des nœuds $d_{E_c}(u)$ est connu, alors il est possible de retrouver le poids des liens (non normalisés) puisque $w_{E_c}(u, v) = d_{E_c}(u)\mathbb{P}_{E_c}(v|u)\, \forall (u, v)$.

Ce chapitre est organisé comme suit. La section 3.1 introduit l'approche proposée, contribution de ce chapitre. Nous y présentons le modèle de régression et discutons des choix faits. Afin de mesurer le potentiel et de se rendre compte des limites de cette approche, nous considérons quelques cas d'étude. En particulier, un premier exemple permettant d'illustrer simplement l'approche est présenté dans la section 3.2, une deuxième application sur des données bibliographiques se concentrant plus sur les méta-chemins – leur interprétation et leurs limites – dans la section 3.3 et une troisième application sur un jeu de données Twitter plus conséquent et complexe dans la section 3.4. Ces illustrations suggèrent qu'il est possible de récupérer et prédire le poids des liens à partir des corrélations présentes entre les liens de types différents d'un HIN. Le fait d'utiliser des méta-chemins comme base permet en outre d'interpréter facilement ces corrélations. Ces différents cas d'études permettent également de lister certaines limites de l'approche utilisée. Nous poursuivons avec une discussion sur l'intérêt des hypergraphes pour la prédiction du poids des liens dans la section 3.5. Nous montrons que ce formalisme permet d'obtenir de meilleurs résultats comparé aux HIN simples. Finalement nous résumons le travail présenté et proposons quelques idées de perspectives dans la section 3.6.

## 3.1. Approche

Le problème de la prédiction du poids des liens de type $E_c$ est vu comme un problème de régression linéaire où la variable à expliquer est la distribution $\mathbb{P}_{E_c}(:|u)$, $\forall u$. Les variables explicatives (VE) sont liées au concept de méta-chemin. Il s'agit des distributions de probabilités obtenues en suivant un méta-chemin $\mathcal{P} \neq E_c$ dont les types de nœuds source et cible sont les même que ceux de $E_c$ : $\mathbb{P}_{\mathcal{P}}(:|u)$, $\forall u$. De cette façon, les VE englobent de l'information structurelle et sémantique.

### 3.1.1. Modèle de régression linéaire

Soient un HIN $H$ et un type de liens cible $E_c$ entre deux types de nœuds $V_s$ et $V_t$. L'objectif est de trouver un ensemble de méta-chemins $\mathcal{E}_{\mathcal{P}}$ et une fonction linéaire $F_{\mathcal{E}_{\mathcal{P}}}$ dépendant de cet ensemble et combinant les distributions de probabilités résultant des marches aléatoires contraintes par ces méta-chemins[1], i.e. $\mathbb{P}_{\mathcal{P}}(:|u)$ avec $\mathcal{P} \in \mathcal{E}_{\mathcal{P}}$, de manière ) ce que $F_{\mathcal{E}_{\mathcal{P}}}$ approxime au mieux le poids normalisé des liens de type $E_c$, i.e. $\mathbb{P}_{E_c}(:|u)$. Plus formellement, pour chaque paire de nœuds $(u, v) \in V_s \times V_t$, nous cherchons $\mathcal{E}_{\mathcal{P}}$ et $F_{\mathcal{E}_{\mathcal{P}}}$ de la forme

$$F_{\mathcal{E}_{\mathcal{P}}}(u, v) := \beta_0 + \sum_{\mathcal{P} \in \mathcal{E}_{\mathcal{P}}} \beta_{\mathcal{P}}\, \mathbb{P}_{\mathcal{P}}(v|u), \tag{3.1}$$

où le vecteur des cœfficients $\boldsymbol{\beta} := [\beta_0, \beta_{\mathcal{P}_1}, \cdots, \beta_{\mathcal{P}_{|\mathcal{E}_{\mathcal{P}}|}}]^{\mathsf{T}}$, à valeurs réelles, quantifie la contribution de chaque méta-chemin à la valeur finale de $F_{\mathcal{E}_{\mathcal{P}}}$. Nous y associons le modèle linéaire suivant

$$\underbrace{\begin{bmatrix} \mathbb{P}_{E_c}(:|u_1)^{\mathsf{T}} \\ \vdots \\ \mathbb{P}_{E_c}(:|u_{|V_s|})^{\mathsf{T}} \end{bmatrix}}_{=:\mathbf{y}} = \underbrace{\underbrace{\begin{bmatrix} 1 & \mathbb{P}_{\mathcal{P}_1}(:|u_1)^{\mathsf{T}} & \cdots & \mathbb{P}_{\mathcal{P}_{|\mathcal{E}_{\mathcal{P}}|}}(:|u_1)^{\mathsf{T}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbb{P}_{\mathcal{P}_1}(:|u_{|V_s|})^{\mathsf{T}} & \cdots & \mathbb{P}_{\mathcal{P}_{|\mathcal{E}_{\mathcal{P}}|}}(:|u_{|V_s|})^{\mathsf{T}} \end{bmatrix}}_{=:\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{\mathcal{P}_{|\mathcal{E}_{\mathcal{P}}|}} \end{bmatrix}}_{=:\boldsymbol{\beta}}}_{=:\mathbf{F}_{\mathcal{E}_{\mathcal{P}}}} + \underbrace{\begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_{|V_s|} \end{bmatrix}}_{=:\boldsymbol{\epsilon}} \tag{3.2}$$

où $\mathbb{P}_{\mathcal{P}}(:|u) = [\mathbb{P}_{\mathcal{P}}(v_0|u), \cdots, \mathbb{P}_{\mathcal{P}}(v_{|V_t|}|u)]$ est le vecteur des probabilités ($\mathbf{y}$ est donc un vecteur colonne de dimension $|V_s|.|V_t|$) et $\boldsymbol{\epsilon} \in \mathbb{R}^{|V_s|.|V_t|}$ est appelé le vecteur (colonne) des erreurs ou des perturbations (i.e. $\boldsymbol{\epsilon}_i \in \mathbb{R}^{|V_t|}$).

En général, ce système est surdéterminé et ne possède donc pas de solution exacte (car système inconsistant). L'objectif est donc de résoudre pour $\boldsymbol{\beta}$ le problème d'optimisation quadratique $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, c'est-à-dire $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$. De façon équivalente, $\hat{\boldsymbol{\beta}}$ minimise la somme

[1] Soit le méta-chemin $\mathcal{P} = V_0 \cdots V_{n-1} \xrightarrow{E_n} V_n$. Pour tout $i \in \{2, ..., n\}$, si $E_i = E_c$, alors le marcheur ne peut retourner sur le nœud de départ $v_0 \in V_0 = V_{i-1}$. Afin d'éviter cela, on corrige *a posteriori* : dans le calcul de l'éq. (1.2), on impose $\mathbb{P}_{E_{i-1}}(v_0|v) = 0$ et on renormalise les autres probabilités pour obtenir $\sum_{v'(\neq v_0)} \mathbb{P}_{E_{i-1}}(v'|v) = 1$, $\forall v \in V_{i-2}$. Cela évite d'utiliser ce que nous cherchons pour trouver ce que nous cherchons.

résiduelle du carré des erreurs (*sum of squared estimate of errors*) SSE $= \|\boldsymbol{\epsilon}\|^2$. Lorsque les colonnes de $\mathbf{X}$ sont linéairement indépendantes, ce problème admet une unique solution donnée par les équations normales $(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$. Ce qui donne l'estimateur $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ : le vecteur de l'hyperplan des moindres carrés. Nous considérons donc tout simplement le problème des moindres carrés ordinaire (*Ordinary Least Squares* OLS).

**Remarque 3.1** (Autre choix de VE)**.** Comme vu au chapitre 2, la valeur $\mathbb{P}_{\mathcal{P}}(v|u)$, représentant la probabilité d'atteindre $v$ partant de $u$ et suivant uniquement des chemins satisfaisant $\mathcal{P}$, peut être vue comme la similarité entre les nœuds $u$ et $v$. Dans la suite, nous l'appelons PCRW. Dès lors, nous pouvons utiliser d'autres mesures de similarité telles que mentionnées dans la section 2.3.1. En particulier, nous rappelons deux mesures qui seront utilisées dans la suite du travail :

- Path Count (PC) [156] : dans sa forme initiale, il s'agit du nombre de chemins connectant deux nœuds. Il ne s'agit donc plus d'une probabilité. Nous considérons ici des chemins pondérés. Plus précisément, $\mathrm{PC}_{\mathcal{P}}(v_n|v_0) = \sum_{p\in\mathcal{P}} \sum_{e_i\in p} w(e_i)$, avec $e_i$ les liens du chemin $p$ ;
- AvgSim [111] : il s'agit de la moyenne arithmétique de deux probabilités résultant des marches aléatoires contraintes par deux méta-chemins, où l'un est l'inverse de l'autre : $\mathrm{AvgSim}_{\mathcal{P}}(v_n|v_0) = [\mathbb{P}_{\mathcal{P}}(v_n|v_0) + \mathbb{P}_{\mathcal{P}^{-1}}(v_0|v_n)]/2$.

De cette façon, la probabilité $\mathbb{P}_{\mathcal{P}}(v_n|v_0)$ peut être remplacée dans l'éq. (3.1) par les mesures de similarité définies ci-dessus. Le cadre proposé est donc tout à fait général et modulable.

### 3.1.2. Sélection de modèles et variables explicatives

#### Critères de sélection

Il existe de nombreux critères de sélection de modèles dans la littérature sur la régression linéaire multiple. Ici, nous envisageons seulement les cœfficients $R^2$ et $R^2_a$. Le cœfficient de détermination $R^2 = 1 - \frac{SSE}{SST} \in [0,1]$ est un score utilisé pour tester la qualité globale du modèle, avec $SST = \sum(y_i - \bar{y})^2$ la somme totale des carrés, i.e. la somme des carrés de la différence entre la variable dépendante et sa moyenne, et $SSE = \sum(y_i - \hat{y}_i)^2$ la somme des carrés des résidus ($\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$). Le $R^2$ représente la proportion de variance dans la variable dépendante qui est prédictible à partir des VE. Il est directement relié aux erreurs et il est évident que minimiser le SSE maximise le $R^2$. Cependant, le défaut principal du $R^2$ est d'être monotone croissant en fonction du nombre de VE. Il sert donc principalement à comparer deux modèles ayant le même nombre de VE.

Un excès de VE a tendance à produire des modèles peu robustes. En effet, si un modèle possède trop de VE, il modélisera en réalité le bruit aléatoire dans les données. Il aura donc un $R^2$ très élevé mais trompeur puisqu'il ne pourra prédire correctement de nouvelles observations. Le modèle

devient dès lors inutilement complexe, peu pertinent et difficile à expliquer. C'est pourquoi nous nous intéressons davantage au cœfficient de détermination ajusté $R_a^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$, où $n$ est le nombre d'observations et $k$ est le nombre de VE. Le $R_a^2$ peut se réécrire comme suit : $R_a^2 = 1 - \frac{(n-1)MSE}{SST}$ où MSE = SSE$/(n-k-1)$ est l'erreur quadratique moyenne. Maximiser $R_a^2$ équivaut donc à minimiser le MSE.

#### Algorithme de sélection

Lorsque l'ensemble $\mathcal{E}_\mathcal{P}$ des VE est *fort* grand, il n'est pas toujours possible (e.g. limite en temps de calcul) de tester les $2^{|\mathcal{E}_\mathcal{P}|}$ modèles possibles afin de déterminer le meilleur au sens du $R_a^2$. Nous avons donc recours à des stratégies de sélection. En particulier, nous optons pour un algorithme de sélection pas-à-pas avec le critère de la $p$-valeur. Il s'agit d'une approche gloutonne mais très simple et intuitive.

Les $p$-valeurs sont utilisées pour tester la significativité de chaque VE $\mathbb{P}_{\mathcal{P}_i}$, $\mathcal{P}_i \in \mathcal{E}_\mathcal{P}$. Étant donnée l'hypothèse nulle $H_0 : \beta_i = 0$ contre l'hypothèse $H_1 : \beta_i \neq 0$, la $p$-valeur est la probabilité d'observer, sous $H_0$, une statistique au moins aussi extrême que la valeur observée. L'hypothèse $H_0$ est rejetée au niveau $\alpha$ si $p \leq \alpha$ en faveur de $H_1$. Sinon, nous rejetons $H_1$ en faveur de $H_0$.

Étant donné $k$ VE (i.e., les distributions de probabilités résultant des marches aléatoires contraintes), la sélection pas-à-pas fonctionne comme suit

- Commencer avec un modèle nul, i.e. aucune VE mais juste une constante. Traditionnellement, il s'agit de la moyenne de la variable à expliquer ;
- Essayer $k$ modèles de régression linéaire avec une seule VE parmi $\mathcal{E}_\mathcal{P}$ et choisir celui qui fournit le meilleur modèle vis-à-vis du critère de sélection, ici maximiser $R_a^2$ ;
- Chercher parmi les VE restantes celle qui, ajoutée au modèle, donne un meilleur résultat, i.e. un $R_a^2$ plus élevé tel que toutes les VE dans le modèle soient significatives, i.e. leur $p$-valeur est en dessous d'un certain seuil $\alpha$ fixé. Itérer cette étape jusqu'à ce qu'il n'y ait plus d'amélioration possible.

### 3.1.3. Tâches descriptives et prédictives

La modélisation statistique peut être profitable pour divers objectifs, parfois complémentaires. Dans ce travail, nous nous intéressons aux fins *descriptives* et *prédictives*. La première tâche vise à rechercher de façon exploratoire des liens entre la variable dépendante $\mathbf{y}$ et d'autres variables, potentiellement explicatives, $\mathbf{X}$. Le but n'est donc pas d'expliquer *toutes* les variables possibles mais bien de se concentrer sur la variable dépendante. Pour la seconde tâche, l'accent est mis sur la qualité des estimateurs et des prédicteurs qui doivent, e.g. minimiser une erreur quadratique

moyenne. Ceci mène à la recherche de modèles économes, i.e. avec un nombre volontairement restreint de VE. Un bon modèle est un modèle qui conduit aux prédictions les plus fiables.

Afin de tester la fiabilité d'une prédiction, nous utilisons dans ce travail la validation croisée Monte Carlo, i.e. une validation répétée de sous-échantillonnages aléatoires [173]. Étant donné un ensemble de $N$ données, la méthode Monte Carlo les divise en un sous-ensemble d'entraînement $s_t$, sur lequel le modèle est entraîné, et un sous-ensemble de test/validation $s_v$, servant à la validation du modèle. Cette procédure est répétée plusieurs fois et les résultats de la prédiction sont ensuite moyennés. Notons que les résultats d'une validation croisée Monte Carlo tendent vers ceux d'une validation croisée exhaustive *leave-p-out* [8] puisque le nombre de divisions aléatoires tend vers l'infini. Un inconvénient de la validation Monte Carlo vient du fait que certaines observations peuvent ne jamais être sélectionnées ou à l'inverse, peuvent être utilisées à chaque division. En outre, les résultats dépendent de la division (variations Monte Carlo). Cependant, il y a des avantages (par rapport à e.g. la validation croisée *k-folds* [8]) puisque le nombre de données dans les ensembles $s_t$ (et donc $s_v$) est indépendant des divisions (*folds*). Cela signifie que Monte Carlo permet d'explorer un peu plus de divisions possibles, bien qu'il soit peu probable de les considérer toutes puisqu'il existe $\binom{N}{|s_t|}$ sous-ensembles d'entraînement uniques.

### 3.1.4. Discussion des choix

Le choix de la méthode OLS pour déterminer le vecteur des cœfficients $\hat{\boldsymbol{\beta}}$ est assez discutable. Cependant, nous le motivons par sa simplicité, son côté intuitif, le fait que l'estimateur obtenu soit le meilleur non biaisé, son emploi toujours d'actualité et répandu et sa puissance d'explication/prédiction.

Nous devons cependant pointer quelques inconvénients. Le premier est celui de la *multicolinéarité*. En ayant en tête que nos prédicteurs sont relatifs aux méta-chemins, nous pouvons imaginer qu'il y ait très certainement des corrélations entre les VE. Cela se remarque facilement lors de l'estimation des paramètres $\hat{\boldsymbol{\beta}}$ qui nécessite le calcul explicite de $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. Quand il y a multicolinéarité, la matrice $\mathbf{X}^\mathsf{T}\mathbf{X}$ est mal conditionnée et cela mène à des estimateurs de variances importantes, voire à des problèmes de précision numérique. Cependant, puisque notre objectif est soit une tâche de description, soit une tâche de prédiction, un remède – certes drastique – est tout simplement de ne pas prendre en compte ces variables[1]. Cela se fait grâce aux procédures de sélection de modèles expliquées précédemment.

Lorsque nous nous attaquons à une tâche prédictive, une autre objection peut venir du fait que l'estimateur OLS est *non biaisé* : l'espérance de $\hat{\boldsymbol{\beta}}$ égale $\boldsymbol{\beta}$ et donc, le biais $Biais(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = 0$. Cela peut conduire au sur-apprentissage (figure 3.1). Cela nous amène à évoquer le compromis "biais-

[1] Si le but était d'*expliquer* toutes les variables, une suppression aussi drastique ne serait pas la bienvenue.

variance". Dit simplement, le biais est la contribution à l'erreur totale des hypothèses simplificatrices intégrées dans le modèle choisi, tandis que la variance est la contribution à l'erreur totale due à la sensibilité de la méthode au bruit dans les données. La méthode OLS est souvent représentée à l'extrémité droite de la figure 3.1 (faible biais, forte variance). Une façon d'y remédier est la *régularisation* avec des modèles tels qu'une *Rigde Regression* (régularisation L2), *Lasso Regression* (régularisation L1). Le problème d'optimisation devient alors

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_p^2, \quad p=1,2,$$

avec $\lambda \geq 0$ le paramètre de régularisation. Ces régularisations pénalisent donc les modèles avec beaucoup de prédicteurs et/ou des cœfficients élevés. L'idée étant que si nous surestimons l'impact d'un prédicteur, i.e. un coefficient élevé, il est probable que nous sur-adaptons. Cependant, il est bien connu que, lorsque le nombre d'observations est bien plus grand que celui des prédicteurs, ce qui est notre cas, l'estimateur OLS tend à avoir peu de variance. Finalement, un inconvénient en pratique du Lasso et d'autres techniques de régularisation est de trouver le coefficient de régularisation $\lambda$ optima [88]. L'utilisation de la validation croisée pour trouver cette valeur peut être aussi coûteuse que les techniques de sélection pas à pas. Pour ces raisons, nous nous contenterons simplement de la méthode OLS.

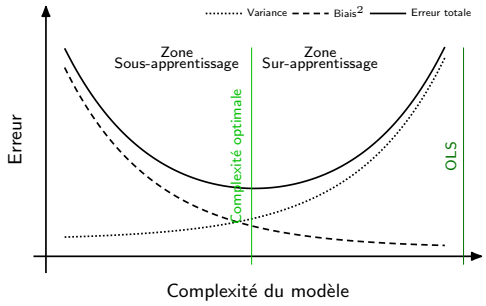



**Figure 3.1.:** Illustration du dilemme "biais-variance" dans la sélection des modèles.

Concernant l'algorithme de sélection pas-à-pas, un fait critiquable est celui d'obtenir un modèle qui n'est pas le meilleur possible, dû au caractère glouton de la méthode. En effet, nous n'obtenons jamais qu'un *optimum local*. Cependant, un premier avantage vient du simple fait qu'observer l'ordre dans lequel les VE sont ajoutées au modèle est souvent informatif sans pour autant noyer le lecteur dans des classements qui feraient perdre la vue d'ensemble. Ensuite, lorsque le nombre de VE envisageables est trop important (potentiellement infini), une méthode gloutonne de sélection est nécessaire. Nous faisons ici le choix d'une sélection pas-à-pas mais des méthodes d'échange[1] ou

[1]Parmi les méthode d'échange, nous pouvons citer l'algorithme de maximisation du $R^2$ qui tente de trouver le

globales[1] auraient également pu être choisies.

Finalement, nous utilisons un modèle de *régression linéaire* car nous sommes intéressés, en plus de la qualité de la prédiction, par son interprétabilité et sa simplicité. Cela exclut, selon nous, les méthodes telles que du *boosting* sur des arbres de décision ou les SVM, qui, bien qu'étant souvent meilleures en terme de prédiction, sont nettement plus abstruses. Enfin, rappelons qu'il n'y a pas de meilleurs modèles dans l'absolu, cela dépend évidemment de l'application et des données utilisées. Par exemple, certaines études ont montré que les prédictions obtenues via des méthodes pas-à-pas ou avec régularisation étaient souvent similaires en pratique [31, 88].

## 3.2. Premier exemple d'application : expliquer et prédire un type de liens

L'objectif de cette section est d'illustrer, sur un simple exemple, l'approche proposée à la section 3.1 afin d'expliquer les poids des liens d'un type ciblé à partir d'autres données présentes dans un HIN.

### 3.2.1. Cas d'étude : la coupe du monde 2014 sur Twitter

#### Présentation des données

Les données utilisées sont un ensemble de tweets relatifs à la coupe du monde de la FIFA 2014. Cet évènement s'est tenu du 12 juin au 13 juillet 2014.

Twitter permet aux twittos, nom générique donné aux entités possédant un compte Twitter et identifiables par le symbole @, de publier de courts messages, i.e. des tweets, comprenant éventuellement des hashtags (H), i.e. une séquence ininterrompue de caractères précédée du symbole #. Dans cette section, nous appelons les twittos les utilisateurs (U).

Twitter permet en outre aux utilisateurs d'interagir via différents types de relation ou actions, nous en considérons trois : retweet RT, reply RP, mention MT. La relation RT signifie qu'un utilisateur diffuse un tweet précédemment publié par un autre. L'action RP est simplement un tweet de réponse à un autre utilisateur en relation avec son tweet précédent. L'action MT, quant à elle, se produit lorsqu'un utilisateur mentionne explicitement un autre dans son message. Finalement, il est possible de *poster* des tweets : il s'agit de la relation post entre un utilisateur et un hashtag.

meilleur modèle pour chaque niveau, i.e. pour chaque nombre de VE. À chaque niveau, l'algorithme sélectionne la VE qui accroît le plus le $R^2$. Il regarde ensuite tous les échanges possibles entre une VE présente dans le modèle et une autre non présente et garde celle qui fournit l'accroissement maximum ; l'algorithme finit lorsqu'il n'est plus possible d'accroître le $R^2$.

[1] Un exemple d'algorithme global est la technique *Leaps-and-Bounds* [52] : il compare tous les modèles possibles et sélectionne celui qui optimise au mieux le $R^2$ (ou tout autre critère de sélection). En outre, cet algorithme écarte certains modèles appartenant à des sous-branches de l'arborescence dont il est possible de savoir *a priori* qu'ils ne sont pas compétitifs.

#### Construction du HIN

À partir de ces relations, nous construisons un HIN avec deux types de nœuds $\mathcal{V}$ ={U, H} et quatre types de liens $\mathcal{E}$ ={RT, RP, MT, post}. Chaque nœud représente donc soit un utilisateur, soit un hashtag. Un lien typé U$\xrightarrow{\text{RT}}$U, U$\xrightarrow{\text{RP}}$U ou U$\xrightarrow{\text{MT}}$U est créé de $u \in$ U à $v \in$ U, si $u$ RT $v$, $u$ RP $v$ ou $u$ MT $v$ et le poids de ce lien correspond au nombre de fois que cette action apparaît dans les données. Pour les relations post, un lien U$\xrightarrow{\text{post}}$H existe entre $u \in$ U et $h \in$ H, si $h$ apparaît dans le tweet de $u$ et le poids du lien correspond au nombre de fois que $u$ écrit $h$ dans un de ses tweets. Notons que nous excluons les hashtags présents dans les posts retweetés puisque dans ces cas, les utilisateurs ne les écrivent pas eux-mêmes. En outre, les considérer provoquerait une corrélation triviale entre e.g. U$\xrightarrow{\text{post}}$H and U$\xrightarrow{\text{RT}}$U$\xrightarrow{\text{post}}$H. Tous les liens sont orientés et pondérés. Étant donné qu'il n'y a qu'un seul type de liens entre U et H, nous notons simplement U→H.

Les données font intervenir 14 000 utilisateurs et 14 000 hashtags uniques. 6 000 utilisateurs sont connectés entre eux par 20 000 liens de type RT, 9 000 par 12 000 de type RP et 12 000 par 61 000 de type MT. Étant donné que les données sont relatives à la coupe monde de la FIFA, les hashtags les plus présents sont ceux faisant référence aux 32 pays impliqués dans la phase finale et en particulier les demi-finalistes : Brésil, Allemagne, Pays-Bas et Argentine ; ainsi que ceux qui se réfèrent directement à l'évènement e.g. #WorldCup, #Brasil, #CM2014.

#### Objectif

Dans la suite de cette section, nous fixons le type de liens "post" (U→H) comme type de liens cible. Cela signifie qu'à partir d'autres types de liens et méta-chemins présents dans le HIN, nous tentons d'expliquer et de prédire le poids normalisé des liens de type "post", i.e. la distribution de probabilités $\mathbb{P}_{\text{post}}(: |u)$, pour tout utilisateur $u$ (figure 3.2).

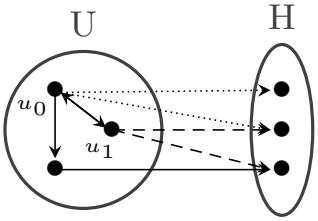



**Figure 3.2:** Illustration de l'objectif U→H. Pour l'utilisateur $u_0$ (resp. $u_1$), l'objectif est de récupérer ou prédire $\mathbb{P}_{\text{post}}(: |u_0)$ (resp. $\mathbb{P}_{\text{post}}(: |u_1)$) : tous les liens sont connus hormis les liens en pointillé court (resp. long).

### 3.2.2. Première régression et interprétation de la solution obtenue

Nous appliquons l'approche proposée (section 3.1) sur l'entièreté du jeu de données FIFA. Les VE données en entrée sont celles associées à tous les méta-chemins de longueur inférieure ou égale

à 4. Une discussion concernant la longueur des méta-chemins est proposée dans la section 3.2.4. Néanmoins, une première motivation pour ces méta-chemins vient du fait qu'au plus le méta-chemin est long, au plus la sémantique est sibylline. En outre, étant donné qu'il y a quatre types de liens dans le HIN, il s'agit de la longueur maximale d'un méta-chemin sans répétition d'un même type de liens. Pour une première illustration, les VE sont calculées à partir de PCRW. Nous comparons par la suite avec d'autres mesures pour calculer les VE.

La figure 3.3 reprend les résultats. Le modèle final (Mod. 5), i.e. le modèle obtenu lorsque l'algorithme s'arrête, contient 5 VE associées à des méta-chemins dont la longueur n'excède pas 3, et une ordonnée à l'origine $\beta_0$ nulle. Ce modèle de régression linéaire est capable de déterminer 71,29% de la distribution des probabilités observées. Pour confirmer la qualité de l'ajustement du modèle, un tracé de densité des probabilités prédites par rapport à celles observées est présenté sur la figure 3.3. La ligne noire représente le cas idéal où les probabilités prédites correspondent aux probabilités observées. La plupart des points tombent sur cette ligne, ce qui tend à valider l'utilisation d'un modèle linéaire.

Nous observons aussi que la meilleure amélioration vis-à-vis du $R_a^2$ vient de l'ajout de la deuxième VE (Mod. 2). Ce modèle est en fait un extremum local puisque, après calculs complémentaires, le modèle à deux VE avec le $R_a^2$ le plus élevé est celui qui fait intervenir les méta-chemins $\mathrm{U} \xrightarrow{\mathrm{RT}} \mathrm{U} \rightarrow \mathrm{H}$ et $\mathrm{U} \xrightarrow{\mathrm{RP}} \mathrm{U} \rightarrow \mathrm{H}$ ($R_a^2 = 0.6116$). Bien que la différence entre le modèle optimal et celui obtenu par l'algorithme soit ténue, cela permet de mettre en évidence deux faiblesses de notre approche : il n'y a aucune garantie de trouver le meilleur modèle et l'ordre de sélection des variables est important. Notons que les deux premières VE font partie des relations les plus directes (méta-chemins de longueur 2), ce qui est intuitif : le voisinage direct d'un utilisateur partage avec celui-ci des sujets d'intérêt communs.

Le dernier méta-chemin $\mathrm{U} \xrightarrow{\mathrm{MT}} \mathrm{U} \xrightarrow{\mathrm{RT}} \mathrm{U} \rightarrow \mathrm{H}$ inclus dans le modèle (Mod. 5) provoque un changement important des autres coefficients. Après exploration, cela est dû à la présence de valeurs aberrantes dans cette VE, i.e. des observations qui diffèrent fortement de la tendance exprimée par les autres observations. Il est bien connu que la méthode OLS y est sensible. En effet, après une identification assez grossière de ces valeurs aberrantes[1] et leur imputation par la moyenne des valeurs de la VE en question, ce méta-chemin ne fait plus partie du modèle et l'algorithme s'arrête après la quatrième itération.

Finalement, nous appliquons la même procédure avec d'autres mesures de similarité. Ainsi qu'illustré dans la table 3.1, les résultats finaux sont différents. Néanmoins, la VE associée au méta-

[1] Les valeurs à l'extérieur de l'intervalle [Q1 - (3*IQR), Q3 + (3*IQR)] avec Q1, Q3 et IQR=Q3-Q1 respectivement le quartile inférieur, supérieur et l'intervalle interquartile.

| Mod. | Méta-chemin | $\hat{\beta}$ | $p$-valeur | $R_a^2$ |
|---|---|---|---|---|
| 0 | Moyenne: 1.8704e-05 | | | 0 |
| 1 | $U \xrightarrow{MT} U \rightarrow H$ | 1.0289 | - | 0.4606 |
| 2 | $U \xrightarrow{MT} U \rightarrow H$ | 0.9391 | 0.0057 | 0.6111 |
| | $U \xrightarrow{RP} U \rightarrow H$ | 0.0052 | 0.0137 | |
| 3 | $U \xrightarrow{MT} U \rightarrow H$ | 0.8464 | 0.0062 | 0.6682 |
| | $U \xrightarrow{RP} U \rightarrow H$ | 0.0335 | 0.0124 | |
| | $U \xrightarrow{RT} U \xrightarrow{RP} U \rightarrow H$ | 0.1077 | 0.0138 | |
| 4 | $U \xrightarrow{MT} U \rightarrow H$ | 0.8114 | 0.0063 | 0.6947 |
| | $U \xrightarrow{RP} U \rightarrow H$ | 0.0362 | 0.0109 | |
| | $U \xrightarrow{RT} U \xrightarrow{RP} U \rightarrow H$ | 0.0766 | 0.0142 | |
| | $U \xrightarrow{RP} U \xrightarrow{MT} U \rightarrow H$ | 0.0676 | 0.0143 | |
| 5 | $U \xrightarrow{MT} U \rightarrow H$ | 0.1974 | 0.0094 | 0.7129 |
| | $U \xrightarrow{RP} U \rightarrow H$ | 0.5556 | 0.0146 | |
| | $U \xrightarrow{RT} U \xrightarrow{RP} U \rightarrow H$ | 0.0650 | 0.0125 | |
| | $U \xrightarrow{RP} U \xrightarrow{MT} U \rightarrow H$ | 0.1591 | 0.0160 | |
| | $U \xrightarrow{MT} U \xrightarrow{RT} U \rightarrow H$ | 0.0074 | 0.0124 | |

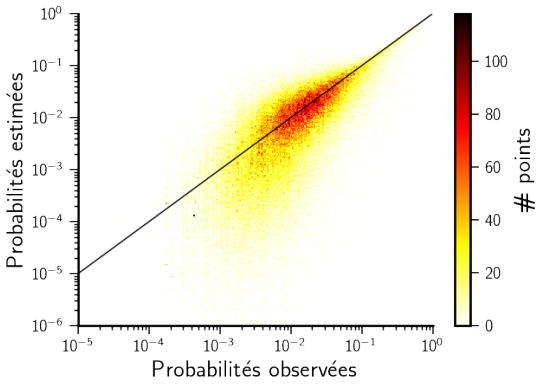



**Figure 3.3.:** Gauche : Résultats des régressions linéaires. Les VE sont construites à partir des PCRW respectant des méta-chemins dont la longueur n'excède pas 4. Le modèle 0 correspond au modèle nul : aucune VE si ce n'est une constante (moyenne de la variable expliquée). Droite : tracé de densité des valeurs observées et estimées pour le modèle final Mod.5 calculé avec la mesure de similarité PCRW. La ligne noire représente l'adéquation parfaite entre les données observées et estimées.

chemin $U \xrightarrow{MT} U \rightarrow H$ est toujours la première à entrer dans le modèle de régression. Les résultats de PCRW et PC sont les plus similaires au niveau des VE présentes. Une explication possible pourrait être le fait que ces deux mesures considèrent un "trajet à sens unique" contrairement à AvgSim qui calcule un "trajet aller-retour". Enfin, il semblerait que ce soit la mesure AvgSim qui offre les meilleurs résultats (i.e. $R_a^2$).

**Table 3.1:** Résumé des modèles descriptifs finaux pour les trois mesures de similarité considérées : PCRW, PC et AvgSim. Pour chaque mesure de similarité, nous indiquons les valeurs du cœfficient $\hat{\beta}$ associé au méta-chemin. Les nombres précédant les cœfficients indiquent l'ordre dans lequel les méta-chemins sont inclus dans le modèle final.

| Méta-chemin | PCRW | PC | AvgSim |
|---|---|---|---|
| $U \xrightarrow{RT} U \rightarrow H$ | | | (5) 0.1629 |
| $U \xrightarrow{RP} U \rightarrow H$ | (2) 0.5556 | (2) 0.2534 | |
| $U \xrightarrow{MT} U \rightarrow H$ | (1) 0.1974 | (1) 0.5550 | (1) 0.5567 |
| $U \xrightarrow{RT} U \xrightarrow{RP} U \rightarrow H$ | (3) 0.0650 | (3) 0.0289 | (3) 0.0380 |
| $U \xrightarrow{RT} U \xrightarrow{MT} U \rightarrow H$ | | (5) 0.0763 | |
| $U \xrightarrow{RP} U \xrightarrow{RP} U \rightarrow H$ | | | (2) 0.3840 |
| $U \xrightarrow{RP} U \xrightarrow{MT} U \rightarrow H$ | (4) 0.1591 | | |
| $U \xrightarrow{MT} U \xrightarrow{RT} U \rightarrow H$ | (5) 0.0074 | (4) 0.0216 | (6) 0.1070 |
| $U \xrightarrow{MT} U \xrightarrow{RP} U \rightarrow H$ | | | (4) 0.1154 |
| $R_a^2$ | 0.7129 | 0.6778 | **0.7649** |

### 3.2.3. Pouvoir prédictif de la solution obtenue

Nous validons la méthode en effectuant une tâche ayant pour but de prédire les poids des liens manquants dans les données FIFA. En d'autres termes, nous tentons de répondre à la question suivante : est-il possible de connaître, de manière quantitative, les hashtags postés par certaines

personnes en sachant ce que font d'autres personnes (i.e. ce que ces dernières postent ainsi que leurs relations avec d'autres individus) ?

Nous effectuons une validation croisée Monte Carlo avec 80% des utilisateurs comme sous-ensemble d'entraînement $s_t$ et obtenons le vecteur $\hat{\boldsymbol{\beta}}$. Ensuite, nous l'utilisons sur le sous-ensemble de validation $s_v$, c'est-à-dire les 20% restants, et nous calculons le $R_a^2$ associé à chaque modèle. Nous réitérons sur dix divisions différentes, i.e. nous créons dix sous-ensembles d'entraînement (et donc de validation).

Les modèles finaux ne comportent pas tous les mêmes VE, cela dépend évidemment des 80% sélectionnés. Le nombre final de VE varie entre cinq et six. Néanmoins, quel que soit le sous-ensemble $s_t$, la VE construite à partir du méta-chemin $\mathrm{U}\xrightarrow{\mathrm{MT}}\mathrm{U}\rightarrow\mathrm{H}$ est toujours choisie lors de la première itération de l'algorithme. Après, il n'y a plus de consensus sur la deuxième VE mais $\mathrm{U}\xrightarrow{\mathrm{RP}}\mathrm{U}\rightarrow\mathrm{H}$ et $\mathrm{U}\xrightarrow{\mathrm{RT}}\mathrm{U}\xrightarrow{\mathrm{RP}}\mathrm{U}\rightarrow\mathrm{H}$ sont toujours en compétition pour la deuxième place. À nouveau, il n'est pas surprenant d'obtenir le méta-chemin $\mathrm{U}\xrightarrow{\mathrm{RP}}\mathrm{U}\rightarrow\mathrm{H}$ puisqu'il relie un utilisateur à ses plus proches voisins dans le HIN. De plus, ce méta-chemin est très faiblement corrélé au méta-chemin $\mathrm{U}\xrightarrow{\mathrm{MT}}\mathrm{U}\rightarrow\mathrm{H}$, sélectionné lors de la première itération de l'algorithme.

Bien que les meilleurs scores $R_a^2$ des modèles finaux atteignent en moyenne 0.7 sur les sous-ensembles d'entraînement $s_t$, nous n'obtenons en moyenne qu'un score de 0.5 pour les sous-ensembles de validation $s_v$ (figure 3.4). L'approche utilisée semble donc atteindre une limite et illustre bien le problème de sur-apprentissage. En effet, même si un modèle correspond mieux à l'ensemble d'entraînement, cela ne signifie pas qu'il donnera le meilleur score de prédiction. Il est donc parfois préférable de considérer un modèle avec moins de VE – et donc un $R_a^2$ inférieur pour l'ensemble d'entraînement – pour parvenir à une meilleure prédiction.

Les résultats présentés sur la figure 3.4 sont obtenus avec les VE construites à partir de PCRW. Étant donné que $R^2_{a,PC} < R^2_{a,PCRW}$ pour la tâche descriptive (table 3.1), nous n'utilisons pas la mesure PC pour la tâche prédictive. Cependant, nous calculons également les scores avec la mesure AvgSim pour laquelle nous obtenons $\langle R^2_{a,pred,AvgSim}\rangle = 0.55$, ce qui est légèrement supérieur au score moyen de PCRW.

### 3.2.4. Équilibre entre qualité et complexité de la solution

Comme déjà mentionné, l'objectif poursuivi est d'obtenir un modèle permettant d'expliquer ou prédire $\forall u$ la distribution $\mathbb{P}_{\mathrm{post}}(:|u)$ qui soit informatif tout en étant interprétable. Cela nous amène à discuter du compromis entre la qualité et la complexité du résultat de la régression linéaire. En particulier, nous nous intéressons aux VE et aux méta-chemins servant à les construire. Pour un HIN donné, le nombre de méta-chemins possibles est infini. Il est donc impératif de définir un

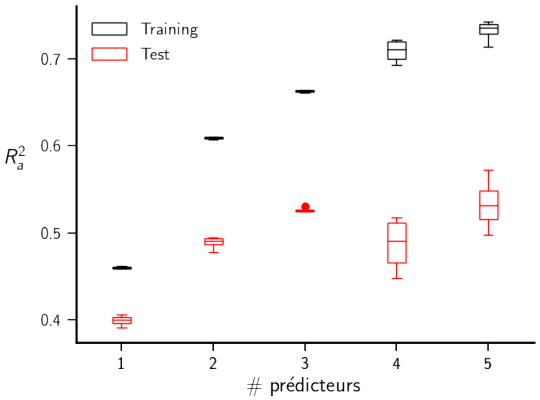



**Figure 3.4:** Boîtes à moustaches des résultats $R_a^2$ des sous-ensembles d'entraînement (noir) et de validation (rouge). Les scores des sous-ensembles d'entraînement augmentent avec le nombre de VE dans le modèle tandis que pour l'ensemble de validation, les scores semblent atteindre un seuil. Ces scores sont obtenus au moyen de validation croisée Monte Carlo sur base de 10 échantillons.

sous-ensemble de méta-chemins potentiels, nécessaires pour calculer les VE. Jusqu'à présent, les méta-chemins choisis pour construire les VE l'ont été de façon tout à fait intuitive, sans aucune quantité mesurable pour défendre le choix. Nous discutons dans cette section la longueur des méta-chemins ainsi que leur répétitions.

### Répétition d'un type de liens dans les méta-chemins

La première propriété à laquelle nous nous intéressons est la *longueur* du méta-chemin. L'objectif est de quantifier, au moyen du $R_a^2$ et sur une tâche descriptive, l'apport d'une VE associée à un méta-chemin de longueur $L$ répétant $l = L - 1$ fois le même type de liens afin d'expliquer la variable dépendante, à savoir les distributions de probabilités $\mathbb{P}_{\text{post}}(: |u)$. Par exemple, pour $l = 2$ et le type de liens RT, le méta-chemin est U $\xrightarrow{\text{RT}}$ U $\xrightarrow{\text{RT}}$ U $\rightarrow$ H, représentant les hashtags postés par les utilisateurs retweetés par les utilisateurs retweetés par l'utilisateur initial. Intuitivement, l'importance d'un méta-chemin décroît en fonction de sa longueur puisque considérer de longs méta-chemins revient à considérer des voisinages plus étendus, d'où des informations plus diffuses. Du point de vue du marcheur aléatoire, cela signifie que ce dernier peut atteindre un nombre beaucoup plus important de nœuds, dont certains sont relativement loin de celui de départ.

Nous pouvons apprécier sur la figure 3.5(a) une tendance décroissante du $R_a^2$ en fonction de la longueur du méta-chemin, venant ainsi corroborer ce propos. De plus, chaque type de liens apporte une quantité d'informations différente et le type MT est le plus informatif pour notre objectif. Cette analyse expose également une caractéristique de la dynamique du type RP : la plupart du temps, les RP impliquent seulement deux utilisateurs [106] (cela s'apparente à une conversation entre deux utilisateurs). Cela se reflète au travers des oscillations des scores $R_a^2$ associés à RP. Les scores associés à un méta-chemin de longueur impaire sont plus bas puisque le marcheur ne peut pas retourner sur le nœud initial lors du pénultième pas de la marche (figure 3.6, cf. section 3.1). Sont aussi reportés en noir les $R_a^2$ lorsque nous ne différencions pas les types de liens (agrégé). Ce score

est inférieur à la moyenne des trois autres scores (types différenciés). Aussi, nous remarquons que le type MT ou RP est plus informatif que l'agrégation, ce qui renforce la pertinence de différencier explicitement les types de liens.

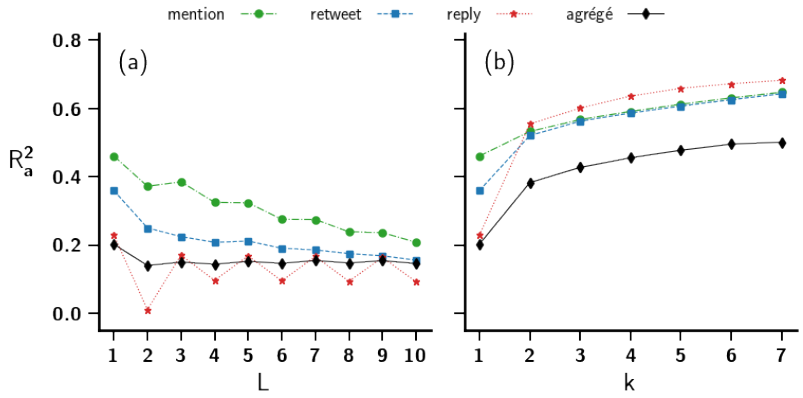



**Figure 3.5.:** (a) Importance de la longueur $L = l + 1$ des méta-chemins. Pour un méta-chemin donné, les $l$ premiers types de liens sont identiques. (b) Importance du nombre $k$ de VE. Pour un modèle de régression donné, ses $k$ VE sont relatives à tous les méta-chemins de longueur au plus $k + 1$ dont les $k$ premiers types de liens sont répétitifs.

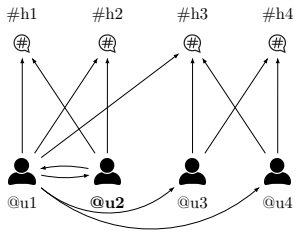



**Figure 3.6.:** Exemple type du cas RP axé sur l'utilisateur $u_2$. Les hashtags postés par $u_2$ sont $h_1$ and $h_2$. Les probabilités résultantes de la marche U→H commençant en $u_2$ sont $[1/2, 1/2, 0, 0]^\mathsf{T}$. Pour les méta-chemins de longueur 2, un marcheur démarrant de $u_2$ suivant le méta-chemin U $\xrightarrow{\text{RP}}$ U → H doit aller, avec probabilité 1 sur $u_1$, et ensuite sur $h_1$, $h_2$ et $h_3$. Les probabilités résultantes sont $[1/3, 1/3, 1/3, 0]^\mathsf{T}$. Pour un méta-chemin de longueur 3, le marcheur ne peut revenir sur $u_2$ après être passé sur $u_1$: il doit aller soit sur $u_3$ soit $u_4$. Mais étant donné que ces derniers ne sont pas reliés à $u_2$ via le type RP, leur hashtags sont plus différents. Les probabilités sont $[0, 0, 1/2, 1/2]^\mathsf{T}$, ce qui est loin de celles obtenues avec U→H: $[1/2, 1/2, 0, 0]^\mathsf{T}$. Conséquemment, le $R_a^2$ est faible (dans ce cas extrême, il est nul). Cependant, pour un méta-chemin de longueur 4, le marcheur peut retourner sur $u_2$ après avoir été sur $u_1$ donc au prochain pas (le troisième pas), le marcheur peut seulement sauter sur $u_1$ qui est un voisin direct de $u_2$. Le raisonnement est le même pour de plus longs méta-chemins : pour des longueurs paires, le marcheur n'est pas affecté par la restriction concernant l'avant dernier pas de la marche, tandis que pour des longueurs impaires, cela a une importance capitale.

### Nombre de variables explicatives et méta-chemins répétitifs

La figure 3.5(b) reprend les $R_a^2$ des modèles lorsque les VE propres à chaque modèle sont relatives au même type de liens mais sont de longueurs variables. Plus précisément, le $R_a^2$ associé à un modèle composé de $k$ VE mesure la qualité du modèle dont les VE sont associées à tous les méta-chemins

de longueur au plus $k+1$ dont les $k$ premiers segments sont répétés, i.e. $k$ répétitions du même type de liens. Par exemple, pour $k=3$ et le type RT, les VE sont U $\xrightarrow{\text{RT}}$ U $\rightarrow$ H, U $\xrightarrow{\text{RT}}$ U $\xrightarrow{\text{RT}}$ U $\rightarrow$ H et U $\xrightarrow{\text{RT}}$ U $\xrightarrow{\text{RT}}$ U $\xrightarrow{\text{RT}}$ U $\rightarrow$ H.

À nouveau, au plus il y a de VE, au plus le score $R_a^2$ est élevé. Néanmoins, cette croissance n'est pas linéaire : la meilleure amélioration se produit lorsque nous combinons les VE associées aux méta-chemins de longueur 1 et 2, indiquant la nécessité de les considérer ensemble. Les scores résultants des types RT et MT sont fort similaires lorsque plus de deux variables sont considérées tandis qu'il y a une claire différence pour une seule variable. Cela signifie que leurs combinaisons respectives ont le même impact en terme de $R_a^2$ malgré leurs sémantiques différentes. À nouveau, le $R_a^2$ pour l'agrégation est bien inférieur aux autres scores.

Pour résumer, étant donné qu'il est toujours désirable d'obtenir un modèle simple en terme d'interprétabilité et de temps de calcul, il y a un compromis entre le $R_a^2$ le plus élevé possible et le coût pour l'atteindre. Les tests menés ici tendent à montrer que considérer des méta-chemins trop longs et trop nombreux n'est pas nécessairement utile dans notre cas. Le gain dans le $R_a^2$ n'en vaut pas la peine compte tenu de la complexité qu'il apporte. Ceci est conforme aux travaux portant sur d'autres objectifs tels que la similarité des nœuds ou le partitionnement : un méta-chemin d'une longueur relativement courte est suffisant pour évaluer la similarité, et un méta chemin plus long peut même détériorer la qualité [147, 157].

## 3.3. Aller plus loin : interprétation et structure des méta-chemins

Dans la section précédente, nous avons déjà pointé l'intérêt de considérer des méta-chemins de longueur restreinte. Nous continuons à examiner le rôle des méta-chemins et ce qu'ils apportent dans l'analyse de données. Pour ce faire, nous considérons un autre cas d'application que sont les publications scientifiques.

### 3.3.1. Cas d'étude : co-publication d'articles scientifiques sur DBLP

Nous nous penchons sur les réseaux bibliographiques scientifiques. Grâce à la diversité des informations contenues dans ces réseaux, e.g. auteurs, papiers, venues, etc, ils sont parfaitement adaptés à une modélisation HIN [142, 156, 157, 175]. Ces réseaux sont utilisés pour de multiples tâches d'analyse de graphes telles que le partitionnement de nœuds et la prédiction de liens, et plus particulièrement de liens de co-publication.

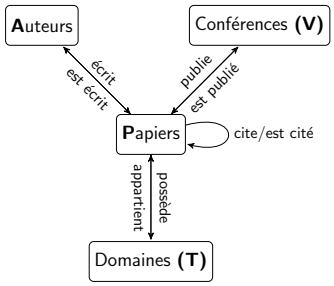
Analyser les relations de co-publication permet de voir, e.g., s'il existe des groupes ou communautés parmi les auteurs, si le fait de participer à une même conférence permet de nouvelles collaborations ou encore si les auteurs collaborent et publient différemment selon leur(s) domaine(s) d'expertise.

### Présentation des données et construction du HIN

Dans cette section, nous utilisons une partie de la bibliographie (de publications scientifiques) numérique informatique DBLP.

Afin de constuire un HIN, quatre types de nœuds sont extraits des données: Auteurs (A), Papiers (P), *Venues*[1] (V), et Domaines (T) ainsi que huit types de liens: écrit/est écrit, publie/est publié, possède/appartient, cite/est cité (figure 3.7(a)) [162]. Les données contiennent 96 000 auteurs avec 1 600 000 relations de co-publication ainsi que 186 000 papiers avec 1 400 000 citations. Les papiers sont classés en neuf domaines distincts et exclusifs tels que repris sur la figure 3.7(b). Ces domaines sont présents dans 92 venues.

**(a)**




**(b)**

| Domaine | #auteur | #papier | $\langle \frac{\#\text{auteur}}{\text{papier}} \rangle$ | #venue |
|---|---|---|---|---|
| Artificial Intelligence (A.I.) | 41538 | 65927 | 2.87 | 23 |
| Computer Graphic: multimedia (C.G.M.) | 25989 | 18877 | 2.92 | 13 |
| Computer Networks (C.N.) | 22374 | 30212 | 3.01 | 9 |
| Database: Data Mining (Database) | 5865 | 9294 | 3.77 | 7 |
| Human Computer Interaction (H.C.I.) | 4660 | 10666 | 3.25 | 5 |
| Information Security (I.S.) | 5298 | 6943 | 3.07 | 6 |
| Interdisciplinary Studies (I.St.) | 46111 | 2614 | 4.16 | 11 |
| Software Engineering (S.E.) | 8147 | 20506 | 3.15 | 8 |
| Theoretical Computer Science (T.C.S.) | 10824 | 21136 | 2.75 | 11 |

**Figure 3.7.:** (a) Schéma d'un réseau bibliographique et (b) Classification des papiers en neuf domaines. Le nombre d'auteurs, de papiers ainsi que de venues présents dans les données pour chaque domaine. Un seul domaine est attribué à chaque papier.

### Objectif

L'objectif est d'expliquer et prédire le poids des liens[2] de type "co-publication" $\mathcal{R}_0$ : A→P←A au moyen d'autres informations présentes dans le HIN. Par abus de notation, nous notons la variable dépendante $\mathbb{P}_{\mathcal{R}_0}$, i.e. le vecteur colonne des probabilités de distributions obtenues par marche aléatoire contrainte par $\mathcal{R}_0$. Pour ce faire, nous sélectionnons un sous-ensemble de méta-chemins parmi l'infinité possible (table 3.2). Ce choix est basé sur leur sémantique et les motivations suivantes[3].

[1] Nous gardons le terme anglais *venue* désignant tantôt une conférence, tantôt un journal.

[2] Il s'agit en fait de chemins satisfaisant $\mathcal{R}_0$ : A→P←A. Dès lors, lorsque nous parlons de co-publication, il s'agit de la projection du graphe APA sur A. Dans cette projection, deux auteurs sont connectés (lien non dirigé) s'ils ont signé un papier ensemble. Le poids de ce lien égale le nombre de papiers co-écrits.

[3] Étant donné qu'il n'existe qu'un seul type de liens entre chaque couple de types de nœuds, le type de liens n'est pas écrit explicitement dans la notation des méta-chemins. Seuls les types de nœuds et l'orientation des liens sont

- A→P→A←P→A signifie que deux auteurs ont chacun co-écrit un papier avec un troisième auteur. Cela correspond à un triangle dans le graphe résultant de la projection de A→P sur A et représente un côté plus "social" ;
- A→P→P←A et A→P←P←A représentent l'intérêt qu'une personne $a$ porte au travail d'une autre $b$. Cela peut être sensé de penser que si $a$ est intéressé par le travail de $b$ et cite son travail, $a$ peut être désireux de collaborer avec $b$ et éventuellement publier un papier avec lui/elle. Il en va de même lorsque les rôles de $a$ et $b$ sont interchangés ;
- A→P→P←P←A signifie que deux auteurs citent le même papier et sont donc potentiellement inspirés par les mêmes idées. Cela peut être une bonne raison d'être coauteurs ;
- A→P←P→P←A est légèrement différent du méta-chemin précédent puisqu'ici, c'est un troisième papier qui cite les travaux de $a$ et $b$. Cela ne signifie cependant pas que $a$ et $b$ travaillent sur les mêmes choses. En conséquence, nous nous attendons à ce que ce méta-chemin soit moins pertinent que le précédent, bien que la structure soit fort proche ;
- A→P→V←P←A et A→P→T←P←A signifient que le papier de $a$ et le papier de $b$ sont publiés dans la même venue ou qu'ils appartiennent au même domaine. Bien qu'une venue puisse rassembler beaucoup de monde, le fait d'être accepté dans une même venue peut être source de collaboration. Il en va de même si deux personnes travaillent sur le même sujet ou dans le même domaine.

| Méta-chemin | Interprétation | Méta-chemin non orienté |
|---|---|---|
| $\mathcal{R}_0$: **A→P←A** | **sont co-auteurs, publient un papier ensemble** | |
| $\mathcal{R}_1$: A→P←A→P←A | partagent un coauteur | $v_A$ |
| $\mathcal{R}_2$: A→P→P←A | cite le papier d'un autre | $v_{PP}$ |
| $\mathcal{R}_3$: A→P←P←A | est cité par le papier d'un autre | |
| $\mathcal{R}_4$: A→P→P←P←A | co-citent le même papier | $v_{PPP}$ |
| $\mathcal{R}_5$: A→P←P→P←A | sont co-cités par le même papier | |
| $\mathcal{R}_6$: A→P→V←P←A | ont un papier dans la même venue | $v_V$ |
| $\mathcal{R}_7$: A→P→T←P→A | ont un papier dans un même domaine | $v_T$ |

**Table 3.2.:** Méta-chemins décrivant quelques notions de proximité entre auteurs. Les méta-chemins non orientés regroupent les méta-chemins qui sont identiques si l'orientation des liens est négligée, i.e. si seuls les types de nœuds des méta-chemins sont pris en compte.

Précisons que pour les méta-chemins dont la structure est A→P→Type nœuds←P→A, les nœuds visités lors du deuxième et quatrième pas de la marche aléatoire doivent être différents. En outre, les méta-chemins sélectionnés, ou du moins une partie, se trouvent être souvent utilisés dans la littérature pour mesurer la similarité entre deux auteurs, e.g. [156–158].

présents.

### 3.3.2. Orientation des liens au sein des méta-chemins

Pour quantifier l'importance de l'orientation[1] des liens dans un méta-chemin, nous comparons les résultats de la régression linéaire obtenus avec *i)* les VE associées aux méta-chemins (modèle 1) ; *ii)* les VE associées aux méta-chemins non orientés, i.e. les méta-chemins qui sont identiques si seuls les types de nœuds des méta-chemins sont pris en compte (modèle 2) ; présentés dans la table 3.2. Les résultats sont présentés dans la table 3.3.

Dans le modèle 2, les méta-chemins (non-orientés) faisant intervenir trois papiers $v_{PPP}$ et les venues $v_V$ ne sont pas sélectionnés par l'algorithme. Pour $v_{PPP}$, cela peut venir du fait que dans le modèle 1, seul $\mathcal{R}_4$ est inclus. En outre, les cœfficients associés sont très faibles, ce qui n'est pas le cas pour $v_{PP}$. En effet, $\mathcal{R}_2$: A→P→P←A intervient avec plus d'importance, expliquant peut-être le fait que $v_{PP}$ soit inclus dans le modèle 2.

Le modèle 2 ne peut expliquer que 59,97% de la variance de $\mathbb{P}_{\mathcal{R}_0}$ alors que le modèle 1 permet d'en expliquer 66,61%. Cette différence n'est pas négligeable et semble signifier que chaque méta-chemin apporte sa propre information. Même si certains semblent proches les uns des autres, vouloir les agréger n'est pas bénéfique pour notre objectif (voir différence dans la sémantique des méta-chemins composant $v_{PPP}$).

En conclusion, il semblerait que le modèle 1 soit préférable au modèle 2 : il est capable d'expliquer plus de variance, la sémantique des VE est très facile à comprendre et ne complexifie que très légèrement le modèle final (légèrement plus de VE que le modèle 2). Il est donc important de prendre en compte l'orientation des liens.

| Méta-chemin | $\hat{\beta}$ | $p$-valeur | Méta-chemin non-orienté | $\hat{\beta}$ | $p$-valeur |
|---|---|---|---|---|---|
| $\mathcal{R}_1$ : A→P←A→P←A | 1.2507 | 0.0038 | $v_A$ | 1.2133 | 0.0028 |
| $\mathcal{R}_2$ : A→P→P←A | 0.9237 | 0.0099 | $v_{PP}$ | 1.8549 | 0.0034 |
| $\mathcal{R}_3$ : A→P←P←A | - | - | | | |
| $\mathcal{R}_4$ : A→P→P←P←A | 0.2813 | 0.0395 | $v_{PPP}$ | - | - |
| $\mathcal{R}_5$ : A→P←P→P←A | - | - | | | |
| $\mathcal{R}_6$ : A→P→V←P←A | 0.1539 | 0.0099 | $v_V$ | - | - |
| $\mathcal{R}_7$ : A→P→T←P←A | - | - | $v_T$ | - | - |
| $R_a^2$ | 0.6661 | | $R_a^2$ | 0.5997 | |

**Table 3.3.:** Comparaison des résultats finaux des modèles linéaires lorsque les méta-chemins et les méta-chemins non orientés sont utilisés pour construire les VE. Les cœfficients $\hat{\beta}$ indiquent la contribution de chaque VE.

[1] Dans le cas présent, l'orientation des liens n'a réellement d'importance qu'entre deux nœuds de même type, i.e. entre deux papiers.

### 3.3.3. Partitionnement thématique des méta-chemins

Nous avons pu voir que le méta-chemin relatif au fait de publier à propos d'un même domaine ($\mathcal{R}_7$: A→P→T←P←A) n'était pas sélectionné par l'algorithme (table 3.3). Le faible nombre de domaines considérés dans ces données, comparé au nombre de papiers, pourrait en partie expliquer cela. En effet, puisqu'un seul domaine est attribué à un papier, le méta-chemin P→T→P génère une "matrice papier-papier" dense[1]. Lors du calcul des marches aléatoires contraintes par $\mathcal{R}_7$, le marcheur peut donc se déplacer sur un trop grand nombre de nœuds lors du passage T→P et au final atterrir sur un nœud $a \in A$ trop lointain par rapport à celui de départ. Ainsi, nous pensons que le méta-chemin $\mathcal{R}_7$ apporte une information trop diffuse pour expliquer A→P←A.

Cependant, l'idée de considérer les domaines n'est pas dénuée de sens puisqu'un auteur intéressé par un domaine l'est souvent pendant un certain temps et a donc le temps de collaborer avec d'autres personnes, qui sont elles-mêmes intéressées par le domaine. Dès lors, nous scindons les données en neufs sous-ensembles, chacun relatif à un domaine. Nous tentons de trouver un modèle de régression linéaire pour expliquer la relation A→P←A à l'intérieur de chaque domaine, i.e. un modèle par domaine. Les VE sont construites à partir des six méta-chemins présentés plus haut ($\mathcal{R}_1 - \mathcal{R}_6$).

En moyenne, nous obtenons un meilleur modèle descriptif que lorsque tous les domaines sont mélangés : $\langle R_a^2 \rangle = 0.697$ (et l'écart-type $\sigma = 0.071$), figure 3.8(a). Cela peut signifier qu'à l'intérieur de chaque domaine, il existe certaines habitudes de collaborations et publications, ce qui fait que nous sommes plus en mesure de les expliquer en isolant les sous-graphes correspondants et ce, malgré les distributions de degrés des graphes de co-publications (i.e. les graphe projeté de APA) assez semblables (figure 3.8(b)). Cependant, pour les domaines Artificial Intelligence (A.I.) et Theoretical Computer Science (T.C.S.), il est plus difficile de trouver un modèle qui explique correctement les données. À titre de comparaison, nous reportons aussi les scores $R_a^2$ obtenus avec PC et AvgSim : PC est toujours surpassé par les deux autres, et ces derniers rivalisent selon le domaine.

Ainsi que mentionné, un intérêt de l'approche proposée est son interprétabilité. Nous nous intéressons donc aux significations des VE du modèle final. Afin de ne pas alourdir inutilement le rapport, seuls les cœfficients pour PCRW sont repris (figure 3.9). Deux VE semblent particulièrement utiles pour notre objectif. En particulier, le méta-chemin $\mathcal{R}_4$: A→P→A←P←A est sélectionné dans chaque domaine : signer un papier avec un même coauteur est le plus utile pour expliquer la relation de coauteurs (triangle sociaux). Le méta-chemin $\mathcal{R}_6$: A→P→V←P←A est inclus dans

[1] Le même commentaire pourrait être fait pour le méta-chemin P→V→P puisque le nombre de venues est également limité, bien que dans une moindre mesure puisqu'un domaine englobe plusieurs venues. Le nombre d'entrées non nulles de la matrice APTPA (pas vraiment le même que celui de PTP mais le résultat final est englobé dans APTPA) est égal à 6 515 232 alors que pour APVPA, ce nombre passe à 3 940 634, ce qui est tout de même 1,6 fois inférieur.

**(a)**

| Domaine | $R^2_{a,PCRW}$ | $R^2_{a,PC}$ | $R^2_{a,AvgSim}$ |
|---|---|---|---|
| A.I. | **0.5914** | 0.5010 | 0.5372 |
| C.G.M. | **0.6358** | 0.5313 | 0.5226 |
| C.N. | **0.6321** | 0.5152 | 0.5730 |
| Database | 0.7349 | 0.6954 | **0.7570** |
| H.C.I. | 0.7723 | 0.7242 | **0.7830** |
| I.S. | 0.7211 | 0.6707 | **0.7919** |
| I.St. | **0.7838** | 0.7037 | 0.6759 |
| S.E. | **0.7222** | 0.6465 | 0.6509 |
| T.C.S. | **0.5796** | 0.4409 | 0.4266 |

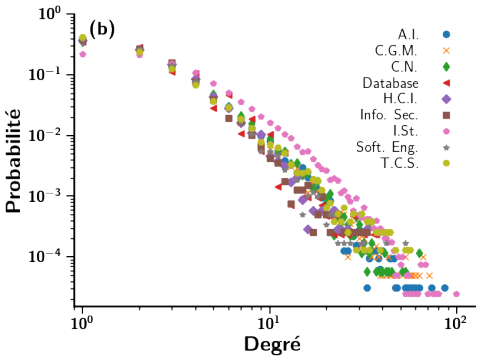



**Figure 3.8.:** (a) Résultats finaux des régressions linéaires appliquées indépendamment sur chaque domaine. Trois mesures de similarité sont utilisées afin de calculer les VE : PCRW, PC et AvgSim. (b) Distributions des degrés du graphe des co-publications présentant une allure assez semblable d'un domaine à l'autre.

7 domaines sur 9. L'information apportée par les venues est donc importante excepté pour les domaines Computer Networks (C.N.) et T.C.S..

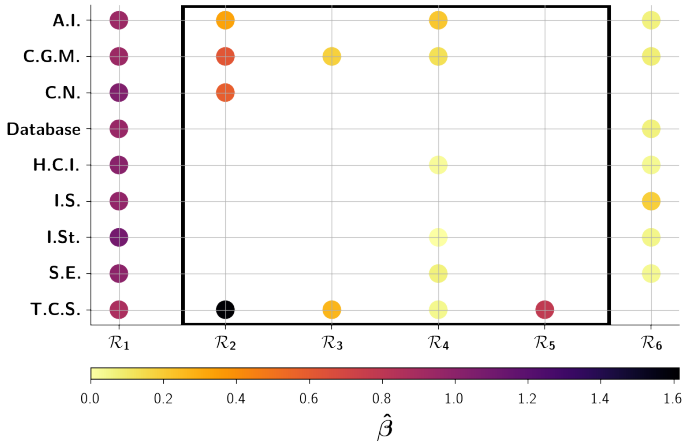



**Figure 3.9:** Cœfficients $\hat{\beta}$ des VE du modèle final obtenu pour chaque domaine, calculés avec la mesure de similarité PCRW. Les VE sont référencées par les méta-chemins présentés dans la table 3.2 associés.

### 3.3.4. Décomposition des méta-chemins en sous-graphes

Bien qu'étant utiles et déjà informatifs, les méta-chemins ne permettent pas de rendre compte de la réelle structure d'un HIN. Afin de préciser ce propos, nous nous intéressons aux cœfficients $\hat{\beta}$ présentés sur la figure 3.9. Les VE encadrées par le rectangle noir sont moins présentes que les deux autres. Cela peut s'expliquer, en partie, par le fait que de nombreuses structures de graphes se cachent derrière un méta-chemin donné. Un point fort intéressant se trouve donc dans les *auto-citations*.

Par exemple, regardons le méta-chemin $\mathcal{R}_2$: A→P→P←A. Le sous-graphe (SG) le plus simple associé est $SG_1$ : quatre nœuds et trois liens (figure 3.10(a)). Cependant il existe d'autres SG dans lesquels il est possible de trouver une chemin satisfaisant $\mathcal{R}_2$ : $SG_2$, $SG_3$ et $SG_4$. Il est évident

que $SG_1$ est plus simple que les trois autres. Par exemple, $SG_1$ est compris dans $SG_2$. Or, un sous-graphe de $SG_2$ est donné par les arcs en bleu (deux auteurs et un papier). Dans ce sous-graphe $SG_2$, les deux auteurs sont donc déjà co-auteurs puisque l'un d'eux se cite lui-même. Il en va de même pour $SG_3$ et $SG_4$ alors que pour $SG_1$, un troisième papier est nécessaire pour être co-auteurs.

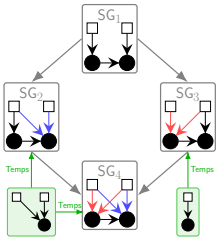
Un problème avec les méta-chemins est que nous ignorons face à quel SG nous nous trouvons. Si seuls les cœfficients des VE importent, cela ne pose pas réellement problème. En revanche, si nous désirons comprendre plus en détails les mécanismes de co-publications, il faudrait être capable distinguer de ces quatre SG.

Par conséquent, un examen des SG voire des motifs serait très intéressant. Les motifs sont les – généralement petits – SG récurrents (i.e. qui semblent être statistiquement significatifs) d'un graphe et sont considérés comme des signatures structurelles du graphe en question [113]. Typiquement, un profil de motifs est construit à partir des fréquences statistiquement validées ($z$-score ou $p$-valeur) des différents motifs observés (figure 3.10(b)). Ce profil peut ensuite être utilisé pour e.g. différencier plusieurs graphes. Cependant, le nombre de motifs possibles croît de manière exponentielle avec le nombre de nœuds considérés. Cela les rend malheureusement assez difficiles à utiliser comme outil descriptif pour les grands graphes. Néanmoins, il est nécessaire d'aller au-delà des méta-chemins et de se tourner vers d'autres structures plus complexes afin de mieux cerner un HIN [135, 150].

À titre d'exemple, nous avons calculé les $z$-scores pour les quatre SG considérés. Pour ce faire, 200 graphes randomisés, à partir des graphes empiriques, sont générés : le degré des nœuds est préservé, mais les liens sont redistribués aléatoirement. Les $z$-scores pour $SG_1$ sont négatifs ($< -3.5$) pour tous les domaines suggérant qu'il est sous-représenté (à l'exception de I.St.). Les $z$-scores des trois autres SG considérés sont tous supérieurs à 10 quel que soit le domaine sauf I.St. Pour ce domaine, ces SG ne semblent pas être des motifs: $|z\text{-score}| < 1$.

**Remarque 3.2** (Temporalité des interactions)**.** Un examen des SG met aussi en avant l'importance du *temps*, i.e. la chronologie des publications. Par exemple, une différence entre $SG_2$ et $SG_3$ est liée au temps. Dans $SG_2$, les deux auteurs $a_1$ et $a_2$ ont d'abord écrit le papier $p_1$ et puis seulement $a_2$ a publié $p_1$, citant $p_2$. La chronologie est différente pour $SG_3$ : $a_2$ a d'abord écrit $p_2$ et ensuite, il a co-écrit $p_1$ avec $a_1$, en citant $p_2$ (panels verts de la figure 3.10(a)). Nous remarquons aussi que $SG_2$ et $SG_4$ peuvent potentiellement résulter d'un même SG. Cependant, les mécanismes sous-jacents sont différents : $SG_4$ suggère qu'un même duo progresse sur un même sujet et publie tandis que $SG_2$ sous-entend que seul $a_1$ poursuit le travail. Considérer le temps complexifie grandement les choses, mais est une piste fort intéressante pour mieux comprendre les mécanismes de collaboration ou tout autre système évoluant dans le temps [80, 102, 122].

**(a)**




**(b)**

| | $SG_1$ | $SG_2$ | $SG_3$ | $SG_4$ |
|---|---|---|---|---|
| A.I. | 1,242 | 117 | 120 | 41 |
| C.G.M. | 10,632 | 731 | 863 | 301 |
| C.N. | 28,732 | 1,050 | 1,011 | 553 |
| Database | 4,333 | 162 | 198 | 90 |
| H.C.I. | 2,582 | 181 | 176 | 66 |
| I.S. | 900 | 57 | 81 | 31 |
| I.Std. | 157 | 1 | 2 | 0 |
| S.E. | 2,097 | 163 | 174 | 61 |
| T.C.S. | 839 | 65 | 59 | 24 |

**Figure 3.10.:** Examen du méta-chemin $\mathcal{R}_2$ : A→P→P←A. (a) Relations entre quatre sous-graphes (SG) pour lesquels il existe un chemin satisfaisant $\mathcal{R}_2$. Les carrés blancs (resp. ronds noirs) représentent les auteurs (resp. papiers). Dans les panels verts, des SG sont explicitement repris afin de mettre en avant la notion du temps, i.e. la chronologie des publications (non exhaustif). (b) Décompte des sous-graphes présents dans les HIN associés à chaque domaine. Les $z$-scores associés (voir texte principal) suggèrent que $SG_1$ est sous-représenté tandis que les autres SG sont sur-représentés, à l'exception du domaine I.St. pour lequel ces SG ne semblent pas être des motifs.

### 3.3.5. Contrôler la qualité des solutions

À l'aide d'outils statistiques, nous regardons plus en détails la qualité des résultats des régressions linéaires obtenus.

#### Explication et intervalles de confiance

En se penchant à nouveau sur les cœfficients des VE, nous remarquons que certains sont particulièrement faibles, pouvant remettre en doute la réelle présence des VE associées dans le modèle de régression. En outre, il est souvent utile d'avoir un intervalle de confiance autour des estimations. Il existe différentes façons de le faire, mais une méthode assez générale consiste à utiliser un bootstrap. Un *bootstrap* est un échantillon aléatoire avec remplacements des données, de même taille que les données d'origine. Il s'agit d'un moyen de générer plusieurs vues des mêmes données. Dans le cas d'une régression linéaire classique, cette méthode est appelée *smooth bootstrap*. En pratique, pour construire un échantillon, nous tirons aléatoirement $n$ données, avec remplacement, dans l'ensemble initial (de taille $n$) des données. Ensuite, pour appliquer le bootstrap : *i)* nous choisissons un nombre d'échantillons (ici, fixé à 1000) ; *ii)* pour chaque échantillon généré, nous cherchons un modèle de régression linéaire et calculons les cœfficients $\hat{\beta}$ et finalement *iii)* nous calculons la moyenne de tous les cœfficients obtenus.

Les résultats pour le domaine "Interdisciplinary Studies (I.St.)" sont présentés sur la figure 3.11. La VE $\mathcal{R}_6$ : A→P→V←P←A possède une variance extrêmement faible, augmentant notre confiance dans notre estimation du cœfficient. Les deux autres VE $\mathcal{R}_1$ : A→P←A→P←A et $\mathcal{R}_4$ : A→P→P←P←A ont une variance un peu plus importante. $\mathcal{R}_4$ est intéressant car les valeurs de cœfficients sont tantôt négatives, tantôt positives (légèrement plus fréquemment positives, figure 3.11(b)), ce qui peut potentiellement indiquer qu'il n'y a pas réellement de relation entre

$\mathcal{R}_4$ et la variable à expliquer, bien que la $p$-valeur associée soit inférieure à 0.05. Cela pointe la difficulté à trouver un véritable bon modèle permettant d'expliquer une variable. De façon générale, cette difficulté vient souvent du degré de qualité des données (en l'occurrence, des VE), mais aussi, comme déjà mentionné, du dilemme "biais-variance". Pour $\mathcal{R}_1$, sa variance ne remet pas en question sa présence dans la modèle de régression.

Les résultats des autres domaines (non montrés) sont analogues : pour les cœfficients extrêmement faibles ($\hat{\beta} < 0.05$), la variance tend à dire qu'il n'y a pas de réelles corrélations tandis que pour les autres valeurs ($\hat{\beta} > 0.05$), le bootstrap ne remet pas en question la présence des VE associées, ni leur cœfficient.

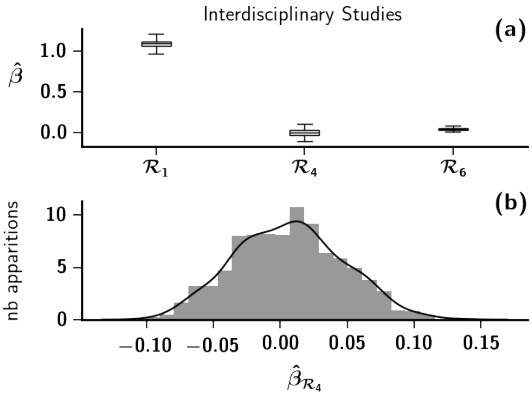



**Figure 3.11:** Analyse des cœfficients du modèle final obtenu pour le domaine "Interdisciplinary Studies (I.St)". (a) Intervalles des cœfficients $\hat{\beta}$ calculés par 1000 bootstraps des données. Le modèle linéaire utilisé est le modèle final à 3 VE. (b) Histogramme des cœfficients $\hat{\beta}_{\mathcal{R}_4}$.

### Prédiction et sur-apprentissage

Dans cette sous-section, nous nous penchons sur une tâche prédictive. Les résultats moyennés des validations croisées Monte Carlo sont reportés dans la table 3.4. Toutes les $p$-valeurs associées aux VE sont inférieures au seuil fixé $\alpha = 0.05$. Étant donné que PC récolte toujours de moins bons résultats que les deux autres mesures (figure 3.8(a)), nous ne calculons que les scores de prédiction pour PCRW et AvgSim.

Pour les domaines Database, Human Computer Interaction (H.C.I.) et Interdisciplinary Studies (I.St.), la prédiction est relativement bonne dans le sens où le $\langle R_a^2 \rangle$ de l'ensemble de validation/test est presqu'aussi élevé que celui de l'ensemble d'entraînement (en moyenne). Pour les autres domaines, la perte de qualité est plus importante, même pour Information Security (I.S.) et Software Engineering (S.E.) qui possèdent un bon $R_a^2$ d'entraînement. Notons qu'à présent, $\langle R^2_{a,test,PCRW} \rangle$ pour Database est supérieur à $\langle R^2_{a,test,AvgSim} \rangle$ : cela peut venir du fait que AvgSim sur-ajuste les données, alors que l'inverse se passe pour I.St. Enfin, pour Theoretical Computer Sciences (T.C.S.), le $R_a^2$ pour la prédiction est fort faible, remettant en question sa réelle pertinence (bien que la $p$-valeur soit

inférieure à 0.05).

| Topics | $\langle R^2_{a,test,PCRW} \rangle$ | $\langle R^2_{a,test,AvgSim} \rangle$ |
|---|---|---|
| All topics | 0.5508 | **0.5742** |
| A.I. | **0.4994** | 0.4290 |
| Comp. Graph. Mult. | **0.5133** | 0.4743 |
| Comp. Net. | **0.5322** | 0.4911 |
| Database | **0.7258** | 0.7057 |
| Hum. Comp. Inter. | 0.7338 | **0.7584** |
| Info. Sec. | 0.6509 | **0.7367** |
| Interdisc. Std. | 0.7440 | **0.7688** |
| Software Eng. | **0.6450** | 0.6130 |
| T.C.S. | **0.3557** | 0.3487 |

**Table 3.4:** Résultats de la récupération de poids pour la cas général (tous domaines confondus) et par domaine. PCRW et AvgSim sont utilisés comme mesures de similarité dans le calcul des VE.

Afin de se convaincre de la pertinence de ces résultats calculés sur les données DBLP, ces derniers sont confrontés à ceux obtenus sur un graphe randomisé : ce graphe préserve certaines propriétés de la la topologie du HIN DBLP – la distribution de degrés – mais réorganise aléatoirement les liens entre les nœuds. L'objectif est de montrer que les distributions de degrés seules ne suffisent pas à engendrer une telle corrélation dans les données et que cette corrélation découle d'autres propriétés topologiques plus complexes. En effet, les résultats obtenus sur de tels graphes randomisés ne sont pas significatifs : aucune VE avec une $p$-valeur inférieure à 0.12 et le cœfficient de corrélation moyenné sur 15 réalisations de graphes aléatoires ne dépasse pas 0.26. Dès lors, bien que les résultats ne soient pas mirobolants, la méthode proposée permet d'une certaine façon, de retrouver les poids des liens manquants.

## 3.4. Analyse empirique d'un exemple complexe

Dans cette section, nous essayons de proposer des outils supplémentaires pour la recherche empirique avec un cas d'étude plus complexe. Nous illustrons tout d'abord l'approche proposée sur un jeu de données plus conséquent. Nous tentons ensuite de corroborer et interpréter les résultats obtenus à l'aide d'autres mesures.

### 3.4.1. Cas d'étude : échange d'information entre individus et médias sur Twitter

#### Présentation des données

Le jeu de données concerne un ensemble de tweets collectés au moyen de l'outil DMI-TCAT [22] avec la requête "migrant OR migrants OR immigrant OR immigrants OR emigrant OR emigrants"

de janvier 2015 à juillet 2016[1]. Ce jeu de données est appelé Migrant par la suite. Seuls les tweets écrits en anglais ou français sont pris en compte (proportion 90% Anglais et 10% Français), les autres langues étant trop peu représentées.

Les sujets abordés dans ces tweets sont beaucoup plus vastes que dans les données FIFA du fait du sujet plus général ainsi que de la période plus étendue (19 mois). Au niveau du contenu et bien que les mots-clefs ne soient pas explicitement liés à une quelconque crise, ces données font principalement référence à la *crise*[2] migratoire européenne et aux évènements ayant lieu en Amérique du Nord et, dans une moindre mesure, en Amérique Centrale et en Asie[3].

Afin d'avoir une petite idée des thèmes abordés dans les tweets, nous construisons le graphe de co-occurrences des hashtags agrégé sur les 19 mois. Deux hashtags sont donc reliés s'ils sont présents dans le même tweet. La méthode de Louvain de détection de communautés est ensuite appliquée [18] afin de comprendre un peu mieux la structure de ce graphe. Cinq communautés ou groupes sémantiques se démarquent. Ces groupes possèdent respectivement 1294, 2017, 1499, 204 et 888 hashtags. Grâce aux nuages de mots représentés aux figures 3.12(a)-3.12(e), nous remarquons que chacun de ces cinq groupes sémantiques comprend au moins un hub : la taille de la police est proportionnelle à la fréquence d'apparition des hashtags. Chacun de ces hubs assume un rôle de "connecteur" entre les différents groupes.

Nous pouvons apprécier sur la figure 3.12(f) qu'au début de la période, le groupe 2, principalement relatif à ce qu'il se passe aux États-Unis d'Amérique, est beaucoup plus important. La plupart des tweets ont au moins un hashtag dans ce groupe sémantique. Au cours du temps, son importance s'amenuise et le groupe 3 prend de plus en plus d'importance. Cela n'est pas étonnant car il fait référence aux évènements politiques marquants en Europe, comme une décision importante d'Angela Merkel : le 31 août 2015, l'Allemagne décide d'ouvrir ses frontières à des centaines de milliers de demandeurs d'asile arrivés par la route des Balkans ; ou le Brexit (vote en juin 2016).

Le diagramme d'accord de la figure 3.12(g) représente les liens entre les groupes sémantiques (les liens intra-communautaires ne sont pas représentés). Le groupe 5 est le moins dense et possède beaucoup de liens avec le groupe 3. Cela n'est pas trop étonnant puisque le groupe 5 est fort lié aux évènements français (#France, #Calais, #Sarkozy, ...) tandis que le groupe 3 est lié principalement aux évènements politiques européens. Un fait intéressant est que le groupe le plus important

[1] Ces données ont été collectées par les partenaires de Université d'Amsterdam du projet ODYCCEUS pour l'étude de cas : "contested european boundaries: refuggees" qui vise à fournir une analyse qualitative et quantitative approfondie des conflits actuels sur les frontières de l'Europe dans le contexte de la crise des réfugiés.

[2] La notion de *crise* est relativement subjective. Cette dernière est généralement employée lorsque le flux migratoire devient très ou trop important et que le(s) pays d'accueil y voi(en)t un problème, réel ou imaginaire.

[3] Le sujet des crises migratoires, très vaste et complexe, dépasse de loin le cadre de ce travail. Pour plus d'informations, voir par exemple [6, 115] et les références cités dans ces articles. En ce qui concerne l'Europe, des sites tels que celui des publications de l'UE sur la crise migratoire (`https://op.europa.eu/fr/web/general-publications/refugee`) ou de la Revue Européenne des Migrations Internationales (`https://journals.openedition.org/remi`) proposent des archives et articles détaillés.

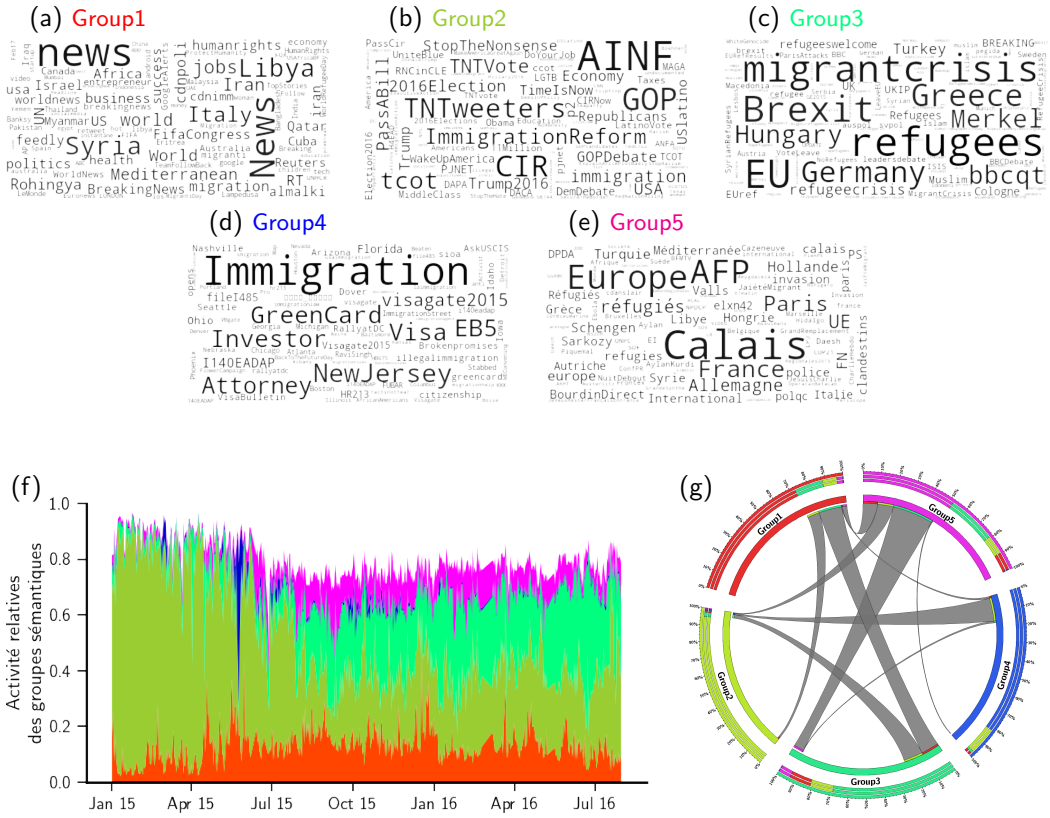



**Figure 3.12.:** (a)-(e) Nuages de mots des cinq groupes sémantiques qui se démarquent, détectés avec l'algorithme de Louvain. (f) Activité relative au temps $t$ : nombre de tweets ayant un hashtag dans le groupe sémantique au temps $t$ divisé par le nombre total de tweets au temps $t$. (g) Diagramme d'accord qui reprend les liens entres les groupes sémantiques. Le code couleur est identique pour chaque panel.

au niveau du nombre de hashtags (groupe 2) n'est pas spécialement celui qui est le plus actif à chaque instant (en juin 2016, le groupe 3 est plus actif grâce au Brexit et ses relations avec l'Union Européenne).

### Construction du HIN

Dans ces données, l'identification de comptes appartenant à des médias est possible et repose sur l'hypothèse que les médias publient nécessairement beaucoup d'urls dont le domaine est le leur et ce, de façon beaucoup plus répétitive que n'importe quel autre compte Twitter. Pus particulièrement la procédure[1] effectuée est la suivante :

- Sélectionner les domaines des urls publiés dans les tweets ;
- Pour chaque compte Twitter : compter le nombre de tweets référençant ces domaines ainsi que les retweets de ces tweets. À partir de ce décompte, sélectionner les domaines les plus

[1] Cette procédure n'a pas été effectuée par moi-même.

référencés et calculer un score de vraisemblance (i.e. probabilité du nombre de références du domaine le plus référencé par rapport à un modèle binomial) ;

- Pour chaque domaine, sélectionner le compte Twitter qui possède le plus haut score de vraisemblance. Nous obtenons finalement le nom de ces comptes, que nous appelons médias dans la suite, ainsi que le domaine associé.

À partir de ces entités et actions, nous construisons un HIN. Chaque nœud représente soit un utilisateur, soit un média, soit un hashtag tandis que les liens représentent les différentes façons qu'ont ces nœuds d'interagir entre eux. Dès lors, le HIN possède trois types de nœuds $\mathcal{V} = \{\mathrm{U}, \mathrm{M}, \mathrm{H}\}$. Les types de liens sont les suivants : RT, RP, MT et post. Ainsi qu'illustré à la figure 3.13, nous avons par exemple RT qui relie tantôt U-U (un utilisateur retweet un autre utilisateur), tantôt U-M (un utilisateur retweet un média). Nous différencions donc ces deux types de RT, ce qui donne finalement huit[1] types de liens $\mathcal{V}$ ={URTU, URPU, UMTU, UpostH, URTM, URPM, UMTM, MpostH}. Un lien est créé de $u$ à $v$ si $u$ RT, RP ou MT $v$ et le poids de ce lien correspond au nombre de fois que cette action apparaît dans les données. Pour les relations post, un lien existe entre $u$ et $h$ si $h$ apparaît dans le tweet de $u$ et le poids du lien correspond au nombre de fois que $u$ écrit $h$. À nouveau, nous excluons les hashtags présents dans les posts retweetés. Tous les graphes sont orientés et pondérés. Étant donné qu'il n'y a qu'un seul type de liens entre U et H, nous notons simplement U→H.

Ce jeu de données comprend 1 800 000 utilisateurs, 1000 médias et 360 000 hashtags uniques. Au niveau des liens, il y a 16 000 000 de RT, 1 800 000 de RP et 17 000 000 de MT.

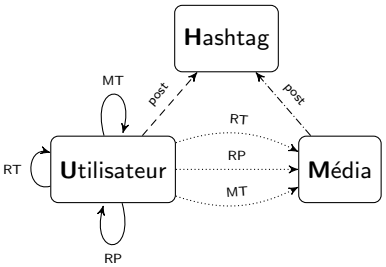



**Figure 3.13:** Schéma d'un HIN créé à partir de données Twitter. Trois types de nœuds: Utilisateurs U, Hashtags H et Média M. Huit types de liens : Retweet RT, Reply RP, Mention MT et Post entre U-U, U-H, U-M et M-H (i.e. les types de nœuds sources et cibles étant différents). Nous décidons de ne pas prendre en compte les liens pour lesquels un média est à l'origine.

### Objectif

Afin d'illustrer l'approche de la section 3.1, nous proposons d'expliquer le type de liens U→H au moyen des autres types de liens et méta-chemins. La qualité des données permettant de différencier les utilisateurs des médias, un autre objectif poursuivi est de voir si les utilisateurs et médias ont le même impact sur la relation U→H.

[1] Puisque nous désirons voir et quantifier l'impact qu'ont les médias sur les utilisateurs, nous décidons de ne pas prendre en compte les liens pour lesquels les médias sont à l'origine.

### 3.4.2. Régressions linéaires au cours du temps

En vue d'expliquer la relation U→H, huit modèles de régression linéaire sont testés. Chaque modèle possède un sous-ensemble de prédicteurs spécifiques, toujours construits à partir des marches aléatoires contraintes. Ces modèles se divisent en cinq catégories selon que la distinction entre les types de nœuds et liens est faite ou non et la longueur des méta-chemins. Les modèles sont numérotés en fonction de leur complexité (figure 3.14). Les deux premiers ne différencient pas les types de liens. Lorsque c'est le cas, le type de liens est écrit au-dessus de la flèche.

À titre d'exemple, concentrons-nous sur le modèle M1. Pour un utilisateur spécifique $u$, le méta-chemin $\mathcal{P}_1 : \mathrm{U} \to \mathrm{U} \to \mathrm{H}$ signifie "les hashtags écrits par les utilisateurs en relation (tous types de liens confondus) avec $u$". Ainsi, $\mathbb{P}_{\mathcal{P}_1}(h|u)$ peut être interprété comme la probabilité qu'a l'utilisateur $u$ de poster le hashtag $h$, étant donné qu'un autre utilisateur en relation avec lui a posté ce hashtag. La signification est la même pour le méta-chemin $\mathcal{P}_2 : \mathrm{U} \to \mathrm{M} \to \mathrm{H}$ et donc $\mathbb{P}_{\mathcal{P}_2}(h|u)$, sauf qu'il se réfère aux médias liés à l'utilisateur $u$ et non à un autre utilisateur.

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{P}_1$ :U → U → H | ✓ | | ✓ | | ✓ | | | |
| $\mathcal{P}_2$ :U → M → H | ✓ | ✓ | | | ✓ | | | |
| $\mathcal{P}_3$ :U $\xrightarrow{\mathrm{RT}}$ U → H | | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_4$ :U $\xrightarrow{\mathrm{RP}}$ U → H | | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_5$ :U $\xrightarrow{\mathrm{MT}}$ U → H | | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_6$ :U $\xrightarrow{\mathrm{RT}}$ M → H | | | ✓ | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_7$ :U $\xrightarrow{\mathrm{RP}}$ M → H | | | ✓ | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_8$ :U $\xrightarrow{\mathrm{MT}}$ M → H | | | ✓ | ✓ | | ✓ | ✓ | ✓ |
| $\mathcal{P}_9$ :U → U → U → H | | | | | ✓ | | | |
| $\mathcal{P}_{10}$ :U → U → M → H | | | | | ✓ | | | |
| $\mathcal{P}_{11}$ :U $\xrightarrow{\mathrm{RT}}$ U $\xrightarrow{\mathrm{RT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{12}$ :U $\xrightarrow{\mathrm{RT}}$ U $\xrightarrow{\mathrm{RP}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{13}$ :U $\xrightarrow{\mathrm{RT}}$ U $\xrightarrow{\mathrm{MT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{14}$ :U $\xrightarrow{\mathrm{RP}}$ U $\xrightarrow{\mathrm{RT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{15}$ :U $\xrightarrow{\mathrm{RP}}$ U $\xrightarrow{\mathrm{RP}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{16}$ :U $\xrightarrow{\mathrm{RP}}$ U $\xrightarrow{\mathrm{MT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{17}$ :U $\xrightarrow{\mathrm{MT}}$ U $\xrightarrow{\mathrm{RT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{18}$ :U $\xrightarrow{\mathrm{MT}}$ U $\xrightarrow{\mathrm{RP}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{19}$ :U $\xrightarrow{\mathrm{MT}}$ U $\xrightarrow{\mathrm{MT}}$ U → H | | | | | | ✓ | | ✓ |
| $\mathcal{P}_{20}$ :U $\xrightarrow{\mathrm{RT}}$ U $\xrightarrow{\mathrm{RT}}$ M → H | | | | | | | ✓ | ✓ |
| $\mathcal{P}_{21}$ :U $\xrightarrow{\mathrm{RT}}$ U $\xrightarrow{\mathrm{MT}}$ M → H | | | | | | | ✓ | ✓ |
| $\mathcal{P}_{22}$ :U $\xrightarrow{\mathrm{RP}}$ U $\xrightarrow{\mathrm{MT}}$ M → H | | | | | | | ✓ | ✓ |
| $\mathcal{P}_{23}$ :U $\xrightarrow{\mathrm{MT}}$ U $\xrightarrow{\mathrm{RT}}$ M → H | | | | | | | ✓ | ✓ |
| $\mathcal{P}_{24}$ :U $\xrightarrow{\mathrm{MT}}$ U $\xrightarrow{\mathrm{MT}}$ M → H | | | | | | | ✓ | ✓ |

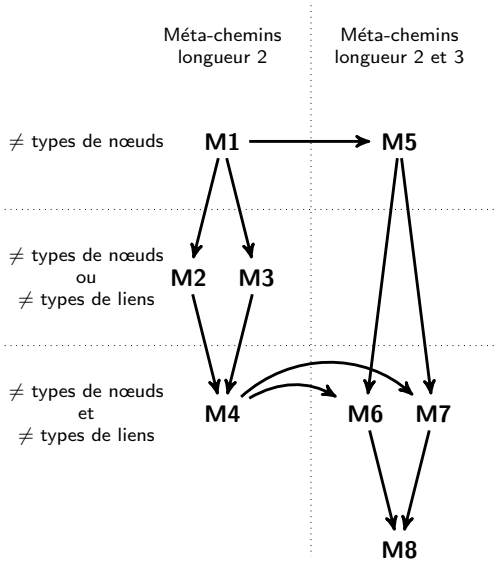



**Figure 3.14.:** Définition des modèles. Le symbole ✓ dans la table signifie que le méta-chemin est inclus dans le sous-ensemble des prédicteurs potentiels du modèle de régression linéaire. $\mathcal{P}_1$ et $\mathcal{P}_2$ ne différencient pas les types de liens. Lorsque qu'il y a effectivement distinction, le type de liens est écrit au-dessus de la flèche. L'organigramme illustre la structure hiérarchique des modèles testés. $M_a \to M_b$ signifie que le modèle $M_b$ englobe le modèle $M_a$, soit parce qu'il contient les méta-chemins de $M_a$ (e.g. $M_1 \to M_5$), soit parce qu'il désagrège un méta-chemin de $M_a$ (différencie les types de liens, e.g. $M_1 \to M_2$). Dit autrement, $M_a \to M_b$ signifie que $M_a$ est un cas particulier de $M_b$. La relation $\to$ est transitive.

L'algorithme glouton est appliqué avec chaque sous-ensemble de méta-chemins en entrée et ce, séparément sur chaque mois de l'ensemble de données Migrants (il y a donc 8 (pour les ensemble prédéfinis de VE) × 19 (nombre de mois) modèles finaux). Cette division temporelle est motivée

par les divers évènements ayant lieu pendant la période inspectée, caractérisés par des hashtags spécifiques (e.g. #ParisAttack, #Brexit, et figure 3.12). Par cette division, nous voulons nous concentrer sur la possibilité d'expliquer la relation entre les U→H sur un "sujet thématique" à l'intérieur du sujet "migrant" général.

Les résultats sont présentés sur la figure 3.15. Le meilleur $R_a^2$ (parmi les huit modèles linéaires finaux obtenus en utilisant l'algorithme glouton avec chaque ensemble de VE prédéfini) sur l'entièreté de la période (les 19 mois) considérée s'élève à seulement 0.46 (ligne pointillée verte sur la figure 3.15(a)), soit une valeur inférieure à la plupart des $R_a^2$ obtenus sur chaque mois. Il est donc avantageux, en moyenne, de diviser la période .

Pour chaque modèle, les cœfficients $R_a^2$ ont une tendance générale décroissante au cours du temps (figure 3.15(a)). Cela signifie qu'à partir de début 2015 jusqu'en 2016, il devient de plus en plus difficile d'expliquer la relation U→H au moyen des autres relations présentes dans le HIN. Il semble donc que les méta-chemins sélectionnés soient insuffisants pour expliquer la cette relation. En d'autres termes, si au début de la période étudiée, il suffit de savoir, pour un utilisateur donné, ce que ses voisins et les médias publient pour savoir ce qu'il publie réellement, il est clair que ce n'est plus le cas au fil du temps.

Concentrons-nous plus en détail sur les modèles M1 et M4 (en raison de leur facilité d'interprétation) et en particulier sur leurs cœfficients. Pour le premier, la VE associée à $\mathcal{P}_1$: U→U→H est sélectionné à chaque mois, soulignant son importance. À l'inverse, la VE associée à $\mathcal{P}_2$: U→M→H n'est pas sélectionnée pour les mois de mars 2015 et mai 2016. Notons au passage que les $R_a^2$ de M1 sont des minima locaux en ces deux mois. En fait, aucun de ces deux mois n'inclut le méta-chemin $\mathcal{P}_2$ car la $p$-valeur associée est supérieure à 0.05 (figure 3.15(b)).

Le modèle M4 est lié à M1 par ce qui suit : il s'agit de la décomposition intuitive des méta-chemins de M1. En effet, le méta-chemin $\mathcal{P}_1$ : U→U→H est divisé en trois en différenciant les types de relation entre les utilisateurs: $\mathcal{P}_3$ : U$\xrightarrow{\text{RT}}$U→H, $\mathcal{P}_4$ : U$\xrightarrow{\text{RP}}$U→H and $\mathcal{P}_5$ : U$\xrightarrow{\text{MT}}$U→H. La même décomposition se produit pour le méta-chemin agrégé relatif aux médias, i.e. $\mathcal{P}_2$ donne lieu à $\mathcal{P}_6$, $\mathcal{P}_7$ et $\mathcal{P}_8$. Ainsi, pour le modèle M4, les méta-chemins liés aux relations U-U sont toujours sélectionnés. Ce qui est vraiment nouveau par rapport à M1 concerne les méta-chemins liés aux relations U-M (figure 3.15(c)). En effet, nous observons que *i)* $\mathcal{P}_7$ n'est jamais sélectionné : les réponses aux médias ne sont jamais pertinentes pour prédire U→H ; *ii)* $\mathcal{P}_3$ est la VE la plus significative pour notre objectif. Le pic du mois de juin pourrait être lié à deux évènements majeurs, comme expliqué dans la section 3.4.4. Enfin, la performance de M4 est presque aussi bonne que celle de M7 et M8 (les deux seuls modèles plus complexes), tout en étant *significativement* meilleure que les cinq autres modèles (les cinq modèles les plus simples). Cela rappelle qu'un modèle doit être un compromis

entre performance, complexité et interprétabilité. Dans cette optique, une option serait M4 comme "modèle à conserver".

Dans ce qui suit, nous détaillons deux idées qui pourraient peut-être expliquer la tendance principale des résultats de la régression, à savoir la diminution générale des scores de $R_a^2$.

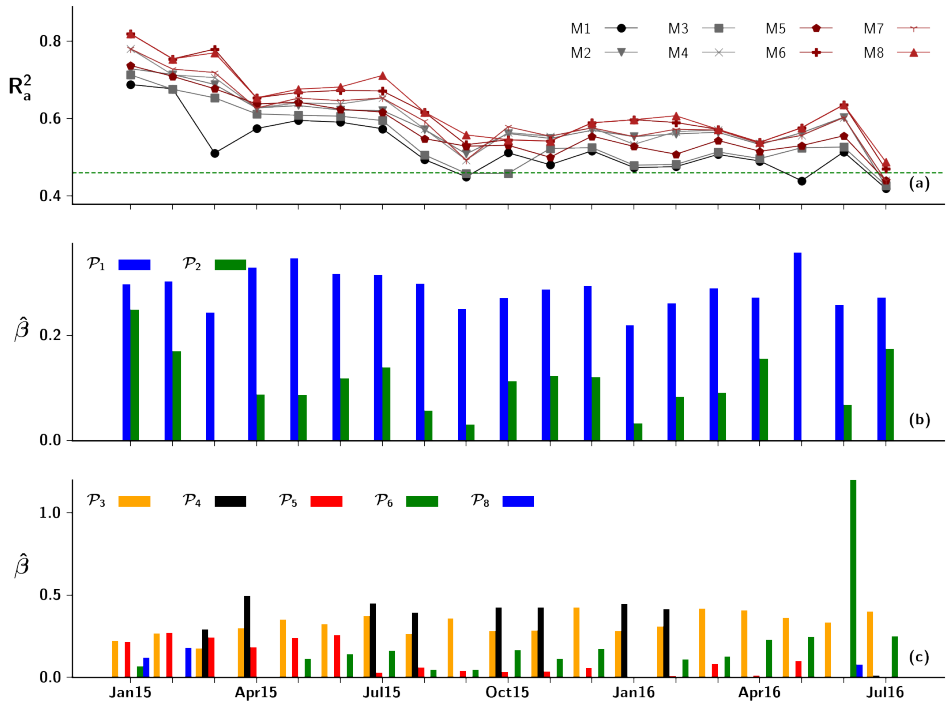



**Figure 3.15.:** Résultats des régressions linéaires. (a) Cœfficients de détermination ajustés $R_a^2$ pour chaque mois au cours du temps. En moyenne, au plus le modèle est complexe, au meilleur est ce $R_a^2$ (numérotés de 1 à 8, figure 3.14). La ligne verte pointillée indique le $R_a^2$ du meilleur modèle obtenu sur toute la période des 19 mois que couvre les données. (b-c) Cœfficients $\hat{\beta}$ des prédicteurs construits à partir de méta-chemins pour chaque mois pour le modèle M1 (b) et M4 (c).

### 3.4.3. Mesures additionnelles pour expliquer le résultat de la régression

Dans cette section, nous tentons d'expliquer, en partie, les résultats des régressions (figure 3.15) à l'aide d'autres mesures sur les données. Comme nous l'avons vu, les scores de $R_a^2$ diminuent au fil du temps. Il semble donc intéressant d'essayer de caractériser quelle quantité ou structure des données est responsable de ce changement.

Nous nous intéressons au nombre de liens entre les utilisateurs ainsi qu'au nombre de hashtags, les deux quantités les plus simples et évidentes. Cependant, ces nombres oscillent de mois en mois et il est difficile d'extraire des informations pertinentes de ces nombres bruts. Par conséquent, nous combinons ces informations en calculant le rapport entre le nombre de liens $n_a$ (tweets, retweets, ...) dont un utilisateur est à l'origine et le nombre de hashtags $n_h$ postés par cet utilisateur, i.e. $n_a/n_h$.

Nous prenons ensuite la moyenne de ce ratio sur tous les individus (toujours noté $n_a/n_h$). Afin que ce ratio serve de mesure de la *diversité* des liens d'un utilisateur (liens avec différents utilisateurs), seuls les liens uniques sont prises en compte : si $u_1$ retweete $n$ fois $u_2$, un seul lien est retenu. Les ratios sont présentés sur la figure 3.17(a). Bien qu'il y ait encore des oscillations, nous pouvons apprécier une tendance croissante de ce ratio au fil du temps. Cela signifie que le nombre de liens entre utilisateurs augmente plus que le nombre de hashtags postés.

Une question se pose alors tout naturellement : étant donné qu'il y a de plus en plus de liens entre les utilisateurs, ces derniers publient-ils les mêmes hashtags ? Si c'est le cas, la régression devrait pouvoir fournir de bons résultats, ou au moins aussi bons, au fil du temps. Une possible façon de répondre à cette question est de s'intéresser à la structure des matrices associées aux marches aléatoires contraintes par les méta-chemins. Plus précisément, les distributions de probabilités résultantes forment une matrice $\mathbf{T}_{\mathcal{P}}$ telle que $\mathbf{T}_{\mathcal{P}}(u, v) = \mathbb{P}_{\mathcal{P}}(v|u)$. Cette matrice est proche de ce que Zhou *et al.* appellent *commuting matrix* [188]. Si les utilisateurs liés écrivent les mêmes hashtags, cela doit aboutir à des groupes d'utilisateurs plus *denses*, identifiables grâce à la matrice $\mathbf{T}_{\text{UH}}$.

L'idée est donc de projeter la matrice $\mathbf{T}_{\text{UH}}$ sur U en calculant la similarité cosinus entre chaque paire d'utilisateurs ; c'est la matrice $proj(\mathbf{T}_{\text{UH}})$. Cette projection relie les utilisateurs entre eux et le poids de leur lien mutuel représente la similarité cosinus entre leur vecteur hashtags. Plus ils publient les mêmes hashtags, plus cette similarité est élevée (proche de 1) et donc, leur poids de lien.

Une technique de partitionnement appropriée est ensuite appliquée : HDBSCAN [110], une variante de DBSCAN [45] mais plus rapide et mieux adaptée pour traiter des données de densité variable, est appliquée aux matrices $\mathbf{T}_{\text{UU}}$ et $proj(\mathbf{T}_{\text{UH}})$ (figure 3.16).

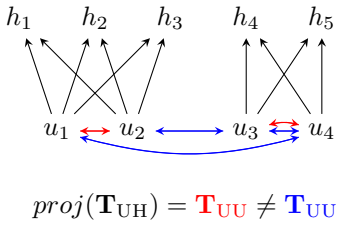



**Figure 3.16:** Motivation de la projection $proj(\mathbf{T}_{\text{UH}})$ et comparaison des matrices $proj(\mathbf{T}_{\text{UH}})$ et $\mathbf{T}_{\text{UU}}$. La comparaison des clusters trouvés – par un algorithme de partitionnement adéquat – dans ces deux matrices peut informer sur le fait que des individus connectés entre eux postent des hashtags communs. Notons que des partitionnements identiques n'impliquent pas que les individus connectés entre eux postent exactement les mêmes hashtags.

Nous obtenons finalement deux partitions des utilisateurs que nous comparons au moyen de l'information mutuelle normalisée $I$.

**Définition 3.1** (Information mutuelle normalisée)**.** Soient deux partitions $\mathscr{P}_\alpha$ et $\mathscr{P}_\beta$ associées aux

matrices $\alpha$ et $\beta$ respectivement. Leur information mutuelle normalisée $I$ est définie par [38]

$$I(\mathscr{P}_\alpha, \mathscr{P}_\beta) = \frac{-2\sum_{i=1}^{C_\alpha}\sum_{j=1}^{C_\beta} N_{ij}\log\left(\frac{N_{ij}N}{N_i N_j}\right)}{\sum_{i=1}^{C_\alpha} N_i \log\left(\frac{N_i}{N}\right) + \sum_{j=1}^{C_\beta} N_j \log\left(\frac{N_j}{N}\right)},$$

où $C_\alpha$ est le nombre de modules de la partition $\mathscr{P}_\alpha$, $N_{ij}$ le nombre de nœuds communs entre le module $i$ de $\mathscr{P}_\alpha$ et le module $j$ de $\mathscr{P}_\beta$, $N_i$ le nombre d'utilisateurs dans le module $i$.

Cet indice se situe dans [0,1] ; des valeurs plus élevées de $I$ indiquent une plus grande similarité entre les partitions.

Les valeurs de $I$ sont présentées sur la figure 3.17(b). Ces valeurs sont relativement basses, suggérant le peu de similitudes entre les partitionnements comparés. En outre, l'aspect erratique suggère qu'aucune amélioration nette ne semble se dessiner au cours du temps. En résumé, cette mesure suggère que les liens créés entre les utilisateurs au fil du temps ne connectent pas des utilisateurs postant les mêmes hashtags ou, du moins, que la proportion de liens créés entre les utilisateurs qui écrivent effectivement les mêmes hashtags est inférieure à celle des utilisateurs qui ne le font pas. Cela offre une possible explication de la baisse des scores $R_a^2$.

Cette procédure peut s'appliquer aux autres matrices liées aux méta-chemins de n'importe quel modèle afin de comparer les partitionnements obtenus avec ceux de $proj(\mathbf{T}_{\text{UH}})$ ($\mathbb{P}_{\text{UH}}$ étant la variable à expliquer dans la régression linéaire). Cela pourrait nous renseigner sur la structure des matrices associées aux VE et ainsi nous éclairer sur les résultats de la régression.

Nous calculons les valeurs d'information mutuelle normalisée[1] pour les modèles M1 et M4 (figures 3.17(c)-3.17(e)). Pour M1, nous obtenons trois paires de partitions pour lesquelles l'information mutuelle normalisée est calculée : $I_1 = I(\mathscr{P}_{\mathbf{T}_{\text{UH}}}, \mathscr{P}_{\mathbf{T}_{\mathcal{P}_1}})$, $I_2 = I(\mathscr{P}_{\mathbf{T}_{\text{UH}}}, \mathscr{P}_{\mathbf{T}_{\mathcal{P}_2}})$ et $I_3 = I(\mathscr{P}_{\mathbf{T}_{\mathcal{P}_1}}, \mathscr{P}_{\mathbf{T}_{\mathcal{P}_2}})$. Nous nous concentrons sur trois points d'intérêt (zones ombragées, figure 3.17(c)) : septembre, où $I_1$ et $I_2$ sont basses et inférieures à $I_3$ ; janvier, où toutes les trois sont élevées mais $I_3$ est toujours supérieure à $I_1$ et $I_2$ ; et juin, où $I_1$ et $I_2$ sont bien plus élevées que $I_3$. Afin d'avoir un $R_a^2$ élevé, ces trois points suggèrent la "nécessité" d'avoir des partitionnements similaires entre $proj(\mathbf{T}_{\text{UH}})$ et les projections des matrices liées aux VE (i.e. $\mathbf{T}_{\mathcal{P}_1}$ et $\mathbf{T}_{\mathcal{P}_2}$) tout en ayant des partitionnements dissemblables entre les VE. En d'autres termes, chaque VE apporte des informations utiles et non redondantes pour notre objectif (des conclusions et des interprétations similaires peuvent être faites pour M4).

En résumé, examiner la structure des matrices liées aux VE peut être utile afin de mieux comprendre les scores $R_a^2$ obtenus : alors que le rapport $n_a/n_h$ croît au cours du temps, les nouvelles

[1] Les résultats sont qualitativement similaires avec l'indice de Rand [177].

relations entre les utilisateurs ne semblent pas nécessairement concerner ceux postant les mêmes hashtags, n'apportant donc pas d'information supplémentaire pour prédire les relations U→H. Nous pensons qu'il s'agit d'une piste potentielle pour expliquer la dégradation générale des résultats de la régression.

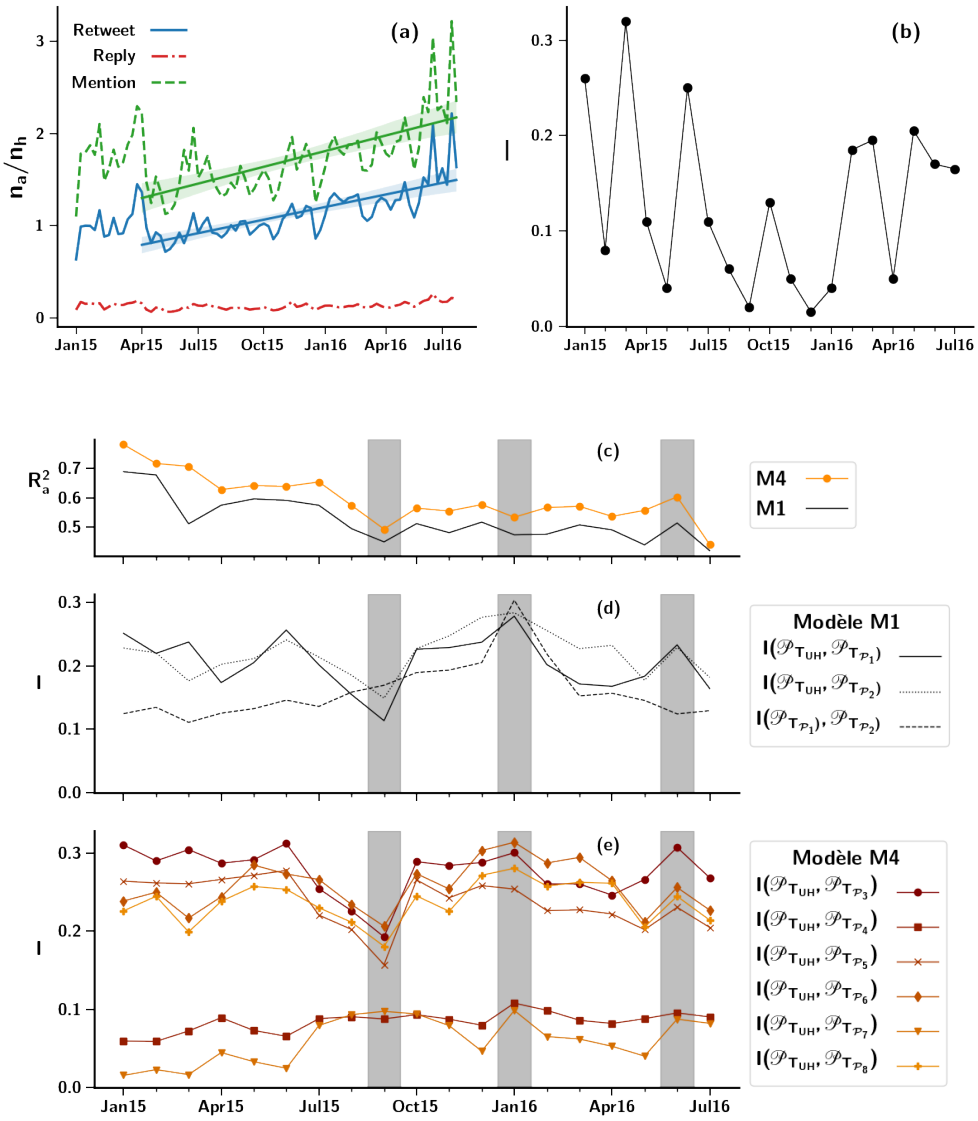



**Figure 3.17.:** Quelques informations complémentaires sur les résultats des régressions linéaires. (a) Ratio entre le nombre de liens $n_a$ pour lesquels un utilisateur est à la source et le nombre de hashtags postés par cet utilisateur $n_h$ (données agrégées par semaine). Pour les retweets et les mentions, les lignes sont interpolées, soulignant la tendance croissante, et les zones ombrées correspondent à l'intervalle de confiance à 95 %. (b) Information mutuelle normalisée I entre les partitions des clusters de $\mathbf{T}_{\text{UU}}$ et $proj(\mathbf{T}_{\text{UH}})$. Les scores sont assez faibles et aucune tendance n'émerge au fil du temps. En combinant les résultats de (a), ceci fournit un début d'explication possible concernant la baisse des scores de $R_a^2$. (c)-(e) Focus sur les modèles M1 et M4. Rappel des scores $R_a^2$ en (c) et similitude entre les groupes des individus à partir des méta-chemins des modèles M1 (noir, (d)) et M4 (ocre, (e)) par rapport aux groupes dans UH. Trois mois (zones grisées) sont sélectionnés pour lesquels les $R_a^2$ associés seraient partiellement expliqués par les valeurs de I.

### 3.4.4. Informations exogènes pour contextualiser et interpréter le résultat de la régression

Dans cette section, nous cherchons dans l'actualité des éléments d'explication empiriques des résultats de la méthode, et en particulier les $R_a^2$ obtenus (figure 3.15(a)). En effet, nous pensons qu'essayer de comprendre ces données sans plonger dans leur signification et/ou leur contexte est voué à l'échec. En d'autres termes, l'analyse ne peut pas se limiter simplement à la structure sous-jacente du réseau, aussi informative soit-elle.

Nous ne proposons pas une analyse exhaustive, ni même une analyse automatisée, car nous ne nous concentrons que sur deux mois, à savoir septembre 2015 et juin 2016. Nous proposons simplement de mettre en évidence les évènements liés à l'actualité de la crise migratoire qui pourraient être à l'origine des scores $R_a^2$ observés lors de ces deux mois. La motivation est la suivante. Si des évènements importants ont eu lieu et ont été grandement médiatisés et débattus pendant cette période, il y a de grandes chances pour que les twittos en aient parlé sur Twitter et donc posté des hashtags en rapport. Grâce aux *trending* hashtags, ces hashtags ont pu avoir une grande visibilité et être utilisés par une grande partie des utilisateurs. À l'inverse, si aucun évènement sortant du lot ne s'est produit, les hashtags sont plus décousus. Il devrait donc être plus aisé de prédire les hashtags postés par les utilisateurs dans le premier cas que dans le second.

Ci-dessous sont repris quelques évènements liées aux crises migratoires durant septembre 2015 et juin 2016. Nous expliquons en quoi, selon nous, ils peuvent fournir des éléments d'explication pour les résultats de la régression linéaire.

- En septembre 2015, de nombreux évènements liés à la crise migratoire ont eu lieu comme par exemple la fermeture des frontières hongroises, des milliers de décès de personnes essayant de rejoindre l'UE par bateau, de nombreuses actions prises par les dirigeants européens. Cependant, aucun de ces évènements n'a été extrêmement repris sur les réseaux sociaux. Ce mois est également caractérisé dans nos données par le deuxième plus faible $R_a^2$ (le pire étant juillet 2016). Une explication possible est la suivante : les utilisateurs ne postent pas de hashtags à propos de chaque évènement de septembre mais les utilisateurs avec lesquels ils sont connectés parlent un peu de tout et postent des hashtags très divers. Ainsi, pour un utilisateur donné, les hashtags postés par ses voisins sont insuffisamment liés/communs avec les siens. Un remède possible pourrait être d'envisager une granularité temporelle plus fine. Par analogie avec un surfeur aléatoire marchant sur le graphe associé, il a beaucoup trop de choix pour les nœuds de destination.
- Le mois de juin 2016 affiche une augmentation du score de $R_a^2$ par rapport aux mois voisins.

Cela pourrait s'expliquer par deux évènements qui ont attiré, plus que tout autre, l'attention des twittos. Le premier est le Congrès "United We Dream 2016" à Houston, du 10 au 12 juin 2016. En quelques mots, c'est le plus grand rassemblement de sans-papiers et de Latinx de la génération du millénaire permettant d'échanger sur leur vision (politique, culturelle, ...) de l'avenir. Ce congrès biennal a été largement relayé sur Twitter par les partisans et les détracteurs de cet évènement. Le second est le référendum sur l'adhésion du Royaume-Uni à l'Union Européenne (Brexit) le 23 juin 2016, particulièrement présent dans les conversations du 23 et 24 juin mais aussi tout au long du mois en général.

Ces deux évènements ont été caractérisés par des hashtags particuliers (#UWDCongress, #undocumented, #Brexit, #EURefResults, #WhatHaveWeDone) utilisés beaucoup plus fréquemment que les autres (sur cette période). Par exemple, #EURefResults et #Brexit sont les deux hashtags les plus présents le 24 juin (jour du résultat final et de la démission du Premier ministre David Cameron) et ce, avec un rapport 10 fois plus important que le troisième hashtags le plus présent (#EUref). De plus, les six hashtags les plus utilisés ce jour-là concernent tous le Brexit. Notons également que le meilleur modèle pour ce mois est M6, le modèle dont les VE sont des méta-chemins de longueur 2 (liées aux utilisateurs et aux médias) et 3 (uniquement celles liées aux utilisateurs). En général, le nombre de *replies* reste assez stable dans le temps ; ce sont principalement les *retweets* et les *mentions* qui fluctuent (figure 3.18). Cependant, fin juin, le nombre de *replies* a également augmenté, ce qui suggère que Brexit a déclenché des dialogues personnels entre les utilisateurs.

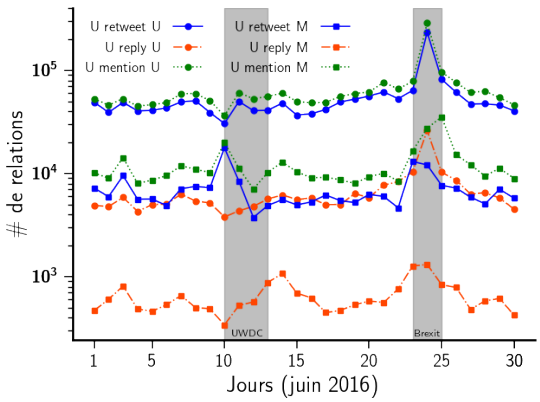



**Figure 3.18:** Nombre de relations typées au mois de juin 2016. Les deux évènements : Congrès "United We Dream 2016" et referendum sur le Brexit sont grisés et affichent des perturbations au niveau des relations.

Ces deux évènements ont rassemblé de nombreux utilisateurs, et tous en ont abondamment parlé au mois de juin. Par conséquent, la probabilité pour un utilisateur donné que ses voisins aient posté les mêmes hashtags est plus grande. En termes de structure, cela correspond à un plus grand nombre de triangles (UUH). Une façon de vérifier si ces évènements sont vraiment liés à l'augmentation du $R_a^2$ serait de se concentrer uniquement sur ces deux périodes (jours

grisés sur la figure 3.18) et de voir s'il y a une nette amélioration du $R_a^2$.

Certes, ces deux faits ne suffisent pas à expliquer les $R_a^2$ tels quels mais fournissent néanmoins quelques détails supplémentaires. Bien qu'il ne s'agisse que d'une première tentative, cela montre que s'intéresser de plus près au contexte dans lequel ont été produites les données peut éclairer la compréhension de la structure de leurs abstractions en graphe.

Pour aller plus loin, nous pourrions également travailler directement sur le graphe de co-occurrences des hashtags et trouver des clusters de hashtags. Cela pourrait d'abord informer sur les différents sujets/hashtags abordés (en examinant concrètement la signification des hashtags) et la taille des clusters pourrait ensuite fournir des informations sur leur importance. De plus, en travaillant avec des séquences de graphes de co-occurrences quotidiennes, nous pourrions également voir l'émergence de sujets limités dans le temps qui permettraient potentiellement une division plus appropriée (pour notre objectif) du temps que celle utilisée dans ce travail, à savoir une division mensuelle.

## 3.5. Intérêt des hypergraphes pour la prédiction de liens

Jusqu'à présent, nous n'avons considéré que des relations dyadiques pour prédire des liens. L'hypothèse sous-jacente est qu'un graphe est un objet adapté pour modéliser les données à analyser. Cependant, un graphe n'est pas toujours l'objet le plus naturel ou évident pour abstraire des données. Par exemple, plusieurs twittos et hashtags peuvent apparaître dans un même tweet. Il n'y a donc, *a priori*, aucune raison pour considérer des liens plutôt que des hyperliens ou hyperarcs. Dans cette section, nous montrons que la modélisation des données joue un rôle majeur dans leur analyse et interprétation. En particulier, nous envisageons des hypergraphes afin de prédire des liens [23, 59, 100] et montrons que pour des certains choix des poids des (hyper)liens et nœuds, les résultats obtenus grâce aux hypergraphes surpassent ceux obtenus grâce aux graphes.

### 3.5.1. Représentation hypergraphe et objectif

Afin de comparer les modélisations HIN et hypergraphes, nous utilisons le jeu de données Twitter Migrant sur la période restreinte de septembre à octobre 2015 ; l'idée est de simplement donner un premier aperçu des hypergraphes en pratique.

#### Types des nœuds et des hyperarcs

Seuls deux types de nœuds sont considérés : les utilisateurs U et les hashtags H. Plusieurs types d'hyperarcs sont également envisagés :

- poste qui correspond à “poste” et à prendre dans un sens très général, i.e. un tel type apparaît lorsqu'un utilisateur mentionne dans son tweet un autre utilisateur ou un hashtag ;
- co-cit qui correspond aux co-citations des utilisateurs, i.e. lorsque deux utilisateurs sont mentionnées dans un même tweet ;
- co-occur qui correspond aux cooccurrences des hashtags ;
- co-appar qui correspond aux co-apparitions des utilisateurs-hashtags.

Les types poste et co-appar correspondent à des hyperarcs tandis que co-cit et co-occur sont des hyperliens. Chaque tweet est associé à au plus quatre hyperarcs pondérés, chacun de type différent. Un exemple est proposé sur la figure 3.19 où un tweet est représenté.

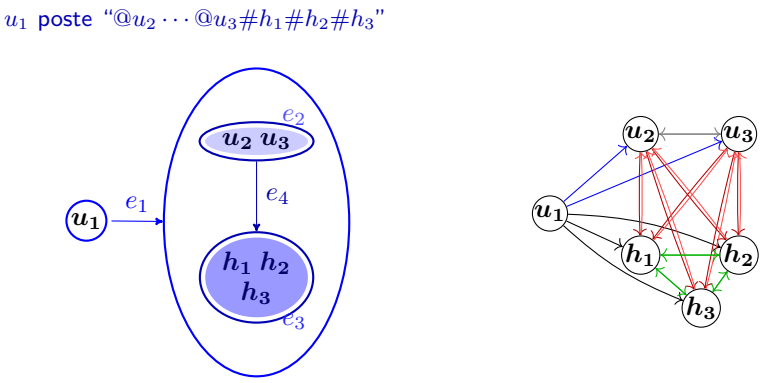



**Figure 3.19.:** Représentation d'un tweet. L'utilisateur $u_1$ poste le tweet contenant “@$u_2$@$u_2$#$h_1$#$h_2$#$h_3$”. Gauche : il en résulte quatre hyperarcs : $e_1 = (\{u_1\}, \{u_2, u_3, h_1, h_2, h_3\})$ de type poste, $e_2 = (\{u_2, u_3\})$ de type co-cit, $e_3 = (\{h_1, h_2, h_3\})$ de type co-occur et $e_4 = (\{u_2, u_3\}, \{h_1, h_2, h_3\})$ de type co-appar. Droite : projection des hyperarcs. Les types de liens sont différenciés par des différentes couleurs. Le type d'hyperarc poste se divise en deux types de liens (car type de nœuds cibles différents). Il en va de même pour co-appar : un type de liens et son inverse.

### Poids des nœuds et des hyperarcs

Ainsi qu'expliqué dans la section 1.2.2, les nœuds et les hyperarcs sont munis de poids. Ces derniers permettent de biaiser les marches aléatoires afin de favoriser certains nœuds, e.g. en fonction de leur degrés. Nous considérons trois cas de pondération :

- $C_1$ : un hypergraphe où tous les nœuds ont le même poids, sans perte de généralité $\lambda_e(v) = 1, \forall e, \forall v$ ;
- $C_2$ : un hypergraphe où les nœuds ont un poids dépendant de leur position dans l'hyperarc (i.e. de leur ordre d'apparition dans le tweet) et de leur type ;
- $C_3$ : projection de l'hypergraphe (voir section 1.2.3, figure 3.19) du cas $C_1$.

Le choix de $C_2$ est motivé par l'hypothèse suivante : les gens écrivent en premier ce qui a le plus d'importance pour eux. Dans ce cas, nous formalisons les hyperarcs sous la forme

$e = (T(e), H(e)) = ([u_1, u_2, ..., u_n], [v_1, v_2, ..., v_m])$. En d'autres mots, un hyperarc n'est plus un couple d'ensembles mais un couple de listes ordonnées. Dans notre cas, l'ordre est celui dans lequel les mots sont écrits dans le tweet. Plus précisément, chaque nœud $u \in e$ possède un type $\phi(u)$. Nous notons $|\phi^T(u)| = |\{v \in T(e) \,|\, \phi(v) = \phi(u)\}|$ le nombre de nœuds de type $\phi(u)$ dans la queue de $e$. De façon analogue, $|\phi^H(u)| = |\{v \in H(e) \,|\, \phi(v) = \phi(u)\}|$ dénote le nombre de nœuds de type $\phi(u)$ dans la tête de $e$. De cette façon, nous définissons le poids d'un nœud $u_j$, $j = 1, ..., n$, dans $T(e)$ comme suit : $\lambda_e(u_j) = |\phi^T(u_j)| - |\{u_l \in T(e) \,|\, \phi(u_l) = \phi(u_j),\, l < j\}|$. De façon similaire, le poids d'un nœud $v_j$, $j = 1, ..., m$, dans $H(e)$ comme suit : $\lambda_e(v_j) = |\phi^H(v_j)| - |\{v_l \in H(e) \,|\, \phi(v_l) = \phi(v_j),\, l < j\}|$. Dans le cas d'un hyperlien (i.e. non orienté), la procédure est la même sauf que $e = [u_1, u_2, ..., u_n]$, i.e. l'hyperlien est une liste ordonnée.

Enfin, aussi bien pour $C_1$ que pour $C_2$, le poids des hyperarcs $e$ correspond à son nombre d'occurrences dans les données multiplié par $\delta^-(e)^\gamma$, avec $\gamma \in \mathbb{R}^+$ un paramètre permettant de régler à volonté l'importance du degré entrant des hyperarcs.

### Hyper-méta-chemins

Les hyper-méta-chemins utilisés dans la suite sont présentés dans la table 3.5.

| Méta-chemin | Interprétation |
|---|---|
| $\mathcal{M}_1$: U $\xrightarrow{\text{poste}}$ U $\xrightarrow{\text{poste}}$ H | hashtags postés par les utilisateurs postés par $u$ |
| $\mathcal{M}_2$: U $\xrightarrow{\text{poste}^{-1}}$ U $\xrightarrow{\text{poste}}$ H | hashtags postés par les utilisateurs ayant mentionné $u$ |
| $\mathcal{M}_3$: U $\xrightarrow{\text{poste}}$ U $\xrightarrow{\text{co-appar}}$ H | hashtags écrits dans un même tweet que les utilisateurs postés par $u$ |
| $\mathcal{M}_4$: U $\xrightarrow{\text{poste}^{-1}}$ U $\xrightarrow{\text{co-appar}}$ H | hashtags écrits dans un même tweet que les utilisateurs ayant mentionné $u$ |
| $\mathcal{M}_5$: U $\xrightarrow{\text{co-cit}}$ U $\xrightarrow{\text{poste}}$ H | hashtags postés par les utilisateurs mentionné dans un même tweet que $u$ |
| $\mathcal{M}_6$: U $\xrightarrow{\text{co-cit}}$ U $\xrightarrow{\text{co-appar}}$ H | hashtags écrits dans un même tweet que les utilisateurs mentionné dans un même tweet que $u$ |
| $\mathcal{M}_7$: U $\xrightarrow{\text{poste}}$ U $\xrightarrow{\text{poste}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags postés par les utilisateurs postés par $u$ |
| $\mathcal{M}_8$: U $\xrightarrow{\text{poste}^{-1}}$ U $\xrightarrow{\text{poste}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags postés par les utilisateurs ayant mentionné $u$ |
| $\mathcal{M}_9$: U $\xrightarrow{\text{poste}}$ U $\xrightarrow{\text{co-appar}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags écrits dans un même tweet que les utilisateurs postés par $u$ |
| $\mathcal{M}_{10}$: U $\xrightarrow{\text{poste}^{-1}}$ U $\xrightarrow{\text{co-appar}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags écrits dans un même tweet que les utilisateurs ayant mentionné $u$ |
| $\mathcal{M}_{11}$: U $\xrightarrow{\text{co-cit}}$ U $\xrightarrow{\text{poste}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags postés par les utilisateurs mentionnés dans un même tweet que $u$ |
| $\mathcal{M}_{12}$: U $\xrightarrow{\text{co-cit}}$ U $\xrightarrow{\text{co-appar}}$ H $\xrightarrow{\text{co-occur}}$ H | hashtags écrits dans un même tweet que les hashtags écrits dans un même tweet que les utilisateurs mentionnés dans un même tweet que $u$ |

**Table 3.5.:** Définition des hyper-méta-chemins.

#### Objectif

L'objectif poursuivi est de montrer que la modélisation des données joue un rôle majeur dans leur analyse et interprétation. Le cas d'application est la prédiction du poids des liens, i.e. interactions dyadiques, en se servant des informations fournies par l'hypergraphe. Bien que cela puisse sembler étrange de toujours considérer des liens comme objectif de prédiction et non des hyperliens ou hyperarcs, c'est également la direction suivie dans e.g. [23, 59, 100].

Nous tentons d'expliquer le poids des liens de types U→H, comme dans la section 3.4. Pour ce faire, nous utilisons la méthode proposée dans la section 3.1 avec les VE associées aux hyper-méta-chemins présentés dans la table 3.5. Dans les cas $C_1$ et $C_2$, les VE sont les distributions de probabilités calculées dans les hypergraphes correspondants : elles sont données par l'éq. (1.8). Dans le cas du HIN, i.e. cas $C_3$, les distributions sont données par l'éq. (1.2), comme c'était le cas dans les sections précédentes.

### 3.5.2. Résultats

Les résultats de la régression linéaire pour retrouver le poids des liens, obtenus par validation croisée avec 10 échantillons (training-test: 80/20%), sont présentés sur la figure 3.20(b). Quel que soit le mois, les résultats obtenus sur les hypergraphes $C_2$ sont toujours meilleurs que ceux obtenus sur le graphe projeté $C_3$, pour toutes les valeurs envisagées[1] de $\gamma$. Le cas $C_1$ n'est pas aussi évident, suggérant que cette représentation des données n'est pas toujours la meilleure pour notre objectif. Aussi, pour chaque mois, $C_2$ surpasse $C_1$, suggérant qu'il peut être intéressant d'introduire une notion d'ordre dans les modélisations. Notons aussi que le choix du poids des nœuds fait ici n'est qu'un exemple mais permet déjà d'améliorer les scores.

Les meilleurs $R_a^2$ sont repris dans la table 3.6. La plupart du temps, les $\gamma$ optimaux sont proches de 1, suggérant qu'il est bénéfique d'accorder légèrement plus d'importance aux hyperarcs de degré plus important.

| | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| septembre | 0.6196 (en $\gamma = 1.2$) | 0.7044 (en $\gamma = 1.6$) | 0.6081 |
| octobre | 0.5297 (en $\gamma = 1$) | 0.5398(en $\gamma = 1.1$) | 0.5135 |

**Table 3.6.:** Resultats pour les régressions linéaires. Pour les cas $C_1$ et $C_2$, seuls les meilleurs $R_a^2$ sont indiqués, accompagnés de la valeur $\gamma$ permettant de les obtenir.

Au niveau des VE sélectionnées dans les modèles finaux, celles faisant apparaître les cooccurrences de hashtags sont beaucoup moins présentes. Une piste d'explication est de s'intéresser aux graphes pondérés de cooccurrences des utilisateurs et hashtags séparément et à leur distribution de degrés. Il y

[1] L'allure des $R_a^2$ étant une courbe concave, il est tentant de dire que pour d'autres valeurs de $\gamma$, $C_3$ surpasse $C_2$.

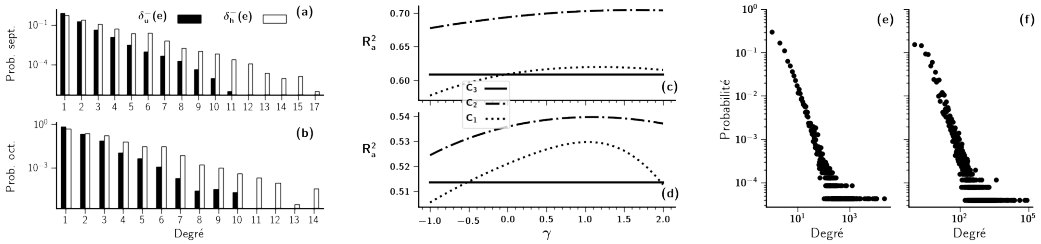



**Figure 3.20.:** Comparaison des distributions des degrés sortants des hyperarcs vers des utilisateurs ($\delta_u^-(e)$, noir) et hashtags ($\delta_h^-(e)$, blanc) pour le mois de septembre (a) et d'octobre (b). Résultats ($R_a^2$) des modèles linéaires pour le mois de septembre (c) et d'octobre (d) en fonction de la valeur du paramètre $\gamma$ lorsqu'on modélise par un hypergraphe avec $\lambda_e(v) = 1$, $\forall e, v$ (trait pointillé) et $\lambda_e(v)$ dépend de l'ordre dans lequel $v$ apparaît dans $e$ (trait long pointillé) et $R^2$ du modèle obtenu par la modélisation graphe (trait plein Distribution de degrés dans les graphes de co-occurrences des utilisateurs (e) et hashtags (f) pour le mois d'octobre).

a plus de hashtags possédant un degré élevé comparé aux utilisateurs (figure 3.20(e)). Une hypothèse est qu'en général, il y a quelques hashtags qui sont très souvent écrits mais accompagnés chaque fois d'un ou plusieurs autres hashtags. En fait, les utilisateurs utilisent un hashtag général, apparaissant par exemple dans les *trending hashtags*, afin de participer au débat général, mais personnalisent leur tweet avec quelque autre hashtag. En revanche, concernant le graphe de cooccurrences des utilisateurs, ces derniers seraient plus souvent connectés avec les mêmes utilisateurs. Cela ne serait pas étonnant puisqu'il s'agit de relations sociales : on préfère toujours interagir avec des personnes proches, que l'on connaît (virtuellement ou non), qu'avec n'importe qui. Dès lors, les relations entre utilisateurs ont plus de sens, pour notre objectif, que les relations entre hashtags.

### 3.5.3. Bilan

Dans cette section, nous avons comparé les résultats obtenus pour la prédiction du poids pondéré des liens lorsque les données sont modélisées par des hypergraphes et des graphes. Les résultats dépendent fortement du poids des hyperarcs et nœuds de l'hypergraphe. Ils suggèrent en outre que considérer plus d'informations présentes dans les données, telle qu'une notion d'ordre pour pondérer les nœuds, peut être utile.

## 3.6. Résumé et perspectives

### Résumé

Dans ce chapitre, nous avons présenté une approche basée sur des marches aléatoires contraintes par des méta-chemins pour prédire le poids de liens manquants dans un HIN. Nous avons expliqué

l'approche de la régression linéaire, sa méthode de résolution ainsi que les raisons qui nous ont menés à ces choix.

Afin de mesurer le potentiel de cette approche, nous avons considéré différents cas d'étude. Les points retenus sont listés ci-dessous :

- Les méta-chemins permettent une première interprétation des données en e.g. permettant de calculer la similarité entre deux nœuds en fonction d'une sémantique particulière. Ils ont l'avantage d'être très simples à définir et à utiliser. Cependant, comme nous avons pu le voir, leur pouvoir de renseignement sur la structure d'un graphe ou sous-graphe est relativement limité. Afin de mieux comprendre l'organisation d'un système, comme par exemple les mécanismes de co-publications, il serait très intéressant de se pencher sur d'autres structures plus complexes telles que des motifs [150]. Voir section 3.3.4 ;
- Lorsqu'il existe un partitionnement explicite ou prédéfini des données, il est quelquefois plus judicieux d'utiliser explicitement ce partitionnement thématique que d'utiliser des méta-chemins pour trouver un partitionnement implicite. Cela peut permettre, par la suite, une meilleure analyse des données. Voir section 3.3.3 ;
- Il est bien souvent difficile de trouver un modèle véritablement bon pour expliquer ou prédire des données. En outre, un modèle ayant pour objectif d'expliquer ou de prédire une variable est un compromis entre complexité et qualité. Dans nos applications, la complexité peut venir du méta-chemin lui-même ou bien d'un trop grand nombre de VE, voir figure 3.5. Un bon modèle descriptif (au sens du $R_a^2$) n'est pas nécessairement un bon modèle prédictif (problème biais-variance, voir figure 3.4). Notons tout de même que même un "mauvais" modèle peut faciliter la compréhension des données analysées ;
- L'approche proposée permet de se faire une première idée des données traitées. Malheureusement, son pouvoir explicatif et prédictif est relativement restreint. Il est dès lors indispensable de recourir à d'autres outils d'analyse et de contextualiser et interpréter les données à l'aide d'éléments exogènes, afin de venir conforter ou non les résultats de la régression. Voir sections 3.4.3 et 3.4.4 ;
- Un point crucial est de bien choisir la représentation/abstraction des données à analyser puisque les résultats en dépendent. Voir figure 3.20(b). Il s'agit en fait des questions basiques qui se posent lorsqu'on désire analyser un système/des données : quelles sont les entités élémentaires, quelles sont les relations qui regroupent ces entités, y a-t-il des dépendances entre ces relations ? Une fois cela identifié : quel formalisme choisir pour représenter ces entités, relations et dépendances ?

## Perspectives

**Intégrer le temps au formalisme des HIN** Généralement, tout évènement est situable dans le *temps*. Avec le développement des plateformes en ligne, des objets connectés et ses capacités de stockage de données, il est de plus en plus aisé de recueillir des données horodatées. Il serait donc intéressant d'intégrer le temps au formalisme des HIN afin d'être capable d'analyser des données temporelles. En outre, considérer le temps est un *premier* pas vers une idée de *cause* et non plus seulement de corrélation.

Depuis quelques années, les chercheurs utilisent des modèles de graphes d'ordre supérieur[1] pour intégrer le temps. Une motivation pour développer ces modèles est de ne plus se limiter aux interactions, mais d'aussi modéliser les relations temporelles (i.e. les chemins temporels) permettant à deux nœuds de s'influencer indirectement.

Nous considérons un graphe temporel représenté par une séquence de liens orientés [98, 141]. Le triplet $(t_i, u_i, v_i)$ représente un lien orienté de $u_i$ vers $v_i$ au temps $t_i$. À l'instar des HIN, les nœuds et liens temporels ont un type.

**Définition 3.2** (Chemin temporel)**.** Un *chemin temporel* de longueur $n$ est une séquence de triplets $p = ((t_0, u_0, v_0), ..., (t_n, u_n, v_n))$ telle que $\forall\, i \in \{0, ..., n-1\}$, $t_i \leq t_{i+1}$ et $v_i = u_{i+1}$. La condition $t_{i+1} - t_i < \tau$, $\tau \in \mathbb{R}^+$ peut être ajoutée afin de limiter le laps de temps entre deux liens consécutifs d'un chemin temporel.

Un chemin temporel $p = ((t_0, u_0, u_1), ..., (t_{n-1}, u_{n-1}, u_n))$ satisfait un méta-chemin $\mathcal{P} = V_0 \xrightarrow{E_1} V_1 \cdots \xrightarrow{E_n} V_n$, noté $p \in \mathcal{P}$, si et seulement si $\forall\, i \in \{0, ..., n-1\}$, $(u_i, u_{i+1}) \in V_i \times V_{i+1}$ et $\psi[(t_i, u_i, v_i)] = E_{i+1}$.

Un intérêt possible des chemins temporels est l'"affinage" des corrélations entre les liens d'un HIN. À titre d'exemple, considérons la figure 3.21. Soit un lien $e_c$ de type $E_c$ entre $u_0 \in V_0$ et $u_3 \in V_3$ au temps $t_a$. Imaginons vouloir expliquer/prédire ce lien $e_c$ à l'aide des chemins satisfaisant les méta-chemins $\mathcal{P}_1 = V_0 \xrightarrow{E_1} V_1 \xrightarrow{E_2} V_2 \xrightarrow{E_3} V_3$ et $\mathcal{P}_2 = V_0 \xrightarrow{E_1} V_1 \xrightarrow{E_4} V_3$. Dans le HIN temporel, seul $\mathcal{P}_1$ est utile puisqu'aucun chemin temporel ne satisfait $\mathcal{P}_2$. En revanche, dans le HIN statique, les deux méta-chemins sont utiles puisque chacun possède une instance de chemin. Cependant, utiliser $\mathcal{P}_2$ est trompeur si on recherche de possibles *causes*.

À partir de ces chemins temporels, une matrice $\mathbf{A}$ est construite avec $A_{ij} = |\{p \in \mathcal{P} \text{ tel que } u_0 = i \text{ et } v_n = j\}|$ le nombre de chemins temporels satisfaisant un méta-chemin $\mathcal{P}$ donné démarrant en $i$ et finissant en $j$. Finalement, $\mathbf{A}$ est rendue stochastique sur les lignes. Il s'agit donc d'une

[1] Notons que le terme "ordre supérieur" est quelque peu ambigu : il est tantôt utilisé pour parler d'interactions temporelles avec mémoire [96, 126, 141], tantôt pour parler de la nature multicouche des systèmes étudiés [40]. Dans cette perspective, il est utilisé en regard du temps.

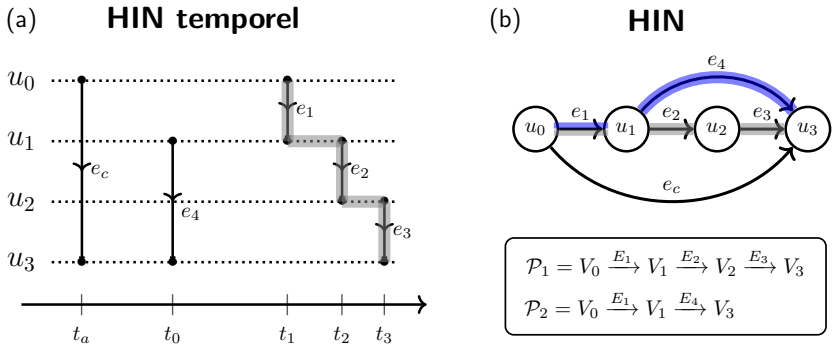



**Figure 3.21.:** Chemins (temporels) satisfaisant un méta-chemin. Soient $\mathcal{P}_1 = V_0 \xrightarrow{E_1} V_1 \xrightarrow{E_2} V_2 \xrightarrow{E_3} V_3$ et $\mathcal{P}_2 = V_0 \xrightarrow{E_1} V_1 \xrightarrow{E_4} V_3$ deux méta-chemins. (a) Dans le HIN temporel, seul un chemin satisfaisant $\mathcal{P}_1$ existe : $p^{temp} = ((t_1, u_0, u_1), (t_2, u_1, u_2), (t_3, u_2, u_3))$ surligné en gris. (b) Dans le HIN statique construit à partir du HIN temporel de (a), i.e. un lien entre deux nœuds $a$ et $b$ existe s'il existe un lien temporel les unissant (le poids du lien égale le nombre de liens temporels entre $a$ et $b$ et le type du lien est le même que le lien temporel dont il est issu), il existe un chemin $p_1^{stat} = u_0 \xrightarrow{e_1} u_1 \xrightarrow{e_2} u_2 \xrightarrow{e_3} u_3$ satisfaisant $\mathcal{P}_1$ (gris) et un chemin $p_2^{stat} = u_0 \xrightarrow{e_1} u_1 \xrightarrow{e_4} u_3$ satisfaisant $\mathcal{P}_2$ (bleu).

version temporelle de *Normalized Path Count* [156]. À l'instar de ce qui a été fait précédemment dans ce chapitre, cette matrice $\mathbf{A}$ pourrait servir de VE pour une régression linéaire.

**Prédire des interactions quelconques** Dans les dernières décennies, de nombreux systèmes complexes ont été analysés, avec succès, grâce au formalisme des graphes. Ces derniers supposent que les relations entre les entités d'un système sont de nature dyadique : les nœuds interagissent par paires. Cependant, dans les situations telles que les conversations entre individus, les réactions chimiques, les réseaux trophiques, etc., une interaction fait intervenir un nombre quelconque d'entités. Afin de modéliser et d'analyser ces interactions, les formalismes des hypergraphes et des complexes simpliciaux se sont imposés comme candidats idéals. Connus depuis longtemps, ces formalismes ont suscité un regain d'intérêt ces dernières années, comme en témoigne le nombre important de récents travaux à ce sujet (voir [13, 165] et les références citées dedans.)

Une perspective de ce travail serait de s'intéresser à la prédiction d'interactions faisant intervenir un nombre quelconque de nœuds. Cela permettrait, par exemple, de prédire un tweet entier ou des collaborations scientifiques, voir [145] (reposant sur la factorisation de tenseurs), [176] (reposant sur des $n$-graphes projetés) et les références citées dedans. Notons que ce problème se distingue de celui de la détection de communautés dont le but est d'identifier des structures à grande échelle dans un graphe.

L'approche que nous avons proposée dans ce chapitre est basée sur des distributions de probabilités. Grâce à ces distributions, nous inférons la structure d'un graphe ou HIN. L'idée est donc, à partir de ces relations dyadiques inférées, de déduire l'hypergraphe qui représente au mieux les données

(figure 3.22).

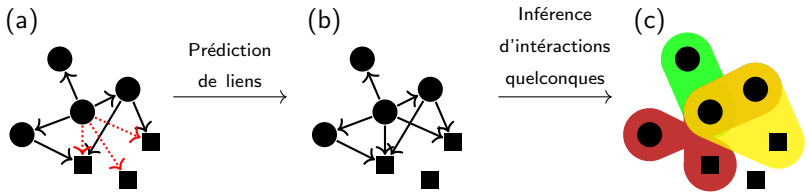



**Figure 3.22.:** Inférence d'interaction d'ordres supérieurs en deux temps. (a) Graphe (ou hypergraphe) avec liens manquants/non observés (en rouge). (b) Récupération et prédiction de liens, e.g. par la méthode proposée à la section 3.1. (c) Inférence d'intéraction d'ordres supérieurs à partir de relations dyadiques.

Sans contraintes supplémentaires, ce problème est *mal défini* (*ill-posed*) dans le sens où il est possible d'inférer plusieurs hypergraphes à partir d'un même graphe. Malgré cela, Young *et al.* ont récemment proposé une approche bayésienne permettant de reconstruire des interactions d'ordres supérieurs (non typées, non orientées et non pondérées) à partir de relations dyadiques [178]. Une telle approche pourrait s'appliquer au HIN prédit par notre méthode.

# 4. Modèle d'évolution du poids des liens

Comme nous l'avons mentionné au début de ce manuscrit, de nombreux systèmes évoluent au cours du temps. Dans ce chapitre, nous nous intéressons aux corrélations entre les liens d'un HIN sous un angle différent du chapitre précédent, à savoir le rôle des corrélations dans l'évolution du poids des liens et ce, afin de proposer un modèle de HIN dynamique. Néanmoins, à l'instar du chapitre 3, le concept de méta-chemin est à la base de la dynamique proposée. À chaque type de liens nous associons un ensemble de méta-chemins. Une combinaison linéaire des distributions de probabilités résultant de marches aléatoires contraintes par ces méta-chemins va permettre de mettre à jour le poids des liens du type en question. L'état futur du HIN ne dépend que de son état présent. Comme application, un modèle d'attention est proposé. Le terme "attention", du latin *attentio* – tension de l'esprit vers quelque chose – est préféré à "opinion", du latin *opinio* – idée que l'on a de quelque chose – car, comme dit, le modèle repose sur des probabilités. Nous décidons d'interpréter ces dernières comme étant la quantité d'attention qu'un individu consacre à un sujet ; nous faisons l'hypothèse d'une capacité d'attention bornée [114] (ici à 1, la somme des probabilités). Il n'empêche que le modèle présenté dans la section 4.2 ressemble fort à – et est inspiré de – certains modèles d'opinions et est au contraire fort peu inspiré de la théorie de l'attention [89, 114, 127].

Ce chapitre exploratoire est organisé comme suit. Nous passons en revue certains modèles de dynamiques d'opinions dans la section 4.1. Le modèle d'évolution du poids des liens, motivé par les modèles d'évolution d'attention (voire d'opinions) des individus, est présenté dans la section 4.2. Un exemple d'application est ensuite proposé dans la section 4.3 : un modèle de dynamique d'attention. Un premier exemple est présenté avec seulement des individus et des sujets d'attention et ensuite, un second exemple dans lequel un média est présent. Finalement, dans la section 4.3.4, nous discutons une extension du modèle proposé aux hypergraphes et résumons le travail réalisé dans la section 4.4.

## 4.1. Modéliser l'influence sociale : un problème difficile

### 4.1.1. Complexité de l'influence sociale

Parmi les systèmes évoluant au cours du temps se trouvent les dynamiques humaines. Par exemple, l'influence sociale fait partie des interactions sociales humaines. Dans de nombreuses situations, les gens modifient leur comportement, leurs opinions, leurs croyances en fonction des individus avec lesquels ils interagissent. Malgré l'abondance de travaux sur ces sujets, l'influence sociale reste un phénomène déroutant, notamment en ce qui concerne la compréhension des interactions complexes entre les processus aux niveaux microscopique et macroscopique. En effet, les mécanismes d'influence sociale peuvent générer des liens micro-macro complexes dans lesquels le résultat des interactions individuelles peut être inattendu et non désiré, ou du moins non intentionnel, du point de vue de l'individu.

À titre d'exemple, rappelons le célèbre modèle de Schelling dont l'objectif est d'illustrer les dynamiques de mobilité résidentielle et de ségrégation raciale et ethnique [139, 140], e.g. un modèle durable de nombreuses grandes villes du monde et en particulier, des USA. Les hypothèses sont que les individus sont répartis en deux groupes, préfèrent avoir un certain pourcentage de leurs voisins du même groupe (50% ou plus), qu'ils ont une vision locale de la population, peuvent détecter la composition de leur voisinage et sont motivés à se déplacer vers l'endroit disponible le plus proche où le pourcentage de voisins communs est acceptable. Grâce à ce modèle très simple, Schelling a montré que les préférences individuelles sur le lieu de résidence se combinent dans des modèles spatiaux agrégés de ségrégation résidentielle, et a ainsi illustré le pouvoir des mécanismes d'interdépendance pour expliquer les liens micro-macros.

Les individus ne sont pas les seuls à avoir une influence sur eux-mêmes. Par exemple, les messages véhiculés par les médias façonnent également l'attitude des gens et peuvent induire un changement d'opinion. Cela signifie que la dynamique humaine n'est pas seulement influencée par le contexte social environnant, mais aussi par les informations apportées par les médias. Cependant, ce n'est pas un flux à sens unique ; en effet, les blogs, les télévisions, les journaux sont affectés par une quantité massive de facteurs individuels et sociaux, par exemple les goûts de leur audience, les objectifs des journaux, la pression sociale.

### 4.1.2. Une physique des opinions

On peut se demander, à juste titre, si la complexité intrinsèque des opinions humaines peut être décrite avec précision par des modèles mathématiques. Certains auteurs affirment que les problèmes sociologiques les plus fondamentaux sont de nature non mathématique [1, 172]. En effet, en raison

de la nature chaotique des opinions humaines, les objectifs de modélisation sont de véritables défis. Cependant, le lien entre la sociologie et la physique remonte à très longtemps [75]. Dans son livre publié en 1896, Comte introduit le concept de "physique sociale" [36]. Elle s'appuie sur des lois déterministes pour étudier la dynamique sociale d'un système. En particulier, les outils développés dans le domaine des statistiques sont bien adaptés à une analyse quantitative des systèmes sociaux, souvent composés d'un grand nombre d'individus [28].

Beaucoup d'efforts ont été consacrés à l'étude des dynamiques d'opinions avec un accent sur les mécanismes élémentaires permettant l'émergence d'un consensus global, ainsi que le rôle de forces internes et externes telles que la pression sociale et les médias. Conséquemment, un nombre important de modèles ont vu le jour [9, 53, 54, 77, 168]. Une notion récurrente dans ces modèles est celle de la confiance limitée (*bounded confidence*) : les individus ne peuvent s'influencer que si leurs opinions sont *suffisamment* proches. Dans ce cas, les individus modifient leurs opinions afin de se rapprocher de l'opinion (e.g. moyenne) des individus engagés (*majority rules*). Les deux modèles les plus connus sont le modèle (asynchrone[1]) de Deffuant-Weisbuch [41, 171] et le modèle (synchrone[1]) de Hegselmann-Krause [74].

Notons que plusieurs chercheurs se sont aussi intéressés à des influences négatives : lorsque des individus possèdent des opinions trop éloignées ou incompatibles, les individus tendent à s'éloigner davantage les uns des autres [86, 131, 138].

Ces dernières années, le rôle des médias dans la formation d'un consensus global a suscité beaucoup d'intérêt. Ces médias peuvent être modélisés par deux stratégies différentes. La première considère les médias comme des entités ayant une opinion immuable, tandis que la seconde envisage les médias avec une approche plus complexe : leur opinion peut varier au cours du temps, de manière indépendante ou bien influencée par leur audience. Cet ajout d'un ou plusieurs médias se fait aussi bien dans des modèles d'opinions discrètes unidimensionnelles [29, 35, 81, 167], multidimensionnelles (typiquement, une adaptation du modèle de dissémination de la culture d'Axelrod [9]) [24, 57, 63–65, 125, 133] que d'opinions continues [27, 108, 164]. Un constat général est le suivant : lorsque le message véhiculé par le(s) média(s) est trop fort ou trop obsessionnel, cela conduit à une fragmentation au sein de la population d'individus. Ce résultat est valable pour les modèles discrets et continus. Les effets médiatiques sont également influencés par la topologie des graphes sociaux et ce, particulièrement lorsque les liens sociaux sont régis par le principe d'homophilie, i.e.

[1] *Synchrone* signifie que tous les individus mettent à jour leur opinon à l'unisson à chaque pas de temps. À l'inverse, une mise à jour *asynchrone* implique qu'à chaque pas de temps, un individu choisi aléatoirement parmi $n$ interagit avec un de ses voisins (pour le modèle de e.g. Deffuant-Weisbuch) et modifie son opinion en fonction, tandis que le reste du système reste fixe. Si l'individu est choisi de façon uniforme, alors après $n$ pas de temps, chaque individu est, en moyenne, mis une fois à jour (en règle générale mais dans le cas de Deffuant-Weisbuch, après $n/2$ pas de temps). Dit autrement, $n$ pas dans un modèle asynchrone correspondent en moyenne à un pas dans un modèle synchrone.

la tendance individuelle à interagir préférentiellement avec des personnes perçues comme semblables. L'homophilie peut induire une fragmentation de l'espace des opinions [58]. La présence de multiples médias n'a été abordée que dans quelques travaux [130, 170]. Il a été constaté par exemple que la concurrence entre les médias induit une fragmentation au niveau des individus.

Les travaux présentés dans le chapitre précédent offrent une classe de modèles pour aborder ces questions. Dans le reste de ce chapitre, nous proposons une dynamique d'évolution du poids des liens d'un HIN. La dynamique se base sur des marches aléatoires contraintes par un méta-chemin afin d'apporter une sémantique (voir chapitre 3) et limitées par une *distance* (explicitée dans la section 4.2) imitant le principe de confiance limitée des modèles d'opinions. Cette dynamique peut servir à définir un modèle d'attention. Ce modèle fait évoluer les distributions d'attention que des individus accordent à certains sujets, lorsque ces individus s'influencent eux-mêmes et sont potentiellement influencés par un unique média dont la distribution d'attention est fixée. Comme nous l'illustrons à l'aide d'exemples dans la section 4.3, ce modèle peut être vu comme une extension de certains modèles d'opinions avec confiance limitée. Dès lors, certaines observations préliminaires présentées dans la suite semblent être en accord avec les résultats cités précédemment, soutenant ainsi la pertinence de notre approche.

## 4.2. Évolution des liens d'un HIN

Soit un HIN $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$ et le schéma associé $S_H = (\mathcal{V}, \mathcal{E}, \nu_s, \nu_t)$. Pour chaque type de liens $E_i \in \mathcal{E}$, nous associons la matrice de transition de probabilité $\mathbf{T}_{\mathbf{E_i}}$ telle que

$$\mathbf{T}_{\mathbf{E_i}}(u, v) = \frac{w_{E_i}(u, v)}{\sum_{v' \in \nu_t(E_i)} w_{E_i}(u, v')}, \tag{4.1}$$

avec $w_{E_i}(u, v)$ le poids des liens de type $E_i$ entre $u$ et $v$ (déf. 1.14). La valeur $\mathbf{T}_{\mathbf{E_i}}(u, v)$ est la probabilité de transition contrainte par un type de liens (éq. (1.1)).

La dynamique du HIN que nous proposons est une dynamique du poids des liens, i.e. $\mathbf{T}_{\mathbf{E_i}}(t)$. D'une certaine façon, l'état du HIN au temps $t$ est déterminé par l'ensemble des matrices $\mathcal{T}(t) = \{\mathbf{T}_{\mathbf{E_i}}(t),\ E_i \in \mathcal{E}\}$. Dans un cadre général, la dynamique du HIN est donnée par

$$\mathcal{T}(t+1) = \mathcal{F}(\mathcal{T}(t)). \tag{4.2}$$

De cette façon, nous supposons que la taille (nombre de nœuds et de liens) du HIN reste inchangée et que seuls les poids des liens, et donc la structure du HIN, évoluent au cours du temps[1].

[1] Nous supposons un HIN complet, mais le poids d'un lien peut être nul.

### 4.2.1. Expression générale du modèle

Dans le chapitre 3, nous avons vu qu'il existait des corrélations entre les poids (normalisés) des liens de différents types d'un HIN. En particulier, pour un type de liens $E_i$, nous avons montré qu'il était possible d'utiliser un ensemble de méta-chemins $\{\mathcal{P}_j\}_i$ permettant de contraindre des marches aléatoires afin d'expliquer/prédire le poids des liens de ce type $E_i$. En particulier, nous avons défini (éq. (3.2))

$$\begin{bmatrix} \mathbb{P}_{E_c}(:|u_1)^\mathsf{T} \\ \vdots \\ \mathbb{P}_{E_c}(:|u_{|V_s|})^\mathsf{T} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{P}_{\mathcal{P}_1}(:|u_1)^\mathsf{T} & \cdots & \mathbb{P}_{\mathcal{P}_{|\mathcal{E}_\mathcal{P}|}}(:|u_1)^\mathsf{T} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbb{P}_{\mathcal{P}_1}(:|u_{|V_s|})^\mathsf{T} & \cdots & \mathbb{P}_{\mathcal{P}_{|\mathcal{E}_\mathcal{P}|}}(:|u_{|V_s|})^\mathsf{T} \end{bmatrix} \begin{bmatrix} \beta_{E_c,0} \\ \vdots \\ \beta_{E_c,\mathcal{P}_{|\mathcal{E}_\mathcal{P}|}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_{|V_s|} \end{bmatrix}.$$

En utilisant la définition des matrices $\mathbf{T}_\mathcal{P}$ dont chaque entrée $\mathbf{T}_\mathcal{P}(u,v) = \mathbb{P}_\mathcal{P}(v|u)$ encode la probabilité d'atteindre $v$ partant de $u$ via une marche aléatoire contrainte par le méta-chemin $\mathcal{P}$ (section 3.4.3), le système ci-dessus peut se réécrire via l'approximation suivante[1]

$$\mathbf{T}_{\mathbf{E_i}} \simeq \sum_{\mathcal{P}_j \in \{\mathcal{P}_j\}_i} \beta_{E_i,\mathcal{P}_j} \mathbf{T}_{\mathcal{P}_\mathbf{j}}, \qquad \beta_{E_i,\mathcal{P}_j} \in \mathbb{R},\ \forall i,j \tag{4.3}$$

Nous nous basons sur l'expression (4.3) pour définir la dynamique du poids des liens d'un HIN. Plus précisément, nous voulons utiliser les méta-chemins non plus pour *expliquer* ou *prédire* un type de liens comme au chapitre 3, mais pour faire *évoluer* un type de liens. Dès lors, nous modifions légèrement l'éq. (4.3) afin d'inclure une notion de temps et considérons la fonction suivante

$$\mathbf{T}^\epsilon_{\mathbf{E_i}}(t+1) = \mathbf{\Lambda}_{\mathbf{E_i}} \sum_{\mathcal{P}_j \in \{\mathcal{P}_j\}_i} \beta_{E_i,\mathcal{P}_j}(t) \mathbf{T}^\epsilon_{\mathcal{P}_\mathbf{j}}(t) + (\mathbf{I} - \mathbf{\Lambda}_{\mathbf{E_i}}) \mathbf{T}^\epsilon_{\mathbf{E_i}}(t), \quad \beta_{E_i,\mathcal{P}_j}(t) \in \mathbb{R},\ \forall i,j \tag{4.4}$$

où $\mathbf{\Lambda}_{\mathbf{E_i}}$ est une matrice diagonale dont les éléments appartiennent à [0,1] et quantifient la volonté ou non de tenir compte de l'état présent. Elle peut être comparée à la matrice des *susceptibilités* à l'influence sociale présente dans les modèles de Friedkin-Johnson [48, 49] ou de DeGroot [42]. Les cœfficients réels $\beta_{E_i,\mathcal{P}_j}(t)$ pondèrent l'importance des matrices $\mathbf{T}^\epsilon_{\mathcal{P}_\mathbf{j}}(t)$. La matrice $\mathbf{T}^\epsilon_{\mathcal{P}_\mathbf{j}}(t)$ est définie à partir du méta-chemin $\mathcal{P}_j$ au temps $t$. À l'instar de la matrice $\mathbf{T}_{\mathcal{P}_\mathbf{j}}$, l'idée est que $\mathbf{T}^\epsilon_{\mathcal{P}_\mathbf{j}}(t)$ encode la *similarité* entre les nœuds du HIN d'après le méta-chemin $\mathcal{P}_j$, au temps $t$. Sa construction[2], et donc l'intérêt d'un paramètre $\epsilon$, sont explicités dans la suite. Enfin, $\mathbf{I}$ est la matrice identité de dimension adéquate.

L'éq. (4.4) signifie que le poids des liens de type $E_i$ au temps $t+1$ dépend du poids des chemins

[1] Les matrices sont évidemment redimensionnées adéquatement. En outre, nous avons vu au chapitre 3 que les modèles finaux n'admettaient jamais d'ordonnée à l'origine, i.e. $\beta_0 = 0$.

[2] Pour un méta-chemin $\mathcal{P}$ de longueur supérieur à 1, $\mathbf{T}_\mathcal{P} \neq \mathbf{T}^\epsilon_\mathcal{P}$

définis à partir des méta-chemins $\{\mathcal{P}_j\}_i$ eu temps $t$ et du poids des liens de type $E_i$ au temps $t$ également. À l'instar de nombreux modèles d'opinions, l'information est intégrée par un mécanisme de moyenne (combinaison convexe pondérée par les cœfficients $\beta_{E_i,\mathcal{P}_j}$) [42, 48, 49].

### 4.2.2. Définition de $\mathbf{T}^{\epsilon}_{\mathcal{P}}(t)$ : marches aléatoires contraintes et limitées

Nous reprenons l'idée de marches aléatoires contraintes par un méta-chemin (section 1.3.1). L'idée est donc que le poids du lien entre deux nœuds $u$ et $v$ évolue en fonction de la probabilité qu'un marcheur a d'atteindre $v$ en partant de $u$.

Rappelons que l'application qui motive la dynamique proposée est un modèle d'évolution d'attention (voire d'opinions) des individus. Dès lors, nous décidons de définir l'évolution des matrices $\mathbf{T}^{\epsilon}_{\mathbf{E_i}}(t)$ à partir du principe de confiance limitée [41, 74]. Cela se traduit par une contrainte supplémentaire sur les marches aléatoires contraintes par un méta-chemin. Dans la suite, ces marches sont appelées *marches aléatoires contraintes et limitées*.

Afin d'introduire cette limitation supplémentaire, nous définissons tout d'abord la distance typée, i.e. une distance entre deux nœuds par rapport à deux types de liens ayant le même type de nœuds cibles.

**Définition 4.1** (Distance typée)**.** Soient deux types de liens $E_1$, $E_2 \in \mathcal{E}$ tels que $\nu_t(E_1) = \nu_t(E_2) = V^* \in \mathcal{V}$. Soient également deux nœuds $u$ et $v$ tels que $u \in \nu_s(E_1)$ et $v \in \nu_s(E_2)$. La $u$-ième ligne de $\mathbf{T}$ se note $\mathbf{T}(u,:)$. La *distance typée* entre $u$ et $v$ selon $E_1$ et $E_2$, notée $d_{E_1,E_2}(u,v)$, est définie par

$$d_{E_1,E_2}(u,v) = \frac{1}{\sqrt{2}}\|\mathbf{T}_{\mathbf{E_2}}(u,:) - \mathbf{T}_{\mathbf{E_2}}(v,:)\|_2. \tag{4.5}$$

**Remarque 4.1** (Normalisation de la distance)**.** Le facteur $\sqrt{2}$ dans l'éq. (4.5) vient du fait que la norme euclidienne $\|\mathbf{T}_{\mathbf{E_2}}(u,:) - \mathbf{T}_{\mathbf{E_2}}(v,:)\|_2$ est comprise entre $[0,\sqrt{2}]$, quelle que soit la dimension des matrices $\mathbf{T}_{\mathbf{E_i}}$. En effet, de manière générale, trouver le maximum de $\|\mathbf{T}_{\mathbf{E_2}}(u,:) - \mathbf{T}_{\mathbf{E_2}}(v,:)\|_2$ revient au problème d'optimisation suivant

$$\begin{aligned} &\max_{\mathbf{x},\mathbf{y}} \|\mathbf{x}-\mathbf{y}\|_2 \\ &\text{sous contrainte : } \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1 \text{ et } \forall i,\, x_i, y_i \in [0,1] \end{aligned}$$

avec $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ ($\mathbf{x},\mathbf{y}$ des lignes d'une matrice stochastique sur les lignes). La norme étant une application convexe, les maxima se situent donc aux extrémités du domaine sur lequel la norme est appliquée. Dans notre cas, les $n$ extrémités du domaine sont imposées par les contraintes et s'expriment comme les vecteurs de la base canonique de $\mathbb{R}^n$, i.e. $e_i = (\delta_{1i}, \delta_{2i}, \cdots, \delta_{ni})$, pour

$i = 1, ..., n$ et où $\delta_{ij}$ désigne le symbole de Kronecker. La norme euclidienne de la différence de n'importe quelle paire de tels vecteurs est égale à $\sqrt{2}$.

**Définition 4.2** (Distances typées induites par un méta-chemin)**.** Soient un HIN $H$ et un méta-chemin[1] $\mathcal{P} : V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{1:2}} V_2 \cdots \xrightarrow{E_{n-1:n}} V_n$ de longueur $n$. Supposons qu'il existe au moins un lien entre chaque paire de nœuds dans le schéma associé au HIN. Les *distances typées induites par* $\mathcal{P}$ sont les suivantes

$$\forall i \in \{0, ..., n-2\},\ d_{E_{i:i+2}, E_{i+1:i+2}}(v_i, v_{i+1}), \quad \forall (v_i, v_{i+1}) \in V_i \times V_{i+1}.$$

Comme dit précédemment, la contrainte supplémentaire des marches aléatoires porte sur les distances typées induites : un marcheur est non seulement contraint par un méta-chemin mais en outre, il faut que la distance typée entre deux nœuds consécutifs de sa marche soit inférieure à un certain seuil

$mathbf{f}\epsilon$, dépendant des types de liens intervenant dans les distances typées et donc de $\mathcal{P}$. Il s'agit d'une *distance typée limitée*, analogue à la confiance limitée. Les résultats d'une telle marche limitée permettent de construire la matrice $\mathbf{T}^{\epsilon}_{\mathcal{P}}$. Nous détaillons le calcul de sa construction dans la suite. Pour un côté plus visuel, la figure 4.1 illustre la démarche.

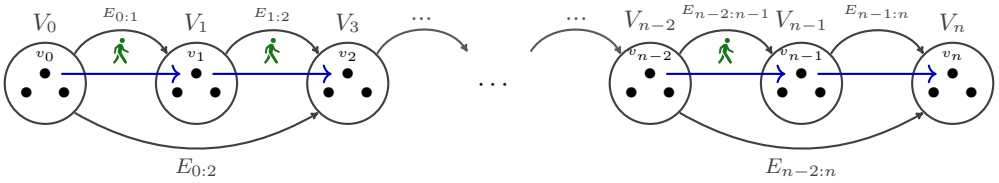
Soient un méta-chemin $\mathcal{P} = V_0 \xrightarrow{E_{0:1}} V_1 ... \xrightarrow{E_{n-1:n}} V_n$ de longueur $n$ et deux nœuds $u \in V_0$ et $v_n \in V_n$. Chaque entrée $\mathbf{T}^{\epsilon}_{\mathcal{P}}(v_0, v_n)$ représente la probabilité d'atteindre $v_n \in V_n$, démarrant de $v_0 \in V_0$, à l'aide d'une marche aléatoire contrainte par $\mathcal{P}$ et limitée par les distances typées induites avec le seuil $\boldsymbol{\epsilon}$. Cette limitation ne porte que sur les $n-1$ premiers pas de la marche (voir définition 4.2 : un méta-chemin de longueur $n$ induit $n-1$ distances typées), le dernier pas est uniquement contraint par le méta-chemin, à savoir $E_{n-1:n}$.

Nous appelons *probabilité contrainte et limitée* la probabilité de passer de $v_0$ à $v_{n-1} \in V_{n-1}$, i.e. un nœud appartenant à l'avant dernier type de nœuds du méta-chemin, contrainte par $\mathcal{P}$ et limites par les seuils $\boldsymbol{\epsilon}$. Ces seuils sont des paramètres et permettent de déterminer si un marcheur peut passer d'un nœud à un autre. Comme déjà mentionné, $\boldsymbol{\epsilon}$ dépend de $\mathcal{P}$, i.e. $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\mathcal{P})$, et est tel que $\boldsymbol{\epsilon} = [\epsilon_{E_{0:2}, E_{1:2}}, ..., \epsilon_{E_{n-2:n}, E_{n-1:n}}] \in \mathbb{R}^{n-1}$. Cette probabilité est notée $\mathbb{P}^{\boldsymbol{\epsilon}}_{\mathcal{P}}(v_{n-1}|v_0)$.

Afin de définir l'expression de $\mathbb{P}^{\boldsymbol{\epsilon}}_{\mathcal{P}}(v_{n-1}|v_0)$, considérons tout d'abord un méta-chemin de longueur 2 : $\mathcal{P} = V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{1:2}} V_2$, deux nœuds $v_0 \in V_0$ et $v_1 \in V_1$ et supposons qu'il existe un[2] type de liens $E_{0:2}$ entre $V_0$ et $V_2$. Dans ce cas, la probabilité contrainte et limitée est

[1] Dans la suite de cette section, les types de liens sont notés $E_{i:j}$ (et non plus indicés par un seul indice comme précédemment) avec $V_i$ (resp. $V_j$) le type de nœuds source (resp. cible). Cela enlève toute ambiguïté dans les types de liens utilisés.

[2] Il se peut qu'il existe plusieurs types de liens entre les nœuds dans $V_0$ et $V_2$. Dans ce cas, $E_{0:2}$ est choisi en fonction du contexte mais quoi qu'il en soit, $E_{0:2}$ existe et est bien défini.




HIN $H = (G, \mathcal{V}, \mathcal{E}, \phi, \psi)$

$\mathcal{P} : V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{1:2}} V_2 \cdots \xrightarrow{E_{n-1:n}} V_n$

$\rightarrow$ : chemin satisfaisant $\mathcal{P}$

🚶 : dépend des distances typées limitées, i.e. $\forall i \in \{0, ..., n-2\}$,

$$d_{E_{i:i+2},E_{i+1:i+2}}(v_i, v_{i+1}) \leq \epsilon_{E_{i:i+2},E_{i+1:i+2}}, \quad \forall (v_i, v_{i+1}) \in V_i \times V_{i+1}$$

**Figure 4.1.:** Explication de la construction de la matrice $\mathbf{T}^{\epsilon}_{\mathcal{P}}$ associée au méta-chemin $\mathcal{P} : V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{1:2}} V_2 \cdots \xrightarrow{E_{n-1:n}} V_n$. Chaque entrée $\mathbf{T}^{\epsilon}_{\mathcal{P}}(u, v)$ représente la probabilité qu'un marcheur, initialement en $u$, atteigne $v$, en étant contraint par $\mathcal{P}$ et limité par les distances typées induites (par $\mathcal{P}$) avec les seuils $\boldsymbol{\epsilon}$ fixés. Les rond noirs représentent les nœuds d'un HIN dont le schéma associé est tracé en gris (types de nœuds et types de liens). Pour plus de clarté, les liens du HIN ne sont pas tracés. Le méta-chemin est la première contrainte de la marche. En bleu, le chemin $P : v_0 \xrightarrow{e_{0:1}} v_1 \cdots \xrightarrow{e_{n-1:n}} v_n$ satisfaisant $\mathcal{P}$. Les icônes des marcheurs représentent la seconde contrainte : le marcheur ne peut emprunter le lien que si la distance typée entre les deux nœuds adjacents au lien en question est inférieure à un seuil fixé $\epsilon$, dépendant des types de liens intervenant dans les distances typées et donc de $\mathcal{P}$. Les deux contraintes interviennent lors des $n-1$ premiers pas de la marche. Pour le dernier, seul le méta-chemin contraint le marcheur ; plus particulièrement, le marcheur doit suivre les liens du type $E_{n-1:n}$.

définie par

$$\mathbb{P}^{\epsilon}_{\mathcal{P}'}(v_1|v_0) = \frac{w_{E_{0:1}}(v_0, v_1)\, \mathbb{1}_{\{d_{E_{0:2},E_{1:2}}(v_0,v_1) \leq \epsilon_{E_{0:2},E_{1:2}}\}}}{\sum_{v' \in V_1} w_{E_{0:1}}(v_0, v')\, \mathbb{1}_{\{d_{E_{0:2},E_{1:2}}(v_0,v') \leq \epsilon_{E_{0:2},E_{1:2}}\}}}, \tag{4.6}$$

avec $\mathbb{1}$ la fonction indicatrice et $\epsilon_{E_{0:2},E_{1:2}} \in \mathbb{R}^+$. Cela signifie que pour qu'un marcheur saute de $v_0$ à $v_1$, la distance typée $d_{E_{0:2},E_{1:2}}(v_0, v_1)$ doit être inférieure à $\epsilon_{E_{0:2},E_{1:2}}$, le seuil associé aux deux types de liens $E_{0:2} et E_{1:2}$ dans la distance typée induite.

Nous pouvons à présent définir la probabilité contrainte par $\mathcal{P}$ et limitée par les distances typées dans le cas général. Il suffit d'appliquer récursivement l'éq. (4.6)

$$\mathbb{P}^{\boldsymbol{\epsilon}}_{\mathcal{P}}(v_{n-1}|v_0) = \sum_{v_{n-2} \in V_{n-2}} \mathbb{P}^{\epsilon}_{\mathcal{P}^{n-2,n}}(v_{n-1}|n-2)\, \mathbb{P}^{\boldsymbol{\epsilon}}_{\mathcal{P}^{0,n-2}}(n-2|v_0), \tag{4.7}$$

avec $\mathbb{P}^{\epsilon}_{\mathcal{P}^{0,2}}(v_2|v_0)$ la base de la récurrence. L'éq. (4.7) est bien définie puisque $\mathcal{P}^{n-2,n}$ est effectivement un méta-chemin de longueur 2. En outre, l'éq. (4.7) se ramène à (4.6) lorsque le méta-chemin est de longueur 2.

Finalement, nous pouvons définir la matrice $\mathbf{T}^{\epsilon}_{\mathcal{P}}$. Il suffit simplement d'ajouter le dernier pas de

la marche, contraint par le type de lien $E_{n-1:n}$, à l'éq.(4.7) comme suit

$$\mathbf{T}^{\epsilon}_{\mathcal{P}}(v_0, v_n) = \sum_{v_{n-1} \in V_{n-1}} \mathbf{T}_{\mathbf{E_{n-1:n}}}(v_{n-1}, v_n)\, \mathbb{P}^{\boldsymbol{\epsilon}}_{\mathcal{P}}(v_{n-1}|v_0). \tag{4.8}$$

L'éq. (4.1) est donc un cas particulier de l'éq. (4.8) lorsque $n = 1$. Aussi, remarquons que lorsque $\epsilon_{E_{i:i+2},E_{i+1:i+2}} \geq 1$, $\forall i \in \{0, ..., n-2\}$, l'éq. (4.8) se ramène à un simple produit de matrices, et donc à une simple marche aléatoire contrainte par un méta-chemin (i.e. sans limitation par une distance typée).

**Exemple 4.1** (Marche aléatoire contrainte par un méta-chemin et limitées par des distances)**.** Considérons un HIN composé de trois types de nœuds $\mathcal{V} = \{V_0, V_1, V_2\}$ et trois types de liens $\mathcal{E} = \{E_{0:1}, E_{0:2}, E_{1:2}\}$ (en gris, figure 4.2) et le méta-chemin $\mathcal{P} : V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{1:2}} V_2$. La limitation des distances ne concerne que le premier pas de la marche (puisque le méta-chemin est de longueur 2). Nous expliquons donc uniquement le passage $V_0 \xrightarrow{E_{0:1}} V_1$. Supposons qu'un marcheur initialement sur un nœud de $V_0$ veuille aller sur $v_1^1 \in V_1$. L'idée est d'autoriser le marcheur à passer du nœud $v_0$ sur lequel il se trouve à $v_1^1$ si et seulement si les liens de $v_0$ vers les nœuds de $V_2$ (le type de nœuds suivant) sont *suffisamment proches* de ceux que $v_1^1$ a avec les nœuds de $V_2$. Dans l'exemple, si le marcheur est initialement sur $v_0^1$, il ne pourra sauter sur $v_1^1$ que si $\epsilon_{E_{0:2},E_{1:2}} > \sqrt{3/2}$ puisque le vecteur des liens de $v_0^1$ vers $V_2$ vaut [1,0,0] et que celui de $v_1^1$ vaut [0,1,1]. En revanche, s'il est positionné sur $v_0^2$, le saut vers $v_1^1$ est toujours permis puisque dans ce cas, le vecteur des liens de $v_0^2$ vers $V_2$ vaut [0,1,1].

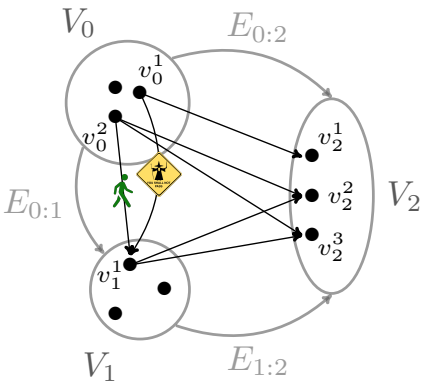



**Figure 4.2:** Illustration d'une marche contrainte par le méta-chemin $\mathcal{P} : V_0 \xrightarrow{E_{0:1}} V_1 \xrightarrow{E_{12}} V_2$ et limitée par des seuils $\epsilon$. $d_{E_{0:2},E_{1:2}}(v_0^1, v_1^1) = \|[1,0,0]-[0,1/2,1/2]\|_2/\sqrt{2} = \sqrt{3/2}$ tandis que $d_{E_{0:2},E_{1:2}}(v_0^2, v_1^1) = \|[0,1/2,1/2] - [0,1/2,1/2]\|_2/\sqrt{2} = 0$. Un marcheur initialement en $v_0^2$ peut se déplacer sans problème sur $v_1^1$. En revanche, pour $\epsilon_{E_{0:2},E_{1:2}} < \sqrt{3/2}$, un marcheur ne peut se déplacer de $v_0^1$ vers $v_1^1$, les vecteurs concernés étant *trop éloignés*. Un fois le marcheur sur $v_1^1$, seul le dernier type de lien de $\mathcal{P}$ contraint la marche : le marcheur peut aussi bien aller sur $v_2^2$ que sur $v_2^3$.

*Exemple concret : modéliser une dynamique d'attention.* Imaginons que $V_0$ et $V_1$ soient des groupes d'individus (e.g. jeunes et vieux, pour avoir une raison de les distinguer) et $V_2$ soit un ensemble de sujets auxquels ils peuvent potentiellement s'intéresser. Les liens entre $V_0$ et $V_1$ sont donc des liens sociaux tandis que ceux entre $V_0$ et $V_2$ et ceux entre $V_1$ et $V_2$ peuvent s'interpréter, e.g. comme l'attention que les individus accordent aux sujets de $V_2$. Dans les modèles classiques de

confiance limitée, les individus de $V_0$ et $V_1$ s'influencent (sur leurs attentions) ssi leurs attentions sont *suffisamment proches*. Dans notre cas, cela revient à imposer que la distance $d_{E_{0:2},E_{1:2}}(v_0, v_1)$ soit inférieure à un seuil $\epsilon_{E_{0:2},E_{1:2}}$. Sur la figure 4.2, $v_0^1$ et $v_1^1$ ne peuvent s'influencer si $\epsilon_{E_{0:2},E_{1:2}} < \sqrt{3/2}$ tandis que $v_0^2$ et $v_1^1$ le peuvent sans limitation.

## 4.3. Modèle d'attention d'individus

L'application proposée dans ce chapitre est un modèle d'attention, terme explicité précédemment. Nous considérons deux types de nœuds $I$ et $S$, qui sont, par extension, des ensembles de nœuds. Le premier est un ensemble de $n_I$ individus tandis que le second est un ensemble de $n_S$ sujets. Nous supposons deux types de liens qui sont aussi, par extension, des ensembles de liens. En particulier, les liens de type **II** relient les individus entre eux et représentent un certain type de relations sociales. Ensuite, les liens de type **IS** lient les individus aux sujets. Ces liens peuvent s'interpréter comme l'attention qu'un individu accorde aux sujets. Ces deux ensembles sont représentés par deux matrices, notées **II** et **IS**, de dimensions $n_I \times n_I$ et $n_I \times n_S$ respectivement. Ces deux matrices sont stochastiques sur les lignes : chaque entrée d'une ligne s'interprète comme la proportion d'attention qu'un individu porte soit à un autre individu (**II**), soit à un sujet (**IS**).

Ces deux matrices sont respectivement mises à jour en fonction des méta-chemins $I \to S \to I$ et $I \to I \to S$ et des seuils $\epsilon_{IS}$ et $\epsilon_{II}$ (qui donnent les matrices[1] **ISI** et **IIS**). Le premier méta-chemin fait référence au fait que les relations sociales peuvent être influencées par l'attention commune qu'ont les individus envers les sujets[2], tandis que le second fait référence au fait que les individus peuvent modifier leur attention en fonction de leurs relations sociales. En formule, nous avons

$$\begin{cases} \mathbf{IS}(t+1) = \mathbf{\Lambda_{IS}}\mathbf{IIS}(t) + (\mathbf{I} - \mathbf{\Lambda_{IS}})\mathbf{IS}(t) \\ \mathbf{II}(t+1) = \mathbf{\Lambda_{II}}\mathbf{ISI}(t) + (\mathbf{I} - \mathbf{\Lambda_{II}})\mathbf{II}(t) \end{cases} \tag{4.9}$$

Pour rappel, les matrices diagonales $\mathbf{\Lambda_{IS}}$ et $\mathbf{\Lambda_{IS}}$ représentent la volonté ou non de tenir compte de l'état présent. Nous considérons le cas particulier de deux sujets, $S = \{s_1, s_2\}$ et un ensemble de $n_I$ individus. Ces $n_I$ individus sont divisés en quatre groupes $g_i$, $i = 1, 2, 3, 4$ de même taille $n_I/4$.

[1] Dans le reste de ce chapitre et afin d'alléger les notations, les matrices **ISI** et **IIS** sont les matrices $\mathbf{T}^{\epsilon}_{I\to S\to I}$ et $\mathbf{T}^{\epsilon}_{I\to I\to S}$ respectivement, définies à la section précédente.

[2] Le type de liens $S \to I$ est l'inverse (déf. 1.5) de $I \to S$ et représente, pour un sujet donné, la répartition de la façon dont il intéresse les individus. Si $I \to S$ est représenté par la matrice stochastique sur les lignes **IS**, $S \to I$ est représenté par la matrice **IS** transposée et renormalisée sur les lignes, notée par abus de notation, $\mathbf{IS}^{\mathsf{T}}$.

Les relations sociales **II** sont toujours initialisées de la façon suivante

$$\mathbf{II}(0) = \begin{matrix} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{matrix} \begin{matrix} \begin{matrix} g_1 & g_2 & g_3 & g_4 \end{matrix} \\ \begin{bmatrix} \mathbf{2/n_I} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{0} \\ \mathbf{0} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{2/n_I} \\ \mathbf{2/n_I} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{0} \\ \mathbf{0} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{2/n_I} \end{bmatrix} \end{matrix} \tag{4.10}$$

où $\mathbf{2/n_I}$ (resp. $\mathbf{0}$) est la matrice de dimensions $n_I/4 \times n_I/4$ dont toutes les entrées valent $2/n_I$ (resp. 0). Cela signifie que les groupes $g_1$ et $g_3$ forment une clique dont les poids des liens sont tous égaux. Il en va de même pour $g_2$ et $g_4$. Dans la suite, la notation $g_{ij}$ est utilisée pour dénoter l'union des groupes $g_i$ et $g_j$.

Dans cette application, les paramètres $\epsilon_{IS}$ et $\epsilon_{II}$ ont une signification plus concrète : ils s'interprètent comme *l'ouverture d'esprit* ou la *tolérance* des individus par rapport à chacune des relations.

Dans la suite, nous nous intéressons à l'évolution et à l'état final[1] de **IS** et **II** suivant (4.9) pour différentes initialisations de **IS**.

### 4.3.1. Deux distributions d'attention initiales

Dans ce premier exemple, les individus portent l'entièreté de leur attention sur un seul sujet, en fonction du groupe auquel ils appartiennent. Plus particulièrement, les individus des groupes $g_1$ et $g_2$ ne s'intéressent qu'au sujet $s_1$, tandis que l'inverse se produit pour les groupes $g_3$ et $g_4$. Cela signifie que la population d'individus est scindée en deux groupes de deux façons différentes : en fonction des relations sociales, $g_{13}$ et $g_{24}$, ou en fonction de l'attention, $g_{12}$ et $g_{34}$. En matrice,

$$\mathbf{IS}(0) = \begin{matrix} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{matrix} \begin{matrix} \begin{matrix} s_1 & s_2 \end{matrix} \\ \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \end{matrix} \tag{4.11}$$

où $\mathbf{1}$ (resp. $\mathbf{0}$) est le vecteur de dimensions $n_I/4 \times 1$ dont les entrées sont 1 (resp. 0).

#### Raisonnement pour la solution triviale

Afin de caractériser l'état d'équilibre, nous détaillons le raisonnement suivi, ce qui facilitera l'analyse des sections suivantes. Pour ce faire, nous nous intéressons aux marches aléatoires contraintes et

[1] Nous *supposons* que le système converge vers un état d'équilibre. La preuve d'une éventuelle convergence dépasse le cadre de ce chapitre.

limitées.

**Première étape** Au niveau du méta-chemin $IIS$, un marcheur ne peut passer de $u$ à $v$ que si $d_{IS,IS}(u,v) \leq \epsilon_{IS}$. Partant de la configuration initiale, un marcheur démarrant d'un nœud de $g_1$ peut aller soit dans $g_1$, soit dans $g_3$. Or, tous les individus de $g_3$ sont tels que leur vecteur $\mathbf{IS} = [0,1]$, alors que ceux de $g_1$ ont $\mathbf{IS} = [1,0]$. Dès lors, à moins de prendre $\epsilon_{IS} = 1$, le marcheur sera contraint de rester dans $g_1$, i.e. le groupe duquel il a démarré. Le même raisonnement s'applique pour les autres groupes. La matrice $\mathbf{IIS}(1)$ sera toujours égale à $\mathbf{IS}(1)$ avec un tel $\mathbf{II}(0)$.

Au niveau du méta-chemin $ISI$, un marcheur ne peut passer de $u$ à $s$ que si $d_{II,SI}(u,s) \leq \epsilon_{II}$. Notons que $\mathbf{IS}^\mathsf{T}$ est stochastique sur les colonnes et non sur les lignes. Dès lors, en normalisant sur les lignes, nous obtenons la structure en blocs suivante (avec les dimensions adéquates)

$$\mathbf{IS}^\mathsf{T}(0) = \begin{bmatrix} \mathbf{2/n_I} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{2/n_I} \end{bmatrix}. \tag{4.12}$$

Dans cette configuration initiale, $d_{II,SI}(u,s) = \frac{1}{\sqrt{2}}[\frac{n_I}{4}(\frac{2}{n_I})^2 \times 2]^2 = \sqrt{1/n_I} < 1$, $\forall u, s$. Dès lors, pour $\epsilon_{II} \geq \epsilon_{II}^m := \sqrt{1/n_I}$, un marcheur peut passer d'un nœud de e.g. $g_1$ vers $s_1$. Une fois sur ce nœud n'ayant plus de contrainte autre que le type de liens, il peut sauter soit sur un nœud de $g_1$, soit de $g_2$. Cela implique qu'au delà de $\epsilon_{II}^m$,

$$\mathbf{II}(1) = \begin{bmatrix} \mathbf{2/n_I} & \mathbf{\Lambda_{II}2/n_I} & \mathbf{(I-\Lambda_{II})2/n_I} & \mathbf{0} \\ \mathbf{\Lambda_{II}2/n_I} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{(I-\Lambda_{II})2/n_I} \\ \mathbf{(I-\Lambda_{II})2/n_I} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{\Lambda_{II}2/n_I} \\ \mathbf{0} & \mathbf{(I-\Lambda_{II})2/n_I} & \mathbf{\Lambda_{II}2/n_I} & \mathbf{2/n_I} \end{bmatrix} \tag{4.13}$$

où $\mathbf{I}$ et $\mathbf{\Lambda_{II}}$ sont respectivement les matrices identité et $\Lambda_{II}$ de dimension $n_I/4 \times n_I/4$.

**Étapes suivantes** En continuant un raisonnement similaire, nous remarquons que les différentes itérations de $\mathbf{IS}(t)$ ne vont pas changer. En effet, comme nous l'avons vu au moyen de la première itération, les poids des liens de $\mathbf{II}(t)$ vont s'amenuiser entre les couples $(g_1, g_3)$ et $(g_2, g_4)$ et se renforcer entre les couples $(g_1, g_2)$ et $(g_3, g_4)$. Cependant, les individus dans ces couples ont exactement le même vecteur $\mathbf{IS}(0)$, i.e. [1,0] et [0,1] respectivement. Dès lors, bien que le marcheur sera capable de se déplacer dans le graphe pour $t > 1$, cela ne va provoquer aucune modification dans la structure (pondérée) de $\mathbf{IS}(t)$. En revanche, les poids des liens de $\mathbf{II}(t)$ vont continuer à évoluer tel que pour tout $t > 0$ (généralisation de l'éq. (4.13)),

$$\mathbf{II}(t) = \begin{bmatrix} \frac{\mathbf{2}}{\mathbf{n_I}} & [\mathbf{I}-(\mathbf{I}-\mathbf{\Lambda_{II}})^t]\frac{\mathbf{2}}{\mathbf{n_I}} & (\mathbf{I}-\mathbf{\Lambda_{II}})^t\frac{\mathbf{2}}{\mathbf{n_I}} & \mathbf{0} \\ [\mathbf{I}-(\mathbf{I}-\mathbf{\Lambda_{II}})^t]\frac{\mathbf{2}}{\mathbf{n_I}} & \frac{\mathbf{2}}{\mathbf{n_I}} & \mathbf{0} & (\mathbf{I}-\mathbf{\Lambda_{II}})^t\frac{\mathbf{2}}{\mathbf{n_I}} \\ (\mathbf{I}-\mathbf{\Lambda_{II}})^t\frac{\mathbf{2}}{\mathbf{n_I}} & \mathbf{0} & \frac{\mathbf{2}}{\mathbf{n_I}} & [\mathbf{I}-(\mathbf{I}-\mathbf{\Lambda_{II}})^t]\frac{\mathbf{2}}{\mathbf{n_I}} \\ \mathbf{0} & (\mathbf{I}-\mathbf{\Lambda_{II}})^t\frac{\mathbf{2}}{\mathbf{n_I}} & [\mathbf{I}-(\mathbf{I}-\mathbf{\Lambda_{II}})^t]\frac{\mathbf{2}}{\mathbf{n_I}} & \frac{\mathbf{2}}{\mathbf{n_I}} \end{bmatrix}, \tag{4.14}$$

où $\mathbf{A}^t$ représente la matrice $\mathbf{A}$ puissance $t$. Les entrées (non nulles) de $(\mathbf{I} - \mathbf{\Lambda_{II}})$ étant inférieures à 1, $(\mathbf{I} - \mathbf{\Lambda_{II}})^t \to \mathbf{0}$, lorsque $t \to \infty$. Nous avons donc bien convergence, et l'état d'équilibre[1], est

$$\mathbf{IS}(\infty) = \mathbf{IS}(0) \qquad \mathbf{II}(\infty) = \begin{bmatrix} \mathbf{2/n_I} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{0} \\ \mathbf{2/n_I} & \mathbf{2/n_I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{2/n_I} \\ \mathbf{0} & \mathbf{0} & \mathbf{2/n_I} & \mathbf{2/n_I} \end{bmatrix}. \tag{4.15}$$

Cela n'est pas surprenant puisque, comme la matrice **IS** ne change pas, **ISI** ne change pas non plus. Or, $\mathbf{II}(t{+}1)$ dépend de $\mathbf{ISI}(t) \equiv \mathbf{ISI}$ et va donc converger vers cette matrice : $\mathbf{II}(\infty) = \mathbf{IS}(0) \times \mathbf{IS}(0)^\mathsf{T}$. Les matrices $\mathbf{\Lambda_{IS}}$ et $\mathbf{\Lambda_{II}}$ vont donc juste impacter le temps de convergence (dans cet exemple, uniquement $\mathbf{\Lambda_{IS}}$).

#### Seuil dépendant du temps

Par simplicité, nous avons considéré que $\epsilon_{II} \geq \epsilon_{II}^m$, seuil à partir duquel un saut du marcheur est possible, est indépendant du temps. Pour le méta-chemin **ISI**, ce seuil minimal décroît avec le temps puisque $\mathbf{II}(t) \to \mathbf{ISI}$. En effet, dans le cas homogène, i.e. $\Lambda_{II}(u) = \lambda$, $\forall u$ et en utilisant les expressions (4.12) et (4.14), on remarque que la distance minimale $d_{min}(t) = \min_{u,s} d^t_{II,SI}(u,s)$ et maximale $d_{max}(t) = \max_{u,s} d^t_{II,SI}(u,s)$ à l'itération $t$ s'expriment comme

$$\begin{aligned} d_{min}(t) &= \frac{1}{\sqrt{2}} \left[ \frac{n_I}{4} \left( \left[1 - (1-\lambda)^t\right] \frac{2}{n_I} - \frac{2}{n_I} \right)^2 + \frac{n_I}{4} \left( (1-\lambda)^t \frac{2}{n_I} \right)^2 \right]^{1/2} \\ &= \frac{1}{\sqrt{n_I}} (1-\lambda)^t \\ d_{max}(t) &= \frac{1}{\sqrt{2}} \left[ \frac{n_I}{4} \left( \frac{2}{n_I} \right)^2 + \frac{n_I}{4} \left( [1 - (1-\lambda)^t] \frac{2}{n_I} \right)^2 + \frac{n_I}{4} \left( [1-\lambda]^t \frac{2}{n_I} - \frac{2}{n_I} \right)^2 + \frac{n_I}{4} \left( \frac{2}{n_I} \right)^2 \right]^2 \\ &= \frac{1}{\sqrt{n_I}} \left[ 1 + (1 - (1-\lambda)^t)^2 \right]^{1/2} \end{aligned}$$

En réalité, la chose intéressante est de voir qu'il suffit en fait prendre $\epsilon_{II}(t) \geq d_{min}(t)$ (avec $d_{min}(t) = \epsilon_{II}^m$ en $t = 0$) pour pouvoir converger vers la solution (4.15).

Finalement remarquons que prendre $\epsilon_{II}(t) \geq d_{max}(t)$ n'a pas d'intérêt en pratique. En effet, soit un individu $u$ de $g_1$ : son attention ne porte que sur $s_1$. Avec $\epsilon_{II}(t) \geq d_{max}(t)$, cela permet, en théorie, au marcheur de se déplacer aussi bien sur $s_1$ que sur $s_2$. Mais en pratique, le marcheur ne peut aller que sur $s_1$, puisque le poids du lien $IS(u, s_2)$ est nul.

#### Bilan

Nous avons démarré d'un ensemble d'individus scindés en deux groupes de deux façons différentes. La première concerne l'attention qu'ils portent à deux sujets, i.e. les groupes $g_{12}$ et $g_{34}$. La seconde

[1] Pour $\mathbf{\Lambda_{II}} = diag(\lambda, ..., \lambda)$, $\lambda \in ]0, 1]$. Si $\lambda = 0$, aucune dynamique et $\mathbf{II}(\infty) = \mathbf{II}(0)$.

est fonction des relations sociales et scinde les individus en $g_{13}$ et $g_{24}$. L'exemple envisagé va provoquer un changement uniquement dans les relations sociales, i.e. les matrices $\mathbf{II}(t)$, pour qu'à l'état d'équilibre, nous obtenions deux groupes distincts $g_{12}$ et $g_{34}$. Ces deux groupes sont des cliques et ne communiquent plus entre eux : il n'y a pas de lien $(u, v) \in g_{12} \times g_{34}$, et concentrent chacun l'entièreté de leur attention sur un seul sujet (différent en fonction du groupe). Cela est donné par la matrice $\mathbf{IS}(0)$, qui reste donc inchangée. Les relations **II** évoluant au cours du temps, le seuil $\epsilon_{II}$ à partir duquel des interactions sont possibles évolue également et décroît.

Plus généralement, l'observation de cet exemple est la suivante : tant que les individus concentrent initialement leur attention sur un seul sujet, i.e. tant que les distributions d'attention initiales sont telles que [0,1] ou [1,0], les relations d'attention **IS** ne vont pas évoluer (à moins de prendre $\epsilon = 1$, ce qui revient à ne pas contraindre par une distance typée les marches aléatoires). Pour les relations sociales **II**, il y aura évolution pour toute initialisation $\mathbf{II}(0)$ telle que $\min_{u,v} d_{IS,IS}(u, v) < 1$ et pour $\epsilon_{II}(t) \geq d_{min}(t), \forall t$. L'état d'équilibre $\mathbf{II}(\infty)$ sera de la forme suivante (à des permutations près)

$$\mathbf{II}(\infty) = \begin{bmatrix} \mathbf{n_1}/\mathbf{n_I} & \mathbf{0} \\ \mathbf{0} & (\mathbf{n_I} - \mathbf{n_1})/\mathbf{n_I} \end{bmatrix},$$

avec $n_1$ le nombre d'individus dont l'attention est [1,0]. Cela signifie que les relations d'attention vont forger les relations sociales : toutes les personnes ayant la même distribution d'attention seront connectées entre elles et n'auront aucun lien avec des personnes ayant une autre distribution d'attention.

### 4.3.2. Trois distributions d'attention initiales

Dans la section précédente, les attentions des individus ne peuvent évoluer, à moins de considérer $\epsilon_{IS} = 1$. Afin de permettre aux individus de modifier leurs attentions, il faut donc initialiser différemment $\mathbf{IS}(0)$. En particulier, il faut qu'il existe au moins un couple $(u, v)$ tel que la distance $d^0_{IS,IS}(u, v)$ soit inférieure à 1.

Nous proposons donc que les individus puissent concentrer leur attention soit sur un seul sujet, soit sur les deux sujets et ce, de façon équivalente. En d'autres termes, les vecteurs initiaux de $\mathbf{IS}(0)$ sont donnés par [1,0], [0,1] ou [1/2,1/2]. Dans le dernier cas, les individus sont dits *éclectiques*.

Le choix de [1/2,1/2] est motivé par le fait qu'au départ, un individu étant informé de l'existence de deux sujets n'a aucun apriori et s'y intéresse avec la même intensité, quel que soit son voisinage. Une autre alternative intuitive aurait été de pondérer en fonction de l'attention du voisinage.

L'objectif est de caractériser l'état final du système en s'intéressant aux nombres de clusters d'individus présents dans $\mathbf{IS}(\infty)$ et $\mathbf{II}(\infty)$ en fonction des paramètres $\epsilon_{IS}$ et $\epsilon_{II}$ constants. Nous

discutons le cas général avec un nombre quelconque d'individus éclectiques et ensuite considérons le cas particulier d'un seul individu éclectique afin d'illustrer plus en détails les propos discutés. Dans le reste du chapitre, $\mathbf{\Lambda_{IS}} = \mathbf{\Lambda_{II}} = \mathrm{diag}(0.5, ..., 0.5)$.

#### Un nombre quelconque d'individus éclectiques : aperçu des états finaux

Nous considérons le vecteur $\mathbf{x} = [x_1, x_2, x_3, x_4]$ où $x_i \in [0, n_I/4]$ représente le nombre d'individus éclectiques dans le groupe $g_i$, $i = 1, ..., 4$. Cela signifie que la distribution d'attention initiale de ces $x_i$ individus est homogène et égale [1/2,1/2].

En suivant le même raisonnement que précédemment, il faut que $\epsilon_{IS} \geq 1/2 =: \epsilon_{IS}^m$ pour initier une modification de $IS$. Cette valeur $\epsilon_{IS}^m$ peut s'interpréter comme *l'ouverture d'esprit minimale* requise pour permettre des interactions **IS**. En revanche, pour les relations sociales **II**, ce seuil dépend des paramètres $x_i$.

Afin de gagner en compréhension, nous introduisons deux quantités qui caractérisent, en partie, le système

$$q = -x_1 - x_2 + x_3 + x_4, \tag{4.16}$$

$$r = x_1 - x_2 - x_3 + x_4. \tag{4.17}$$

Ces quantités peuvent être vues comme des mesures macroscopiques de symétrie : $q$ (resp. $r$) mesure la symétrie des modifications effectuées entre les groupes $g_{12}$ et $g_{34}$ (resp. $g_{14}$ et $g_{23}$). Cela permet, entre autres, d'exprimer $\mathbf{IS}^\mathsf{T}(0)$ comme

$$\mathbf{IS}^\mathsf{T}(0) = \begin{array}{c} \begin{matrix} \scriptstyle \frac{n_I}{4}-x_1 & \scriptstyle x_1 & \scriptstyle \frac{n_I}{4}-x_2 & \scriptstyle x_2 & \scriptstyle \frac{n_I}{4}-x_3 & \scriptstyle x_3 & \scriptstyle \frac{n_I}{4}-x_4 & \scriptstyle x_4 \end{matrix} \\ \left[\begin{matrix} \frac{1}{n_I+q} & \frac{2}{n_I+q} & \frac{1}{n_I+q} & \frac{2}{n_I+q} & \frac{1}{n_I+q} & 0 & \frac{1}{n_I+q} & 0 \\ \frac{1}{n_I-q} & 0 & \frac{1}{n_I-q} & 0 & \frac{1}{n_I-q} & \frac{2}{n_I-q} & \frac{1}{n_I-q} & \frac{2}{n_I-q} \end{matrix}\right] \begin{matrix} \scriptstyle 1 \\ \scriptstyle 1 \end{matrix} \\ \begin{matrix} \underbrace{\qquad\qquad}_{g_1} & \underbrace{\qquad\qquad}_{g_2} & \underbrace{\qquad\qquad}_{g_3} & \underbrace{\qquad\qquad}_{g_4} \end{matrix} \end{array} \tag{4.18}$$

où les annotations en gris indiquent les dimensions.

**Définition des régions de convergence** Par le même raisonnement que précédemment, et en utilisant les expressions (4.11) et (4.18), les quatre distances possibles[1] $d_{II,SI}$ en $t = 0$ s'expriment comme

$$d_{II,SI}^{(1)} = \frac{1}{\sqrt{2}} \left[ \frac{(x_1+x_3)(n_I+2q)^2 + (x_2+x_4)n_I^2 + (n_I-4x_1)q^2 + (n_I-4x_2)n_I^2 + (n_I-4x_3)(n_I+q)^2}{n_I^2(n_I+q)^2} \right]^{1/2}$$

$$d_{II,SI}^{(2)} = \frac{1}{\sqrt{2}} \left[ \frac{(x_1+x_3)n_I^2 + (x_2+x_4)(n_I+2q)^2 + (n_I-4x_1)n_I^2 + (n_I-4x_2)q^2 + (n_I-4x_4)(n_I+q)^2}{n_I^2(n_I+q)^2} \right]^{1/2}$$

$$d_{II,SI}^{(3)} = \frac{1}{\sqrt{2}} \left[ \frac{(x_1+x_3)(n_I-2q)^2 + (x_2+x_4)n_I^2 + (n_I-4x_1)(n_I-q)^2 + (n_I-4x_3)q^2 + (n_I-4x_4)n_I^2}{n_I^2(n_I-q)^2} \right]^{1/2}$$

[1] Il existe deux lignes uniques dans $\mathbf{II}(0)$, ce qui implique $2 \times 2$ distances.

$$d^{(4)}_{II,SI} = \frac{1}{\sqrt{2}} \left[ \frac{(x_1+x_3)n_I^2 + (x_2+x_4)(n_I-2q)^2 + (n_I-4x_2)(n_I-q)^2 + (n_I-4x_3)n_I^2 + (n_I-4x_4)q^2}{n_I^2(n_I-q)^2} \right]^{1/2}. \quad (4.19)$$

Pour $n_I$ fixé, les distances minimales $\epsilon^m_{II} = \min d^{(k)}_{II,SI}$ et maximales $\epsilon^M_{II} = \max d^{(k)}_{II,SI}$ varient en fonction de $\mathbf{x}$. Néanmoins, les seuils minimaux $\epsilon^m_{IS}$ et $\epsilon^m_{II}$ définissent quatre régions de convergence A, B, C et D, différenciées par les teintes de gris sur la figure 4.3 :

$$\begin{aligned}
\text{A}: &\quad 0 < \epsilon_{IS} < \epsilon^m_{IS} \text{ et } 0 < \epsilon_{II} < \epsilon^m_{II} \\
\text{B}: &\quad 0 < \epsilon_{IS} < \epsilon^m_{IS} \text{ et } \epsilon^m_{II} \leq \epsilon_{II} \\
\text{C}: &\quad \epsilon^m_{IS} \leq \epsilon_{IS} \text{ et } 0 < \epsilon_{II} < \epsilon^m_{II} \\
\text{D}: &\quad \epsilon^m_{IS} \leq \epsilon_{IS} \text{ et } \epsilon^m_{II} \leq \epsilon_{II}.
\end{aligned}$$

À l'intérieur des régions B, C et D, nous distinguons, en fonction des paramètres du système, des sous-régions SB, SC et SD (surfaces à motifs). Ces sous-régions existent car les seuils minimaux ne sont pas assez élevés pour déclencher des suites de modifications aussi bien dans **II** que dans **IS**. Notons que les sous-régions SC et SD dépendent des deux matrices **IS** et **II**, contrairement à SB qui n'est fonction que de **II**.

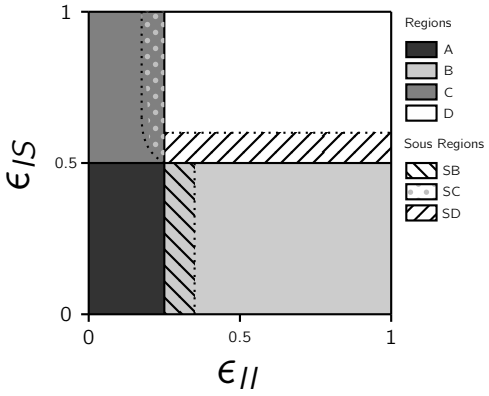



**Figure 4.3:** Schéma différentes régions de convergence caractérisées en fonction de $\mathbf{x}$, $i = 1, ..., 4$, dans le plan $(\epsilon_{II}, \epsilon_{IS})$. Le vecteur $\mathbf{x}$ détermine **IS**(0) qui ne peut contenir, au plus, que trois distributions d'attention différentes : [1,0], [0,1] et [1/2,1/2]. **II**(0) est toujours initialisée par l'éq. (4.10), i.e. individus divisés en deux groupes (cliques) $g_{13}$ et $g_{24}$.

**Définition des sous-régions de convergence** Nous n'avons pu effectuer une analyse exhaustive de tous les cas possibles. Néanmoins, afin d'avoir un aperçu un *peu* plus précis, les quantités $q$ et $r$ (éqs (4.16) et (4.17)) permettent de définir trois familles de cas : (FS) $q = r = 0$ : *fortement* symétrique ; (fS) $q \neq r$ avec $q \times r = 0$ : *faiblement* symétrique et ; (NS) $q \times r \neq 0$ : *non* symétrique. Pour chacun de ces cas, les distances $d^{(k)}_{II,SI}$ adoptent des valeurs différentes qui vont impacter les sous-régions SB, SC et SD.

Nous nous intéressons tout d'abord à la sous-région SB. Pour le cas FS, les quatre distances

$d^{(k)}_{II,SI}$ sont identiques. Cela signifie que la sous-région SB n'existe pas : une transition directe se fait de la région A à B. Dans le cas fS, deux valeurs différentes existent

$$\begin{aligned}
q = 0,\, r < 0 \quad &\Rightarrow \quad d^{(1)}_{II,SI} = d^{(4)}_{II,SI} < d^{(2)}_{II,SI} = d^{(3)}_{II,SI} \\
q = 0,\, r > 0 \quad &\Rightarrow \quad d^{(2)}_{II,SI} = d^{(3)}_{II,SI} < d^{(1)}_{II,SI} = d^{(4)}_{II,SI} \\
r = 0,\, q < 0 \quad &\Rightarrow \quad d^{(3)}_{II,SI} = d^{(4)}_{II,SI} < d^{(1)}_{II,SI} = d^{(2)}_{II,SI} \\
r = 0,\, q > 0 \quad &\Rightarrow \quad d^{(1)}_{II,SI} = d^{(2)}_{II,SI} < d^{(3)}_{II,SI} = d^{(4)}_{II,SI}.
\end{aligned}$$

Cela signifie que SB existe, sans subdivision. Enfin, dans le cas NS, deux cas sont à différencier. Lorsque $n_I/4 > 4$, il existe des vecteurs particuliers $\mathbf{x}$ pour lesquels seules trois valeurs uniques existent pour les distances, divisant SB en deux zones. Plus précisément, quatre situations existent

- $d^{(1)}_{II,SI} < d^{(2)}_{II,SI} = d^{(4)}_{II,SI} < d^{(3)}_{II,SI}$
- $d^{(2)}_{II,SI} < d^{(1)}_{II,SI} = d^{(3)}_{II,SI} < d^{(4)}_{II,SI}$
- $d^{(3)}_{II,SI} < d^{(2)}_{II,SI} = d^{(4)}_{II,SI} < d^{(1)}_{II,SI}$
- $d^{(4)}_{II,SI} < d^{(1)}_{II,SI} = d^{(3)}_{II,SI} < d^{(2)}_{II,SI}$.

Finalement, pour le reste des vecteurs $\mathbf{x}$ possibles, toutes les distances sont différentes, définissant ainsi trois zones dans SB, comme montré numériquement dans la section suivante avec le cas particulier $\mathbf{x} = [1, 0, 0, 0]$.

Comme déjà mentionné et contrairement à SB, les deux autres sous-régions SC et SD ne dépendent pas uniquement des distances $d^{(k)}_{II,SI}$ mais également de l'interaction générale des relations **IS** et **II**. En conséquence, il est beaucoup plus ardu de déterminer les limites exactes de ces sous-régions : nous ne les analysons donc pas analytiquement dans ce travail.

**Caractérisation des régions en fonction de l'état d'équilibre** Dans la région A, aucune dynamique n'est possible, les états finaux sont identiques aux initiaux. Les individus restent campés sur leurs positions initiales, aussi bien au niveau de leur attention **IS** que de leurs relations sociales **II**, et rien ni personne ne peut les faire changer de comportement.

En ce qui concerne la région B\SB, les relations sociales **II** vont se calquer sur les relations d'attention initiales **IS**(0), ces dernières n'évoluant pas. Le graphe associé à **II** sera connexe, mais des liens plus forts uniront les personnes ayant la même distribution d'attention. Les individus recherchent des relations sociales en accord avec leurs centres d'intérêt.

La région C\SC est très riche en comportements. Les relations sociales **IS** vont évoluer en

fonction de $\mathbf{x}$, influencées par les relations sociales initiales $\mathbf{II}(0)$, ces dernières n'évoluant pas. Nous insistons sur le fait que SC est très variable en fonction de $\epsilon_{II}, \epsilon_{II}$ et $\mathbf{x}$, l'analyse qui suit n'est qu'un premier aperçu. Nous n'analysons que la zone de C où seules des modifications de **IS** se produisent. Dans ce cas, trois types de configurations finales $\mathbf{IS}(\infty)$ sont possibles

- 1 cluster : lorsque $\mathbf{x}$ est de la forme *(i)* $[a, a, b, b]$ ou *(ii)* $[a, b, a, b]$ avec $ab \neq 0$. Cela signifie que les modifications des vecteurs d'attention initiaux se font en accord avec les individus qui ont *(i)* la même distribution d'attention (dans les groupes initiaux $g_i$) ou *(ii)* les mêmes relations sociales. Dans ces cas, étant donné la symétrie de $\mathbf{x}$ et la structure des relations sociales **II** (éq. (4.10)), une "communication globale" est possible entre les individus. Ces derniers vont s'accorder sur l'attention à porter aux sujets, jusqu'à obtention d'un consensus.
- 3 clusters : lorsque qu'un seul $x_i$ est non nul ou $\mathbf{x} = [a, b, a, b]$, $a = 0$ ou $b = 0$. Tout d'abord, regardons le cas d'un seul $x_i$ non nul. Afin d'illustrer, posons sans perte de généralité $i = 1$. Rappelons que les relations sociales lient les groupes $g_{13}$ et $g_{24}$. Dès lors, les individus des groupes $g_1$ et $g_3$ (i.e. le groupe avec lequel il est lié socialement) vont s'accorder sur l'attention à donner aux sujets. Les individus des deux groupes restant vont camper sur leurs positions d'origine, i.e. [1,0] ou [0,1]. Dans le cas $\mathbf{x} = [a, b, a, b]$, $a = 0$ ou $b = 0$, le raisonnement est similaire. Les individus des groupes possédant des individus éclectiques vont s'accorder sur l'attention à donner aux sujets, tandis que les deux autres groupes vont garder chacun leur propre attention initiale. Ce cas laisse supposer que les individus ont toujours une certaine "indépendance de caractère" : des individus reliés socialement ne vont pas nécessairement adopter la même distribution d'attention.
- 2 clusters : toutes les autres configurations de $\mathbf{x}$. Les relations sociales **II** vont imposer l'attention **IS** des individus, les clusters de $\mathbf{IS}(\infty)$ étant ceux de $\mathbf{II}(0)$. Les individus tiennent à leurs relations sociales et changent leurs attentions pour être en accord avec leurs connaissances.

La région (D\SD) ∪ SC est assez évidente : toutes les relations évoluent de façon imbriquée. Un consensus est atteint dans $\mathbf{IS}(\infty)$ pour lequel les individus accordent autant d'importance au deux sujets. En outre, tous les individus forment une clique où chaque individu interagit avec la même intensité avec tous les autres individus (et lui-même). Notons aussi que cette partie D\SD semble faiblement réaliste car elle suppose une trop grande ouverture d'esprit des individus.

Finalement, les sous-régions SB et SD ne sont pas analysées car il existe un trop grand nombre de cas possibles en fonction des paramètres du système.

**Analyse approfondie d'un cas particulier : un seul individu éclectique**

Nous nous concentrons sur le cas particulier d'un seul individu éclectique. Ce choix est motivé par le fait qu'il s'agit de la plus petite modification possible dans l'initialisation (par rapport à l'éq. (4.11)). Cette simple modification a des conséquences sur l'état final du système (4.9). Afin d'illustrer, nous considérons une population de $n_I = 20$ individus et sans perte de généralité, nous fixons l'individu 1 (du groupe $g_1$) comme individu éclectique, i.e. $\mathbf{x} = [1, 0, 0, 0]$.

Afin de caractériser l'état final du système, nous nous intéressons ici aux partitions des matrices $\mathbf{IS}(\infty)$ et $\mathbf{II}(\infty)$. Bien que jusqu'à présent, nous nous soyons beaucoup concentrés sur le nombre de clusters d'individus (individus ayant les mêmes vecteurs $\mathbf{IS}(\infty)$ ou $\mathbf{II}(\infty)$), les informations apportées par les partitions[1] sont plus parlantes puisque derrière un même nombre de clusters se cachent différentes partitions. Les propos qui suivent sont illustrés à la figure 4.4, avec $n_I = 20$.

Les distances minimales entre les vecteurs des matrices $\mathbf{IS}(0)$ et $\mathbf{II}(0)$ sont les suivantes (éq. (4.19))

$$\epsilon^m_{IS} := \min_{u,v} d^0_{IS,IS}(u, v) = 1/2 \tag{4.20}$$

$$\epsilon^m_{II} := \min_{u,s} d^0_{IS,SI}(u, s) = \frac{1}{\sqrt{2}} \left[ \frac{2n_I^2 - n_I - 2}{n_I(n_I + 1)^2} \right]^{1/2}. \tag{4.21}$$

Ces deux valeurs sont les seuils à partir desquels les relations sociales et attentions peuvent évoluer. Ces seuils définissent quatre régions dans le plan $(\epsilon_{II}, \epsilon_{IS})$, délimitées par les droites *frontières* $\epsilon_{II} = \epsilon^m_{II}$ et $\epsilon_{IS} = \epsilon^m_{IS}$ comme illustré sur la figure 4.4 (lignes grises en pointillé).

Les comportements qualitatifs et "interprétation" discutés dans la sous-section précédente pour les régions A, B\SB, C\SC et D\SD sont bien sûr d'application ici. Nous nous concentrons donc plus particulièrement sur les sous-régions SB, SC et SD.

Pour le vecteur $\mathbf{x}$ envisagé, SB est divisé en trois parties : les quatre distances (4.19) sont différentes et valent (illustration avec $n_I = 20$)

$$\epsilon^m_{II} = d^{(3)}_{II,SI} \simeq 0.21 < d^{(2)}_{II,SI} \simeq 0.2204 < d^{(4)}_{II,SI} \simeq 0.2216 < d^{(1)}_{II,SI} = \epsilon^M_{II} \simeq 0.2321. \tag{4.22}$$

Puisque $\mathbf{x} = [1, 0, 0, 0]$, la sous-région SC (sous-région dans laquelle il y a des modifications **II** et **IS**) se divise elle-même en deux zones (par rapport à $\mathbf{IS}(\infty)$). La première zone est caractérisée par l'existence de deux clusters d'individus au niveau des relations $\mathbf{IS}(\infty)$ (voir discussion précédemment, partition N$^o$2). Ce cas est très intéressant car l'individu 1 va finir par avoir la même attention finale que les individus 16:20 avec lesquels il n'a aucun lien social initialement. Cela est dû au

[1] Une partition d'un ensemble $X$ est un ensemble de parties non vides de $X$, deux à deux disjointes, dont l'union est $X$.

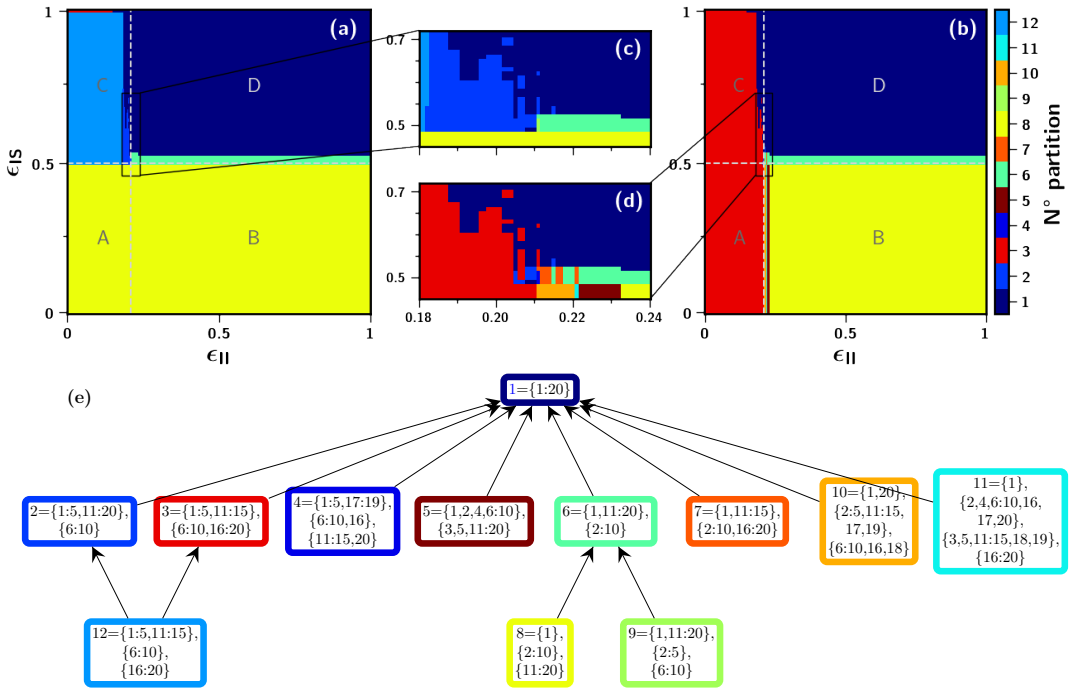



**Figure 4.4.:** Partitions des individus dans les matrices finales (a) $\mathbf{IS}(\infty)$ et (b) $\mathbf{II}(\infty)$, en fonction de $(\epsilon_{II}, \epsilon_{IS})$. Les *frontières* délimitant les quatre zones A, B, C et D sont tracées en pointillé gris. Dans les encarts, agrandissements sur les valeurs $[0.18, 0.25] \times [0.48, 0.72]$ pour (c) $\mathbf{IS}(\infty)$ et (d) $\mathbf{II}(\infty)$. (d) Diagramme de Hasse des partitions représentées sur les panels (a)-(d). La notation $i : j$ signifie $i, i+1, ..., j-1, j$ et le code couleur est identique pour tous les panels. Les résultats sont obtenus avec $n_I = 20$.

léger déséquilibre induit en choisissant l'individu 1 comme éclectique : initialement, il y a $n_I/2 - 1$ individus tels que $\mathbf{IS}(0) = [1, 0]$ et $n_I/2$ individus tels que $\mathbf{IS}(0) = [0, 1]$. Ces derniers vont avoir plus d'influence sur le système, faisant évoluer de façon non symétrique les relations sociales **II** pour qu'au final, la co-évolution des relations **IS** et **II** fasse que les groupes $g_1$ et $g_{34}$ s'accordent sur l'attention à donner aux sujets.

La seconde zone est constituée de valeurs $\epsilon_{II}$ très proches de la frontière $\epsilon_{II}^m$ où $\epsilon_{IS}$ plus grand que pour la première zone. Dans ce cas, les deux types de relations vont mutuellement s'influencer pour converger vers une solution *homogène*. Cela signifie que tous les individus ont la même distribution d'attention $[(n_I - 1)/n_I, (n_I + 1)/n_I]$ et que les liens sociaux entre eux ont tous la même importance. Dit autrement, tous les individus vont porter exactement la même attention aux deux sujets (*consensus*) et cette attention est donnée par la moyenne des attentions initiales. Au niveau social, les individus forment une clique et s'accordent tous autant d'importance (partition N°1). Dans ce cas, même si au départ les individus n'ont pas une ouverture d'esprit suffisamment grande, la co-évolution des relations fait qu'ils vont finalement atteindre un consensus.

Comme vu à la section précédente, tout couple $(\epsilon_{II}, \epsilon_{IS}) \geq (\epsilon_{II}^m, \epsilon_{IS}^m)$ permet au marcheur de se

déplacer le long des méta-chemins (région D). Cependant, il sera limité pour $\epsilon_{IS} \leq \epsilon_{IS}^{crit} = \frac{2\lambda+n_I}{2n_I}$ (sous-région SD, illustration $\epsilon_{IS}^{crit} = 0.525$). Au-delà de cette valeur critique, i.e. $(\epsilon_{II}, \epsilon_{IS}) \geq (\epsilon_{II}^{m}, \epsilon_{IS}^{crit})$, le système converge vers la solution *homogène* évoquée précédemment. Finalement, pour $\epsilon_{IS} \in [\epsilon_{IS}^{m}, \epsilon_{IS}^{crit}]$ et $\epsilon_{II} > \epsilon_{II}^{m}$, l'individu 1 va adopter la même distribution d'attention **IS** que la majorité, i.e. [0,1], tandis que les vecteurs associés aux autres individus ne vont pas changer (partition N$^o$6, figure 4.4). En ce qui concerne le graphe **II**, il va devenir connexe sans nécessairement que tous ses liens soient de mêmes poids, et sans être composé de deux clusters distincts connectés par un seul nœud.

#### Bilan

Dans cette sous-section, nous avons envisagé le cas de trois distributions d'attention initiales afin de permettre des modifications dans les relations **IS** via la méthode proposée. En particulier, nous avons repris le cas de la section précédente (section 4.3.1) à la différence que certains individus (dits éclectiques) démarrent avec une attention initiale homogène (i.e. [1/2,1/2]). Nous avons grossièrement discuté du cas général, à savoir un nombre quelconque d'individus éclectiques. Cela a permis une première caractérisation dans le plan $(\epsilon_{II}, \epsilon_{IS})$ du système (4.9) à l'état d'équilibre. Nous nous sommes ensuite attardés sur un cas particulier d'un seul individu éclectique. Ce cas avait pour objectif de montrer comment une légère modification (par rapport à l'éq.(4.11)) dans l'initialisation de **IS** engendrait des comportements différents à l'état d'équilibre dans le plan $(\epsilon_{II}, \epsilon_{IS})$, caractérisés à l'aide des partitions d'individus.

### 4.3.3. Influence d'un média

#### Modèle

Dans cette section, nous introduisons un média, modélisé comme un agent connecté à tous les individus et qui ne change jamais sa distribution d'attention. L'attention des individus est éventuellement influencée par ce média toutes les $T$ itérations. $T$ est appelée la *période* du média. Nous n'envisageons pas de faire évoluer les relations sociales en fonction du média. En d'autre termes, cela signifie que le méta-chemin I→M→S est utilité pour faire évoluer **IS** tandis que I→M→I n'intervient pas pour les relations sociales **II**. La distribution de l'attention du média entre les deux

sujets s'exprime comme $\mathbf{MS} = [\alpha, 1-\alpha]$, $\alpha \in [0,1]$. Le système (4.9) devient

$$\begin{cases} \mathbf{IS}(t+1) = \mathbf{\Lambda_{IS}} \left[\beta_{IIS}(t)\, \mathbf{IIS}(t) + \beta_{IMS}(t)\, \mathbf{IMS}(t)\right] + (\mathbf{I} - \mathbf{\Lambda_{IS}})\, \mathbf{IS}(t) \\ \mathbf{II}(t+1) = \mathbf{\Lambda_{II}}\, \mathbf{ISI}(t) + (\mathbf{I} - \mathbf{\Lambda_{II}})\, \mathbf{II}(t) \end{cases} \tag{4.23}$$

$$\text{avec} \quad (\beta_{IIS}(t), \beta_{IMS}(t)) \quad = \begin{cases} (1/2, 1/2) & \text{si mod } (t,T) = 0 \\ (1,0) & \text{sinon} \end{cases}$$

#### Résultats

L'objectif est d'analyser (numériquement) l'impact d'un tel média en s'intéressant à l'état de convergence du système (4.23). Plus précisément, trois quantités sont analysées : le nombre de clusters d'individus partageant la même distribution d'attention finale ; l'efficacité du média $M_{eff}$, définie comme le pourcentage d'individus qui adoptent finalement la même distribution d'attention que lui ; le nombre d'itérations nécessaires pour que le système converge.

Dans ce qui suit, la matrice d'attention **IS** est initialisée aléatoirement (loi uniforme U([0,1]). La matrice initiale des relations sociales est toujours donnée par l'expression (4.10) : les individus sont organisé en deux cliques. À nouveau, nous considérons $\mathbf{\Lambda_{IS}} = \mathbf{\Lambda_{II}} = diag(0.5)$. Nous fixons $\alpha = 0.1$. Les analyses suivantes sont fonction de $T$, $\epsilon_{IS}$ et $\epsilon_{II}$. Le nombre d'individus est fixé à $n_I = 40$.

**Nombre de clusters et efficacité du média** Regardons d'abord le cas des relations sociales statiques, i.e. $\epsilon_{II} = 0$. La figure 4.5(a) illustre le nombre de clusters d'individus partageant la même distribution finale d'attention. Sans surprise, la tendance générale est une diminution du nombre de clusters lorsque $\epsilon_{IS}$ croît puisque cela signifie que les individus ont une ouverture d'esprit plus grande. Lorsqu'il n'y a qu'un seul cluster, le média possède 100% d'efficacité[1] (figure 4.5(c)). Nous remarquons aussi que l'efficacité du média dépend de la période $T$. Il est préférable pour le média d'intervenir épisodiquement pour laisser le temps aux individus de s'influencer mutuellement. En effet, si le média est omniprésent ($T = 1$), les individus dont l'attention initiale est proche du média vont effectivement converger vers lui, tandis que ceux dont l'attention initiale est éloignée vont converger vers la moyenne de leur attention initiale. Cela rappelle le problème de *surexposition* [2] dont un exemple est le cas du reciblage publicitaire en ligne : les consommateurs remarquent qu'ils sont "poursuivis" sur Internet par un article qu'ils ont déjà acheté. Cette répétition agace les gens, qui finissent éventuellement par "bloquer" les annonces ou publicités. Néanmoins, le média doit être un minimum présent pour influencer la population. Si $T$ est trop grand, le média risque d'avoir un impact au début du processus (figures 4.6(c)), mais finalement perdre toute influence car les

[1] Le cas particulier d'un seul cluster d'individus avec 100% d'efficacité du média pourrait théoriquement être obtenu si la moyenne des attentions initiales égale l'attention du média.

individus vont s'influencer trop rapidement et devenir inatteignables pour le média (figures 4.6(c) et 4.6(d)).

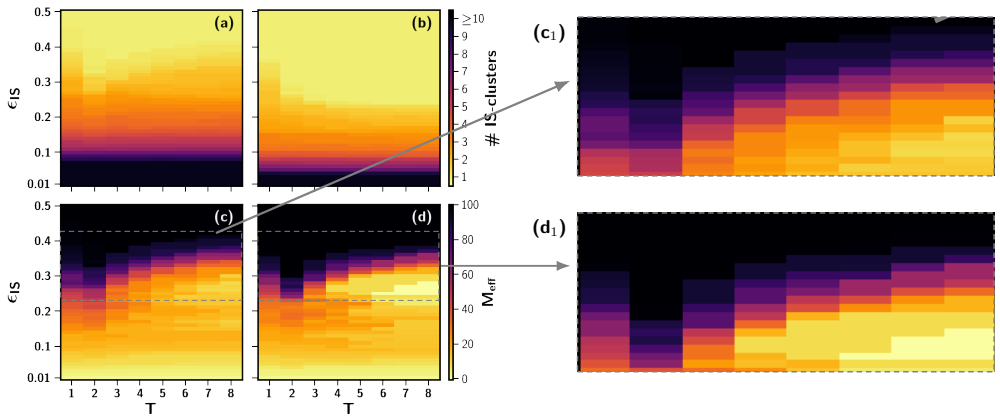



**Figure 4.5.:** Nombre de clusters d'individus en fonction de la matrice IS, i.e. individus partitionnés en fonction de l'attention qu'ils accordent aux deux sujets, lorsque (a) $\epsilon_{II} = 0$ (les liens sociaux sont fixés et immuables) et (b) $\epsilon_{II} = 1$ (les liens sociaux s'adaptent sans contrainte) ; et efficacité du média $M_{eff}$, i.e. pourcentage d'individus ayant la même répartition d'attention que le média à l'état de convergence, lorsque (c) $\epsilon_{II} = 0$ et (d) $\epsilon_{II} = 1$. Tous les panels sont obtenus avec $n_I$=40 et moyennés sur 1000 simulations. Sur les panels ($c_1$) et ($d_1$), agrandissements afin d'apprécier l'impact de la période $T$ du média.

Dans le cas où les interactions sociales évoluent, un plus grand nombre de configurations $(T, \epsilon_{IS})$ permettent d'obtenir un seul cluster d'individus partageant la même attention (i.e. cluster suivant **IS**, figure 4.5(c)). Cela n'est pas étonnant puisque tous les individus peuvent s'influencer et vont donc s'accorder sur une distribution d'attention commune. Cependant, dans ce cas, un consensus au niveau des attentions ne signifie pas que le média possède 100% d'efficacité. Il arrive même qu'il n'en ait aucune, i.e. le consensus est différent de la distribution du média. Cela arrive lorsque le média intervient trop rarement et que $\epsilon_{IS}$ se trouve approximativement dans [0.25,0.3], i.e. une faible ouverture d'esprit des individus. Dans ce cas, les individus convergent vers la moyenne des distributions initiales (figure 4.6(h)). Nous observons toujours que, lorsque le média est trop présent ($T = 1$), la population se fragmente (figure 4.6(e)).

Notons tout de même que la région de l'espace $(T, \epsilon_{IS})$ avec $\epsilon_{IS} < 0.2$ est celle qui ressemble le plus à une "vraie" situation car elle correspond plus à la "véritable" ouverture d'esprit d'une population et ce, quel que soit le paramètre $\epsilon_{II}$. Notons également que dans cette région, l'efficacité du média fluctue peu et principalement en fonction de $\epsilon_{IS}$. La période $T$ a peu d'influence sur cette efficacité comme le montre précisément la figure 4.7(a) (comparé e.g. aux zones zoomées, figures 4.5(c) et 4.5(d)). En général, le média ne peut influencer que les individus initialement proches de lui. Son efficacité est donc restreinte. Deux comportements d'utilisateurs sont possibles : fragmentation de la population ou formation de deux clusters (figure 4.7). Le premier cas fait

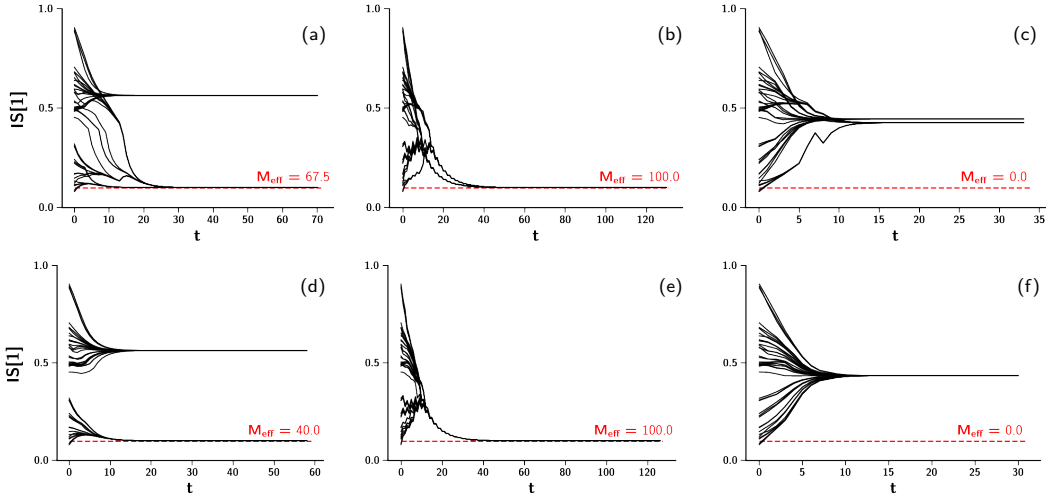



**Figure 4.6.:** Évolution type du système (4.23) en fonction de $T$ et $\epsilon_{II}$. La valeur $\epsilon_{IS}$ est fixée à 0.3. Seule l'attention portée au premier sujet est représentée, IS[1]. La ligne rouge représente l'attention du média pour le sujet visualisé, $\alpha = 0.1$, et son efficacité à l'état final est indiquée. (a),(b) et (c) correspondent au cas où il n'y a aucune modification dans les relations II, i.e. $\epsilon_{II} = 0$, et avec $T$=1, 2 et 8 resp. (d),(e) et (f) correspondent au cas où il n'y a des modifications dans les relations II, i.e. $\epsilon_{II} = 1$, et avec $T$=1, 2 et 8 resp.

penser à une situation où les individus connaissent bien les sujets ainsi que l'importance qu'ils ont pour eux en général. Ils ont une vision claire de ce qui les intéresse et ne modifient dès lors que très peu leurs attentions. Le second cas fait plutôt penser à une situation où les individus connaissent peu les sujets. Ils sont plus influençables et il leur faut plus de temps avant de savoir comment répartir leur attention. Le média a une plus grande influence dans ce cas et parvient même à imposer son attention à des individus initialement éloignés en terme d'attention.

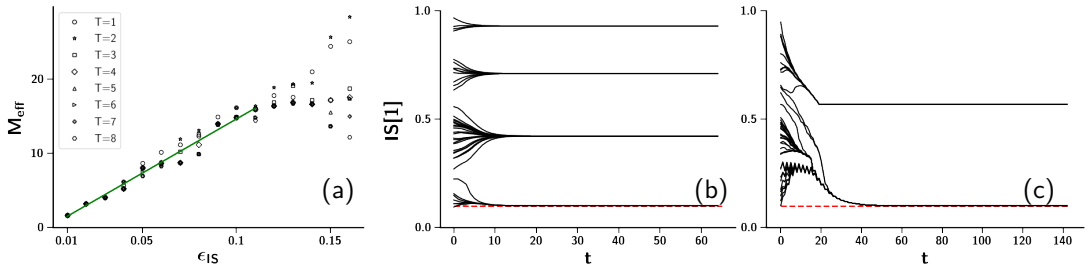



**Figure 4.7.:** (a) Efficacité du média en fonction de $\epsilon_{IS} \in [0, 0.15]$ lorsque $\epsilon_{II} = 0$ (pas d'évolution des relations sociales). Pour $\epsilon_{IS} < 0.1$, $M_{eff}$ ne dépend que très peu de $T$ et semble être corrélé linéairement à $\epsilon_{IS}$ (ajustement ligne verte $M_{eff} = 1.46 + 150\,\epsilon_{IS}$) ; Exemple d'évolution du système (4.23) avec $T = 3$, $\epsilon_{II} = 1$ et (a) $\epsilon_{IS} = 0.1$ et (b) $\epsilon_{IS} = 0.2$. Seule l'attention porté au premier sujet est représentée, IS[1]. La ligne rouge représente l'attention du média pour le sujet visualisé, $\alpha = 0.1$.

**Temps de convergence** Nous regardons le nombre d'itérations pour atteindre l'état d'équilibre (figure 4.8). Au plus la période $T$ est élevée, au plus le temps pour converger est important, les autres paramètres étant égaux. Cela est assez intuitif puisque au plus $T$ est grand, au plus la perturbation est lente à arriver. Pour $T$ fixé, le nombre d'itérations dépend du seuil $\epsilon_{IS}$ et donc de l'efficacité du média. Lorsque ce dernier n'a aucune efficacité, le système converge rapidement : les individus moyennent leur attention, ce qui est très rapide puisqu'aucune perturbation n'intervient. Finalement, nous n'observons que peu de différences entre $\epsilon_{II} = 0$ et $\epsilon_{II} = 1$, figures 4.8(a) et 4.8(b) respectivement. La différence se marque principalement pour de faibles ouvertures d'esprit $\epsilon_{IS}$.

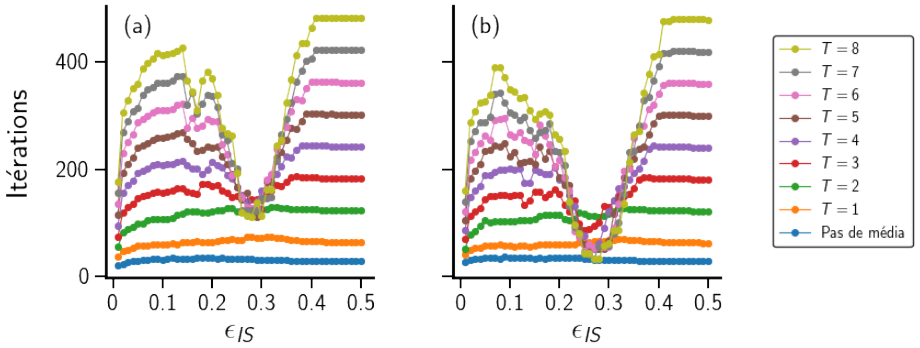



**Figure 4.8.:** Nombre d'itérations pour atteindre l'état de convergence en fonction de $\epsilon_{IS}$ pour différentes valeurs de $T$. L'attention pour le premier sujet du média est $\alpha = 0.1$. (a) $\epsilon_{II} = 0$ et (b) $\epsilon_{II} = 1$. Les résultats sont obtenus avec $n_I$=40 et moyennés sur 1000 simulations.

#### Bilan

Dans cette sous-section, nous avons introduit un média dont l'attention fixe peut potentiellement influencer l'attention des individus de façon périodique. Le résultat principal est qu'en fonction de la période du média et de l'ouverture d'esprit des individus, le média a plus ou moins d'efficacité. Ainsi que relevé dans la littérature, lorsque le média est trop présent, il provoque un fragmentation de la population. Lorsqu'il intervient trop peu souvent, il n'a d'influence qu'au début du processus mais perd toute efficacité au bout d'un moment. Pour des périodes intermédiaires, le média peut atteindre 100% d'efficacité, cela dépend de l'ouverture d'esprit des individus. Finalement, le temps de convergence dépend de l'ouverture d'esprit des individus $\epsilon_{IS}$ (et dans une moindre mesure $\epsilon_{II}$) et de la période $T$ du média, et donc, implicitement de son efficacité $M_{eff}$.

### 4.3.4. Discussion d'une extension du modèle aux hypergraphes

Dans ce chapitre, nous avons envisagé une dynamique des liens d'un HIN dyadique, i.e. un HIN où les liens ne relient que deux nœuds à la fois. Cependant, ainsi que déjà mentionné, un graphe

dyadique n'est pas toujours la représentation la plus intuitive ni la plus adaptée lorsqu'il s'agit de représenter des relations sociales ; nous avons dès lors évoqué les hypergraphes. Dans cette section, nous discutons brièvement d'une possible extension du modèle de dynamique des liens aux hypergraphes et l'impact que cela pourrait avoir sur l'évolution du modèle d'attention proposé (éq. (4.9)).

Dans cette discussion, nous proposons un simple exemple qui est une extension directe du cas analysé dans la section 4.3.1 avec seulement deux distribution d'attention [1,0] et [0,1], et nous nous intéressons uniquement à la modélisation (multi-adiques) des relations sociales **II**. Par conséquent, les relations **IS** restent dyadiques. La figure 4.9 illustre quelques cas possibles ; en particulier, ceux dont $|T(e)|, |H(e)| \in \{1, n_I/4\}$ par rapport aux groupes $g_i$. Cela signifie qu'une personne ou un groupe complet peut soit interagir avec une seule autre personne, soit avec un groupe complet. Dans la suite, les liens *intra-groupes* sont ceux relatifs aux relations à l'intérieur de chaque groupe $g_i$, $i = 1, ..., 4$ tandis que les liens *inter-groupes* font référence aux relations entre les groupes $g_{13}$ et $g_{24}$, connectés socialement dans la situation de base (i.e. HIN dyadique, éq. (4.10)).

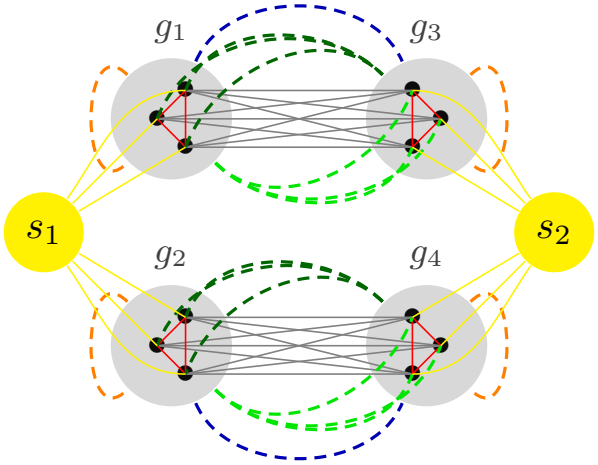



**Figure 4.9:** Plusieurs possibilités de représentations des relations sociales **II** en fonction des quatre groupes d'individus prédéfinis $g_i$, $i = 1, ..., 4$. Les possibles liens sont rouges et gris, traits fins continus. Les possibles hyperliens sont oranges, vert et bleus, traits épais discontinus. En jaune, les relations "de base" (liens) pour la relation **IS**.

Ainsi que mentionné, nous nous intéressons à l'impact de la représentation choisie (liens ou hyperliens) sur l'évolution du système (4.9). En particulier, la représentation va définir la matrice des relations $\mathbf{II}(0)$. Pour rappel, cette matrice représente également la probabilité de passer d'un individu à l'autre suivant le méta-chemin I $\rightarrow$ I. Dans le cas des hypergraphes, cette matrice de probabilités de transition est donnée par la formule générale $\mathbf{P} = \mathbf{D_v^{-1}} \mathbf{H}^+ \mathbf{W} \mathbf{D_{ve}^{-1}} (\mathbf{H}^-)^\mathsf{T}$ (éq. (1.3)). Afin de proposer une illustration, nous nous plaçons dans le cadre le plus simple et fixons le poids des nœuds égale à 1, i.e. $\lambda_e(v) = 1, \forall e, \forall v$. Dans ce cas, $\mathbf{D_{ve}} = \mathbf{D_e}$. En utilisant la formule des probabilités de transition pour calculer $\mathbf{II}_\mathcal{H}(0)$ (indice $\mathcal{H}$ pour indiquer l'hypergraphe), nous remarquons que dans certains cas, la matrice $\mathbf{II}_\mathcal{H}(0)$ est identique à $\mathbf{II}(0)$. Cela vient des symétries

des (hyper)liens et des poids des hyperarcs que nous avons définis (pour rappel, $\omega(e) = \tilde{\omega}(e)[\delta(e)]^{\gamma}$, section 1.2.2). En particulier, soit $v \in V$. Si $\forall e \in E$ tq $v \in T(e)$: $|T(e)| = k_1$ et $|H(e)| = k_2$, avec $k_1, k_2 \in \mathbb{Z}$, alors considérer des hyperarcs n'aura pas d'effet dans le calcul des probabilités de transition au départ de $v$. Mais si cette condition n'est pas vérifiée, alors l'utilisation des hyperarcs va avoir un impact.

Afin d'illustrer ce propos, considérons la situation suivante pour laquelle $\mathbf{II}_{\mathcal{H}}(0) \neq \mathbf{II}(0)$ : à l'intérieur de chaque groupe, il y a un seul hyperlien reliant l'intégralité du groupe avec lui-même tel que $|T(e)| = |H(e)| = n_I/4$ tandis qu'entre les groupes $g_1$ et $g_3$, et $g_2$ et $g_4$, les nœuds sont tous reliés de façon dyadique, i.e. ce sont des hyperliens tels que $|T(e)| = |H(e)| = 1$. Cela revient en fait à de simples liens inter-groupes. La figure 4.9 reprend cette situation avec les hyperliens intra-groupes oranges et les (hyper)liens inter-groupes gris. Une possible interprétation est la suivante. On part de la situation de base (graphe dyadique, section 4.3.1) où les individus des groupes $g_1$ (resp. $g_2$) ont tous un lien avec les individus de $g_2$ (resp. $g_4$). La situation décrite par l'hypergraphe est la suivante : les individus connectés et ayant la même attention interagissent en groupe (e.g. les individus ont des points communs et préfèrent partager leurs intérêts en groupe), tandis que lorsqu'ils sont connectés mais ont des attentions différentes, ils interagissent deux à deux (e.g. les individus sont intéressés par des choses différentes et afin de ne pas avoir de pression sociale, ils préfèrent interagir deux à deux).

L'objectif de l'analyse effectuée ici est simplement de voir l'impact du paramètre $\gamma$ , intervenant dans la définition du poids des hyperliens : $\omega(e) = \tilde{\omega}(e)[\delta(e)]^{\gamma}$. L'importance de la taille de l'hyperarc est donc réglée grâce à $\gamma$. Dans cette discussion, nous considérons $\tilde{\omega}(e) = 1$, $\forall e$. Pour rappel, nous reprenons l'application la plus basique, à savoir seulement deux distributions d'attention initiales [1,0] et [0,1]. Il n'y a donc aucune modification des relations **IS**. La figure 4.10 illustre les solutions $\mathbf{II}(\infty)$ en fonction de $\epsilon_{II}$ et $\gamma$.

Tout d'abord, $\mathbf{II}_{\mathcal{H}}(0) = \mathbf{II}(0)$ pour $\gamma = 1$. Cela signifie que pour cette valeur de $\gamma$, le fait de considérer des hyperliens intra-groupes ne se distingue pas, dans le calcul des pois des relations sociales initiales $\mathbf{II}_{\mathcal{H}}(0)$, du fait de considérer de simples liens intra-groupes. C'est également pour cette valeur que le nombre d'itérations pour converger vers l'état d'équilibre est minimal. Ensuite, lorsque $\gamma > 1$ (resp. $< 1$ ), les liens intra- (resp. inter-) groupes sont plus importants puisque le poids d'un hyperlien est donné par $\omega(e) = [\delta(e)]^{\gamma}$ et que les hyperliens intra-groupes comportent plus de nœuds que les hyperliens inter-goupes. Lorsqu'il y modification des relations sociales, la solution est toujours la même quel que soit $\gamma$ : les distributions d'attention façonnent les relations sociales **II** étant donné que les relations d'attention **IS** ne varient pas (matrice $M_2$ dans la figure 4.10). Le seuil $\epsilon_{II}^{m}$, i.e. l'ouverture d'esprit minimale pour déclencher les modifications,

dépend de $\gamma$. En particulier,

$$\epsilon_{II}^{m}(\gamma) = \frac{1}{4\sqrt{2}n_I d^2}\left[\left(2d - 4\left(\frac{n_I}{2}\right)^{\gamma}\right)^2 + (n_I 2^{\gamma})^2 + 4d^2\right]^2, \quad \text{avec } d = \left(\frac{n_I}{2}\right)^{\gamma} + 2^{\gamma}\frac{n_I}{4}. \tag{4.24}$$

Cette fonction possède un minimum en $\gamma = 1.68$, correspondant à la configuration initiale $\mathbf{II}_{\mathcal{H}}(0)$ permettant le plus facilement de modifier les relations sociales. Une façon d'interpréter cette fonction est la suivante : lorsque les individus accordent beaucoup plus d'importance aux individus d'un autre groupe ($\gamma \ll 1$), individus avec lesquels ils interagissent deux à deux et ont des attentions différentes, l'ouverture d'esprit minimale $\epsilon_{II}^{m}$ doit être plus élevée pour qu'ils modifient leurs relations sociales : les liens dyadiques, plus personnels et plus nombreux que leur relation en groupe (un seul hyperlien) résistent et il faut une grande ouverture d'esprit des individus pour que l'esprit de groupe domine. À l'inverse, lorsque les individus, interagissant en groupe partageant la même attention, s'accordent plus d'importance ($\gamma \gg 1$), alors l'ouverture d'esprit minimale $\epsilon_{II}^{m}$ n'aura pas besoin d'être aussi élevée que dans le cas précédent car ils sont entraînés plus facilement par l'esprit de groupe.

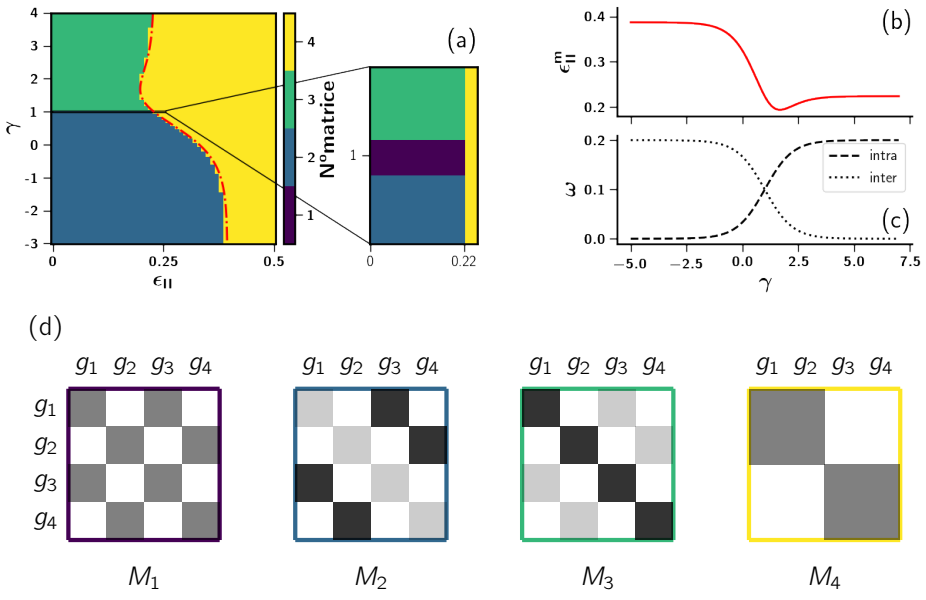



**Figure 4.10.:** (a) États finaux de la matrice $\mathbf{II}(\infty)$ dans le plan $(\epsilon_{II}, \gamma)$ lorsque des hypergraphes sont considérés (voir texte principal pour la topologies des hypergraphes). Ces états sont référencés par un numéro de matrice. Les matrices correspondantes sont schématiquement représentées en (d). Les matrices initiales $\mathbf{II}(0)$ sont retrouvées en $\epsilon_{II} = 0$. Lorsque $\gamma = 1$, on retrouve le cas d'un HIN simple. Lorsque $\gamma > 1$ (resp. $< 1$ ), les liens intra- (resp. inter-) groupes sont plus importants. Lorsqu'il y a modification, l'état final est indépendant de $\gamma$. Le seuil $\epsilon_{II}^{m}$ minimal pour engendrer des modifications dans $\mathbf{II}$ dépend de $\gamma$ : la solution analytique $\epsilon_{II}^{m}(\gamma)$ est représentée en rouge et est reprise en (b) pour plus de clarté (éq. (4.24)). (c) Poids des hyperliens $\omega(e)$ intra- et inter-groupes. Pour $\gamma = 1$, tous les poids sont égaux et $\mathbf{II}_{\mathcal{H}}(0) = \mathbf{II}(0)$. (d) Matrices types (structure en blocs) avec numéro correspondant en (a). Au plus la zone est gris foncé, au plus le poids des hyperliens est élevé.

En conclusion, en fonction de la façon dont s'organisent les relations (dyadiques ou non) et l'importance (i.e. le poids caractérisé par le paramètre $\gamma$) que chaque individu accorde aux autres, la valeur de l'ouverture d'esprit minimale de l'ensemble de la population pour engendrer des modifications sociales **II** varie. Dans cette illustration, nous avons fixé la topologie de l'hypergraphe et avons brièvement discuté de l'impact de $\gamma$. Néanmoins, ce simple exemple permet d'illustrer le rôle de la représentation non-dyadique d'un système. Des résultats similaires peuvent être tout aussi facilement obtenus avec les situations décrites sur la figure 4.9. Notons que l'exemple considéré se concentre uniquement sur une modification des relations sociales **II**. Il serait intéressant de modifier légèrement $\mathbf{IS}(0)$ (et définir $\mathbf{IS}_{\mathcal{H}}(0)$) afin d'avoir un système d'hypergraphes en co-évolution. Finalement, notons que les cas envisagés (figure 4.9) sont tous "symétriques". D'autres cas intéressants à considérer sont les hypergraphes dont les hyperarcs ont des tailles variables, e.g. fonctions des individus et des groupes, et qui représenteraient des situations plus complexes des réseaux réels.

## 4.4. Résumé et perspectives

Dans ce chapitre exploratoire, nous avons proposé un modèle d'évolution du poids des liens d'un HIN. Ce modèle se base sur des marches aléatoires contraintes par des méta-chemins et limitée par des distances entre les nœuds. Nous avons ensuite proposé un cas d'application, à savoir les dynamiques d'attention d'un ensemble d'individus.

Nous avons envisagé une population formée de quatre groupes d'individus. Ces individus sont liés entre eux par des liens sociaux **II** et sont caractérisés par la quantité d'attention **IS** qu'ils portent à deux sujets différents. Dans la vie réelle, ces deux relations peuvent s'influencer, e.g. les individus vont accorder leur attention en imitant celle de leur voisinage ou bien renforcer leurs contacts avec les personnes ayant les mêmes centres d'intérêt. Dès lors, nous avons considéré une dynamique où ces deux relations s'influencent, basée sur le principe bien connu de confiance limitée. L'approche envisagée se différencie des modèles d'opinions "classiques" en ce sens qu'il s'agit d'une dynamique de liens, et non une dynamique d'état des nœuds.

L'objectif était de caractériser l'état d'équilibre en fonction de l'ouverture d'esprit des individus par rapport aux sujets ($\epsilon_{IS}$) et aux individus ($\epsilon_{II}$).

Pour ce faire, nous avons imposé une structure des relations sociales en deux groupes. Nous avons envisagé le cas d'un nombre quelconque d'individus éclectiques, i.e. distribuant initialement leur attention équitablement entre deux sujets (les autres individus étant obnubilés par un seul), ce qui nous a permis d'avoir un premier aperçu des différents états finaux possibles. En particulier,

l'analyse dans le plan $(\epsilon_{II}, \epsilon_{IS})$ a tout d'abord permis de définir quatre régions principales, délimitées par l'ouverture d'esprit minimale nécessaire pour déclencher des modifications dans les relations **II** et **IS**. À l'intérieur de certaines d'entre elles, des sous-régions ont pu être identifiées. Nous avons ensuite illustré ces propos sur le cas particulier d'un seul individu éclectique.

Finalement, nous avons introduit au sein du modèle un média possédant toujours la même valeur d'attention. Le résultat principal est qu'en fonction de la période de diffusion du média et de l'ouverture d'esprit des individus, le média a plus ou moins d'efficacité à convaincre les individus à adopter la même attention que lui.

Ce modèle d'attention permet d'explorer des mécanismes sociaux potentiellement plausibles. Les hypothèses et configurations du modèle étant très simplistes, il serait intéressant de l'approfondir en le complexifiant. Par exemple, considérer plus de sujets se ferait assez simplement, même si les résultats macroscopiques seraient certes très différents. Une autre idée est d'envisager de faire évoluer l'attention du média **MS**, représentant e.g. sa ligne éditoriale, en fonction de son audience et inversement, i.e. tous les individus ne sont pas d'office sujet à l'influence du média. Cela rendrait le modèle un peu plus réaliste. Concrètement, cela reviendrait à prendre en compte le méta-chemin $\mathrm{M} \to \mathrm{I} \to \mathrm{S}$ pour l'évolution de **MS** et par exemple, les méta-chemins $\mathrm{I} \to \mathrm{I} \to \mathrm{M}$ et $\mathrm{I} \to \mathrm{S} \to \mathrm{M}$ pour l'évolution de **IM**. De façon générale, l'ajout de méta-chemins (pertinents) dans l'éq. (4.4) revient à prendre en compte plus de relations et de corrélations, de manière plus ou moins directe. Finalement, notons qu'il ne s'agit que d'un exemple d'application du modèle général d'évolution du poids des liens d'un HIN. D'autres applications et analyses théoriques peuvent découler de ce modèle général comme par exemple des processus de diffusion dans des HIN basés sur des marches aléatoires contraintes et limitées.

# Conclusion

## Résumé

Le travail présenté dans ce manuscrit a pour but d'investiguer les corrélations entre les liens des HINs. Nous nous sommes concentrés sur deux cas d'application particuliers. Le premier est la prédiction structurelle du poids des liens tandis que le second est la modélisation de l'évolution du poids des liens. Pour chacun de ces objectifs, l'hypothèse sous-jacente est la suivante : lorsqu'un HIN représente – ou a pour but de représenter – des données réelles, il existe des corrélations entre les liens de ce HIN pour la simple et bonne raison qu'en général, les évènements de ce monde sont corrélés les uns aux autres.

Dans un premier temps, chapitre 1, nous avons introduit les définitions et concepts utiles à la compréhension du reste du travail. En particulier, nous avons motivé les raisons d'utiliser le formalisme des HIN. Ensuite, les définitions relatives aux HIN et hypergraphes ont été présentées avec un accent sur la notion de méta-chemin, notion centrale de ce manuscrit. Les marches aléatoires contraintes par des méta-chemins ont été discutées en détails, dans les HIN et les hypergraphes.

Dans les deux chapitres suivants – chapitres 2 et 3 – nous nous sommes intéressés au problème de prédiction du poids des liens dans un HIN.

Tout d'abord, nous avons rapporté une partie de la littérature pertinente à ce sujet. Elle peut se décomposer en deux grandes parties : prédiction du poids des liens dans les graphes et prédiction de l'existence des liens dans les HINs. Grâce à ce bref parcours des méthodes existantes, nous avons pu identifier quelques éléments ayant fait leurs preuves pour résoudre le problème générique de la prédiction de liens (que ce soit dans les graphes ou dans les HINs), et ce, afin de proposer une méthode permettant de récupérer et de prédire le poids des liens dans un HIN.

Nous avons ensuite présenté notre approche basée sur les méta-chemins et plus précisément, sur les distributions de probabilités résultantes de marches aléatoires contraintes par des méta-chemins. L'approche proposée repose tout simplement sur la combinaison linéaire de telles distributions. À travers de nombreux exemples, nous avons pu évaluer la validité de l'approche proposée et pointer une partie de ses inconvénients et limites. En particulier, différentes questions ont été abordées et

les principaux points retenus sont les suivants : des méta-chemins trop longs ou trop répétitifs ne sont pas bénéfiques pour notre objectif, les méta-chemins ont un pouvoir explicatif limité (e.g. ils ne permettent pas de rendre compte de la structure du HIN), un seul modèle pour expliquer des données est souvent fort peu (dans notre cas, ça l'est clairement) et il faut se tourner vers d'autres méthodes d'analyse ou se plonger dans la signification des données afin de mieux les cerner.

L'importance de la représentation des données a également été abordée. Il convient toujours d'identifier ce que l'on recherche, voire ce que l'on désire montrer, afin de choisir la meilleure modélisation *a priori* des données à analyser. Dans ce travail, certaines données ont été représentées par un graphe dans lequel les interactions ne font intervenir que deux entités, ou un hypergraphe dans lequel une interaction concerne un nombre quelconque d'entités, afin de comparer leurs avantages et inconvénients. Il a été montré que pour des choix judicieux des poids de nœuds et des liens, considérer des interactions de plusieurs entités était bénéfique pour prédire le poids des liens.

Finalement, la seconde idée de la thèse a été approfondie dans le chapitre 4 : tout lien, toute relation évolue dans le temps. Par conséquent, nous avons proposé un modèle d'évolution du poids des liens. À nouveau, le concept de méta-chemin est au centre de la dynamique proposée : le poids des liens n'évolue pas de manière arbitraire mais bien en fonction d'autres relations présentes dans le HIN. Afin de donner un exemple concret de ce modèle, nous avons envisagé un modèle d'attention, où les individus, liés par des liens sociaux, s'influencent afin de modifier leur distribution d'attention sur des sujets. En outre les liens sociaux peuvent éventuellement évoluer, en fonction des distributions d'attention. Nous avons pu observer qu'en fonction de l'ouverture d'esprit des individus, les relations sociales pouvaient imposer les distributions d'attention ou inversement. Il existe aussi des valeurs d'ouverture d'esprit (ni trop faibles, ni trop élevées) pour lesquelles les états finaux sont plus complexes, reflétant peut-être des situations plus réalistes. Afin d'aller un peu plus loin, un média a été ajouté au modèle et une première analyse de son efficacité a été proposée. Le principal résultat est qu'en fonction de l'ouverture d'esprit des individus et de la période du média, ce dernier a plus ou moins d'influence sur la population. Bien que motivé par des observations empiriques, ce travail exploratoire se base uniquement sur des simulations.

## Retour sur les questions de recherche

Ce travail portait sur les possibles manifestations des corrélations et des interdépendances dans des HIN. À l'aide de la première application – récupération et prédiction du poids des liens – nous avons montré qu'il y avait effectivement des corrélations (statistiques) entre certains liens de types différents. Ces corrélations ont pu être mesurées, interprétées et ont permis de récupérer,

dans une certaine mesure, le poids des liens. La méthode proposée, reposant sur l'hypothèse de corrélations entre les liens, possède des limites (voir chapitre 3 pour plus de détails). Ces limites proviennent soit de méthodes statistiques inadaptées (ou du moins, faiblement adaptées), soit du fait que l'hypothèse de base n'est que *partiellement* correcte. Néanmoins, cette application permet d'aborder le problème de la prédiction du poids des liens d'un HIN en terme de corrélations de liens.

La seconde application est plutôt une tentative de répondre à la question : est-ce que les corrélations peuvent servir à modéliser des dynamiques de co-évolution ? En se basant sur les résultats du chapitre 3, à savoir qu'un type de liens est linéairement corrélé à d'autres types de liens, nous avons proposé un modèle de co-évolution du poids des liens. Afin d'illustrer ce modèle très général, un modèle d'attention a été proposé. Les simulations effectuées ont fourni des résultats similaires à d'autres présents dans la littérature (plutôt relatifs aux dynamiques d'opinions). Il n'empêche que cela nous conforte dans l'idée que faire évoluer le poids des liens au moyen de combinaisons linéaires du poids de liens et de chemins de types différentes est une piste intéressante.

Bien évidemment, ces deux applications ne répondent que partiellement aux questions posées en début de travail et les réponses apportées ne sont que superficielles (e.g. la méthode n'est appliquée que sur trois jeux de données). Néanmoins, ce travail a permis de nous poser des questions plus générales, que nous discutons dans la section suivante.

## Perspectives

Lorsqu'on traite d'un sujet général tel que les corrélations entre les liens dans un HIN, nous ne pouvons nous concentrer que sur un ou deux points précis. Il est dès lors naturel d'envisager de nouvelles investigations qui pourraient venir compléter la compréhension des résultats présentés dans ce travail. Des perspectives propres à chaque partie ont été déjà été évoquées. Nous proposons ici des perspectives plus globales et à plus long terme.

Tout au long de ce travail, nous n'avons considéré que des corrélations linéaires. Dans le cas de la prédiction du poids des liens, nous avons motivé ce choix par la volonté d'un modèle facilement interprétable, et avons dès lors exclu des méthodes d'apprentissage automatique ou d'apprentissage profond qui sont, en général, trop complexes à interpréter. Pourtant, ces dernières années, nombre de travaux ont montré que l'analyse de quantités massives de données avec des méthodes d'apprentissage statistique, sans comprendre les variables, a tendance à produire de meilleures prédictions que l'approche théorique qui tente de modéliser la façon dont les variables sont liées les unes aux autres.

Derrière ce choix se cache en réalité la vision que l'on a souvent de la modélisation statistique. Le statisticien Leo Breiman distinguait deux cultures [20] : *i)* la culture de la modélisation des données, qui suppose que la nature est une "boîte noire" dans laquelle les variables sont associées de manière stochastique, et le travail des modélisateurs consiste à identifier le modèle qui correspond au mieux à ces associations sous-jacentes ; *ii)* la culture de la modélisation algorithmique, qui suppose que les associations dans la boîte noire sont trop complexes pour être décrites par un modèle simple, et le travail des modélisateurs est d'utiliser l'algorithme qui peut le mieux estimer la sortie à partir des variables d'entrée, sans s'attendre à ce que les vraies associations sous-jacentes des variables à l'intérieur de la boîte noire puissent être comprises.

Actuellement, l'industrie de la science des données attend principalement des scientifiques qu'ils se concentrent sur des modèles prédictifs, sans pour autant comprendre précisément les mécanismes qui mènent à ces prédictions. Des concours tels que celui de Netflix[1] ou les compétitions de Kaggle[2] mettent bien en avant l'importance qu'on accorde, en général, à la prédiction par rapport à celui de la compréhension et de l'explication, bien que de plus en plus de personnes veulent de l'interprétabilité.

Nous pensons qu'avoir à choisir entre une de ces deux visions n'est pas une bonne approche. Il n'y a aucune raison pour que nous ne puissions en choisir qu'une et il existe de nombreuses possibilités pour fusionner les deux cultures. Ainsi, rendre les modèles d'apprentissage automatique plus interprétables est un objectif à long terme qui mérite qu'on y consacre beaucoup d'efforts. De plus en plus de travaux s'attellent à cette tâche [17, 61, 161].

Cette remarque est très générale et est donc difficile à appliquer en pratique. En ce qui concerne le travail présenté dans ce manuscrit, i.e. une simple régression linéaire minimisant la somme des carrés des résidus, un premier pas serait de rester dans des algorithmes dits "white-box" mais un peu moins contraints que les régressions linéaires, e.g. les arbres de décision (dans ce cas, *regression tree*). Par rapport aux régressions linéaires, ces derniers ont l'avantage de supporter des non-linéarités et de mieux gérer la colinéarité des données, tout en étant facilement interprétables et rapides à exécuter. Les VE associées aux méta-chemins peuvent être prises comme *features* des arbres de décision, permettant d'intuitivement prendre en compte la sémantique de ces variables. Afin d'aller un peu plus loin, nous pourrions aussi utiliser des modèles de substitution (*surrogate models*). Un modèle de substitution est un modèle interprétable qui est formé pour approximer les prédictions d'un modèle "black-box" (e.g. réseaux de neurones, *gradient boosting*). Dit autrement, l'idée est de tirer des conclusions sur le modèle "black-box" en interprétant le modèle de substitution (en

[1] `https://www.netflixprize.com/`
[2] `https://www.kaggle.com/competitions`

général, il s'agit d'un modèle "white-box" tels que les régressions linéaires ou les arbres de décision).

Ainsi que déjà mentionné au début de ce travail, corrélation n'implique pas causalité (*cum hoc sed non propter hoc*). Il est certes déjà informatif de connaître et de quantifier les corrélations au sein d'un système ou plus précisément, les corrélations présentes dans la représentation d'un système, mais être capable d'identifier les causes et effets est bien plus informatif. Étant donné que la question d'une "cause" est profondément philosophique, nous ne proposons qu'une idée concernant la "causalité prédictive". Le concept de chemins temporels évoquée dans les perpectives du chapitre 3, et directement applicable à la méthode proposée, va dans ce sens. Bien que des chemins respectant le temps ne soient pas suffisants pour parler de cause (*post hoc, ergo propter hoc*), il s'agit néanmoins d'une première étape.

Finalement, avec une remarque d'un autre ordre, être capable d'identifier les causes et conséquences dans des données dépend souvent de la représentation que l'on s'en fait, ainsi que des outils d'analyse utilisés. Par exemple, certaines informations cruciales peuvent être perdues lors de la modélisation (dans les données utilisées au chapitre 3, du fait de la modélisation HIN, nous ne pouvions dire, e.g. si deux hashtags faisaient partie d'un même tweet ou non). Avec l'explosion de la science des réseaux ces dernières décennies dans toute sorte de disciplines, le nombre de représentations, approches, méthodes d'analyses, vocabulaires, etc., ont également explosé. Il est dès lors bien souvent ardu de s'y retrouver, surtout lorsqu'on sort de sa discipline. Par exemple, dans le chapitre 1, nous avons vu que les graphes *multiplexes* étaient référencés par beaucoup d'autres termes. À l'inverse, nous avons vu qu'un même terme pouvait faire référence à deux choses complètement différentes : les interactions d'*ordres supérieurs* (*higher-order interactions*) font aussi bien référence aux interactions avec mémoire (e.g. des chemins) qu'aux interactions impliquant un nombre quelconque d'entités (e.g. complexes simpliciaux, hyperliens). Dès lors, comprendre les hypothèses de chaque formalisme, ce qui les différencie les uns des autres, mais aussi ce qui les relie et ce qu'ils impliquent et ce, afin d'utiliser le formalisme et outils *a priori* les plus adaptés lorsqu'on désire analyser des systèmes et de fournir un unique "langage" serait un travail colossal, extrêmement précieux mais aussi fort peu probable. Néanmoins, cette remarque invite à une réflexion sur le développement de la science des réseaux, et l'importance de la communication et de l'interdisciplinarité dans la recherche en général.

# Références


1. Abbott, D. The Reasonable Ineffectiveness of Mathematics [Point of View]. *Proceedings of the IEEE* **101,** 2147 –2153 (Oct. 2013) (cité en page 82).
2. Abebe, R., Adamic, L. A. & Kleinberg, J. M. Mitigating Overexposure in Viral Marketing. in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018* (eds McIlraith, S. A. & Weinberger, K. Q.) (AAAI Press, 2018), 241–248 (cité en page 102).
3. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social networks* **25,** 211–230 (2003) (cité en page 28).
4. Aicher, C., Jacobs, A. Z. & Clauset, A. Learning latent block structure in weighted networks. *Journal of Complex Networks* **3,** 221–248. issn: 2051-1329 (2014) (cité en pages 25 et 32).
5. Aleta, A. & Moreno, Y. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics* **10,** 45–62 (2019) (cité en pages 6 et 7).
6. Alisic, E. & Letschert, R. M. Fresh eyes on the European refugee crisis. in *European journal of psychotraumatology* (2016) (cité en page 60).
7. Allard, A., Noël, P.-A., Dubé, L. J. & Pourbohloul, B. Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics. *Phys. Rev. E* **79.** issn: 1550-2376 (2009) (cité en page 7).
8. Arlot, S., Celisse, A., *et al.* A survey of cross-validation procedures for model selection. *Statistics surveys* **4,** 40–79 (2010) (cité en page 41).
9. Axelrod, R. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution* **41,** 203–226 (1997) (cité en page 83).
10. Bahulkar, A., Szymanski, B. K., Chan, K. & Lizardo, O. Coevolution of a multilayer node-aligned network whose layers represent different social relations. *Computational social networks* **4,** 11 (2017) (cité en page 6).
11. Barrat, A., Barthelemy, M. & Vespignani, A. *Dynamical processes on complex networks* (Cambridge university press, 2008) (cité en page 5).
12. Battiston, F., Nicosia, V. & Latora, V. Efficient exploration of multiplex networks. *New Journal of Physics* **18,** 043035 (2016) (cité en page 6).
13. Battiston, F. *et al.* Networks beyond pairwise interactions: structure and dynamics. *Physics Reports* (2020) (cité en page 78).
14. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353,** 163–166 (2016) (cité en page 10).
15. Berge, C. Graphs and hypergraphs (1973) (cité en page 11).
16. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A. & Pedreschi, D. Foundations of multidimensional network analysis. in *2011 international conference on advances in social networks analysis and mining* (2011), 485–489 (cité en page 6).
17. Bibal, A. & Frénay, B. Interpretability of machine learning models and representations: an introduction. in *ESANN* (2016) (cité en page 114).
18. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (2008) (cité en page 60).

19. Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Physics Reports* **544.** The structure and dynamics of multilayer networks, 1 –122. issn: 0370-1573 (2014) (cité en page 7).
20. Breiman, L. *et al.* Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* **16,** 199–231 (2001) (cité en page 114).
21. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30.** Proceedings of the Seventh International World Wide Web Conference, 107 –117. issn: 0169-7552 (1998) (cité en page 15).
22. Bruns, A., Weller, K., Borra, E. & Rieder, B. Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management* (2014) (cité en page 59).
23. Bu, J. *et al.* Music recommendation by unified hypergraph: combining social media information and music content. in *Proceedings of the 18th ACM international conference on Multimedia* (2010), 391–400 (cité en pages 11, 34, 71, et 74).
24. Candia, J. & Mazzitello, K. I. Mass media influence spreading in social networks with community structure. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P07007 (2008) (cité en page 83).
25. Cao, B., Kong, X. & Philip, S. Y. Collective prediction of multiple types of links in heterogeneous information networks. in *2014 IEEE International Conference on Data Mining* (2014), 50–59 (cité en pages 23 et 28).
26. Cardillo, A. *et al.* Modeling the multi-layer nature of the European Air Transport Network: Resilience and passengers re-scheduling under random failures. *The European Physical Journal Special Topics* **215,** 23–33 (2013) (cité en page 6).
27. Carletti, T., Fanelli, D., Grolli, S. & Guarino, A. How to make an efficient propaganda. *EPL (Europhysics Letters)* **74,** 222 (2006) (cité en page 83).
28. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81,** 591–646 (Apr. 2009) (cité en page 83).
29. Castelló, X., Eguíluz, V. M. & Miguel, M. S. Ordering dynamics with two non-excluding options: bilingualism in language competition. *New Journal of Physics* **8,** 308 (2006) (cité en page 83).
30. Chen, J., Gao, H., Wu, Z. & Li, D. Tag Co-occurrence Relationship Prediction in Heterogeneous Information Networks. in *2013 International Conference on Parallel and Distributed Systems* (2013), 528–533 (cité en page 28).
31. Chen, J. *et al.* A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International* **130,** 104934. issn: 0160-4120 (2019) (cité en page 43).
32. Chen, W., Castillo, C. & Lakshmanan, L. V. S. *Information and Influence Propagation in Social Networks* (2013) (cité en page 29).
33. Chitra, U. & Raphael, B. J. *Random walks on hypergraphs with edge-dependent vertex weights* 2019. arXiv: 1905.08287 [cs.LG] (cité en page 17).
34. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (2008) (cité en pages 22 et 25).
35. Colaiori, F. & Castellano, C. Interplay between media and social influence in the collective behavior of opinion dynamics. *Phys. Rev. E* **92,** 042815 (4 2015) (cité en page 83).
36. Comte, A., Martineau, H. & Harrison, F. *The Positive Philosophy of Auguste Comte Bohn's philosophical library* **vol. 3** (G. Bell & sons, 1896) (cité en page 83).
37. Cui, Y., Zhang, L., Wang, Q., Chen, P. & Xie, C. Heterogeneous Network Linkage-weight Based Link Prediction in Bipartite Graph for Personalized Recommendation. *Procedia Computer Science* **91,** 953 –958. issn: 1877-0509 (2016) (cité en page 21).

38. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005,** P09008–P09008 (2005) (cité en page 67).

39. Davis, D., Lichtenwalter, R. & Chawla, N. V. Supervised methods for multi-relational link prediction. *Social network analysis and mining* **3,** 127–141 (2013) (cité en page 28).

40. De Domenico, M., Granell, C., Porter, M. A. & Arenas, A. The physics of spreading processes in multilayer networks. *Nature Physics* **12,** 901–906 (2016) (cité en page 77).

41. Deffuant, G., Neau, D. & Amblard F.and Weisbuch, G. Mixing beliefs among interacting agents. *Advances in Complex Systems (ACS)* **03,** 87–98 (2000) (cité en pages 83 et 86).

42. DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association* **69,** 118–121 (1974) (cité en pages 85 et 86).

43. Dong, Y. *et al.* Link prediction and recommendation across heterogeneous social networks. in *2012 IEEE 12th International conference on data mining* (2012), 181–190 (cité en page 29).

44. Ducournau, A. & Bretto, A. Random walks in directed hypergraphs and application to semi-supervised image segmentation. *Computer Vision and Image Understanding* **120,** 91–102 (2014) (cité en pages 12 et 17).

45. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters a Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, Oregon, 1996), 226–231 (cité en page 66).

46. Estrada, E. & Rodríguez-Velázquez, J. A. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications* **364,** 581–594. issn: 0378-4371 (2006) (cité en page 11).

47. Fang, Y. *et al.* Semantic proximity search on graphs with metagraph-based learning. in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)* (2016), 277–288 (cité en page 26).

48. Friedkin, N. & Johnsen, E. C. Social Influence Networks and Opinion Change. *Advances in Group Processes*, 1–19 (1999) (cité en pages 85 et 86).

49. Friedkin, N. E. & Johnsen, E. C. *Social influence network theory: A sociological examination of small group dynamics* (Cambridge University Press, 2011) (cité en pages 85 et 86).

50. Fronczak, A. & Fronczak, P. Biased random walks in complex networks: The role of local navigation rules. *Physical Review E* **80.** issn: 1550-2376 (2009) (cité en page 15).

51. Fu, C. *et al.* Link Weight Prediction Using Supervised Learning Methods and Its Application to Yelp Layered Network. *IEEE Transactions on Knowledge and Data Engineering* **30,** 1507–1518 (2018) (cité en page 31).

52. Furnival, G. M. & Wilson, R. W. Regressions by leaps and bounds. *Technometrics* **16,** 499–511 (1974) (cité en page 43).

53. Galam, S. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. *Journal of Mathematical Psychology* **30,** 426 –434. issn: 0022-2496 (1986) (cité en page 83).

54. Galam, S. Minority opinion spreading in random geometry. *The European Physical Journal B-Condensed Matter and Complex Systems* **25,** 403–406 (2002) (cité en page 83).

55. Gallo, G., Longo, G., Pallottino, S. & Nguyen, S. Directed hypergraphs and applications. *Discrete applied mathematics* **42,** 177–201 (1993) (cité en pages 11 et 12).

56. Galton, F. Kinship and correlation. *The North American Review* **150,** 419–431 (1890) (cité en page 2).

57. Gandica, Y., Charmell, A., Villegas-Febres, J. & Bonalde, I. Cluster-size entropy in the Axelrod model of social influence: Small-world networks and mass media. *Physical Review E* **84,** 046109 (2011) (cité en page 83).

58. Gargiulo, F. & Gandica, Y. The Role of Homophily in the Emergence of Opinion Controversies. *Journal of Artificial Societies and Social Simulation* **20,** 8. issn: 1460-7425 (2017) (cité en page 84).

59. Gerow, A., Lou, B., Duede, E. & Evans, J. Proposing ties in a dense hypergraph of academics. in *International Conference on Social Informatics* (2015), 209–226 (cité en pages 34, 71, et 74).

60. Ghoshal, G., Zlatić, V., Caldarelli, G. & Newman, M. E. Random hypergraphs and their applications. *Phys. Rev. E* **79,** 066118 (6 2009) (cité en page 11).

61. Gilpin, L. H. *et al.* Explaining explanations: An overview of interpretability of machine learning. in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (2018), 80–89 (cité en page 114).

62. Gomes, G., Rao, V. & Neville, J. *Community detection over a heterogeneous population of non-aligned networks* 2019. arXiv: 1904.05332 [cs.SI] (cité en page 6).

63. González-Avella, J. C., Cosenza, M. G. & Tucci, K. Nonequilibrium transition induced by mass media in a model for social influence. *Phys. Rev. E* **72,** 065102 (6 2005) (cité en page 83).

64. González-Avella, J. C. *et al.* Local versus global interactions in nonequilibrium transitions: A model of social dynamics. *Phys. Rev. E* **73,** 046119 (4 2006) (cité en page 83).

65. González-Avella, J. C., Cosenza, M. G., Eguíluz, V. M. & San Miguel, M. Spontaneous ordering against an external field in non-equilibrium systems. *New Journal of Physics* **12,** 013010 (2010) (cité en page 83).

66. Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* **78,** 1360–1380 (1973) (cité en page 25).

67. Greening Jr, B. R., Pinter-Wollman, N. & Fefferman, N. H. Higher-order interactions: understanding the knowledge capacity of social groups using simplicial sets. *Current Zoology* **61,** 114–127 (2015) (cité en page 10).

68. Grover, A. & Leskovec, J. Node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, San Francisco, California, USA, 2016), 855–864. isbn: 9781450342322 (cité en page 33).

69. Guisheng, Y., Wansi, Y. & Yuxin, D. A New Link Prediction Algorithm: Node Link Strength Algorithm. in *2014 IEEE Symposium on Computer Applications and Communications* (2014), 5–9 (cité en page 26).

70. Gupta, M., Kumar, P. & Bhasker, B. DPRel: a meta-path based relevance measure for mining heterogeneous networks. *Information Systems Frontiers*, 1–17 (2017) (cité en page 26).

71. Han, J. Mining heterogeneous information networks by exploring the power of links. in *International Conference on Discovery Science* (2009), 13–30 (cité en page 6).

72. Hasan, M. A. & Zaki, M. J. in *Social Network Data Analytics* (ed Aggarwal, C. C.) 243–275 (Springer US, Boston, MA, 2011). isbn: 978-1-4419-8462-3 (cité en page 26).

73. He, J., Bailey, J. & Zhang, R. Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks. in *International Conference on Database Systems for Advanced Applications* (2014), 141–155 (cité en page 27).

74. Hegselmann, R. & Krause, U. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation* **5,** 1–24 (2002) (cité en pages 83 et 86).

75. Hobbes, T. & Gaskin, J. *Leviathan* isbn: 9780192834980 (Oxford University Press, 1996) (cité en page 83).

76. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: First steps. *Social networks* **5,** 109–137 (1983) (cité en pages 25 et 32).

77. Holley, R. A. & Liggett, T. M. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. *The Annals of Probability* **3,** 643–663. issn: 00911798 (1975) (cité en page 83).

78. Hou, Y. & Holder, L. B. *Link weight prediction with node embeddings* (cité en page 33).

79. Hou, Y. & Holder, L. B. Deep learning approach to link weight prediction. in *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), 1855–1862 (cité en page 33).

80. Hsieh, H.-P. & Li, C.-T. Mining temporal subgraph patterns in heterogeneous information networks. in *2010 IEEE Second International Conference on Social Computing* (2010), 282–287 (cité en page 56).

81. Hu, H. Competing opinion diffusion on social networks. *Royal Society Open Science* **4** (2017) (cité en page 83).

82. Huang, Z. *et al.* Meta structure: Computing relevance in large heterogeneous information networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1595–1604 (cité en page 26).

83. Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nature communications* **10,** 1–9 (2019) (cité en page 10).

84. Ibe, O. C. *Elements of Random Walk and Diffusion Processes* 1st. isbn: 1118618092 (Wiley Publishing, 2013) (cité en page 15).

85. Ide, K., Zamami, R. & Namatame, A. Diffusion Centrality in Interconnected Networks. *Procedia Computer Science* **24.** 17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013, 227 –238. issn: 1877-0509 (2013) (cité en page 6).

86. Jager, W. & Amblard, F. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory* **10,** 295–303 (2005) (cité en page 83).

87. Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N. & Perc, M. Link prediction in multiplex online social networks. *Royal Society open science* **4,** 160863 (2017) (cité en page 29).

88. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning* (Springer, 2013) (cité en pages 42 et 43).

89. Johnston, W. A. & Dark, V. J. Selective Attention. *Annual Review of Psychology* **37,** 43–75 (1986) (cité en page 81).

90. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Physical Review E* **83.** issn: 1550-2376 (2011) (cité en page 33).

91. Kim, M. & Leskovec, J. The network completion problem: Inferring missing nodes and edges in networks. in *Proceedings of the 2011 SIAM International Conference on Data Mining* (2011), 47–58 (cité en page 22).

92. Kivelä, M. *et al.* Multilayer networks. *Journal of Complex Networks* **2,** 203–271. issn: 2051-1310 (July 2014) (cité en pages 6 et 7).

93. Klamt, S., Haus, U.-U. & Theis, F. Hypergraphs and Cellular Networks. *PLOS Computational Biology* **5,** 1–6 (May 2009) (cité en page 10).

94. Kumar, S., Spezzano, F., Subrahmanian, V. S. & Faloutsos, C. Edge Weight Prediction in Weighted Signed Networks. in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (2016), 221–230 (cité en page 34).

95. Kunegis, J. & Lommatzsch, A. Learning Spectral Graph Transformations for Link Prediction. in *Proceedings of the 26th Annual International Conference on Machine Learning* (Association for Computing Machinery, Montreal, Quebec, Canada, 2009), 561–568. isbn: 9781605585161 (cité en page 34).

96. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nature physics* **15,** 313–320 (2019) (cité en page 77).

97. Lao, N. & Cohen, W. W. Relational retrieval using a combination of path-constrained random walks. *Machine learning* **81,** 53–67 (2010) (cité en page 26).

98. Latapy, M., Viard, T. & Magnien, C. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining* **8,** 61 (2018) (cité en page 77).

99. Leydesdorff, L. & Vaughan, L. Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and technology* **57,** 1616–1628 (2006) (cité en page 27).

100. Li, D., Xu, Z., Li, S. & Sun, X. Link Prediction in Social Networks Based on Hypergraph. in *Proceedings of the 22nd International Conference on World Wide Web* (Association for Computing Machinery, Rio de Janeiro, Brazil, 2013), 41–42. isbn: 9781450320382 (cité en pages 34, 71, et 74).

101. Li, J., Zhang, L., Meng, F. & Li, F. Recommendation Algorithm based on Link Prediction and Domain Knowledge in Retail Transactions. *Procedia Computer Science* **31.** 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014, 875 –881. issn: 1877-0509 (2014) (cité en page 21).

102. Li, Y., Lou, Z., Shi, Y. & Han, J. Temporal motifs in heterogeneous information networks. in *MLG Workshop KDD* (2018) (cité en page 56).

103. Liao, H., Zeng, A. & Zhang, Y.-C. Predicting missing links via correlation between nodes. *Physica A: Statistical Mechanics and its Applications* **436,** 216 –223. issn: 0378-4371 (2015) (cité en page 26).

104. Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *Journal of the American society for information science and technology* **58,** 1019–1031 (2007) (cité en pages 22 et 26).

105. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390,** 1150 –1170. issn: 0378-4371 (2011) (cité en page 26).

106. Macskassy, S. A. On the study of social interactions in twitter. in *Sixth International AAAI Conference on Weblogs and Social Media* (2012) (cité en page 48).

107. Martínez, V., Berzal, F. & Cubero, J.-C. A Survey of Link Prediction in Complex Networks. *ACM Comput. Surv.* **49.** issn: 0360-0300 (Dec. 2016) (cité en page 26).

108. Martins, T. V., Pineda, M. & Toral, R. Mass media and repulsive interactions in continuous-opinion dynamics. *EPL (Europhysics Letters)* **91,** 48003 (2010) (cité en page 83).

109. Mayfield, M. M. & Stouffer, D. B. Higher-order interactions capture unexplained complexity in diverse communities. *Nature ecology & evolution* **1,** 1–7 (2017) (cité en page 10).

110. McInnes, L. & Healy, J. Accelerated Hierarchical Density Based Clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42 (2017) (cité en page 66).

111. Meng, X., Shi, C., Li, Y., Zhang, L. & Wu, B. Relevance measure in large-scale heterogeneous networks. in *Asia-Pacific Web Conference* (2014), 636–643 (cité en pages 27 et 39).

112. Mikolov, T., Yih, W.-t. & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Atlanta, Georgia, June 2013), 746–751 (cité en page 33).

113. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002) (cité en page 56).

114. Moore, T. & Zirnsak, M. Neural Mechanisms of Selective Visual Attention. *Annual Review of Psychology* **68.** PMID: 28051934, 47–72 (2017) (cité en page 81).

115. Morice, A. Situation actuelle des migrations internationales: réalités et controverses. *L'information psychiatrique* **91,** 207–215 (2015) (cité en page 60).

116. Najari, S., Salehi, M., Ranjbar, V. & Jalili, M. Link prediction in multiplex networks based on interlayer similarity. *Physica A: Statistical Mechanics and its Applications* **536,** 120978. issn: 0378-4371 (2019) (cité en page 29).

117. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67,** 026126 (2 2003) (cité en page 7).

118. Newman, M. *Networks* (Oxford university press, 2018) (cité en pages 5 et 15).

119. Newman, M. J. A measure of betweenness centrality based on random walks. *Social Networks* **27,** 39 –54. issn: 0378-8733 (2005) (cité en page 15).

120. Noh, J. D. & Rieger, H. Random Walks on Complex Networks. *Physical Review Letters* **92.** issn: 1079-7114 (2004) (cité en page 15).

121. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22,** 1345–1359 (2009) (cité en page 29).

122. Paranjape, A., Benson, A. R. & Leskovec, J. Motifs in Temporal Networks. in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, Cambridge, United Kingdom, 2017), 601–610. isbn: 9781450346757 (cité en page 56).

123. Patania, A., Petri, G. & Vaccarino, F. The shape of collaborations. *EPJ Data Science* **6,** 18 (2017) (cité en page 10).

124. Pearson, K. *The grammar of science* (Walter Scott, London, 1892) (cité en page 2).

125. Peres, L. R. & Fontanari, J. F. The media effect in Axelrod's model explained. *EPL (Europhysics Letters)* **96,** 38004 (2011) (cité en page 83).

126. Petrovic, L. V. & Scholtes, I. *Counting Causal Paths in Big Times Series Data on Networks* 2019. arXiv: 1905.11287 [cs.SI] (cité en page 77).

127. Plude, D. J., Enns, J. T. & Brodeur, D. The development of selective attention: A life-span overview. *Acta psychologica* **86,** 227–272 (1994) (cité en page 81).

128. Pons, P. & Latapy, M. Computing Communities in Large Networks Using Random Walks. in *Computer and Information Sciences - ISCIS 2005* (eds Yolum, p., Güngör, T., Gürgen, F. & Özturan, C.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005), 284–293 (cité en page 15).

129. Pujari, M. & Kanawati, R. Link prediction in multiplex networks. *Networks & Heterogeneous Media* **10,** 17 (2015) (cité en page 29).

130. Quattrociocchi, W., Caldarelli, G. & Scala, A. Opinion dynamics on interacting networks: media competition and social influence. *Scientific Reports* **4** (2014) (cité en page 84).

131. Radillo-Díaz, A., Pérez, L. A. & del Castillo-Mussot, M. Axelrod models of social influence with cultural repulsion. *Physical Review E* **80,** 066107 (2009) (cité en page 83).

132. Rezaeipanah, A., Ahmadi, G. & Matoori, S. S. A classification approach to link prediction in multiplex online ego-social networks. *Social Network Analysis and Mining* **10,** 27 (2020) (cité en page 29).

133. Rodríguez, A. H. & Moreno, Y. Effects of mass media action on the Axelrod model with social influence. *pre* **82,** 016111 (July 2010) (cité en page 83).

134. Rodriguez, M. A. & Shinavier, J. Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics* **4,** 29–41 (2010) (cité en page 6).

135. Rossi, R. A. *et al. Heterogeneous Network Motifs* 2019. arXiv: 1901.10026 [cs.SI] (cité en page 56).

136. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105,** 1118–1123. issn: 0027-8424 (2008) (cité en page 15).

137. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **290,** 2323–2326 (2000) (cité en page 33).

138. Salzarulo, L. A Continuous Opinion Dynamics Model Based on the Principle of Meta-Contrast. *Journal of Artificial Societies and Social Simulation* **9,** 13. issn: 1460-7425 (2006) (cité en page 83).

139. Schelling, T. C. Models of segregation. *The American Economic Review* **59,** 488–493 (1969) (cité en page 82).

140. Schelling, T. C. Dynamic models of segregation. *Journal of mathematical sociology* **1,** 143–186 (1971) (cité en page 82).

141. Scholtes, I. *et al.* Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature communications* **5,** 1–9 (2014) (cité en page 77).

142. Shakibian, H. & Charkari, N. M. Mutual information model for link prediction in heterogeneous complex networks. *Scientific Reports* **7** (2017) (cité en pages 28 et 50).

143. Shakibian, H. & Charkari, N. M. Statistical similarity measures for link prediction in heterogeneous complex networks. *Physica A: Statistical Mechanics and its Applications* **501,** 248 –263. issn: 0378-4371 (2018) (cité en page 28).

144. Shakibian, H., Charkari, N. M. & Jalili, S. Multi-kernel one class link prediction in heterogeneous complex networks. *Applied Intelligence* **48,** 3411–3428 (2018) (cité en page 29).

145. Sharma, A., Srivastava, J. & Chandra, A. *Predicting Multi-actor collaborations using Hypergraphs* 2014. arXiv: 1401.6404 [cs.SI] (cité en pages 34 et 78).

146. Sharma, S. & Singh, A. An efficient method for link prediction in weighted multiplex networks. *Computational social networks* **3,** 7 (2016) (cité en page 29).

147. Shi, C., Kong, X., Huang, Y., Philip, S. Y. & Wu, B. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering* **26,** 2479–2492 (2014) (cité en pages 27 et 50).

148. Shi, C., Li, Y., Zhang, J., Sun, Y. & Philip, S. Y. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* **29,** 17–37 (2016) (cité en pages 6, 9, et 15).

149. Skardal, P. S. & Adhikari, S. Dynamics of Nonlinear Random Walks on Complex Networks. *Journal of Nonlinear Science* **29,** 1419–1444. issn: 1432-1467 (2018) (cité en page 15).

150. Spitz, A. *et al.* Heterogeneous Subgraph Features for Information Networks. in *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)* (Association for Computing Machinery, Houston, Texas, 2018). isbn: 9781450356954 (cité en pages 56 et 76).

151. Srilatha, P. & Manjula, R. Similarity index based link prediction algorithms in social networks: A survey. *Journal of Telecommunications and Information Technology* (2016) (cité en page 26).

152. Starnini, M., Baronchelli, A., Barrat, A. & Pastor-Satorras, R. Random walks on temporal networks. *Phys. Rev. E* **85,** 056115 (5 2012) (cité en page 15).

153. Sun, Y. & Han, J. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* **3,** 1–159 (2012) (cité en pages 6 et 15).

154. Sun, Y. & Han, J. Mining heterogeneous information networks: a structural analysis approach. *Acm Sigkdd Explorations Newsletter* **14,** 20–28 (2013) (cité en pages 1 et 6).

155. Sun, Y. *et al.* RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (Association for Computing Machinery, Saint Petersburg, Russia, 2009), 565–576. isbn: 9781605584225 (cité en page 1).

156. Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. Co-author relationship prediction in heterogeneous bibliographic networks. in *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), 121–128 (cité en pages 26, 27, 28, 39, 50, 52, et 78).

157. Sun, Y., Han, J., Yan, X., Yu, P. S. & Wu, T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* **4,** 992–1003 (2011) (cité en pages 27, 50, et 52).

158. Sun, Y., Han, J., Aggarwal, C. C. & Chawla, N. V. When Will It Happen? Relationship Prediction in Heterogeneous Information Networks. in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, Seattle, Washington, USA, 2012), 663–672. isbn: 9781450307475 (cité en page 52).

159. Symeonidis, P., Tiakas, E. & Manolopoulos, Y. Transitive Node Similarity for Link Prediction in Social Networks with Positive and Negative Links. in *Proceedings of the Fourth ACM Conference on Recommender Systems* (ACM, Barcelona, Spain, 2010), 183–190. isbn: 978-1-60558-906-0 (cité en page 26).

160. Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* **107,** 13636–13641 (2010) (cité en page 6).

161. Tamagnini, P., Krause, J., Dasgupta, A. & Bertini, E. Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. in *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (Association for Computing Machinery, Chicago, IL, USA, 2017). isbn: 9781450350297 (cité en page 114).

162. Tang, J. *et al.* ArnetMiner: Extraction and Mining of Academic Social Networks. in *KDD'08* (2008), 990–998 (cité en page 51).

163. Tang, L., Liu, H., Zhang, J. & Nazeri, Z. Community Evolution in Dynamic Multi-Mode Networks. in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, Las Vegas, Nevada, USA, 2008), 677–685. isbn: 9781605581934 (cité en page 7).

164. Timothy, J. J. *How does propaganda influence the opinion dynamics of a population ?* Mar. 2017. arXiv: 1703.10138 [physics.soc-ph] (cité en page 83).

165. Torres, L., Blevins, A. S., Bassett, D. S. & Eliassi-Rad, T. *The why, how, and when of representations for complex systems* 2020. arXiv: 2006.02870 [cs.SI] (cité en page 78).

166. Varga, I. Weighted multiplex network of air transportation. *The European Physical Journal B* **89,** 139 (2016) (cité en page 6).

167. Vazquez, F., Krapivsky, P. L. & Redner, S. Constrained opinion dynamics: freezing and slow evolution. *Journal of Physics A: Mathematical and General* **36,** L61 (2003) (cité en page 83).

168. Vazquez, F., Eguíluz, V. M. & Miguel, M. S. Generic Absorbing Transition in Coevolution Dynamics. *Phys. Rev. Lett.* **100,** 108702 (10 2008) (cité en page 83).

169. Wang, P., Xu, B., Wu, Y. & Zhou, X. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* **58,** 1–38 (2015) (cité en page 21).

170. Wang, S. & Fu, X. Opinion Dynamics on Online-offline Interacting Networks: Media Influence and Antagonistic Interaction. in *Proceedings of the 8th ACM International Workshop on Hot Topics in Planet-scale mObile Computing and Online Social neTworking* (ACM, Paderborn, Germany, 2016), 55–60. isbn: 978-1-4503-4344-2 (cité en page 84).

171. Weisbuch, G., Deffuant, G., Amblard, F. & Nadal, J.-P. Meet, discuss, and segregate! *Complexity* **7,** 55–63 (2002) (cité en page 83).

172. White, L. A. Sociology, Physics and Mathematics. *American Sociological Review* **8,** 373–379. issn: 00031224 (1943) (cité en page 82).

173. Xu, Q.-S. & Liang, Y.-Z. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **56,** 1–11 (2001) (cité en page 41).

174. Yang, Y., Chawla, N., Sun, Y. & Hani, J. Predicting links in multi-relational and heterogeneous networks. in *2012 IEEE 12th international conference on data mining* (2012), 755–764 (cité en pages 6 et 29).

175. Yao, K., Mak, H. F., *et al.* PathSimExt: revisiting PathSim in heterogeneous information networks. in *International Conference on Web-Age Information Management* (2014), 38–42 (cité en pages 27 et 50).

176. Yoon, S.-e., Song, H., Shin, K. & Yi, Y. How Much and When Do We Need Higher-order Informationin Hypergraphs? A Case Study on Hyperedge Prediction. *Proceedings of The Web Conference 2020* (2020) (cité en pages 34 et 78).

177. Youness, G. & Saporta, G. Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée* **52,** 97–120 (2004) (cité en page 67).

178. Young, J.-G., Petri, G. & Peixoto, T. P. *Hypergraph reconstruction from network data* 2020. arXiv: 2008.04948 [cs.SI] (cité en page 79).

179. Yu, X., Gu, Q., Zhou, M. & Han, J. Citation Prediction in Heterogeneous Bibliographic Networks. in *SDM* (SIAM / Omnipress, 2012), 1119–1130. isbn: 978-1-61197-282-5 (cité en page 28).

180. Zhang, M., Cui, Z., Jiang, S. & Chen, Y. Beyond Link Prediction: Predicting Hyperlinks in Adjacency Space. in *AAAI* (2018) (cité en page 34).

181. Zhao, J. *et al.* Prediction of links and weights in networks by reliable routes. *Scientific reports* **5,** 12261 (2015) (cité en pages 25, 30, 31, et 32).

182. Zhong, E., Fan, W., Zhu, Y. & Yang, Q. Modeling the Dynamics of Composite Social Networks. in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, Chicago, Illinois, USA, 2013), 937–945. isbn: 9781450321747 (cité en page 6).

183. Zhou, D., Huang, J. & Schölkopf, B. Learning with Hypergraphs: Clustering, Classification, and Embedding. in *Proceedings of the 19th International Conference on Neural Information Processing Systems* (MIT Press, Canada, 2006), 1601–1608 (cité en pages 17 et 34).

184. Zhou, K., Michalak, T. P., Waniek, M., Rahwan, T. & Vorobeychik, Y. Attacking Similarity-Based Link Prediction in Social Networks. in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, Montreal QC, Canada, 2019), 305–313. isbn: 978-1-4503-6309-9 (cité en page 26).

185. Zhou, Q. & Leydesdorff, L. The normalization of occurrence and o-occurrence matrices in bibliometrics using Cosine similarities and Ochiai coefficients. *Journal of the Association for Information Science and Technology* **67,** 2805–2814 (2016) (cité en page 27).

186. Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *The European Physical Journal B* **71,** 623–630. issn: 1434-6036 (2009) (cité en page 31).

187. Zhou, Y., Huang, J., Sun, H. & Sun, Y. Recurrent meta-structure for robust similarity measure in heterogeneous information networks. *arXiv preprint.* arXiv:1712.09008 (2017) (cité en page 26).

188. Zhou, Y. *et al.* A semantic-rich similarity measure in heterogeneous information networks. *Knowledge-Based Systems* **154,** 32–42 (2018) (cité en page 66).

189. Zhu, B., Xia, Y. & Zhang, X.-J. Weight prediction in complex networks based on neighbor set. *Scientific reports* **6,** 38080 (2016) (cité en page 31).

190. Zhu, X., Tian, H., Cai, S. & Zhou, T. Erratum: Predicting missing links via significant paths. *EPL (Europhysics Letters)* **108,** 49901 (2014) (cité en page 26).

191. Zhu, X., Tian, H., Cai, S., Huang, J. & Zhou, T. Predicting missing links via significant paths. *EPL (Europhysics Letters)* **106,** 18008 (2014) (cité en page 26).