Modélisation d'une variable aléatoire à l'aide d'un réseau

Mémoire

Heikel Jarras

Maîtrise en statistique - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

Modélisation d'une variable aléatoire à l'aide d'un réseau

Mémoire

Heikel Jarras

Sous la direction de:

Louis-Paul Rivest, directeur de recherche
Antoine Allard, codirecteur de recherche

Résumé

Le domaine de l'assurance regorge de toutes sortes de données. Avec des milliers, voire des millions de clients, les compagnies d'assurance ont su emmagasiner un nombre impressionnant d'informations. À partir de celles-ci, elles sont en mesure de développer plusieurs modèles qui leur permettent d'anticiper le comportement de leur clientèle. Elles ont maintenant à leur disposition des modèles qui permettent d'estimer le temps restant avant qu'un client n'abandonne une police d'assurance de dommages. Une compagnie d'assurance souhaite cependant approfondir ses connaissances et améliorer ses prévisions en étudiant l'influence des relations entre les clients sur l'abandon d'une police d'assurance. Certaines données descriptives des clients sont disponibles ainsi que cinq fichiers qui lient les individus à des identifiants de groupe. Ces derniers sont utilisés pour créer des réseaux représentant les relations qui existent entre les clients de la compagnie.

L'objectif de ce mémoire est donc d'explorer les données réseaux et de comprendre l'impact que les relations peuvent avoir sur certaines variables, plus particulièrement sur l'abandon d'une police d'assurance de dommages. Des statistiques descriptives en lien avec les réseaux, comme le nombre de liens entre deux individus qui abandonnent ou l'assortativité, permettent rapidement de savoir s'il est pertinent de continuer l'exploration ou non. Par la suite, un test de permutation permet de mieux comprendre l'influence des relations sur le fait qu'un client abandonne ou non. Puis, pour terminer, un modèle statistique qui permet d'estimer une matrice de covariance à partir des relations d'un réseau est présenté.

Abstract

The insurance sector is full of all kinds of data. With thousands, if not millions, of customers, insurance companies have accumulated a substantial amount of information. From this information, they can develop several models that allow them to anticipate their customer's behavior. They now have models that allow them to estimate the remaining time before a customer cancels their insurance policy. However, an insurance company wishes to deepen their understanding, and improve predictions by studying the influence of relationships between clients on the cancellation of damage insurance policies. Some descriptive data on the customers is available, as well as five files linking individuals to groups. This is how the networks are created.

The objective of this thesis is therefore to explore network data and understand the influence that relationships can have on certain descriptive variables, and more specifically on the cancellation of a damage insurance policy. Descriptive statistics related to networks, such as the number of links between two individuals who cancel or assortativity, quickly allow us to know if it is relevant to continue the exploration or not. Then, the permutation test allows us to better understand the influence of relationships on the cancellation of the insurance policy. Finally, a statistical model that allows us to estimate a covariance matrix from a network is presented.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Au nom d'Allah le tout miséricordieux le très miséricordieux. Louange à Allah qui a rendu possible la réalisation de ce mémoire.

Je tiens à remercier personnellement M. Louis-Paul Rivest qui a été un directeur de recherche exemplaire et qui m'a donné la possibilité de travailler sur ce projet de recherche. Sa disponibilité, sa capacité à expliquer les concepts statistiques et ses commentaires positifs ont été des ingrédients primordiaux pour la réalisation de ce mémoire. Son ouverture à explorer de nouvelles méthodes a été d'une grande aide et je lui en suis reconnaissant.

Je souhaite également remercier mon codirecteur de recherche, M. Antoine Allard. Ce mémoire n'aurait pas pu être réalisé sans son expertise sur les réseaux. Sa patience, sa disponibilité à m'expliquer des concepts théoriques et ses commentaires constructifs ont été d'une immense aide et je le remercie énormément.

J'aimerai également remercier M. Thierry Duchesne, responsable du projet, qui, grâce à son travail, a rendu possible cette recherche. Je ne peux passer sous le silence également les collaborateurs de la compagnie d'assurance qui ont toujours été présents pour répondre à mes questions.

Ce mémoire est l'aboutissement de plusieurs années d'études à l'Université Laval et j'en profite pour remercier tout le corps professoral avec qui j'ai partagé les classes de cours.

J'aimerai dire un gros merci à mes amis et toutes les personnes qui, de près ou de loin, ont rendu possible ce travail.

Finalement, j'aimerai également remercier du fond du cœur ma famille particulièrement ma mère, mon père et mes deux sœurs. Ce mémoire n'aurait pas pu être possible sans leur soutien et l'amour qu'ils me donnent au quotidien.

Introduction

Dans le domaine de l'assurance, la concurrence est rude. Chaque compagnie fait preuve de créativité pour offrir les produits les plus avantageux et ainsi gagner de nouvelles parts de marché. Étant ardemment gagné, les compagnies d'assurance doivent aussi prendre soin de chaque client au risque de le voir quitter pour un concurrent direct. Pour freiner cela, les compagnies d'assurance tentent de comprendre les causes qui peuvent mener à l'abandon d'un produit par un de ses clients. Elles cherchent également à prédire le temps restant avant qu'un abandon se produise. Elles peuvent ainsi se concentrer sur les clients les plus à risque de quitter en leur proposant par exemple des produits plus adaptés à leur réalité.

Lorsque l'on veut prédire le temps restant avant la survenue d'un évènement, on parle alors d'analyse de survie. C'est un domaine largement étudié en statistique depuis plusieurs décennies (Klein and Moeschberger, 2003). Une des avancées les plus importantes en analyse de survie est le modèle de Cox (Cox, 1972). Celui-ci est, encore à ce jour, un des plus utilisés lorsqu'on veut analyser le temps restant avant qu'un évènement se produise.

Dans certains cas, la variable de survie étudiée peut être dépendante d'une ou plusieurs variables longitudinales. Sa valeur varie alors dans le temps. C'est souvent le cas en médecine, par exemple, lorsqu'on calcule des taux de certaines composantes dans le corps. Le modèle de Cox n'est pas adapté pour bien composer avec ces variables. Un autre type de modèle, nommé modèle conjoint, est alors mieux équipé pour gérer les variables longitudinales. Introduit vers la fin des années 90 (Wulfsohn and Tsiatis, 1997), il permet de modéliser une variable de survie ainsi qu'une ou plusieurs variables longitudinales en même temps. Au cours de la dernière décennie, les travaux du professeur Dimitris Rizopoulos (Rizopoulos, 2012) ont permis d'offrir une meilleure visibilité à la modélisation conjointe avec notamment le développement de la librairie JM (Rizopoulos, 2010) dans le langage R, rendant ainsi son utilisation simple et accessible.

Au-delà des variables de survie et longitudinales, la compagnie d'assurance possède également des données permettant de créer des réseaux de clients. C'est le sujet qui est traité dans ce mémoire. L'objectif est d'offrir une bonne introduction au concept de réseau et plus particulièrement à la structure de réseau biparti. Certains tests statistiques sont présentés pour permettre de comprendre l'influence que les relations peuvent avoir sur certaines variables me-

surées sur les nœuds d'un réseau. Puis, un modèle statistique permettant l'estimation d'une matrice de covariance à partir de la matrice d'adjacence d'un réseau est également introduit. Les travaux de ce mémoire visent à donner une piste de solution quant à l'inclusion des réseaux dans les modèles conjoints de données longitudinales et de survie.

Le concept de réseau est étudié depuis de nombreuses décennies. Les premiers travaux étaient reliés à des concepts sociologiques durant les années 1930 lorsque Moreno, 1934 présente le concept de sociogramme. Par la suite, durant les années, 60 et 70, Granovetter, 1973 étudie l'impact que peuvent avoir les liens faibles dans la diffusion de l'information et des occasions d'emploi. Voyant l'importance que peuvent avoir les réseaux, des chercheurs de différents domaines comme l'informatique, la physique, l'économie, etc. commencent à s'y intéresser. L'avènement de l'internet et la plus grande capacité de calcul des ordinateurs permettent également un tournant dans l'étude des réseaux. Mislove et al., 2007 étudient les réseaux en ligne créés par des sites populaires comme Flickr et YouTube.

Le domaine d'application des réseaux étant très vaste il n'est pas nécessaire de toujours mettre en relation des humains. Les travaux de Barabási and Oltvai, 2004 en biologie permettent de mieux comprendre les interactions entre les molécules et les cellules. Le domaine de l'économie a également connu son lot de recherche. Easley and Kleinberg, 2010 abordent les réseaux dans le contexte des marchés économiques. Ils examinent les mécanismes de formation, les effets des réseaux sur les comportements individuels et sur le fonctionnement des marchés. Ils explorent également les différences entre les réseaux et les autres formes d'interaction économique, telles que les contrats et les institutions.

Les différents domaines d'application influencent les propriétés des réseaux tout comme le type de réseau analysé. Dans ce mémoire, une emphase particulière est mise sur les réseaux bipartis. La particularité d'un réseau biparti est qu'il est constitué de deux types de nœuds. Un premier type représentant des individus ou des entités, et un second type représentant des groupes auxquels ils appartiennent. Son utilisation devient très pratique pour représenter les relations existantes entre différents éléments d'un système complexe. Les travaux de Newman, 2010 explorent les propriétés des réseaux bipartis et proposent certaines méthodes pour les étudier, comme la projection unimodale. Ce type de réseau peut lui aussi être utilisé dans différents domaines et celui de l'assurance en fait partie. Il peut aider à mettre en relation des clients en les liant à une situation géographique, un lieu de travail, un numéro de téléphone, etc.

Pour être en mesure de bien comprendre la structure d'un réseau et les relations qui le constituent, plusieurs métriques ont été développées. La densité, la centralité, la transitivité, la distribution des degrés, les composantes sont toutes des notions étudiées par Barabási, 2016 et qui permettent d'avoir une meilleure compréhension du réseau analysé. Les travaux de Erdős and Rényi, 1959 sur les réseaux aléatoires ont fourni les fondements théoriques de ces

concepts, ce qui mena par la suite à une meilleure compréhension des réseaux complexes.

La mesure d'assortativité est une autre mesure qui caractérise un réseau. Elle donne une idée sur l'influence que peuvent avoir les relations d'un réseau sur la valeur d'une variable aléatoire (Newman, 2003). Pour prendre en considération les poids des liens entre les individus, Arcagni et al., 2021 proposent une petite variante à la formule d'assortativité et Yuan et al., 2021 font le même genre de travail, mais spécifiquement pour un réseau orienté.

Après l'étude des réseaux, l'un des objectifs de ce mémoire est d'être en mesure d'extraire de l'information d'un réseau pour être, par la suite, exploité dans un modèle prédictif. Pour ce faire, l'approche choisie est d'estimer une matrice de covariance à partir d'un réseau. Ce type de méthode est assez nouveau, mais plusieurs recherches ont été faites durant les dernières décennies. Schäfer and Strimmer, 2005 développent une méthode assurant l'obtention d'une matrice définie positive. L'estimation de la matrice de covariance est faite à partir du lemme de Ledoit-Wolf (Ledoit and Wolf, 2003). Une autre démarche en deux temps est proposée par Chen et al., 2018. Tout d'abord, il faut détecter les structures de réseau latentes à partir de la matrice de corrélation d'échantillon à l'aide d'une optimisation pénalisée, puis une régularisation de la matrice de covariance est effectuée en utilisant les informations de topologie des réseaux détectées. Plus récemment, Lan et al., 2018 ont travaillé sur un modèle statistique permettant d'estimer la matrice de covariance à partir de la matrice d'adjacence d'un réseau. Cette méthode permet de réduire considérablement le nombre de paramètres à estimer. Ce modèle est notamment étudié en profondeur dans ce mémoire.

Ce présent mémoire est divisé en cinq chapitres. Le premier chapitre présente les principaux aspects théoriques en lien avec les réseaux en se concentrant sur les réseaux bipartis et aléatoires. Le second contient des tests statistiques permettant de déterminer l'influence que peut avoir un réseau sur une variable aléatoire discrète. Le troisième chapitre analyse l'approche proposée par Lan et al., 2018 pour estimer une matrice de covariance à partir d'une matrice d'adjacence. Le quatrième chapitre présente une approche alternative au modèle de Lan et al., 2018 ainsi qu'une introduction à la modélisation conjointe. Une conclusion sous forme de discussion fait un retour sur les analyses faites.

Chapitre 1

Les réseaux

Les réseaux sont des structures utilisées lorsque certaines relations existent dans les données. Ils permettent non seulement la visualisation de ces relations, mais également de prendre en considération cette dépendance lors de l'analyse des données. En général, les réseaux sont constitués de nœuds qui représentent des entités, ou des individus, ainsi que de liens, appelés arêtes, qui représentent les relations entre ces entités, ou ces individus. Les domaines d'application de ces structures sont évidemment très vastes et le lecteur intéressé par une bonne introduction peut se référer à la section 1.3 de l'essai de Sylvain-Morneau, 2021. Dans ce chapitre, les réseaux bipartis et aléatoires sont étudiés en profondeur. Une dernière section traite également des données réseau analysées dans ce mémoire.

1.1 Réseau biparti

Le réseau biparti est une structure constituée uniquement de deux types de nœuds. Le premier type représente généralement une entité, nommée l'ensemble $\mathcal{N}$, et le second représente un groupe auquel elle est liée nommé l'ensemble $\mathcal{Q}$ (Newman, 2018). De plus, une autre particularité du réseau biparti réside dans les connexions qui peuvent uniquement être présentes si elles mettent en relation deux nœuds de types différents. Donc, dans l'ensemble des liens du réseau ξ, il existe uniquement des connexions de type (n_i, q_j) où n_i indique la $i^{\text{ième}}$ entité et q_j indique le $j^{\text{ième}}$ groupe. Pour la suite de cette section et pour faciliter la compréhension, l'ensemble $\mathcal{N}$ représente des individus alors que l'ensemble $\mathcal{Q}$ représente des groupes.

Posons une population de 8 individus pouvant être membre de 7 groupes différents. Dans ce cas, les ensembles décrits précédemment sont les suivants : $\mathcal{N} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ et $\mathcal{Q} = \{A, B, C, D, E, F, G\}$. Deux individus partagent un lien, dans la projection unimodale présentée plus loin, s'ils font partie du même groupe. Le tableau 1.1 présente l'information de départ du réseau biparti.

Tableau 1.1 – Jeu de données fictif d'un réseau biparti

ID individu	ID groupe
1	A
1	B
1	E
2	B
2	E
3	D
4	A
4	E
5	D
5	F
6	C
6	F
7	C
8	G

Le tableau 1.1 met en évidence les différents groupes desquels fait partie chaque individu. Chaque ligne représente un lien dans le réseau. Avec cette représentation, il est évident que les liens du réseau biparti sont constitués des deux types nœuds (ID individu et ID groupe).

Les relations qui existent dans le réseau biparti peuvent être exprimées à l'aide d'une matrice d'incidence $\mathbf{B}$ qui est de taille $|\mathcal{Q}| \times |\mathcal{N}|$. L'élément b_{ij} prend la valeur 1 si la l'individu j fait partie du groupe i et 0 autrement. Dans l'exemple précédent, la taille de la matrice d'incidence est de 7×8. La matrice d'incidence est présentée au tableau 1.2.

Tableau 1.2 – Matrice d'incidence $\mathbf{B}$

	1	2	3	4	5	6	7	8
A	1	0	0	1	0	0	0	0
B	1	1	0	0	0	0	0	0
C	0	0	0	0	0	1	1	0
D	0	0	1	0	1	0	0	0
E	1	1	0	1	0	0	0	0
F	0	0	0	0	1	1	0	0
G	0	0	0	0	0	0	0	1

La matrice d'incidence est utile lorsqu'on s'intéresse à la taille des groupes ou bien au nombre de groupes desquels chaque individu fait partie. Il suffit donc de faire la somme des lignes pour avoir la taille de chaque groupe (éléments de l'ensemble $\mathcal{Q}$), et la somme des colonnes pour obtenir le nombre de groupes auxquels appartient chaque individu (éléments de l'ensemble $\mathcal{N}$). L'équation (1.1) présente la définition mathématique pour calculer la taille d'un groupe selon la matrice d'incidence $\mathbf{B}$,

$$t_q = \sum_{n=1}^{|\mathcal{N}|} b_{qn}, \tag{1.1}$$

où $|\mathcal{N}|$ représente la taille de l'ensemble $\mathcal{N}$. Au premier coup d'œil, certains sujets semblent être en relation puisqu'ils font partie d'un même groupe. La figure 1.1 est une représentation visuelle du réseau et il devient beaucoup plus simple de voir les liens qui existent. À l'aide de cette représentation, il est évident que les liens existent uniquement entre deux nœuds de types différents puisqu'ils unissent des individus avec des groupes.

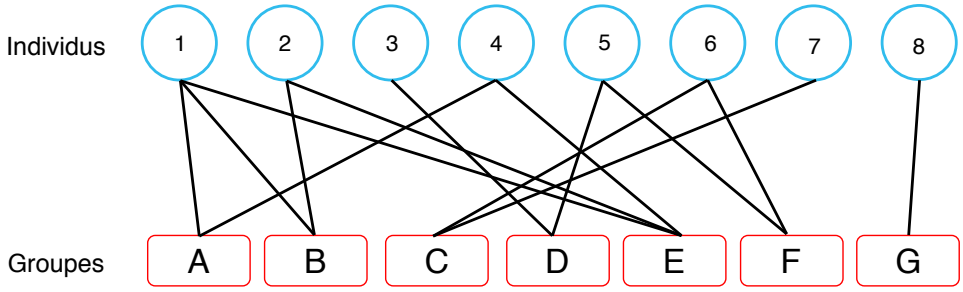


FIGURE 1.1 – Réseau biparti

1.1.1 Poids d'un lien

Comme il est mentionné au début de cette section, deux individus partagent une relation, dans la projection du réseau biparti, s'ils font partie d'un même groupe. Jusqu'à présent, avec le tableau 1.2 et la figure 1.1, il est possible de savoir qu'elles sont les paires d'individus qui sont en relation. Pour pousser l'analyse un peu plus loin et pour avoir une meilleure compréhension des interactions qui existent entre deux individus, il est intéressant de quantifier la force de ces relations. Celles-ci peuvent être définies de plusieurs façons. Dans ce mémoire, la formule utilisée est la somme de l'inverse des tailles des groupes partagés par deux personnes. Toujours à partir de la matrice d'incidence $\mathbf{B}$, le poids d'un lien entre les individus i et j s'écrit,

$$w_{ij} = \sum_{q=1}^{|\mathcal{Q}|} \frac{b_{qi} b_{qj}}{t_q}, \tag{1.2}$$

où t_q est la taille du groupe q dans le réseau biparti calculé à partir de l'équation (1.1) et $|\mathcal{Q}|$ la taille de l'ensemble $\mathcal{Q}$. Dans cette équation, chaque groupe est considéré comme étant indépendant, d'où la somme sur le nombre de groupes du réseau. L'intuition derrière cette formule est que plus la taille d'un groupe est grande, plus la force de la relation entre deux individus devrait être faible. Prenons par exemple un groupe composé de plus de 80 individus et un autre de seulement deux personnes. En général, deux individus pris au hasard dans le premier groupe devraient avoir une relation plus faible que les deux sujets du second groupe qui se côtoient constamment. Revenons maintenant à l'exemple étudié dans cette section. Le

tableau 1.3 présente tous les liens qui existent entre les sept individus ainsi que la force de ces liens calculée à partir de l'équation (1.2).

Tableau 1.3 – Force des liens entre les individus

ID Individu i	ID Individu j	w_{ij}
1	2	$1/2 + 1/3 = 5/6$
1	4	$1/2 + 1/3 = 5/6$
2	4	$1/3$
3	5	$1/2$
5	6	$1/2$
6	7	$1/2$

À partir du tableau 1.3, il est possible d'avoir une meilleure idée des relations qui existent entre les individus du jeu de données. La force du lien qui unit deux individus y est présente, ce qui permet de comparer les différentes relations d'un individu. Par exemple, le lien que partage l'individu 2 avec le 1 semble plus fort que celui qu'il partage avec l'individu 4.

1.1.2 Matrice d'adjacence et degré

La représentation des liens qui existent dans un réseau entre individus peut également être faite avec une matrice d'adjacence $\mathbf{A}$. Cette matrice est toujours carrée puisque le nombre de colonnes et le nombre de lignes représentent le nombre d'individus dans l'ensemble $\mathcal{N}$ du réseau. Donc, contrairement à la matrice d'incidence $\mathbf{B}$ la matrice d'adjacence $\mathbf{A}$ ne considère qu'un seul type de nœud. Dans le cas d'un réseau biparti, puisque les liens entre les individus sont bidirectionnels, la matrice d'adjacence est symétrique. En général, si l'objectif est uniquement de représenter les liens du réseau, l'élément a_{ij} prend la valeur 1 si un lien existe entre les individus i et j et 0 sinon. Il est également possible de mettre les poids des relations calculées au tableau 1.3. Le tableau 1.4 présente la matrice d'adjacence du réseau.

Tableau 1.4 – Matrice d'adjacence $\mathbf{A}$

	1	2	3	4	5	6	7	8
1	0	1	0	1	0	0	0	0
2	1	0	0	1	0	0	0	0
3	0	0	0	0	1	0	0	0
4	1	1	0	0	0	0	0	0
5	0	0	1	0	0	1	0	0
6	0	0	0	0	1	0	1	0
7	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	0

À partir de la matrice d'adjacence, il devient très simple de calculer les degrés des nœuds d'un réseau. Le degré d'un nœud représente le nombre de nœuds qui lui sont connectés (Newman,

2018). Pour obtenir cette mesure, il suffit de prendre la somme des lignes ou des colonnes de la matrice $\mathbf{A}$. Comme la matrice est symétrique, les résultats sont identiques. Le degré d de l'individu i peut donc être obtenu avec l'équation suivante,

$$d_i = \sum_{j=1}^{J} a_{ij}, \tag{1.3}$$

où J représente le nombre de lignes ou le nombre de colonnes de la matrice $\mathbf{A}$. En utilisant l'équation (1.3) et le tableau 1.4 on peut facilement trouver le degré des 8 individus constituant notre réseau.

Tableau 1.5 – Degré des individus

ID individu	Degré
1	2
2	2
3	1
4	2
5	2
6	2
7	1
8	0

À partir du concept de degré, il est facile d'établir des équations pour la somme ainsi que la moyenne des degrés.

Posons m le nombre total de liens présents dans un réseau. S'il est non orienté, comme c'est le cas pour les réseaux bipartis, chaque lien comprendra deux extrémités. Il y a donc au total $2m$ extrémités dans le réseau. Cette quantité est également équivalente à la somme des degrés des nœuds dont l'équation s'écrit,

$$2m = \sum_{i=1}^{n} d_i = \sum_{ij} a_{ij}. \tag{1.4}$$

Pour simplifier la notation, la lettre n est utilisée à la place de $|\mathcal{N}|$ pour désigner la taille de l'ensemble $\mathcal{N}$. À l'aide de ce résultat, la moyenne des degrés peut s'écrit comme étant,

$$c = \frac{1}{n} \sum_{i=1}^{n} d_i = \frac{2m}{n}. \tag{1.5}$$

Dans le réseau du tableau 1.1, le nombre total de liens m est donc de 6, la somme des degrés 12 et la moyenne des degrés c est de $\frac{12}{8} = 1.5$.

1.1.3 Densité

Les résultats précédents permettent maintenant de développer le concept de densité. Dans un réseau constitué de n nœuds, le nombre total de connexions possibles est de $\binom{n}{2} = \frac{1}{2}n(n-1)$. La densité d'un réseau se définit comme étant la proportion de toutes les connexions possibles qui existent réellement. Son équation peut s'écrire ainsi,

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{c}{n-1} \approx \frac{c}{n}. \tag{1.6}$$

La dernière approximation peut être faite lorsqu'on considère des réseaux avec un nombre de nœuds n très élevé. La mesure de densité ρ prend donc des valeurs entre 0 et 1. On peut voir cette proportion comme étant la probabilité qu'une paire de nœuds prise au hasard soit connectée dans le réseau.

L'équation (1.6) est celle utilisée en général dans la littérature. Par contre, pour un réseau biparti, le nombre de liens total est un peu différent puisqu'un lien peut seulement exister entre des nœuds de types différents comme il a été vu à la section 1.1. Il est défini par la multiplication des tailles des ensembles $\mathcal{Q}$ et $\mathcal{N}$. La densité s'écrit donc ainsi pour un réseau biparti,

$$\rho = \frac{m}{|\mathcal{Q}| \times |\mathcal{N}|}. \tag{1.7}$$

Un réseau est défini comme étant dense lorsque la densité demeure non nulle alors que le nombre de nœuds tend vers l'infini. Dans le cas contraire, où ρ tend vers 0 au fur et à mesure que le nombre de nœuds augmente, le réseau est alors dit creux. Le nombre de 0 dans la matrice d'adjacence A est alors très élevé par rapport au nombre total de connexions possibles.

L'augmentation du nombre de nœuds n dans un réseau peut être fait lorsqu'on travaille avec des réseaux simulés ou synthétiques. Par contre, si l'on travaille avec un réseau représentant un concept réel du monde, il peut être compliqué d'ajouter des nœuds au réseau. Toutefois, il existe certains domaines où il est possible d'augmenter la taille de n. Par exemple, lorsque le concept de l'Internet est étudié sur une longue période de temps, certains sites web, représentant les nœuds, peuvent s'ajouter au réseau. Sa taille à la fin de l'étude est donc plus grande que celle du début.

La majorité des réseaux qui sont étudiés ont tendance à être creux. En particulier les réseaux décrivant un concept social. Un réseau représentant les amis d'un grand nombre de personnes est évidemment creux. Le nombre d'amis d'un individu dépend du temps qu'il peut investir dans ses relations. Dans ce cas, lorsque le nombre de nœuds tend vers l'infini, la densité devient nulle.

1.1.4 Projection unimodale

La figure 1.2 représente une projection unimodale du réseau faite sur les individus. Une projection est le fait de passer d'une représentation bimodale d'un réseau comme à la figure 1.1 à une représentation unimodale ne comportant qu'un seul type de nœud. Elle permet de visualiser directement les interactions entre les individus. Il est évident que trois composantes distinctes forment le réseau. Parmi celles-ci, l'individu 8 se retrouve isolé puisqu'il ne partage aucun lien avec les autres membres du réseau. Dans ce cas, l'individu est dit orphelin. Les individus 1, 2 et 4 forment une composante et les individus 3, 5, 6 et 7 en forment une autre. Le concept de composante est analysé plus en détail à la sous-section 1.1.5. Le même exercice peut également être fait sur les groupes. Dans ce cas, les groupes qui partagent au moins un individu commun sont en relation. Un groupe est orphelin si les individus qui le forment ne sont présents dans aucune autre groupe.

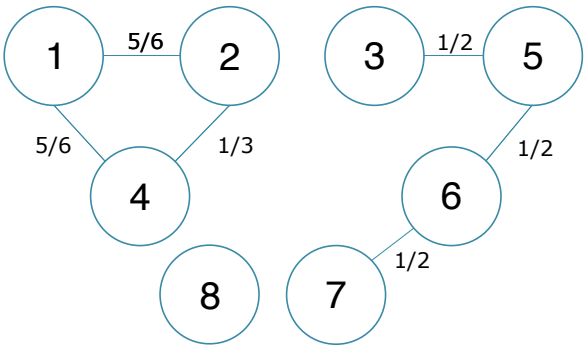


FIGURE 1.2 – Projection unimodale sur les individus

Le tableau 1.6, représente la projection unimodale sur les individus du réseau biparti. Chaque ligne représente un sujet et on y retrouve les identifiants des groupes dont il fait partie. Ce tableau est très utile pour certains calculs numériques puisqu'il permet d'avoir accès rapidement aux groupes d'un individu.

Tableau 1.6 – Projection unimodale sur les individus

ID individu	ID groupe
1	A, B, E
2	B, E
3	D
4	A, E
5	D, F
6	C, F
7	C
8	G

Encore une fois, la projection peut également se faire sur les groupes. Dans ce cas, chaque ligne représente un groupe et on y retrouve tous les individus qui le constituent. Cette représentation

est également très utile puisqu'elle permet de voir rapidement quels sont les individus qui font partie d'un même groupe. Grâce au réseau biparti, il est possible, à partir du tableau 1.1, de trouver les liens qui existent entre les différents sujets du jeu de données, d'attribuer une force à leur relation et de présenter l'information sous différentes formes ayant chacune leurs avantages.

1.1.5 Composantes d'un réseau

Il existe plusieurs façons de décrire un réseau et l'analyse des composantes fait certainement partie des plus intéressantes. Pour bien comprendre ce qu'est une composante, définissons d'abord ce qu'est un chemin. Dans un réseau, un chemin est une séquence de nœuds telle que chaque paire de nœuds consécutifs dans la séquence est reliée par un lien. Sur la figure 1.2, il est possible de passer du nœud 1 au nœud 2 de façon directe ou en passant par le nœud 4. Il existe alors deux chemins $1 \rightarrow 2$ et $1 \rightarrow 4 \rightarrow 2$ entre ces deux individus. Maintenant, posons le chemin le plus court entre deux nœuds A et B comme étant celui qui contient le plus petit nombre de liens parmi tous les chemins possibles. Selon cette définition, le chemin le plus court pour passer du nœud 1 au nœud 2 est le chemin direct puisqu'il ne contient qu'un seul lien. Pour la suite du mémoire, le terme « chemin » fera toujours référence au chemin le plus court à moins d'indication contraire. Le lecteur intéressé à avoir plus d'informations sur la recherche de chemins dans un réseau peut consulter l'algorithme de Dijkstra, 1959, qui a pour objectif de trouver tous les chemins les plus courts entre les nœuds d'un réseau.

Après avoir compris ce concept, nous pouvons maintenant définir celui de composante. Une composante est un sous-groupe des nœuds d'un réseau tel qu'il existe au moins un chemin entre chaque paire de nœuds formant ce sous-groupe. De plus, il n'est pas possible d'ajouter un autre nœud du réseau tout en préservant cette propriété. Pour bien comprendre ce concept, analysons la figure 1.3. Elle représente les composantes présentes dans le réseau de la figure 1.2.

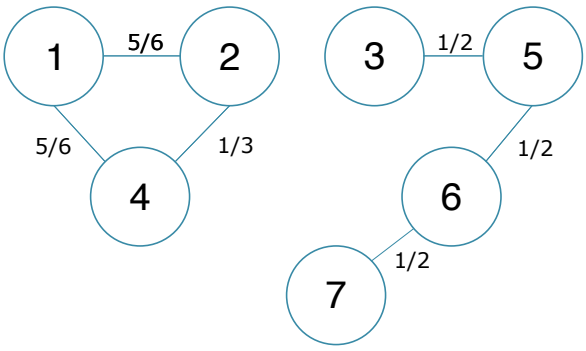


FIGURE 1.3 – Composantes du réseau

Le réseau est formé de deux composantes. L'individu 8 est exclu puisqu'il forme une com-

posante de taille 1. Les composantes peuvent être vues comme une partition de l'ensemble des individus $\mathcal{N}$. L'ensemble des composantes χ du réseau peut être écrit comme suit : $\chi = [\{1, 2, 4\}, \{3, 5, 6, 7\}]$. La notion de composante est très importante pour comprendre la structure d'un réseau. Si celui-ci est formé d'une seule composante, il est dit connexe. Cela signifie que tous les individus du réseau sont liés et ils peuvent tous potentiellement s'influencer. Par contre, si un réseau est formé de plus d'une composante, il est alors dit non connexe. Il existe alors certaines paires d'individus non liés puisqu'il n'y a pas de chemin les unissant. Par exemple, l'individu 3 peut potentiellement influencer le comportement de l'individu 5, dans le réseau, mais pas celui de l'individu 4 puisqu'ils ne font pas partie de la même composante. Pour être en mesure de quantifier cette connectivité, introduisons la quantité suivante,

$$1 - \frac{\text{Nombre de composantes du réseau}}{|\mathcal{Q}'|}. \tag{1.8}$$

Au dénominateur, $|\mathcal{Q}'|$ représente le nombre de groupes minimaux tels que la structure des liens du réseau ne change pas. En d'autres mots, l'ensemble $\mathcal{Q}'$ exclut les groupes formés uniquement avec une partie ou l'entièreté des individus d'un autre groupe. Ces groupes ainsi que les groupes orphelins sont ignorés, dans l'équation (1.8), puisqu'ils n'apportent aucune information supplémentaire sur les liens entre les individus. De plus, ils viendraient augmenter la valeur de $|\mathcal{Q}'|$ de façon injustifiée et augmenteraient la connectivité du réseau.

En reprenant le réseau de la figure 1.3, le nombre de composantes au numérateur est de 2 comme il est mentionné précédemment. Dans le réseau de la figure 1.1, les groupes A et B sont constitués des mêmes individus que le groupe E. Ils ne sont donc pas inclus dans le calcul de connectivité. La taille de $\mathcal{Q}'$ au dénominateur est donc de 4 puisque le groupe G est également ignoré étant orphelin. La connectivité a donc une valeur de $1 - \frac{2}{4} = \frac{1}{2}$. L'intuition derrière cette formule réside dans le ratio entre le nombre de composantes et le nombre de groupes dans l'ensemble $\mathcal{Q}'$. Dans un réseau, il ne peut pas y avoir plus de composantes que le nombre de groupes qui le compose. S'ils sont égaux, cela signifie que chaque groupe est indépendant des autres. Aucun individu ne fait partie de plus qu'un groupe. Le ratio est donc de 1 et la connectivité est 0. Au contraire, si les groupes sont tous reliés les uns aux autres et qu'aucun d'entre eux n'est isolé, le réseau est alors formé d'une seule composante. Le ratio est alors faible et la connectivité tend vers 1. Évidemment, cette formule fonctionne bien lorsque la taille de $\mathcal{Q}'$ est grande. Dans le cas extrême où le réseau est composé d'un seul groupe, la connectivité est nulle puisqu'il contient une composante également. Par contre, pour un réseau biparti, ce cas de figure est assez rare.

Il est intéressant de remarquer que si la connectivité d'un réseau est nulle la structure du réseau sera comme celle d'une grappe qui est un concept bien connu dans le domaine de la statistique (Kaufman and Rousseeuw, 2009).

La décomposition d'un réseau en composantes donne une bonne idée de sa structure. Il existe certaines statistiques qui permettent d'avoir une meilleure compréhension des composantes. Parmi celles-ci, il y a la longueur moyenne des chemins d'une composante. Une petite valeur signifie qu'en général se rendre d'un nœud à un autre se fait assez rapidement. Une autre mesure qui permet de mieux comprendre la composante est le diamètre. Il représente le chemin le plus long parmi tous les chemins les plus courts dans une composante. Si le diamètre est également petit, cela signifie que tous les nœuds d'une composante sont à une petite distance les uns des autres. En reprenant les deux composantes de la figure 1.2, il est possible de calculer les deux statistiques définies précédemment. Pour la composante formée des nœuds $1, 2$ et 4 tous les chemins qui existent sont de longueur 1, la moyenne des chemins ainsi que le diamètre des deux composantes ont également une valeur de 1. Pour la seconde composante, les longueurs des 6 chemins sont présentes dans le tableau 1.7. La moyenne de la longueur des chemins est donc de $\frac{10}{6} = 1.67$ et le diamètre est de 3.

Tableau 1.7 – Longueur des chemins de la seconde composante de la figure 1.3

Chemin	Longueur
$3 \leftrightarrow 5$	1
$3 \leftrightarrow 6$	2
$3 \leftrightarrow 7$	3
$5 \leftrightarrow 6$	1
$5 \leftrightarrow 7$	2
$6 \leftrightarrow 7$	1

Les individus de la première composante sont plus proches les uns des autres que ceux de la seconde composante. Passer d'un nœud à un autre se fait donc, en général, plus rapidement puisque la moyenne de la longueur des chemins est plus faible.

1.1.6 Assortativité

Les premières sections du chapitre 1 présentent des concepts qui permettent de comprendre la structure d'un réseau. Toutefois, l'étude d'un réseau devient très intéressante si les liens qui le constituent peuvent en partie expliquer les valeurs d'une variable externe au réseau prises par les nœuds. C'est ce qui est calculé par la mesure d'assortativité. Ce coefficient prend des valeurs entre -1 et 1 et permet de savoir si les nœuds connectés d'un réseau ont tendance, pour une certaine variable, à prendre des valeurs similaires. Ce concept s'approche de celui de la corrélation intra grappe bien connue en statistique (Rabe-Hesketh and Skrondal, 2008). La première formule proposée par Newman, 2002, également nommée corrélation degré à degré, permet de savoir si les nœuds connectés ont tendance à avoir des degrés similaires. Une seconde formule (Newman, 2003) permet de calculer le coefficient d'assortativité sur n'importe quelle variable liée aux nœuds.

Pour mesurer l'assortativité dans un réseau, la matrice d'adjacence ainsi que les valeurs pour la variable à analyser sont nécessaires. L'équation peut être écrite sous cette forme,

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\bar{X})(X_j-\bar{X})A_{ij}}{\sum_{i=1}^{n}d_i(X_i-\bar{X})^2}, \tag{1.9}$$

où X_i est la valeur de la variable à analyser pour l'individu i et $\bar{X}$ est une moyenne pondérée par les degrés suivant la formule suivante,

$$\frac{\sum_{i=1}^{n}d_iX_i}{\sum_{i=1}^{n}d_i}. \tag{1.10}$$

Pour être en mesure de bien comprendre le concept d'assortativité, l'âge des individus du réseau fictif est présenté au tableau 1.8. Cette variable représente une caractéristique unique de chaque nœud.

Tableau 1.8 – Âge des individus du réseau

Individu	Âge
1	21
2	23
3	45
4	21
5	40
6	45
7	43
8	37

La moyenne pondérée $\bar{X}$ a une valeur de 32.33. Le reste de l'information nécessaire au calcul est présente au tableau 1.9.

Tableau 1.9 – Informations pour le calcul de l'assortativité

Lien	X_i	X_j	$X_i-\bar{X}$	$X_j-\bar{X}$	$(X_i-\bar{X})(X_j-\bar{X})$
$1\circ 2$	21	23	-11.33	-9.33	105.708
$1\circ 4$	21	21	-11.33	-11.33	128.368
$2\circ 4$	23	21	-9.33	-11.33	105.708
$3\circ 5$	45	40	12.67	7.67	97.179
$5\circ 6$	40	45	7.67	12.67	97.179
$6\circ 7$	45	43	12.67	10.67	135.189

En appliquant la formule (1.9), le coefficient d'assortativité obtenu est de 0.9557.

La valeur très proche de 1 signifie que dans le réseau les individus qui sont liés ont tendance à avoir un âge similaire. En regardant l'âge des individus dans les deux composantes, c'est

effectivement le cas. Pour la première, l'âge des trois individus varie de 21 à 23 ans, alors que pour la deuxième l'âge des quatre individus varie de 40 à 45 ans.

Si la valeur de l'assortativité est négative, cela signifie que deux personnes liées dans le réseau ont tendance à avoir des âges très différents. Si l'assortativité est proche de 0, cela signifie que les liens du réseau n'influencent pas vraiment la variable étudiée.

Dans le calcul de l'assortativité, la formule (1.9) considère tous les liens comme étant égaux. Par contre, nous avons vu précédemment qu'il est possible de leur attribuer une force à l'aide de l'équation (1.2). Il est également possible de calculer l'assortativité en prenant en considération la force des liens. Les liens les plus forts ont alors plus d'influences sur le coefficient d'assortativité (Yuan et al., 2021). Le tableau 1.3 montre que les liens les plus forts sont ceux partagés entre les individus $1, 2$ et 4 qui forment la première composante. Comme leurs âges sont très corrélés, l'assortativité calculée en considérant les forces de lien devrait être un peu plus élevée que celle calculée précédemment. Le calcul peut être fait à l'aide de la fonction `assortment.continuous` de la librairie assortnet (Farine, 2016) dans R. La valeur obtenue est de 0.9606. La prise en considération des forces de liens présentes dans le réseau a donc fait augmenter la valeur de l'assortativité calculée sur l'âge des individus.

1.1.7 Transitivité

Une notion très importante également, lorsqu'on étudie des réseaux, est celle de transitivité. Elle est également appelée coefficient de « clustering ». Elle vient mesurer le niveau de regroupement des nœuds. En d'autres mots, c'est la probabilité que deux nœuds soient en relation sachant qu'ils ont un voisin en commun. Au niveau mathématique, la transitivité est le fait que si l'égalité $a = b$ existe et $b = c$ existe alors $a = c$ est également vraie.

Au niveau réseau, la relation $a = b$ est équivalent à dire que l'individu a est lié à l'individu b. Donc, une relation est dite transitive lorsque pour tout individu a lié à b et b lié à c, cela implique que a et c sont également liés. Dans un réseau social, c'est l'équivalent de dire « l'ami de mon ami est également mon ami ». En général, dans un réseau, le fait que a et b soient liés et que b et c le soient également n'implique pas directement que a et c soient en relation. Par contre, la probabilité que a partage un lien avec c est plus élevée que la probabilité qu'il soit lié à tout autre nœud sélectionné au hasard.

Pour être en mesure de quantifier le niveau de transitivité dans un réseau, suivons la méthodologie suivante. Posons a et b étant liés ainsi que b et c. Il existe alors un chemin de longueur 2 $a \circ b \circ c$, où $\circ$ représente un lien. Dans le cas où a et c sont également en relation, le chemin est alors dit fermé formant une boucle de longueur 3. Ce type de relation est aussi appelé un triangle dû à la forme qu'elle engendre. Le niveau de transitivité est donc défini comme étant la proportion de chemins de longueur 2 qui sont également fermés. La formule mathématique est assez simple et elle s'écrit comme suit.

$$C = \frac{\text{Nombre de chemins de longueur 2 fermés}}{\text{Nombre de chemins de longueur 2}} \tag{1.11}$$

Le coefficient de « clustering » peut prendre des valeurs entre 0 et 1. Aux extrêmes, si $C = 0$, cela implique qu'aucun chemin de longueur 2 n'est fermé. Si $C = 1$, cela signifie que chaque composante du réseau forme ce qu'on appelle une clique. Cela signifie que tous les individus formant une composante partagent un lien entre eux.

Calculons le niveau de transitivité dans le réseau du tableau 1.1. Comme nous l'avons vu précédemment, les composantes de ce réseau sont les suivantes.

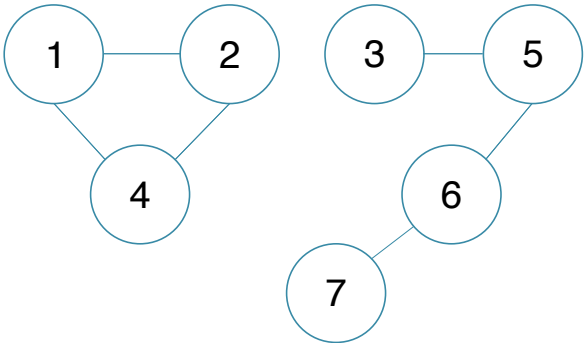


FIGURE 1.4 – Composantes du réseau

Dans un réseau, chaque chemin a une direction. Dans la figure 1.4, le chemin $1 \circ 2 \circ 4$ est différent du chemin $4 \circ 2 \circ 1$. C'est pour cette raison que dans la première composante composée des individus $1, 2$ et 4, il y a 6 chemins de longueur 2 qui sont tous fermés. Dans la seconde composante, il y a 4 chemins de longueur 2 ($3 \circ 5 \circ 6$, $6 \circ 5 \circ 3$, $5 \circ 6 \circ 7$, $7 \circ 6 \circ 5$), mais aucun d'entre eux n'est fermé. Le coefficient de « clustering » pour notre réseau est donc $C = \frac{6}{10} = 0.6$.

1.2 Réseau Bernoulli

Les réseaux peuvent se présenter sous plusieurs formes. Un réseau ayant des liens reliant plus que deux nœuds est appelé hypergraphe. Un arbre est un autre type de réseau qui est connecté, non orienté et qui ne contient aucune boucle. Une boucle représente un lien partant d'un nœud et revenant sur celui-ci. Le lecteur intéressé à explorer les différents types de réseau peut se référer au livre de Newman, 2018.

Il est également possible de simuler un réseau. Il existe plusieurs façons de procéder. Parmi les plus populaires et les plus étudiées, il y a les réseaux aléatoires introduits par les mathématiciens hongrois Paul Erdős et Alfréd Rényi en 1960 (Erdős et al., 1960). Le premier modèle introduit, nommé $G(n, m)$, fixe le nombre de nœuds n et le nombre de liens m. Dans cette sous-section, nous nous intéressons plus particulièrement au deuxième type de modèle dans

lequel on fixe uniquement le nombre de nœuds n. Le nombre de liens n'est pas fixe, mais on attribue plutôt une probabilité p pour que celui-ci existe entre deux individus. Ce modèle est nommé $G(n,p)$. Dans les cas extrêmes, où p est égale à 0 ou à 1, le nombre de liens m dans le réseau est respectivement de 0 et de $\frac{n(n-1)}{2}$. Il existe alors un lien entre chaque paire de nœuds possible.

Pour faire preuve de plus de rigueur, les réseaux aléatoires sont définis en fonction de la fonction de densité de tous les réseaux possibles et non en termes d'un réseau unique. La fonction de densité de l'ensemble des réseaux $G(n,p)$ avec n nœuds et probabilité p est définie ainsi,

$$P(G) = p^m \left(1-p\right)^{\binom{n}{2}-m} \quad G \in G(n,p), \tag{1.12}$$

où m représente le nombre de liens dans le réseau G.

On remarque que cette fonction est une distribution binomiale où le nombre total de liens possible est de $\binom{n}{2}$. La probabilité pour qu'un réseau aléatoire ait exactement m liens suit donc la loi binomiale $\left(\binom{n}{2}, p\right)$ suivante,

$$P(G_m) = \binom{\binom{n}{2}}{m} p^m \left(1-p\right)^{\binom{n}{2}-m}. \tag{1.13}$$

C'est pour cela que ce type de réseau est également appelé réseau aléatoire Bernoulli. Il peut également porter le nom de réseau aléatoire Poisson en référence à la distribution que suivent les degrés des nœuds comme il sera démontré plus tard. Dans cette section, nous étudierons certaines caractéristiques en lien avec les réseaux aléatoires.

1.2.1 Moyenne du nombre de liens et des degrés

Une des spécificités d'un réseau Bernouilli repose sur le fait que le nombre de liens m n'est pas connu au départ. Comme mentionné auparavant, il faut uniquement poser la probabilité que celui-ci existe. Avec cette information, il devient assez simple de calculer l'espérance du nombre de liens. Il suffit simplement de multiplier cette probabilité p par le nombre de liens possibles dans le réseau comme le montre l'équation qui suit,

$$\mathbb{E}\left[m\right] = \mu_m = \binom{n}{2} p. \tag{1.14}$$

Cette fonction peut maintenant être utilisée pour calculer l'espérance des degrés du réseau. Pour rappel, le degré d'un nœud est défini selon l'équation (1.3) et il représente le nombre de nœuds lui étant relié. Si le nombre de liens m est connu, la moyenne des degrés est alors $c = 2m/n$. Pour un réseau Bernoulli, m est inconnu par contre nous connaissons maintenant

son espérance. Il suffit alors de l'employer dans le calcul pour trouver l'espérance des degrés du réseau $G(n,p)$ comme suit,

$$\mathbb{E}[c] = \mathbb{E}\left[\frac{2m}{n}\right] = \frac{2}{n}\mathbb{E}[m] = \frac{2}{n}\binom{n}{2}p = \frac{2}{n}\frac{n!}{2\,(n-2)!}p = (n-1)p. \tag{1.15}$$

L'espérance des degrés pour chaque nœud est donc le nombre de nœuds restant dans le réseau multiplié par la probabilité p d'avoir un lien avec ceux-ci.

1.2.2 Distribution des degrés

Comme il est mentionné précédemment, la distribution des degrés dans un réseau Bernoulli suit une loi binomiale. Voyons l'explication mathématique qui nous donne ce résultat. Par définition, un réseau $G(n,p)$ contient exactement n nœuds. Si un nœud est choisi au hasard, il y a une probabilité p qu'il soit connecté à un des $n-1$ nœuds restants. Le degré d d'un nœud étant équivalent au nombre de connexions qu'il a, la probabilité qu'un nœud soit en relation avec d nœuds spécifiques est donc $p^d\,(1-p)^{n-1-d}$. Par contre, en prenant d nœuds quelconques dans le réseau il faut tenir compte qu'il y a $\binom{n-1}{d}$ façons de les sélectionner. La probabilité qu'un nœud soit connecté à d autres nœuds est donc la suivante,

$$p_d = \binom{n-1}{d} p^d\,(1-p)^{n-1-d}\,. \tag{1.16}$$

La distribution des degrés d suit bien une loi binomiale. En statistique, il est connu que la loi binomiale tend vers une loi de Poisson lorsque n est très grand et que le paramètre p est très petit. La preuve est disponible à l'annexe A.3. La distribution des degrés d'un réseau Bernoulli suit donc une loi de Poisson lorsque sa taille est très grande et peut s'écrire ainsi,

$$p_d = e^{-c}\frac{c^d}{d!}, \tag{1.17}$$

où c est la moyenne des degrés dans le réseau.

1.2.3 Transitivité d'un réseau aléatoire

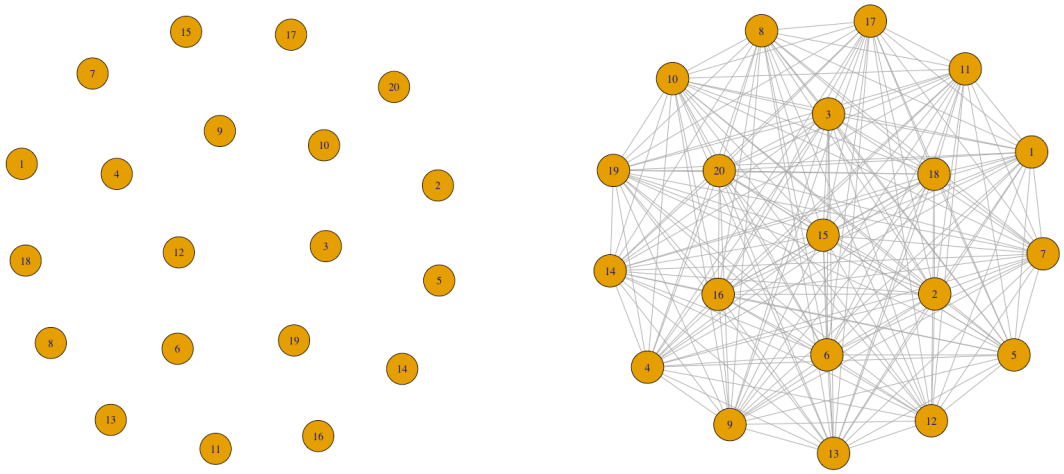
Précédemment, à la sous-section 1.1.7, nous avons introduit la notion de transitivité qu'on peut également appeler coefficient de « clustering » . L'équation (1.11) permet de calculer ce coefficient. Comme il a été mentionné, la transitivité dans un réseau est la probabilité que deux individus ayant un voisin commun soient également en relation. Une des caractéristiques des réseaux Bernoulli est que ce coefficient soit exactement égal à la probabilité p fixée au départ puisqu'elle représente la probabilité qu'un lien existe entre deux individus. À partir de l'équation (1.15), il est possible de réécrire l'équation (1.11) pour les réseaux Bernoulli ainsi,

$$C = p = \frac{c}{n-1}. \tag{1.18}$$

1.2.4 Composante géante

Lorsqu'un réseau $G(n, p)$ est simulé, il est très fréquent de retrouver une composante qui soit beaucoup plus grande que les autres. Celle-ci porte le nom de composante géante et c'est ce phénomène qui est discuté dans cette sous-section.

Lors de l'initialisation du réseau $G(n, p)$, il faut poser la probabilité p qu'un lien existe entre deux individus. À l'extrême, il y a deux situations possibles. Une première pour laquelle $p = 0$ où le réseau n'a aucun lien. On est alors en présence de n individus orphelins indépendants les uns des autres. Dans la seconde situation où $p = 1$, le réseau est formé d'une seule composante unissant toutes les paires d'individus ensembles formant ainsi une clique. La figure 1.5 présente visuellement les deux situations décrites.



(a) Réseau aléatoire avec p=0 (b) Réseau aléatoire avec p=1

FIGURE 1.5 – Réseau aléatoire avec p=0 et p=1

Au-delà du nombre de liens dans les deux réseaux, il existe une différence fondamentale sur la taille des composantes. Bien évidemment, dans le cas où $p = 0$ la taille de la plus grande composante est de 1. Alors que dans le cas où $p = 1$ la plus grande et seule composante du réseau est formée de n individus. Dans cette situation, la taille de la plus grande composante est donc proportionnelle à n contrairement au premier cas. C'est ce principe qui permet de définir le concept de composante géante. On dit d'une composante qu'elle est géante lorsque sa taille grandit proportionnelle à la taille du réseau n.

Dans plusieurs situations, le concept de composante géante est très important pour que le réseau soit fonctionnel. Prenons par exemple un réseau téléphonique. Si l'on veut être en

mesure d'offrir une bonne communication entre les différentes personnes du réseau, il faut qu'une composante géante le domine pour regrouper une majorité des abonnés même s'il reste certains individus non reliés à celle-ci.

Comme mentionné au début de cette section, le fait de passer la valeur de p d'un extrême à l'autre rend la taille de la plus grande composante du réseau dépendante de n. On pourrait être porté à croire que le changement se fait de façon continue au fur et à mesure que p augmente. Ce n'est cependant pas le cas. La transition se fait plutôt de façon brusque lorsque p atteint une certaine valeur dite critique.

Commençons par explorer la probabilité qu'un nœud i choisi au hasard ne fasse pas partie de la composante géante. Cela implique que i ne soit lié à aucun nœud faisant partie de cette composante.

Explorons d'abord la probabilité que i ne fasse pas partie de la composante géante par l'intermédiaire d'un nœud j. Cette probabilité peut donc être séparée en 2 parties. Une première situation où le nœud i n'est pas connecté au nœud j et cette probabilité peut s'écrire comme $1-p$. Dans la seconde situation, i et j sont liés, mais j ne fait pas partie de la composante géante. Cette probabilité est donc égale à pu où u représente la fraction des nœuds du réseau ne faisant pas partie de la composante géante. En regroupant ces deux situations, la probabilité que i ne soit pas lié à la composante géante via j est donc de $1-p+pu$. Comme il y a $n-1$ nœuds dans le réseau autres que le nœud i, la probabilité que celui-ci ne fasse pas partie de la composante géante est donc,

$$u = (1-p+pu)^{n-1}\,. \tag{1.19}$$

Encore une fois, en employant la formule (1.15), il est possible de remplacer p dans la formule précédente pour obtenir,

$$u = \left[1 - \frac{c}{n-1}\,(1-u)\right]^{n-1}\,. \tag{1.20}$$

Continuons à développer la formule en prenant le logarithme de chaque côté et en employant le développement de Taylor. Pour rappel, les séries de Taylor permettent de développer $\log(1-x)$ comme étant la série $-x - \frac{x^2}{2} - \frac{x^3}{3} + O(x^3)$ (Taylor, 1715). En supposant encore une fois que n est très grand on a,

$$\ln u = (n-1)\ln\left[1 - \frac{c}{n-1}(1-u)\right] \simeq -(n-1)\frac{c}{n-1}(1-u) = -c(1-u). \tag{1.21}$$

On peut par la suite enlever le logarithme en prenant l'exponentielle des deux côtés,

$$u = e^{-c(1-u)}. \tag{1.22}$$

Précédemment, u a été défini comme étant la proportion des nœuds du réseau ne faisant pas partie de la composante géante. Il est donc possible de poser $S = 1 - u$ comme étant la proportion des nœuds du réseau faisant partie de la composante géante. L'équation précédente peut donc être réécrite,

$$S = 1 - e^{-cS}. \tag{1.23}$$

À l'aide de l'équation (1.23), on voit que la proportion des nœuds du réseau constituant la composante géante dépend de l'espérance des degrés c. Toute chose étant égale, la valeur de S augmente lorsque celle de c grandit également. Cela est logique puisque si le degré moyen augmente, chaque individu aura donc plus de liaisons avec d'autres nœuds et la probabilité qu'il fasse partie de la composante géante grandit.

L'équation (1.23) ne permet cependant pas d'obtenir une solution unique pour la valeur de S. La seule solution évidente est lorsque $S = 0$. Il est toutefois possible de regarder le comportement de cette équation selon certaines valeurs de c, en posant $y = 1 - e^{-cS}$. C'est ce qui est présenté à la figure 1.6.

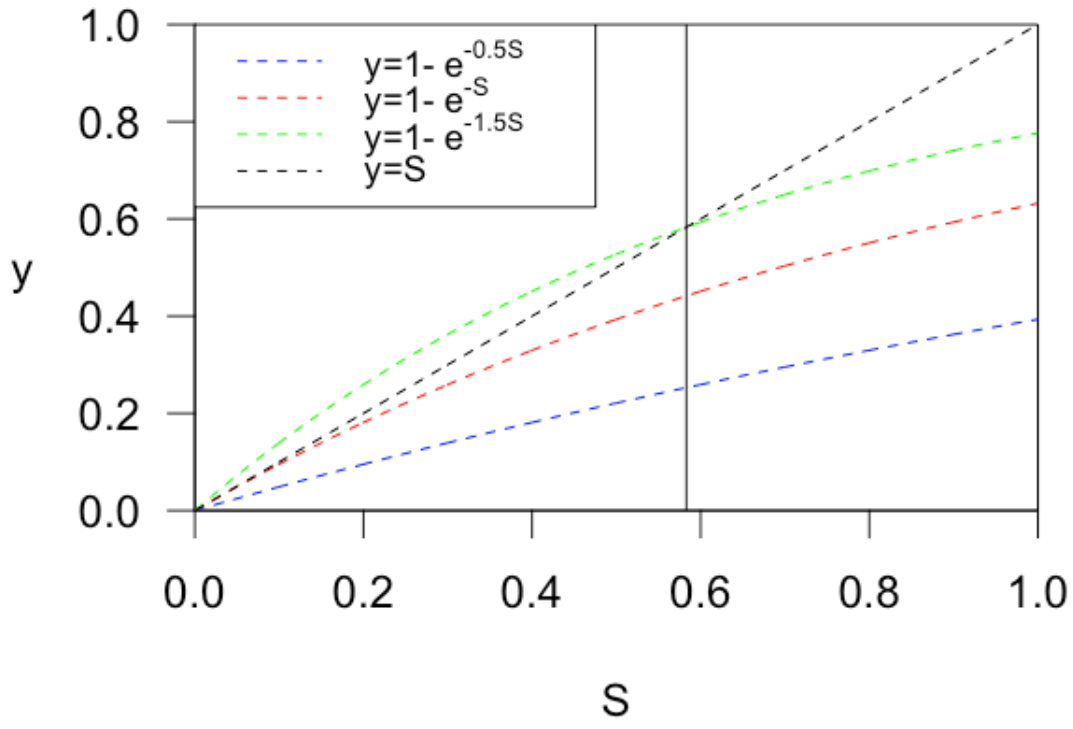


FIGURE 1.6 – Courbe de S selon différentes valeurs de c

La ligne pointillée noire représente l'équation $y = S$. Donc, chaque point d'intersection avec cette ligne représente une solution pour l'équation (1.23) selon la valeur de c choisie. Comme S représente la proportion des nœuds du réseau faisant partie de la composante géante, si sa valeur est de 0, alors cela signifie qu'il n'y a pas de composante géante. C'est ce qu'on constate sur la courbe pour $c = 0.5$ puisque la seule solution possible est $S = 0$. Par contre, plus la valeur de c augmente, plus on approche d'une solution où $S > 0$. La courbe verte avec $c = 1.5$ en est un exemple. Deux points de rencontre existent entre celle-ci et la ligne pointillée noire. Le premier évidemment lorsque $S = 0$ et le second à $S = 0.583$. Cela signifie que lorsque la moyenne de degrés d'un réseau Bernoulli c a une valeur de 1.5, il existe une composante géante regroupant 58.3% des nœuds.

Il existe donc une transition qui se fait entre $c = 0.5$ et $c = 1.5$ pour qu'un réseau soit constitué d'une composante géante. En réalité, elle se produit sur la courbe rouge lorsque $c = 1$. Plus précisément quand les dérivés par rapport à S des lignes noire et rouge se rencontrent à $S = 0$. On a donc,

$$\begin{aligned} \frac{d}{dS}\left(1 - e^{-cS}\right) &= \frac{d}{dS}S, \\ ce^{-cS} &= 1. \end{aligned} \tag{1.24}$$

Il suffit de remplacer S par 0 dans l'équation (1.24) et on obtient $c = 1$. La transition entre un réseau Bernoulli sans composante géante et un avec une composante géante se fait donc lorsque la valeur de c est égale à 1. Le réseau Bernoulli est donc constitué d'une composante géante dès que la moyenne des degrés c est supérieure à 1.

Nous avons vu précédemment que $p = c/(n-1)$. On peut donc affirmer maintenant qu'une composante géante existe dans le réseau Bernoulli lorsque,

$$p > \frac{1}{n-1}. \tag{1.25}$$

1.2.5 Désavantages du réseau aléatoire

L'introduction aux réseaux Bernoulli a permis de voir leur importance dans l'étude des réseaux. Tout d'abord, ils sont très simples à définir. De plus, les calculs faits sur un réseau aléatoire sont peu complexes et ils respectent plusieurs propriétés des réseaux présents dans le monde réel. Cependant, tout n'est pas parfait et il existe certaines différences entre un réseau Bernoulli et ceux décrivant un système réel.

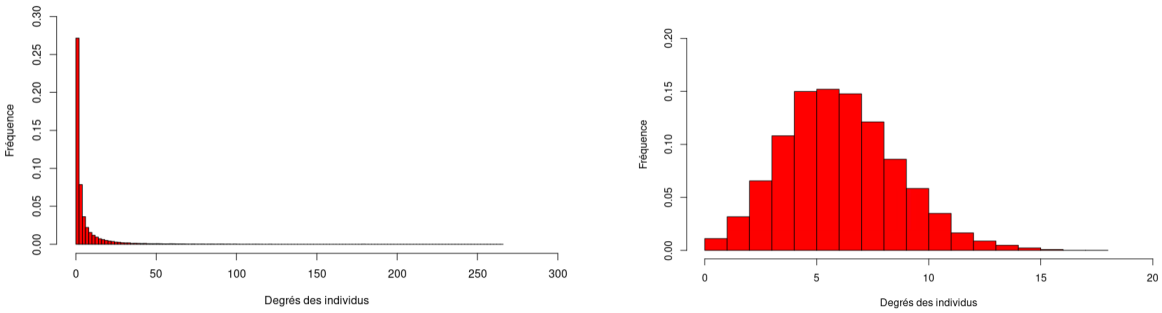
L'équation (1.18) permet de calculer le coefficient de « clustering » pour un réseau Bernoulli. Il est facile de remarquer que lorsque $n \to \infty$ la mesure de transitivité C tend vers 0. Cependant, les vrais réseaux ont tendance à avoir un coefficient de « clustering » qui varie entre 0.01 et

0.5 (Newman, 2018). En observant également les coefficients de « clustering » en lien avec les cinq réseaux fournis par la compagnie d'assurance au tableau 1.18, on voit qu'ils sont très élevés, les valeurs variant entre 0.7478 et 0.9996. Il est clair que les réseaux Bernoulli diffèrent énormément des réseaux bipartis fournis par la compagnie d'assurance au niveau de la transitivité.

Une autre différence majeure est présente au niveau de la corrélation entre les degrés de nœuds voisins. Dans la réalité, il existe une forte corrélation entre les degrés des nœuds voisins. Par contre, dans les réseaux Bernoulli, ce n'est pas le cas puisque les liens ont été placés de façon aléatoire. En utilisant encore les réseaux fournis par la compagnie d'assurance, dans le tableau 1.17, on remarque que l'assortativité est très élevée lorsqu'elle est calculée sur les degrés. Les valeurs pour les cinq réseaux varient entre 0.41 et 0.99. Cela signifie que les individus liés dans ces réseaux ont tendance à avoir un degré similaire. Le même coefficient calculé sur le réseau Bernouilli de 1000 nœuds, utilisé à la section 3.2, obtient une valeur de 0.0012. Il n'y a donc presque aucune corrélation entre le fait que deux individus soient reliés et leur degré respectif.

Dans la vraie vie, il y a un phénomène de « communauté » qui tend à regrouper des groupes de personnes ensemble. Ceux-ci sont donc très reliés entre eux et ils tendent naturellement à avoir un degré similaire. Cela peut expliquer pourquoi la corrélation est très forte entre les degrés des individus connectés. C'est un autre aspect qui n'est pas retrouvé dans les réseaux Bernoulli.

On peut continuer l'analyse sur les degrés des deux types de réseaux pour se rendre compte d'une autre grande différence. En observant la distribution des degrés d'un vrai réseau et d'un réseau Bernoulli à la figure 1.7 il est clair qu'ils ne suivent pas la même loi. La majorité des nœuds d'un réseau décrivant un système réel ont tendance à avoir un degré très faible. Par contre, à l'extrême, certains nœuds, représentant le « centre d'attention » du réseau, ont un degré très élevé, ce qui rend la distribution des degrés très asymétrique vers la droite qui semble suivre une loi exponentielle. Par contre, nous avons vu, à l'équation (1.17) que la distribution des degrés d'un réseau Bernoulli suit une loi de Poisson. L'asymétrie vers la droite est donc beaucoup moins accentuée comme le montre la figure 1.7. Donc, même si la moyenne des degrés pour les deux réseaux est similaire (6.588 pour le réseau 2 fourni par la compagnie d'assurance et 5.068 pour le réseau Bernouilli), il est évident que la distribution de leurs degrés ne suit pas la même loi.



(a) Distribution des degrés du réseau 2 fourni par la compagnie d'assurance

(b) Distribution des degrés d'un réseau Bernoulli

FIGURE 1.7 – Distribution des degrés

1.3 Données réseau

Après avoir couvert certains concepts théoriques importants en lien avec les réseaux bipartis et Bernoulli, il est temps de regarder un peu plus les données analysées dans ce mémoire. Cette section présente une description des données qui sont fournies par une compagnie d'assurance canadienne.

1.3.1 Description des données

Les données analysées se présentent sous deux formes ; des données réseau et une base de données longitudinales. Cette dernière contient de l'information sur chaque client prise sur une période de trois ans. Des variables explicatives qui décrivent certaines caractéristiques des biens qui sont assurés ainsi que des variables en lien avec le profil de l'assuré sont disponibles. La variable d'intérêt est l'annulation d'une police d'assurance. Quant aux données réseau, elles se présentent sous la forme de cinq réseaux différents. Chacun d'eux met en relation les différents assurés selon un critère particulier. Pour faire l'analogie avec l'exemple de la section 1.1, nous pouvons imaginer que chacun des cinq réseaux représente un ensemble de groupes différents mettant en relation des individus. Dans les données analysées, ces groupes sont représentés par un identifiant d'informations. Donc, pour chaque réseau, si deux individus partagent un identifiant similaire, un lien les unit. L'objectif est donc de comprendre et de quantifier l'influence que peuvent avoir les relations présentes dans les réseaux sur le fait qu'une personne annule ou non sa police d'assurance.

Pour avoir une analyse cohérente, la première étape consiste à enlever tous les individus orphelins. La même définition vue à la section 1.1 s'applique ici. Les individus orphelins sont ignorés dans cette étude puisqu'ils n'apportent que très peu, voir aucune information. La deuxième étape consiste à conserver uniquement les individus qui sont dans la base de données longitudinales et qui sont en relation avec au moins une autre personne dans les données réseau. Sur les 1 000 000 individus présents au départ, 600 000 sont conservés à la suite de ces deux

étapes. Le tableau 1.10 permet d'avoir une idée de la répartition des individus dans les cinq réseaux différents. Pour bien comprendre l'information qui s'y retrouve, analysons la deuxième ligne. La combinaison (1,0,0,0,0) représente tous les individus qui font partie d'au moins un groupe dans le premier réseau et qui ne sont pas en relation dans les quatre autres.

Tableau 1.10 – Répartition des individus dans les différents réseaux. Pour une combinaison, la valeur 0 signifie qu'un individu ne fait pas partie d'un groupe pour ce réseau. S'il fait partie d'un groupe ou plus, la valeur 1 apparaît.

Réseau 1	Réseau 2	Réseau 3	Réseau 4	Réseau 5	Frequence
0	0	0	0	0	400 000
1	0	0	0	0	5000
0	1	0	0	0	430 000
0	0	1	0	0	30 000
0	0	0	1	0	10 000
0	0	0	0	1	3000
1	1	0	0	0	4000
1	0	1	0	0	1000
1	0	0	1	0	2000
1	0	0	0	1	70
0	1	1	0	0	30 000
0	1	0	1	0	10 000
0	1	0	0	1	50 000
0	0	1	1	0	7000
0	0	1	0	1	100
0	0	0	1	1	100
1	1	1	0	0	900
1	1	0	1	0	2000
1	1	0	0	1	500
1	0	1	1	0	2000
1	0	1	0	1	40
1	0	0	1	1	50
0	1	1	1	0	5000
0	1	1	0	1	3000
0	1	0	1	1	1000
0	0	1	1	1	70
1	1	1	1	0	2000
1	1	1	0	1	100
1	1	0	1	1	200
1	0	1	1	1	70
0	1	1	1	1	600
1	1	1	1	1	200

$*$ Les chiffres réels ont été arrondis pour des raisons de confidentialité et ce pour l'intégralité du mémoire.

Parmi les 600000 individus constituant les données réseaux, 478000 ne font partie que d'un

seul réseau. On constate également qu'une grande proportion des personnes présentes dans le deuxième réseau ne font partie d'aucun autre réseau, près de 80%. Pour le cinquième réseau, ce pourcentage est de seulement 5% alors que pour les trois autres réseaux, il varie entre 24% et 37%. En fait, 94% des personnes présentent dans le réseau 5 sont également dans le réseau 2. Au total, seulement 200 individus sont présents dans les cinq réseaux. La combinaison la plus rare est celle des réseaux 1, 3 et 5 pour laquelle il existe seulement 40 individus. Les 400000 individus ne faisant partie d'aucun réseau sont les individus orphelins ou ceux ne figurant pas dans les données longitudinales.

Après avoir vu la répartition des individus dans les différents réseaux, le tableau 1.11 présente certaines statistiques descriptives en lien avec chacun des cinq réseaux.

Tableau 1.11 – Statistiques descriptives sur les cinq réseaux. Les résultats sur la dernière ligne *Combiné* sont obtenus en regroupant les cinq réseaux ensemble. Pour certaines colonnes, comme le nombre d'individus ou le nombre de liens, le total ne représente pas la somme des cinq réseaux puisque les individus ou les liens peuvent revenir dans plusieurs réseaux différents. La colonne pourcentage représente la proportion des individus totale faisant partie des différents réseaux.

	Nb individus	Pourcentage	Nb de groupes	Nb de liens dans le réseau	Densité ρ ($\times 10^{-4}$)
Réseau 1	20 130	3.36%	11 703	31 175	1.56
Réseau 2	539 500	89.92%	223 293	1 834 351	0.12
Réseau 3	82 080	13.68%	23 562	56 331 395	158.11
Réseau 4	42 290	7.05%	16 696	10 949 348	86.67
Réseau 5	59 100	9.85%	34 701	32 459	0.18
Combiné	600 000		309 955	69 121 840	3.54

À première vue, les cinq réseaux semblent avoir des caractéristiques différentes. Premièrement, le nombre d'individus que contient chacun diffère considérablement. Le réseau 1 avec 20 130 individus est presque 27 fois plus petit que le deuxième qui en a 539 500. La taille des trois autres réseaux varie entre 42 290 et 82 080 personnes. Quant au nombre de groupes formant chaque réseau, l'écart y est encore élevé. Le deuxième réseau contient cette fois-ci presque 20 fois plus de groupes que le premier avec respectivement 223 293 et 11 703 groupes. Il est également celui qui contient la plus grande proportion de la population totale avec plus de 89% des individus. Quant au troisième réseau, il est le plus grand en termes de liens avec 56 331 395 relations. Le premier et cinquième réseau ont un nombre de liens similaire avec un peu plus de 30 000 relations. Le quatrième contient quant à lui dix fois plus de liens que le deuxième. Pour ce qui est de la densité des réseaux, le troisième est encore une fois celui qui est le plus dense. Il comprend un peu plus de 1% de tous les liens possibles. Les quatre autres réseaux ont une densité inférieure à 1%. Comme il a été expliqué précédemment à la sous-section 1.1.3, on peut affirmer que les cinq réseaux sont clairsemés. Comme ils représentent des relations sociales entre individus, cela n'est pas vraiment surprenant.

Ayant une idée générale de la composition de chaque réseau, il est intéressant de pousser l'exploration un peu plus loin et de calculer certaines statistiques en lien plus spécifiquement avec les groupes. Le tableau 1.12 permet d'avoir une idée de la répartition de la taille des groupes dans les différents réseaux calculée à l'aide de l'équation (1.1). En général, peu importe le réseau, les groupes sont de petite taille. Pour les réseaux 1, 4 et 5, au moins 75% des groupes mettent en relation 2 individus uniquement. Pour le deuxième et troisième, au moins 75% des groupes sont formés de 3 personnes. Cependant, les tailles maximales des groupes diffèrent un peu plus. Le troisième et quatrième réseaux contiennent des groupes de taille 4677 et 4452 respectivement. Tandis que pour les deux premiers réseaux, les plus gros groupes unissent entre 178 et 181 individus. Le dernier réseau semble assez différent des autres puisque le plus grand groupe est formé de 5 personnes uniquement. Les groupes volumineux permettent d'associer de nombreuses personnes. Par contre, ces liens créés ont une force très faible selon l'équation (1.2) puisque la taille du groupe est très élevée. Les plus petits groupes, quant à eux, associent moins d'individus, mais ils créent des liens plus forts.

Tableau 1.12 – Statistiques sur la taille des groupes dans chaque réseau. Q1 représente le premier quartile et Q3 le troisième.

	Min	Q1	Médiane	Q3	Max	Moyenne	Nb de groupes
Réseau 1	2	2	2	2	178	2.13	11 703
Réseau 2	2	2	2	3	181	3.03	223 293
Réseau 3	2	2	2	3	4677	5.99	23 562
Réseau 4	2	2	2	2	4452	3.48	16 696
Réseau 5	2	2	2	2	5	2.04	34 701

Continuons maintenant l'analyse avec une approche orientée sur les individus. Le tableau 1.13 présente la distribution du nombre de groupes dont fait partie chaque sujet dans les différents réseaux. Dans les cinq réseaux, sauf le troisième, 75% des individus n'appartiennent qu'à un seul groupe. Encore une fois, la différence devient plus accentuée au niveau du maximum. Pour le réseau 3, au moins un individu fait partie de 124 groupes alors que pour les autres réseaux les maximums varient entre 8 et 11. Le fait qu'un individu soit lié à plus d'un groupe permet d'obtenir une structure de réseau plus complexe. Deux personnes peuvent donc être reliées par un voisin commun, ce qui permet de créer de plus longs chemins. Si chaque individu ne fait partie que d'un seul groupe, le réseau est alors formé de simples grappes de deux personnes ou plus.

Tableau 1.13 – Statistiques sur le nombre de groupes auxquels appartiennent les individus dans chaque réseau

	Min	Q1	Médiane	Q3	Max	Moyenne	Nb d'individus
Réseau 1	1	1	1	1	11	1.25	20 130
Réseau 2	1	1	1	1	10	1.21	539 500
Réseau 3	1	1	1	2	124	1.67	82 080
Réseau 4	1	1	1	1	8	1.16	42 290
Réseau 5	1	1	1	1	8	1.17	59 100

Les tableaux précédents offrent une bonne idée de la composition de chaque réseau que ce soit au niveau de la taille des groupes ou de la récurrence des individus dans ceux-ci. Cependant, comme il a été mentionné précédemment à la section 1.3.1, il est possible grâce à l'équation (1.2) d'attribuer une force au lien qui unit deux individus. C'est ce qui est traité au tableau 1.14, qui présente la distribution des forces de liens.

Tableau 1.14 – Statistiques sur la distribution des forces de liens dans chaque réseau

	Min	Q1	Médiane	Q3	Max	Moyenne	Nb de liens
Réseau 1	0.0056	0.0056	0.0056	0.5	3.08	0.21	31 175
Réseau 2	0.0055	0.0232	0.0588	0.1667	2.25	0.12	1 834 351
Réseau 3	0.0002	0.0002	0.0002	0.0004	12.23	0.001	56 331 395
Réseau 4	0.0002	0.0002	0.0002	0.0002	3.83	0.002	10 949 348
Réseau 5	0.2	0.5	0.5	0.5	3.5	0.55	32 459

À première vue, les réseaux 3 et 4 ont des forces de lien qui sont assez faibles. Cela est compréhensible au vu des résultats présentés au tableau 1.12. Ces deux réseaux contiennent les groupes ayant les plus grandes tailles. Le nombre de combinaisons entre deux individus augmente alors considérablement et selon l'équation (1.2) la force de ces liens est très faible due à la taille du groupe qui est très grande. Le troisième réseau contient cependant le lien avec le poids le plus fort avec une valeur de 12.23. Quant au réseau 5, on observe peu de variation dans les forces des liens qui le composent. Cela est normal puisque la grandeur des groupes varie également très peu pour ce réseau.

Précédemment, nous avons vu avec l'équation (1.3) comment calculer le degré d'un individu. Cette mesure est très importante lorsqu'on essaie d'étudier la structure d'un réseau. Elle permet d'avoir une idée sur le nombre de relations qu'a chacun des individus. Le tableau 1.15 présente certaines statistiques en lien avec la distribution des degrés dans les cinq réseaux.

Tableau 1.15 – Statistiques sur la distribution des degrés des individus dans chaque réseau

	Min	Q1	Médiane	Q3	Max	Moyenne	Nb d'individus
Réseau 1	1	1	1	1	178	3.128	20 130
Réseau 2	1	1	2	6	266	6.588	539 500
Réseau 3	1	5	94	2662	13 952	1335	82 080
Réseau 4	1	1	2	38	5817	435.7	42 290
Réseau 5	1	1	1	1	5	1.077	59 100

Dans la formation des réseaux, chaque paire d'individus partageant une information commune est mise en relation. Les résultats du tableau 1.15 sont donc évidemment similaires à ceux du tableau 1.12. Les réseaux 3 et 4 sont ceux ayant les groupes les plus grands. Ils ont donc également les nœuds ayant les plus grands degrés. Le réseau 3 se distingue plus particulièrement puisqu'au moins 75% des nœuds ont un degré plus grand ou égal à 2662. Au niveau des trois autres réseaux, les degrés sont en général plus petits. Par contre, en regardant les maximums, les réseaux 1 et 2 ont des degrés beaucoup plus élevés que le réseau 5. Dans ce dernier, le degré maximal est de 5. La constitution de ce réseau semble donc encore une fois être assez différente des quatre autres.

1.3.2 Composantes

Tableau 1.16 – Statistiques sur les composantes dans chaque réseau

	Nb compo	Connectivité (%)	Taille minimale	Taille maximale	Taille moyenne	Longueur moy des chemins	Diamètre
Réseau 1	9217	6.08	2	179	2.16	1.02	4
Réseau 2	114 977	48.30	2	216 858	4.84	10.48	40
Réseau 3	4043	82.46	2	74 175	20.88	2.88	12
Réseau 4	11 647	26.73	2	21 825	4.32	3.76	17
Réseau 5	28 999	7.00	2	6	2.08	1.04	3

La sous-section 1.1.5 présente le concept de composantes. Leur étude est très importante puisqu'elle oriente les analyses faites par la suite. Les statistiques en lien avec les composantes sont présentes au tableau 1.16 et permettent de mieux comprendre comment chaque réseau est formé. La connectivité est obtenue à partir de l'équation (1.8). Le troisième réseau a une valeur très élevée comparativement aux autres réseaux. Cela signifie que les groupes ont tendance à être plus reliés dans ce réseau que dans les autres. Le tableau 1.13 explique la forte connectivité du réseau 3. Comme mentionné précédemment, certains individus font partie de 124 groupes. Le nombre de groupes reliés augmente alors rapidement, ce qui a pour conséquence d'augmenter également la connectivité. Le premier et le cinquième réseau ont une connectivité plus faible avec des valeurs respectives de 6% et 7%. Cela signifie qu'il n'y a presque aucune relation entre les différents groupes du réseau. Cela peut se voir aux figures 1.8 et 1.9. Chacune d'elles représente un échantillon des composantes du réseau 1 et du réseau

3 respectivement et chaque nœud représente un individu et un trait symbolise un lien. Si le nœud est rouge, cela signifie que l'individu a abandonné sa police d'assurance. Ce concept est étudié au chapitre 2. Sur la figure 1.8, représentant l'échantillon du réseau 1, les groupes semblent bien indépendants les uns des autres. Alors que la figure 1.9, représentant l'échantillon du réseau 3, présente certaines relations qui permettent d'unir des groupes ensembles. On y retrouve une composante qui unit 135 individus. Ces deux images permettent donc de mieux comprendre pourquoi la connectivité du réseau 3 est plus élevée que celle du réseau 1.

La moyenne des chemins permet d'avoir une idée sur la distance qui sépare deux points individus dans le réseau en général. Pour les réseaux 1 et 5 cette valeur est assez faible. Comme ces réseaux sont assez déconnectés et que les tailles de leurs composantes sont petites, il est normal que la longueur des chemins unissant deux individus soit faible. Quant aux réseaux 3 et 4 la longueur moyenne des chemins augmente légèrement. La différence est particulièrement marquée pour le deuxième réseau avec une valeur de 10.48. Une plus grande connectivité du réseau et des tailles plus élevées des composantes peuvent expliquer cela. Pour terminer, le diamètre permet de savoir la distance du plus long chemin du réseau. Encore une fois, le même raisonnement peut être fait que précédemment puisque les réseaux 2 et 4 sont ceux avec les plus grands diamètres. Le réseau 2 est à nouveau celui avec la plus grande valeur qui est de 40. Tous les résultats présentés au tableau 1.16 ont été obtenus à partir de la fonction `components` de la librairie igraph Csardi and Nepusz, 2006. Le lecteur intéressé à en savoir plus sur l'utilisation de la librairie igraph pour l'analyse des réseaux peut se référer à l'annexe B.1.

La taille maximale des composantes permet de savoir le nombre maximal d'individus reliés dans un réseau. Avec une composante regroupant 74 175 personnes, le réseau 3 permet de relier presque 88% de ses nœuds. Pour les réseaux 2 et 4 ce sont environ 40% des individus pour lesquels il existe un chemin entre eux. Encore une fois, les valeurs sont beaucoup plus petites pour les réseaux 1 et 5 puisque leurs plus grandes composantes ont une taille respective de 179 et 6.

L'analyse des composantes est intéressante puisqu'elle donne une idée plus claire de la structure des réseaux. Les réseaux 1 et 5 avec une faible connectivité et des composantes de petite taille peuvent être vus comme des grappes. Alors qu'un réseau ayant une forte connectivité et des composantes de grande taille comme le troisième possède une structure plus complexe.

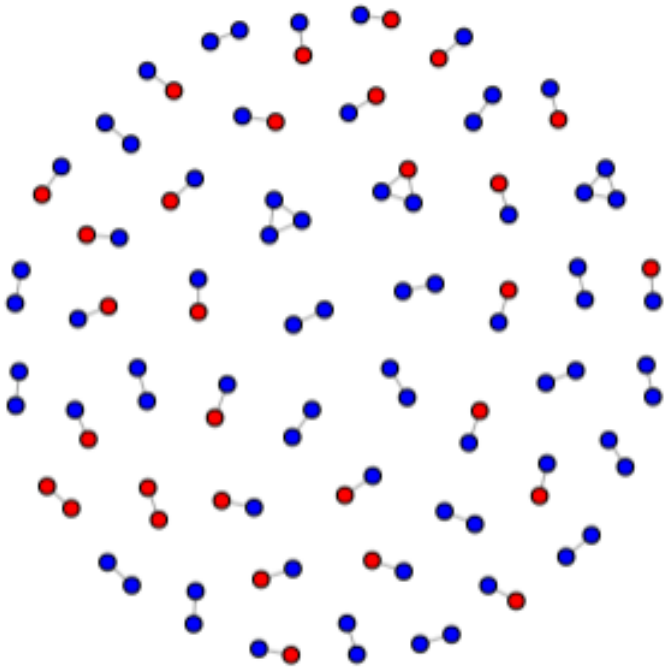


FIGURE 1.8 – Échantillon du réseau 1

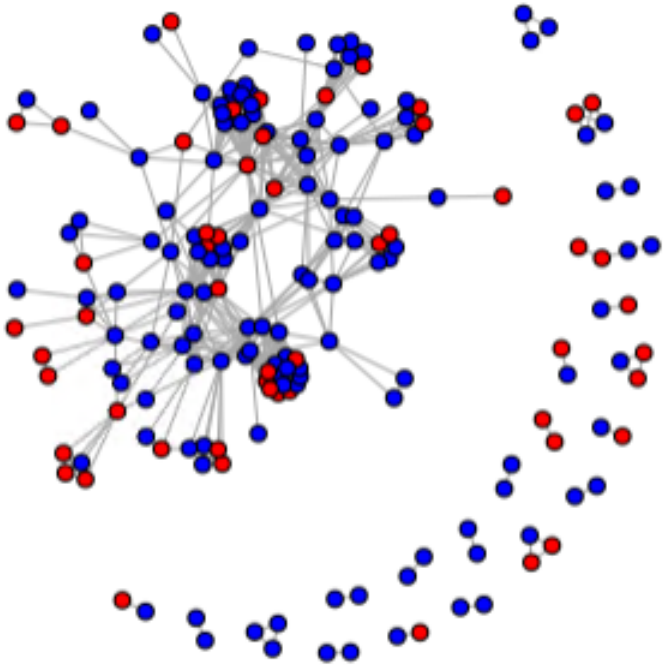


FIGURE 1.9 – Échantillon du réseau 3

1.3.3 Assortativité

La sous-section 1.1.6 offre une introduction au concept d'assortativité. Ce coefficient permet de savoir si les nœuds ayant des caractéristiques similaires sont généralement plus en relation dans un réseau ou non. Si c'est le cas, cette information est très importante pour la suite de l'analyse puisque cela signifie que les relations d'un réseau influencent une variable externe au réseau.

Dans les données analysées, chaque nœud représente un client de la compagnie d'assurance. Pour chacun d'eux, certaines caractéristiques sont disponibles. L'assortativité a été calculée sur quatre d'entre elles. Le degré, l'abandon de l'assurance dommages ou non, le nombre d'années comme client de la compagnie d'assurance et le nombre d'interactions qu'il a eues avec celle-ci.

Tableau 1.17 – Valeurs des coefficients d'assortativité selon chaque réseau

	Degrés	Abandon	Nb années	Nb interactions
Réseau 1	0.99	-0.01	0.03	0.009
Réseau 2	0.88	0.02	0.06	0.03
Réseau 3	0.41	-0.00009	0.0003	0.00001
Réseau 4	0.92	-0.0002	-0.0002	-0.00001
Réseau 5	0.74	0.08	0.05	-0.11

Pour les trois caractéristiques externes au réseau, on voit que l'assortativitié est assez faible. Les réseaux 3 et 4 sont ceux avec les assortativités les plus faibles. Le réseau 5 a une assortativité négative en lien avec le nombre d'interactions. Cela signifie que dans ce réseau, si un individu interagit beaucoup avec la compagnie d'assurance, les individus qui lui sont liés ont tendance à moins interagir. Par contre, dans ce même réseau l'assortativité est positive lorsqu'on l'a calcule pour l'abandon. Donc, dans le réseau 5, si une personne abandonne sa police d'assurance, les personnes en relation avec elle ont aussi tendance à annuler leur police. La même interprétation peut être faite dans le réseau 2 en lien avec la variable du nombre d'années d'ancienneté dans la compagnie. Ces corrélations restent toutefois assez faibles puisqu'elles varient entre 6% et 11% en valeur absolue.

1.3.4 Transitivité

La sous-section 1.1.7 présente le concept de transitivité. Comme mentionné précédemment il peut également être appelé coefficient de « clustering » . Celui-ci mesure la proportion de nœuds connectés qui ont également un voisin commun. Le tableau 1.18 présente les résultats selon chacun des cinq réseaux.

Tableau 1.18 – Coefficient de « clustering » dans chaque réseau

	Coefficient de « clustering »
Réseau 1	0.9996
Réseau 2	0.8776
Réseau 3	0.8623
Réseau 4	0.9877
Réseau 5	0.7478

Pour chacun des cinq réseaux, le coefficient de « clustering » est très élevé. Pour les réseaux 1 et 4, il est presque de 1, ce qui signifie que presque tous les individus partageant un voisin sont également en relation.

En général, les réseaux représentant un concept social, comme dans notre cas, ont tendance à avoir un coefficient de « clustering » plus élevé que 0.1 (Newman, 2018). Par contre, ici la mesure est très élevée. Cela peut être expliqué par la façon dont les relations dans les réseaux ont été construites. Dans un groupe, chaque individu en faisant partie se voit attribuer un lien avec les autres individus du groupe. Donc, pour les groupes de plus de 3 individus, les chemins de longueur 2 deviennent automatiquement fermés. En d'autres mots, si deux individus partagent un voisin commun, ils sont également en relation. En fait, la seule façon d'obtenir un chemin de longueur 2 non fermé est donc en ayant un individu faisant partie de plusieurs groupes. Des chemins de longueur 2 non fermés sont alors créés entre les individus des différents groupes par l'intermédiaire de ce voisin commun. Cependant, comme mentionné précédemment au tableau 1.13 cette situation n'arrive que rarement, puisque tous les réseaux, sauf le troisième, ont au moins 75% des personnes qui ne sont présentes que dans un seul groupe. Cela peut donc expliquer les mesures de coefficient de « clustering » très élevées des cinq réseaux étudiés.

Chapitre 2

Étude de l'impact d'un réseau sur une variable aléatoire discrète

Le chapitre précédent offre une bonne introduction aux concepts de base des réseaux bipartis. Les différents types de projection permettent de regrouper l'information sous diverses formes et les graphiques développés offrent une représentation visuelle des liens qui existent entre les individus. De plus, il a été possible d'attribuer un poids à ces relations. La deuxième section a également permis d'avoir une idée sur les données réseau fournies par la compagnie d'assurance. Ce deuxième chapitre couvre certains tests qui permettent d'évaluer l'influence que peuvent avoir les relations entre les individus sur une variable aléatoire. Comme mentionné au chapitre 1, la variable étudiée est binaire et représente l'abandon d'une police d'assurance par un individu. Regardons d'abord quelques statistiques descriptives en lien avec cette variable. C'est ce que présente le tableau 2.1.

Tableau 2.1 – Statistique d'abandon de la police d'assurance selon le réseau

	Nb individus	Nb de liens	Taux d'abandon de la police	Pourcentage de liens entre 2 abandons
Réseau 1	20 130	31 175	29.14%	6.21%
Réseau 2	539 500	1 834 351	31.31%	12.14%
Réseau 3	82 080	56 331 395	26.93%	7.08%
Réseau 4	42 290	10 949 348	27.84%	7.53%
Réseau 5	59 100	32 459	29.55%	10.30%

Le taux d'abandon de la police d'assurance semble être légèrement différent d'un réseau à l'autre. Le deuxième réseau est celui où les individus abandonnent le plus leur police. Les réseaux 1 et 5 et 3 et 4 ont des taux semblables avec des valeurs respectives de 29% et 27%. La différence est plus remarquable sur le pourcentage de liens entre deux personnes qui abandonnent. Les réseaux 2 et 5 ont des pourcentages respectifs de 12.14% et 10.30% alors qu'il

varie entre 6% et 7% pour les autres réseaux. Au vu des résultats présents dans le tableau 2.1, le réseau semble influencer l'abandon de la police d'assurance. C'est ce qui est traité dans la suite du chapitre. Pour ce faire, quatre méthodes statistiques sont présentées. La première est un test de permutation, puis ce sont les tests de Kruskal-Wallis et de Wilcoxon et la dernière est une régression logistique.

2.1 Test de permutation

2.1.1 Notions théorique

Le test de permutation est basé sur une méthode de ré-échantillonnage. Il fait partie de la catégorie des tests exacts puisqu'il permet d'obtenir la distribution empirique de la statistique de test sous l'hypothèse nulle H_0. Pour ce faire, la statistique de test est calculée selon toutes les M permutations possibles des données. Un des principaux avantages réside dans le fait que le test ne demande pas de connaître la distribution théorique des données analysées (Berry et al., 2021).

Pour bien comprendre le fonctionnement du test, considérons un exemple simple où un réseau est composé de plusieurs groupes. Simplifions les choses et disons que nous voulons évaluer si la taille moyenne des hommes μ_1 est similaire à celle des femmes μ_2 ou non. Les hypothèses nulle et alternative du test sont donc les suivantes :

$$H_0 : \mu_1 = \mu_2 \qquad vs \qquad H_1 : \mu_1 \neq \mu_2$$

De plus, supposons qu'il y a 10 individus de chaque sexe dans le réseau. Le test de permutation se déroule donc selon les cinq étapes suivantes :

1. Calculer la statistique observée t_0 en soutraillant la moyenne des tailles des femmes à celle des hommes.
2. Placer les 20 individus ensemble et en attribuer aléatoirement 10 au groupe des hommes et 10 au groupe des femmes.
3. Calculer la nouvelle statistique du test t en soustraillant la moyenne des tailles du groupe des femmes à celle des hommes.
4. Recommencer les étapes 2 et 3 M fois.
5. La distribution empirique est alors obtenue en ordonnant les M résultats simulés de la statistique de test t.

Comme il est mentionné précédemment, le nombre de simulations M correspond au nombre de permutations possibles des données. Il est donné par l'équation suivante,

$$M = \binom{n}{r}, \tag{2.1}$$

où n représente la taille de la population et r la taille d'un des groupes. Dans cet exemple, le nombre de combinaisons possibles est donc de $C_{20}^{10} = 184\,756$. Par la suite, la valeur p peut-être calculée selon l'équation suivante,

$$P\left(t \leq t_0 | H_0\right) = \frac{\text{nombre de } t \text{ valeurs} \leq t_0}{M}, \tag{2.2}$$

et comparée au seuil choisi, généralement 0.05. Pour un test bilatéral, il faut multiplier la valeur p par 2. Si la probabilité obtenue avec l'équation (2.2) est plus grande que 0.5, il faut d'abord la soustraire à 1 avant de la multiplier par 2 pour savoir si l'hypothèse nulle est rejetée ou non.

La valeur M obtenue à l'équation (2.1) peut devenir extrêmement grande si la taille de la population n est plus élevée. Dans l'exemple précédent, 184 756 combinaisons sont possibles pour former deux groupes de 10 individus. Si M devient trop élevé, il devient alors presque impossible au niveau computationnel d'utiliser toutes les combinaisons pour calculer la distribution empirique. Il faut alors utiliser un test de permutation Monte-Carlo. Le processus est exactement le même, mais un échantillon L des M combinaisons possibles est utilisé. L'équation (2.2) peut être réécrite ainsi,

$$P\left(t \leq t_0 | H_0\right) = \frac{\text{nombre de } t \text{ valeurs} \leq t_0}{L}. \tag{2.3}$$

La valeur L doit rester assez élevée pour bien approximer la distribution empirique de la statistique de test. Howell, 2007 propose 100 000 permutations, alors que Johnston et al., 2007 proposent jusqu'à 1 000 000 permutations.

2.1.2 Application : lien entre les réseaux et l'abandon de la police

Cette méthodologie peut être appliquée sur les réseaux fournis par la compagnie d'assurance. L'objectif est d'évaluer l'influence du réseau sur le fait qu'un individu abandonne sa police d'assurance. Pour ce faire, il faut se concentrer sur les liens unissant deux individus ayant annulé leur police. L'intuition derrière est de voir si la probabilité qu'une personne annule sa police augmente lorsqu'elle est en relation avec un individu qui a annulé la sienne. L'annulation de la police est une variable aléatoire X qui suit une loi de Bernoulli avec une probabilité de succès p spécifique à chaque réseau. La probabilité p est disponible à la troisième colonne du tableau 2.1.

Les hypothèses du test s'écrivent donc ainsi :
H_0 : Les relations du réseau n'influencent pas le fait qu'un individu abandonne sa police d'assurance.
vs

H_1 : Les relations du réseau influencent le fait qu'un individu abandonne sa police d'assurance.

Pour commencer, il faut calculer le nombre d'individus ayant annulé leur police d'assurance dans le réseau *somme annulation*. Puis, les cinq étapes décrites précédemment peuvent être adaptées de cette façon :

1. Calculer la statistique observée t_0 qui représente le nombre de liens partagés par deux individus qui ont abandonné leur police d'assurance.
2. Simuler une pseudo variable d'abandon de la police d'assurance pour tous les individus du réseau. C'est un vecteur de 0 et de 1 où le nombre de 1 est égal à *somme annulation*.
3. Calculer la nouvelle statistique du test t en évaluant le nombre de liens partagés par les *somme annulation* individus ayant abandonnés la police selon le vecteur simulé.
4. Recommencer les étapes 2 et 3 100 000 fois.
5. La distribution empirique est alors obtenue en ordonnant les 100 000 résultats simulés de la statistique de test t.

Comme il est mentionné dans l'énumération précédente, 100 000 simulations ont été effectuées et les résultats sont présentés dans le tableau 2.2.

Tableau 2.2 – Résultats des tests de permutation selon chaque réseau. La statistique observée est le nombre de liens entre deux individus qui abandonnent. Le seuil observé est deux fois le taux de statistiques simulées inférieures à la statistique observée calculé selon l'équation (2.3).

	Statistique observée	Seuil observé	Probabilité marginale d'abandon	Probabilité conditionnelle d'abandon
Réseau 1	1936	0.001	29.14%	21.31%
Réseau 2	222 693	0	31.31%	38.77%
Réseau 3	3 990 232	0.14	26.93%	26.29%
Réseau 4	824 230	0.492	27.84%	27.05%
Réseau 5	3344	0	29.55%	34.86%

Le test de permutation permet de voir si les relations dans le réseau influencent le comportement des individus en lien avec la variable analysée. Comme ce test est bilatéral, il peut tester si le comportement de deux personnes liées est similaire ou opposé. En d'autres mots, si le nombre de relations qui unissent deux personnes ayant annulé leur police d'assurance est similaire à la distribution empirique des 100 000 simulations, le réseau n'influence pas vraiment la variable réponse. D'après les résultats du tableau 2.2, les réseaux 3 et 4 sont les seuls qui n'influencent pas l'annulation de la police d'assurance. La probabilité conditionnelle d'abandon calculée selon l'équation (2.4) donne un premier indice sur l'influence possible du réseau. Pour les réseaux 1, 2 et 5, cette probabilité est assez différente de la probabilité marginale d'abandon, ce qui est un indicateur que le réseau influence la variable réponse. Le test de

permutation vient ensuite confirmer cette hypothèse pour ces trois réseaux au seuil significatif de 5%.

$$\text{Probabilité conditionnelle} = \frac{\text{Probabilité de lien entre 2 personnes qui abandonnent}}{\text{Probabilité marginale d'abandon}} \tag{2.4}$$

Pour les réseaux ayant rejeté l'hypothèse nulle H_0, l'analyse peut être poussée un peu plus loin pour mieux comprendre l'influence qu'une relation peut avoir sur l'abandon de la police. Le réseau 1 est utilisé ici. Le graphique 2.1, présente la distribution empirique obtenue à partir des 100 000 simulations. La ligne verticale bleue correspond à la valeur de la statistique observée qui est de 1936. En étudiant le graphique, il est possible de déduire que le réseau fait diminuer le nombre de liens partagés par deux personnes ayant abandonné leur police. La grande majorité des valeurs de la distribution empirique sont supérieures à la statistique observée t_0. Le hasard prédit que le nombre de liens partagés par deux personnes ayant abandonnées leur police devrait tourner aux alentours de 2500 alors que la statistique observée à une valeur de 1936. Cette interprétation est cohérente avec les résultats obtenus au tableau 2.2 puisque la probabilité conditionnelle d'abandon du réseau 1 est plus petite que la probabilité marginale d'abandon. Donc, dans le réseau 1, si deux individus sont en relation et que l'un d'entre eux annule sa police d'assurance, la probabilité que l'autre annule également diminue.

Au contraire, pour les réseaux 2 et 5, les conclusions sont inversées. La probabilité conditionnelle d'abandon est supérieure à la probabilité marginale d'abandon. Donc, la statistique t_0 est en général plus élevée que le nombre de liens partagés par deux personnes ayant abandonné leur police prédit par le hasard. Ce qui signifie que pour les réseaux 2 et 5 le fait qu'un individu annule sa police d'assurance fait augmenter la probabilité d'abandon des individus auxquels il est lié.

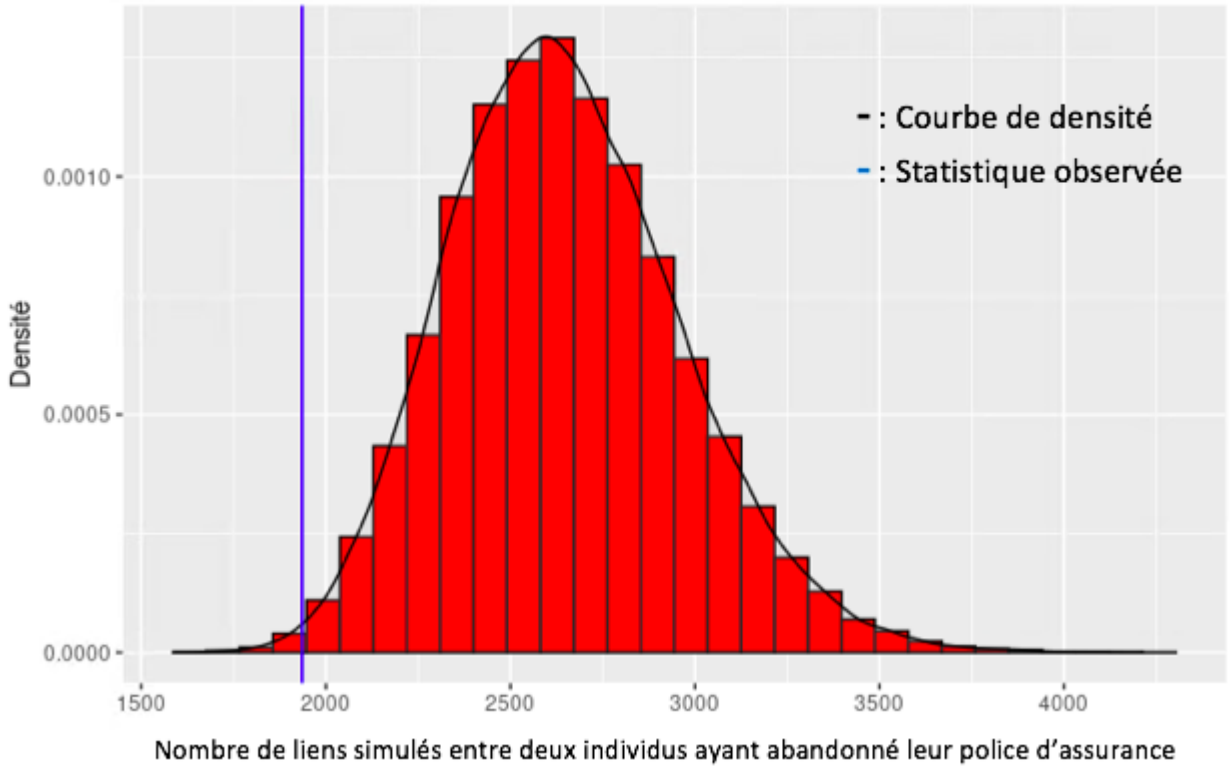


FIGURE 2.1 – Distribution empirique du test de permutation pour le réseau 1

2.2 Étude de la relation entre la force des liens du réseau et l'abandon de la police

Test de Kruskal-Wallis

Le test de permutation effectué à la section précédente permet de comprendre l'influence des relations d'un réseau sur la probabilité d'abandon des individus. Toujours en prenant en considération cette variable d'annulation de la police d'assurance, il est possible de diviser les liens des réseaux étudiés en différentes catégories. Chacune d'elles est caractérisée par le nombre d'individus ayant abandonné leur police d'assurance dans un lien. Comme chaque relation unit deux personnes, il y a trois catégories possibles. La première où aucun individu n'abandonne, la deuxième où seulement un individu abandonne et la troisième où les deux individus abandonnent leur police d'assurance. Nous avons vu, à la section 1.1, que l'équation (1.2) permet d'attribuer à chaque lien une force. Encore une fois, dans le but de mieux comprendre le lien entre les réseaux et la variable réponse, l'objectif de cette section est d'étudier l'influence que peut avoir le nombre d'abandons dans un lien sur la force de celui-ci. Ces analyses permettent aussi de déterminer s'il faut utiliser la force du lien pour représenter les interactions entre les individus dans les analyses futures. En général, l'ANOVA est utilisée pour ce type d'analyse.

Par contre, l'hypothèse de normalité n'est pas respectée sur les distributions des forces de liens. Le test non paramétrique de Kruskal-Wallis est donc employé (Kruskal and Wallis, 1952). Ce test de rang permet de comparer la distribution empirique de deux échantillons ou plus, puis de déterminer s'ils proviennent de la même distribution ou non.

Avant d'effectuer le test, une analyse visuelle peut être faite en regardant la distribution des forces de liens selon le nombre d'abandons pour le second réseau à la figure 2.2. Cette figure correspond à l'information qui est déjà disponible au tableau 1.3 séparée selon le nombre d'abandons dans un lien. Le point rouge représente la moyenne des forces de liens.

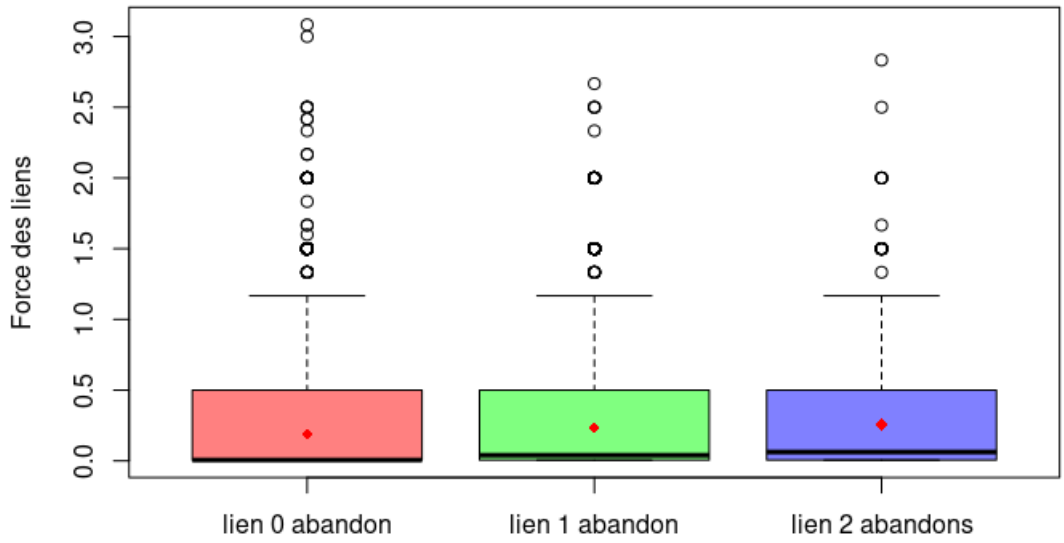


FIGURE 2.2 – Boîte à moustache des forces de lien pour le réseau 2 selon le type de lien

À première vue, les distributions semblent assez similaires. Il ne semble pas y avoir de grandes différences entre les différents quartiles entre les trois types de lien. L'analyse graphique ne permet pas de conclure franchement sur un effet du nombre d'abandons dans un lien sur sa force.

Avant de procéder au test de Kruskal-Wallis, le tableau 2.3 présente les tailles d'échantillon des différents réseaux.

Tableau 2.3 – Taille d'échantillon des types de lien selon chaque réseau

	Nombre de liens avec 0 abandon	Nombre de liens avec 1 abandon	Nombre de liens avec 2 abandons
Réseau 1	17 295	11 944	1936
Réseau 2	801 733	809 925	222 693
Réseau 3	30 333 001	22 008 162	3 990 232
Réseau 4	5 763 388	4 361 730	824 230
Réseau 5	16 664	12 451	3344

Dans les cinq réseaux, on semble retrouver les mêmes tendances. Le nombre de liens avec 2 abandons est toujours plus petits que les deux autres types de lien. De plus, ces liens qui contiennent 0 ou 1 abandon représentent la grande majorité des relations avec des proportions variant entre 88% et 94% de la taille du réseau.

Les résultats des tests de Kruskal-Wallis sont disponibles au tableau 2.4. La dernière colonne fournit la valeur p du test et permet de savoir s'il est significatif au seuil de 5%. Par contre, les trois premières colonnes sont plus intéressantes au niveau de l'analyse. Elle comporte la moyenne de la force de lien pour les trois types de lien possible. Si le test est significatif, il est alors possible de voir s'il existe une corrélation quelconque entre le nombre d'abandons dans un lien et sa force.

Le test de Kruskal-Wallis est significatif pour tous les réseaux sauf le troisième. Pour les réseaux 1 et 5, la corrélation est positive entre le nombre d'abandons dans un lien et sa force. Cela signifie qu'en général, si deux individus liés abandonnent leur police d'assurance, la force qui les unit est plus grande que s'ils venaient à conserver leur police. L'abandon de l'un des individus peut donc avoir été influencé par l'autre personne. C'est l'inverse pour les réseaux 2 et 4, puisque la corrélation est négative. Donc, dans ces réseaux, deux individus qui conservent leur police d'assurance ont un lien plus fort que s'ils venaient à l'abandonner.

Tableau 2.4 – Moyenne des forces de lien selon le type de lien et valeur p du test de Kruskal-Wallis

	Moyenne force de lien avec 0 abandon	Moyenne force de lien avec 1 abandon	Moyenne force de lien avec 2 abandons	Valeur p
Réseau 1	0.1911	0.2364	0.2566	2.2x10-16
Réseau 2	0.1292	0.1199	0.1156	2.2x10-16
Réseau 3	0.0010	0.0010	0.0010	0.8454
Réseau 4	0.0018	0.0019	0.0017	2.2x10-16
Réseau 5	0.5428	0.5584	0.5845	2.2x10-16

Test de Wilcoxon

Pour pousser l'analyse un peu plus, le test de Wilcoxon peut être employé (Wilcoxon, 1945). Contrairement au test de Krukal-Wallis, qui compare les trois types de lien ensemble, le test

de Wilcoxon va les comparer deux à deux. Encore une fois, l'analyse peut débuter avec une représentation graphique de la distribution empirique de la force des liens pour un réseau quelconque. C'est la visualisation offerte par la figure 2.3 qui compare la distribution empirique des forces de lien avec 0 et 2 abandons pour le deuxième réseau. Les deux distributions semblent suivre la même trajectoire. Par contre, la courbe des liens avec 0 abandon semble presque toujours au-dessus de celle des liens avec 2 abandons. En d'autres mots, lorsque la force de lien augmente la probabilité cumulative des liens avec 0 abandon est plus élevée que celle des liens avec 2 abandons. Donc, les forces de liens des liens avec 0 abandon semblent peut-être être plus faibles que les liens avec 2 abandons pour la force du lien. Par contre, la force maximale des liens sans abandon est plus élevée. Ces éléments suggèrent que la distribution des deux forces de lien est différente.

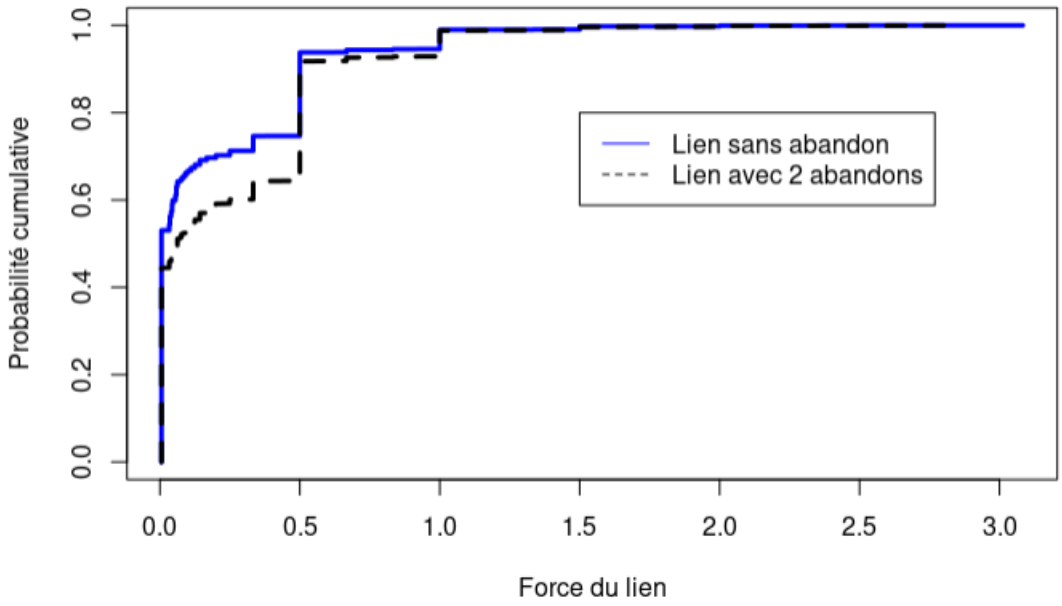


FIGURE 2.3 – Distribution empirique des forces de lien avec 0 et 2 abandons pour le réseau 2

Pour les trois combinaisons possibles des types de lien, le test de Wilcoxon offre toujours des résultats significatifs pour les réseaux 1, 2, 4 et 5 au seuil de 5%. Les différences de moyenne des forces de lien du tableau 2.4 sont donc significatives et les conclusions précédemment faites concernant les corrélations entre le nombre d'abandons dans un lien et sa force sont confirmées.

2.3 Régression logistique

Les tests non paramétriques de la section précédente ont pour objectif de voir l'influence du nombre d'annulations entre deux individus sur la force du lien qui les unit. Dans cette section,

un modèle de régression logistique est ajusté pour évaluer l'impact que peut avoir la force d'un lien sur le nombre d'annulations qui le compose. C'est donc l'effet inverse qui est étudié.

La régression logistique fait partie de la famille des modèles linéaires généralisés. Une introduction à ce type de modèle est disponible à la section 4.2. Pour utiliser un modèle de régression logistique, la variable réponse Y_i doit suivre une loi binomiale (n_i, π_i). En général, n_i prend la valeur 1 et Y_i suit alors une loi Bernoulli avec probabilité de succès π_i. L'espérance de la variable Y_i est donc égale à la probabilité de succès π_i. En régression logistique, la fonction de lien utilisée est la fonction logit définie ainsi (Duchesne, 2020),

$$\text{logit}\,(Y) = \ln\left(\frac{Y}{1-Y}\right). \tag{2.5}$$

En utilisant l'équation (4.6) et celle de la fonction logit (2.5), il est possible d'écrire l'équation du modèle de régression logistique ayant p variables explicatives ainsi Kleinbaum et al., 2002,

$$\begin{aligned} \ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 x_{i1} + ... + \beta_0 x_{ip'}, \\ \frac{\pi_i}{1-\pi_i} &= \exp\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_0 x_{ip'}\right). \end{aligned} \tag{2.6}$$

Le rapport $\frac{\pi_i}{1-\pi_i}$ est appelé cote et l'interprétation des coefficients de la régression logistique est souvent faites à partir de celui-ci. Pour que le modèle de régression logistique soit valide, quatre postulats doivent être respectés.

1. Linéarité $g\left(E\left[\mathbf{Y};\mathbf{X}\right]\right) = \mathbf{X}\beta$ où g est la fonction de lien logit
2. Homoscédasticité des termes d'erreurs
3. Indépendance des observations
4. Y_i suit une loi Bernoulli

L'homoscédasticité signifie que la variance des termes d'erreur est constante. Le modèle utilisé dans cette analyse est assez simple. L'unité d'observation est une combinaison lien-individu auquel est associé la variable Y en lien avec l'individu. Chaque lien est alors présent 2 fois dans le jeu de données qui est de taille $2m$. Pour bien comprendre la forme du jeu de données, le tableau 2.5 montre un exemple produit à partir des trois premières lignes du tableau 1.3.

Tableau 2.5 – Exemple de jeu de données pour la régression linéaire à partir d'un échantillon du tableau 1.3

Lien	Individu	Force du lien	Abandon de la police
1	1	5/6	1
1	2	5/6	0
2	1	5/6	1
2	4	5/6	0
3	2	1/3	0
3	4	1/3	0

L'objectif étant de voir l'influence de la force d'un lien sur la probabilité d'annulation dans ce lien, il y a donc seulement une variable explicative qui est la force de lien. L'équation (2.6) peut donc être réécrite de cette façon,

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \beta_1 \text{force de lien}_i\right), \tag{2.7}$$

où π_i représente la probabilité d'annulation de la police d'assurance dans un lien du réseau. Les résultats sont présentés au tableau 2.6.

Tableau 2.6 – Estimations du coefficient associé à la force de lien selon chaque réseau

	Coefficient β_1	Écart-type	Valeur p
Réseau 1	0.3701	0.027	0
Réseau 2	-0.3743	0.008	0
Réseau 3	-0.0221	0.017	0.181
Réseau 4	0.0619	0.021	0.004
Réseau 5	0.4153	0.038	0

Le réseau 3 est donc le seul pour lequel la force d'un lien ne semble pas influencer le nombre d'annulations pour une paire d'individus liés par le réseau puisque le coefficient en lien avec la variable explicative n'est pas significativement différent de 0. Pour les réseaux 1, 4 et 5 la valeur du coefficient β_1 est positive et donc une augmentation de la force d'un lien augmente la probabilité d'observer une annulation. Plus précisément, en prenant le réseau 5, si la force d'un lien augmente d'une unité, la cote $\frac{\pi_i}{1-\pi_i}$ est multipliée par $\exp 0.4153 = 1.5148$. Pour le deuxième réseau, la corrélation entre la probabilité d'abandon dans un lien et sa force est négative puisque l'estimation du coefficient β_1 est négative.

2.4 Bilan

Au cours de ce chapitre, quatre méthodes statistiques ont été présentées pour essayer de comprendre l'impact qu'un réseau peut avoir sur une variable aléatoire discrète. Le test de permutation a permis d'évaluer si les liens d'un réseau influencent le comportement des individus quant à l'abandon de leur police d'assurance. Suite à l'étude, cela semble être le cas pour les réseaux 1, 2 et 5. Cela suppose que le fait de tenir compte des relations présentes dans ces réseaux permettrait de mieux prédire l'annulation d'une police d'assurance d'un individu. Un des points forts de ce test réside dans le fait qu'aucun postulat n'est fait sur les données analysées. Par contre, le test développé considère uniquement les liens de longueur 1 présents dans le réseau. Normalement, ces relations sont les plus fortes. Par contre, un individu pourrait en influencer un autre même s'ils partagent seulement un chemin de longueur 2 ou plus et le test de permutation n'en tient pas compte.

Par la suite, le test non paramétrique de Kurskal-Wallis a permis d'étudier l'influence que pouvait avoir le nombre d'annulations dans un lien sur sa force. Trois types de lien ont été étudiés selon le nombre d'annulations 0, 1 ou 2. Le test a donné des résultats très intéressants. Seuls les types de lien du réseau 3 ne semblaient pas influencer la force des liens entre deux individus. Un des avantages du test est qu'il permet de comparer plus de deux groupes en même temps. Par contre, on ne peut pas savoir directement quels sont les groupes qui sont différents si le test est significatif. Le test de Wilcoxon permet de compléter cette analyse en comparant les groupes deux à deux. Ce test a permis de conclure que la force de lien était significativement différente entre les trois types de lien dans les réseaux 1, 2, 4 et 5. De plus, il a été possible d'extraire des corrélations entre le type de lien et sa force. Pour les réseaux 2 et 4 la corrélation est négative et pour les réseaux 1 et 5 elle est positive. Ce test avait également comme objectif de voir si la force de lien calculée par l'équation (1.2) est une bonne mesure et s'il fallait en tenir compte pour modéliser la variable d'abandon de la police d'assurance. Au vu des résultats, calculer la force d'un lien au lieu de le considérer comme une variable binaire peut être une bonne idée pour tous les réseaux sauf le troisième. Encore une fois, l'avantage de ces tests réside dans le fait qu'il n'y a aucun postulat à respecter sur les données. Par contre, comme ces tests utilisent la distribution empirique des données, un trop faible nombre d'observations peut entrainer des résultats approximatifs. Une autre limite aux tests de Kruskal-Wallis et Wilcoxon est que même si une différence significative est détectée, il n'est pas possible d'attribuer une force à l'influence du facteur sur la variable réponse.

C'est pour cela que la régression logistique est la dernière méthode présentée dans ce chapitre. Le modèle régressif permet d'attribuer une force à l'impact que peut avoir un facteur sur une variable réponse. Cette fois-ci, c'est la probabilité d'abandon qui a été modélisée à partir de la force des liens. Tout comme pour le test de permutation, les liens de longueur plus grande que 1 n'ont pas été étudiés ce qui limite quelque peu les résultats puisqu'ils auraient pu avoir une influence sur la probabilité d'abandon. Selon la régression logistique, la force d'un lien a de

l'influence sur la probabilité d'abandon de la police d'assurance pour tous les réseaux sauf le troisième. La régression linéaire et le test de Wilcoxon offrent des corrélations similaires entre la probabilité d'abandon et la force de lien pour les réseaux 1 et 5 (corrélation positive) et pour le réseau 2 (corrélation négative). Par contre, pour le quatrième réseau, la corrélation obtenue avec la régression linéaire est positive alors qu'elle est négative avec le test de Wilcoxon.

En conclusion, les différentes analyses permettent de mieux comprendre l'influence d'un réseau sur la variable d'abandon de la police d'assurance sous différents angles. Les résultats changent légèrement d'une méthode à une autre. Par contre, en général, les réseaux 1, 2 et 5 sont ceux qui semblent avoir le plus d'influence sur cette variable, étant ceux qui ressortent les plus souvent dans les conclusions des différentes analyses. Quant au réseau 3, il semble être celui qui influence le moins le fait qu'un individu abandonne sa police d'assurance puisqu'aucun test effectué sur lui n'est significatif. Le test de Wilcoxon et la régression logistique ont montré que la force d'un lien définie avec l'équation (1.2) peut être utile pour modéliser la variable d'abandon de la police d'assurance dans les réseaux $1, 2, 4$ et 5.

Chapitre 3

Modélisation d'une variable continue à partir d'un réseau

Les précédents chapitres ont permis de constater que l'information contenue dans un réseau peut influencer une variable aléatoire. Dans certains cas, il peut donc être intéressant d'introduire cette information dans la modélisation d'une variable aléatoire. L'idée principale est d'extraire à partir du réseau une matrice de covariance qui peut être utilisée par la suite dans un modèle prédictif. Le défi est d'obtenir une matrice qui résume bien l'information du réseau en plus de s'assurer qu'elle soit au minimum semi-définie positive (Strang, 2006). L'article de Lan et al., 2018 propose une approche simple et intuitive pour estimer une matrice de covariance à partir d'un réseau. C'est cette méthode qui est présentée dans ce chapitre.

3.1 Approche de Lan et al., 2018

3.1.1 Introduction

En 2018, Lan et al., 2018 ont publié une approche qui permet d'estimer, à partir d'un réseau, une matrice de covariance. L'intuition derrière l'approche est que le réseau explique une partie de la variable réponse Y et une autre partie est expliquée par des facteurs externes. Reprenons le réseau présenté à la figure 1.2. Posons une nouvelle variable réponse Y qui représente le nombre d'amis de chaque individu. Cette variable Y peut être décomposée en deux parties. La première représente les amis d'un individu faisant partie du réseau. Par exemple, si les groupes du réseau représentent des équipes sportives, cela peut être vu comme étant les amitiés qui se créent lors de la pratique d'un sport. Quant à la seconde, elle représente les amis qu'un individu a à l'extérieur du réseau. Donc, cela peut contenir les relations créées au travail, à l'école, dans la famille, etc. L'influence que le réseau peut avoir sur la variable Y est simple à comprendre. Plus un individu a de relations dans le réseau (il fait partie d'une équipe nombreuse, il fait partie de plus d'une équipe), plus la probabilité qu'il ait beaucoup d'amis augmente.

Il devient alors nécessaire de prendre en considération la structure du réseau lors de la modélisation de cette variable Y. Le modèle proposé par Lan et al., 2018 permet d'estimer la matrice de variance covariance Σ, qui n'est pas observée en pratique, comme étant une fonction polynomiale de la matrice d'adjacence A. L'estimation de la matrice Σ qui, en général, est un problème de haute dimensionnalité, est réduit à un problème d'estimation à faible dimensionnalité. Il suffit effectivement d'estimer uniquement les coefficients de l'équation polynomiale. De plus, la méthode requiert uniquement l'observation de la matrice A, ce qui est souvent le cas en pratique comme pour le réseau des fictifs au tableau 1.4.

3.1.2 Modélisation

La matrice d'adjacence A représente les relations directes qui existent dans un réseau. Si deux individus i et j partagent un lien l'élément a_{ij} prend la valeur 1 et 0 sinon. Introduisons, maintenant, la notion de degré à la matrice d'adjacence. La matrice A^k représente alors la $k^{\text{ième}}$ puissance de la matrice A. La valeur prise par l'élément $a_{ij}^{(k)}$ représente le nombre de chemins de longueur k entre les deux individus i et j. En prenant le carré de la matrice d'adjacence du tableau 1.4 on obtient la matrice suivante.

Tableau 3.1 – Matrice d'adjacence au carré A^2

	1	2	3	4	5	6	7	8
1	2	1	0	1	0	0	0	0
2	1	2	0	1	0	0	0	0
3	0	0	1	0	0	1	0	0
4	1	1	0	2	0	0	0	0
5	0	0	0	0	2	0	1	0
6	0	0	1	0	0	2	0	0
7	0	0	0	0	1	0	1	0
8	0	0	0	0	0	0	0	0

Donc, toutes les valeurs présentes dans la matrice 3.1 représentent le nombre de chemins de longueur 2 entre les différents individus.

Par la suite, Lan et al., 2018 posent un premier modèle pour estimer la matrice de variance covariance. Celui-ci ne demande en entrée qu'une matrice identité de dimension $n \times n$ ainsi que la matrice d'adjacence et deux coefficients estimés,

$$cov(\mathbb{Y}|A) = \Sigma(A) = \beta_0 I_n + \beta_1 A, \tag{3.1}$$

où $\mathbb{Y}$ représente le vecteur des valeurs de la variable Y introduite précédemment. Dans cette première version, seulement les chemins de longueur 1 sont utilisés pour modéliser la matrice

Σ. Il est possible de généraliser la formule pour permettre de prendre en considération des chemins de différentes longueurs (k). L'équation généralisée est donc la suivante,

$$cov(\mathbb{Y}|A) = \Sigma(A) = \sum_{k=0}^{K} \beta_k A^k = \beta_0 I_n + \beta_1 A + ... + \beta_K A^K, \tag{3.2}$$

où K est le degré du polynôme. En d'autres mots, c'est la longueur maximale des chemins du réseau à considérer dans l'estimation de la matrice Σ.

3.1.3 Interprétation du modèle

En regardant l'équation (3.2), on retrouve une fonction semblable à celle d'une régression linéaire. L'ordonnée à l'origine représentée par le coefficient β_0 vient multiplier la matrice identité qui représente l'information externe au réseau. Quant aux coefficients $(\beta_1, ..., \beta_K)$ qui multiplent les matrices $\left(A, ..., A^K\right)$, ils représentent l'information pouvant être expliquée par le réseau.

Degré 1

Prenons le cas d'un modèle de degré 1 s'écrivant comme l'équation (3.1). Dans ce modèle, seulement les paires d'individus ayant un lien direct sont considérées. Dans ce cas, le coefficient attribué à β_0 représente la variance de la variable réponse $Var(\mathbb{Y})$. Pour ce qui est de β_1, son coefficient représente la valeur de la covariance entre deux individus liés directement par le réseau. Il devient donc possible d'estimer la corrélation entre deux individus selon les deux cas de figure suivants :

- $cor(Y_i, Y_j) = \frac{\beta_1}{\beta_0}$ $\quad i$ et j sont liés dans le réseau
- $cor(Y_i, Y_j) = 0$ $\quad$ sinon.

Degré 2

Prenons maintenant le cas d'un modèle de degré 2. À partir de l'équation (3.2), le modèle s'écrit sous la forme suivante,

$$cov(\mathbb{Y}|A) = \beta_0 I_p + \beta_1 A + \beta_2 A^2.$$

La variance des individus s'écrit maintenant ainsi, $Var(Y_i) = \beta_0 + \text{degré i} * \beta_2$. Le calcul de la covariance entre deux individus se complique lui aussi quelque peu. Il faut considérer les quatre cas de figure suivant :

- $cov(Y_i, Y_j) = \beta_1$ — i et j partagent uniquement un lien direct
- $cov(Y_i, Y_j) = \beta_1 + \beta_2 * [A^2]_{ij}$ — i et j partagent un lien direct et au moins un lien de longueur 2
- $cov(Y_i, Y_j) = \beta_2 * [A^2]_{ij}$ — i et j ne partagent aucun lien direct, mais au moins un lien de longueur 2
- $cov(Y_i, Y_j) = 0$ — sinon

Dans le deuxième cas de figure, $[A^2]_{ij}$ représente l'élément ij de la matrice d'adjacence A au carré. Comme mentionné précédemment, c'est le nombre de chemins entre i et j de longueur 2.

La corrélation entre deux individus peut maintenant être présentée encore une fois selon les trois situations suivantes :

- $cor(Y_i, Y_j) = \frac{\beta_1}{\sqrt{(\beta_0+d_i*\beta_2)*(\beta_0+d_j*\beta_2)}}$ — i et j partagent uniquement un lien direct
- $cor(Y_i, Y_j) = \frac{\beta_1+\beta_2*[A^2]_{ij}}{\sqrt{(\beta_0+d_i*\beta_2)*(\beta_0+d_j*\beta_2)}}$ — i et j partagent un lien direct et au moins un lien de longueur 2
- $cor(Y_i, Y_j) = \frac{\beta_2*[A^2]_{ij}}{\sqrt{(\beta_0+d_i*\beta_2)*(\beta_0+d_j*\beta_2)}}$ — i et j ne partagent aucun lien direct, mais au moins un lien de longueur
- $cor(Y_i, Y_j) = 0$ — sinon

Au niveau du dénominateur représentant la variance entre i et j, d_i et d_j représentent les degrés respectifs de ces individus dans la matrice d'adjacence A. Pour rappel, le degré d'un individu peut être obtenu à l'aide de la matrice d'adjacence A en sommant sur la ligne ou la colonne respective comme il est présenté à l'équation (1.3).

Peu importe le nombre de degrés choisi pour le modèle, la valeur de la corrélation entre deux individus est très intéressante puisqu'elle permet de comprendre l'influence que l'un peut avoir sur l'autre. Supposons une corrélation positive, cela signifie que plus la valeur de Y est élevée pour l'individu i, plus celle de l'individu j tendra à être élevée également. Au contraire, si la corrélation est négative alors plus la valeur de Y est élevée pour l'individu i, plus celle de l'individu j tendra à être faible cette fois-ci. Puis, si la corrélation est nulle, cela signifie que les individus i et j ne s'influencent pas sur la valeur que prend la variable étudiée Y.

Il y a un parallèle intéressant à faire entre les corrélations obtenues à partir du modèle de Lan et al., 2018 et le coefficient d'assortativité présenté à la sous-section 1.1.6. Pour rappel, le coefficient d'assortativité mesure à quel point les liens qui existent dans un réseau permettent d'expliquer les valeurs prises par une variable aléatoire. Au degré 1, l'équation (3.1) permet donc de calculer une assortativité théorique avec la formule suivante,

$$\text{assortativité théorique} = \frac{\beta_1}{\beta_0}. \tag{3.3}$$

3.1.4 Espace des paramètres

Le modèle de Lan et al., 2018 permet, à partir de l'information contenue dans un réseau, de formuler une matrice de covariance. Cependant, cette matrice doit être définie positive. Pour

être en mesure de respecter cette caractéristique, une condition s'impose sur l'estimation des paramètres β du modèle qui va restreindre l'espace des valeurs qu'ils peuvent prendre.

Posons $\lambda_{min}(A)$ et $\lambda_{max}(A)$ les valeurs propres minimale et maximale de la matrice A. Définissons $\lambda^*_{max}(A) = max\left\{|\lambda_{min}(A)|, |\lambda_{max}(A)|\right\}$ comme étant la valeur propre absolue maximale. On oblige par la suite $\lambda^*_{max}(A)$ a être bornée supérieurement $\lambda^*_{max}(A) \leq a_{max}$, a_{max} étant une constante positive. L'espace des paramètres peut donc être construit comme suit,

$$\Theta = \left\{ \beta : \sum_{k=0}^{K} \beta_k \lambda^k > 0 \quad \forall\, \lambda\, \epsilon\, [-a_{max}, a_{max}] \right\}, \tag{3.4}$$

où K est le degré du polynôme. En respectant cette condition sur l'estimation des β, on s'assure que la matrice de variance covariance obtenue par l'équation (3.2) est définie positive.

3.1.5 Estimation des paramètres

Cette sous-section présente quatre algorithmes permettant d'estimer les paramètres du modèle de Lan et al., 2018.

Moindres carrés

Pour être en mesure d'estimer les paramètres de l'équation (3.2), il faut transformer la variable Y. La première étape est de centrer le vecteur de la variable réponse. La deuxième étape consiste, par la suite, à multiplier le vecteur centré par la matrice des vecteurs propres Q de la matrice d'adjacence A. On a alors $\tilde{\mathbb{Y}} = Q^\top \mathbb{Y}_{\text{centré}}$. Le nouveau vecteur $\tilde{\mathbb{Y}}$ suit donc une distribution de moyenne 0 et de matrice de variance covariance $\sum_{k=0}^{K} \beta_k D^k$ où D est la matrice diagonale composée des valeurs propres de A. On peut facilement montrer que $E\left(\tilde{\mathbb{Y}}^2\right) = \sum_{k=0}^{K} \beta_k \lambda^k$. L'espérance du vecteur $\tilde{\mathbb{Y}}$ au carré est donc une fonction linéaire des valeurs propres de la matrice d'adjacence A. Il suffit alors de modéliser $\tilde{\mathbb{Y}}^2$ en fonction des valeurs propres λ de la matrice A pour trouver les coefficients β. Pour ce faire, la méthode des moindres carrés peut-être utilisée en minimisant l'équation suivante,

$$\hat{\beta} = \arg\min_{\beta\, \exists\, \Theta} \sum_{i=0}^{n} \left(\tilde{Y}^{(i)2} - \sum_{k=0}^{K} \beta_k \lambda^k \right)^2. \tag{3.5}$$

Elle peut être employée en modélisant $\tilde{\mathbb{Y}}^2$ dans une régression linéaire classique.

Maximum de vraisemblance de Lan et al., 2018

Le deuxième algorithme permettant d'estimer les β est celui du maximum de vraisemblance présenté par Lan et al., 2018. Dans la procédure, il postule d'abord que la moyenne de la variable aléatoire $\mu_{\mathbb{Y}}$ est égale à la moyenne empirique de ses valeurs $m_{\mathbb{Y}}$. Puis, suite à la

transformation décrite précédemment avec la méthode des moindres carrés, on obtient $\mathbb{Y} \sim N_n(0, \beta_0 I + \beta_1 A)$. Il suffit de minimiser l'équation suivante,

$$\hat{\beta} = \arg\min_{\beta \exists \Theta}(-\mathcal{L}(\beta)) = \arg\min_{\beta \exists \Theta} np\log(2\pi) + \sum_{i=1}^{n}\left\{\tilde{Y}^{(i)2}\left(X^{\top}\beta\right)^{-1} + \log\left(X^{\top}\beta\right)\right\}. \quad (3.6)$$

Dans cette équation, le vecteur X est de longueur $K+1$. La première colonne est toujours composée de valeur 1 et les suivantes sont les valeurs propres $\lambda, \lambda^2, ..., \lambda^K$.

Maximum de vraisemblance

L'algorithme du maximum de vraisemblance peut aussi être utilisé en considérant la dépendance des valeurs de $\mathbb{Y}$. Posons la décomposition de A en valeur propre comme étant $A = QD(\lambda_i)Q^T$. Posons également le vecteur $\mathbb{Y}$ de taille $n \times 1$ suivant une loi $N_n(\mu\mathbf{1}, \beta_0 I + \beta_1 A)$. La principale différence avec l'approche de Lan et al., 2018 est au niveau de la moyenne de $\mathbb{Y}$ qui est estimée ici en fonction des paramètres β_0 et β_1,

$$f(y|\mu, \beta_0, \beta_1) = \frac{1}{(2\pi)^{\frac{n}{2}}|\beta_0 I + \beta_1 A|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(y-\mu\mathbf{1})^T(\beta_0 I + \beta_1 A)^{-1}(y-\mu\mathbf{1})\right). \quad (3.7)$$

Pour déterminer le μ qui maximise la vraisemblance, il faut l'écrire comme une fonction de β_0 et β_1,

$$\mu_\beta = \frac{1^T(\beta_0 I + \beta_1 A)^{-1} y}{1^T(\beta_0 I + \beta_1 A)^{-1} 1} = \frac{1^T Q(\beta_0 I + \beta_1 D(\lambda_i))^{-1} Q^T y}{1^T Q(\beta_0 I + \beta_1 D(\lambda_i))^{-1} Q^T 1} = \frac{\sum_{i=1}^{n}\frac{q_i y_{qi}}{\beta_0 + \beta_1\lambda_i}}{\sum_{i=1}^{n}\frac{q_i}{\beta_0+\beta_1\lambda_i}}. \quad (3.8)$$

La moyenne de $\mathbb{Y}$ est une moyenne pondérée par l'inverse de la variance. L'idée est que plus la variance d'une observation est grande, plus sont poids est faible dans le calcul puisqu'il y a plus d'incertitude reliée à cette valeur. La précédente formule fait intervenir deux vecteurs $q = (q_1, ..., q_n)^T$ et $y_q = (y_{q1}, ..., y_{qn})^T$. Ils ne dépendent pas des paramètres inconnus et peuvent être calculés une seule fois ainsi :

$q = Q^T 1$ et $y_q = Q^T y$.

Ensuite, pour être en mesure dévaluer l'équation (3.7) on simplifie le terme dans l'exponentielle ainsi,

$$-\frac{1}{2}(y-\mu\mathbf{1})^T(\beta_0 I + \beta_1 A)^{-1}(y-\mu\mathbf{1}),$$

$$= y^T Q(\beta_0 I + \beta_1 D(\lambda_i))^{-1} Q^T y - 2\mu_\beta y^T Q(\beta_0 I + \beta_1 D(\lambda_i))^{-1} Q^T 1 + \mu_\beta^2 1^T Q(\beta_0 I + \beta_1 D(\lambda_i))^{-1} Q^T \mathbf{1},$$

$$= y^T Q \left(\beta_0 I + \beta_1 D(\lambda_i)\right)^{-1} Q^T y - \mu_\beta^2 1^T Q \left(\beta_0 I + \beta_1 D(\lambda_i)\right)^{-1} Q^T 1,$$

$$= \sum_{i=1}^{n} \frac{y_{qi}^2}{\beta_0 + \beta_1 \lambda_i} - \frac{\left(\sum_{i=1}^{n} \frac{q_i y_{qi}}{\beta_0 + \beta_i \lambda_i}\right)^2}{\sum_{i=1}^{n} \frac{q_i^2}{\beta_0 + \beta_1 \lambda_i}}.$$

Finalement, moins 2 fois la log-vraisemblance s'écrit ainsi,

$$-2\log\left\{f\left(y|\mu,\beta_0,\beta_1\right)\right\} = \log\left[\sum_{i=1}^{n}\left(\beta_0+\beta_1\lambda_i\right)\right] + \sum_{i=1}^{n}\frac{y_{qi}^2}{\beta_0+\beta_1\lambda_i} - \frac{\left(\sum_{i=1}^{n}\frac{q_i y_{qi}}{\beta_0+\beta_1\lambda_i}\right)^2}{\sum_{i=1}^{n}\frac{q_i^2}{\beta_0+\beta_1\lambda_i}}. \quad (3.9)$$

Cette fonction dépend uniquement des paramètres β_0 et β_1.

Assortativité

Il a été discuté à la fin de la sous-section 3.1.3 du parallèle qui existe entre la mesure d'assortativité et le modèle de Lan et al., 2018. Ce lien permet de développer une façon d'estimer les paramètres d'un modèle de degré 1.

Cette méthode des moments permet d'estimer les coefficients sans recourir à un algorithme particulier. Dans le cas où le modèle est de degré 1, le coefficient de β_0 est égal à la variance de $\mathbb{Y}$. Il est donc facile de calculer cette valeur. De plus, la corrélation entre Y_i et Y_j est égale au coefficient d'assortativité théorique comme il a été vu à l'équation (3.3).

Le seul élément inconnu est le β_1 qui peut facilement être isolé ainsi,

$$\beta_1 = \text{assortativité théorique} * \beta_0. \quad (3.10)$$

L'estimation des paramètres du modèle de Lan et al., 2018 au degré 1 se fait assez simplement en utilisant le vecteur $\mathbb{Y}$ et la mesure d'assortativité.

3.1.6 Application

Reprenons le réseau du chapitre 1 et calculons la matrice de covariance avec l'approche de Lan et al., 2018. Posons le vecteur $\mathbb{Y} = (15, 8, 16, 5, 3, 4, 7, 13)^\top$ représentant le nombre d'amis de chaque individu du réseau. La première étape est de centrer le vecteur $\mathbb{Y}$. La moyenne étant de 8.875, en la soustrayant de chaque élément, le vecteur prend maintenant les valeurs suivantes $(6.125, -0.875, 7.125, -3.875, -5.875, -4.875, -1.875, 4.125)^\top$. La deuxième étape consiste à faire la multiplication matricielle des vecteurs propres de A par le vecteur Y centré. On a donc,

$$\tilde{\mathbb{Y}} = Q^\top \mathbb{Y} =$$

$$\begin{bmatrix} 0.577 & 0.577 & 0 & 0.577 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.372 & 0 & 0.602 & 0.602 & 0.372 & 0 \\ 0 & 0 & -0.602 & 0 & -0.372 & 0.372 & 0.602 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.602 & 0 & -0.372 & -0.372 & 0.602 & 0 \\ 0.816 & -0.408 & 0 & -0.408 & 0 & 0 & 0 & 0 \\ 0 & 0.707 & 0 & -0.707 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.372 & 0 & -0.602 & 0.602 & -0.372 & 0 \end{bmatrix} \times \begin{bmatrix} 6.125 \\ -0.875 \\ 7.125 \\ -3.875 \\ -5.875 \\ -4.875 \\ -1.875 \\ 4.125 \end{bmatrix} = \begin{bmatrix} 0.794 \\ -4.514 \\ -5.042 \\ 4.125 \\ 7.154 \\ 6.940 \\ 2.121 \\ 3.947 \end{bmatrix}$$

L'étape suivante est de mettre au carré les éléments du vecteur $\tilde{\mathbb{Y}}$. Le nouveau vecteur est donc $\tilde{\mathbb{Y}}^2 = (0.630, 20.380, 25.419, 17.016, 51.182, 48.167, 4.500, 15.581)^\top$.

Avant d'estimer les paramètres β, il faut regarder l'espace des paramètres à respecter pour les estimations. Les valeurs propres de la matrice d'adjacence sont les suivantes :

$$(2, 1.618, 0.618, 0, -0.618, -1, -1, -1.618)^\top .$$

La valeur de $\lambda^*_{max}(A)$ est donc de 2. La borne supérieure des valeurs propres est donc $a_{max} = 2$. Le développement de l'équation (3.4) aux deux extrémités $-a_{max}$ et a_{max} offre les deux inégalités suivantes :

1. $\beta_0 - 2\beta_1 > 0$
2. $\beta_0 + 2\beta_1 > 0$

Ces deux inégalités représentent en réalité des plans dans un espace en deux dimensions et ils peuvent être visualisés sur l'image 3.1.

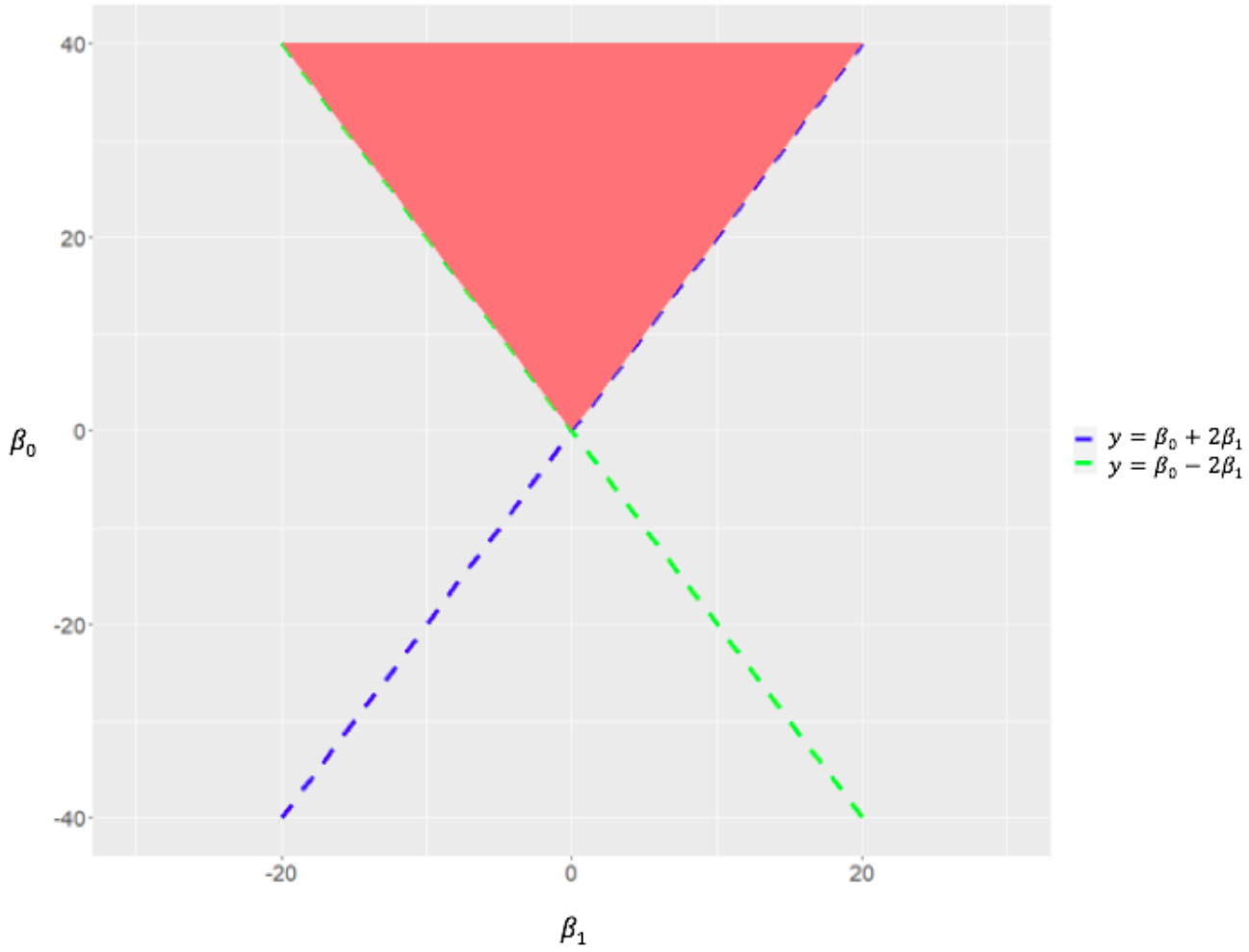


FIGURE 3.1 – Espace des paramètres

L'étape suivante consiste à effectuer une régression linéaire pour estimer les coefficients β qui sont utilisés pour calculer la matrice de covariance comme dans l'équation (3.1). La variable dépendante est le vecteur $\tilde{\mathbb{Y}}^2$ et la variable indépendante est le vecteur des valeurs propres de la matrice d'adjacence $\mathbb{X} = (2, 1.618, 0.618, 0, -0.618, -1, -1, -1.618)^\top$.

Suite à la régression, la valeur de β_0 est estimée à 22.859 et celle de β_1 à -4.964. Il suffit alors de reprendre l'équation (3.1) pour estimer la matrice de variance covariance Σ. Le calcul est le suivant,

$$22.859 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} - 4.964 \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} =$$

$$\begin{bmatrix} 22.859 & -4.964 & 0 & -4.964 & 0 & 0 & 0 & 0 \\ -4.964 & 22.859 & 0 & -4.964 & 0 & 0 & 0 & 0 \\ 0 & 0 & 22.859 & 0 & -4.964 & 0 & 0 & 0 \\ -4.964 & -4.964 & 0 & 22.859 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4.964 & 0 & 22.859 & -4.964 & 0 & 0 \\ 0 & 0 & 0 & 0 & -4.964 & 22.859 & -4.964 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4.964 & 22.859 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 22.859 \end{bmatrix}$$

Les valeurs estimées des paramètres β_0 et β_1 satisfont la condition de la figure 3.1 ce qui assure que la matrice engendrée est définie positive.

En analysant de plus près la matrice de covariance obtenue, on se rend compte que la diagonale représente l'effet hors réseau représenté par le coefficient β_0. Puis, l'effet réseau est représenté par le coefficient β_1 et il est attribué à chaque paire d'individus partageant un chemin de longueur 1. Cela était attendu puisque dans l'approche seuls les chemins de longueur 1 ont été considérés. Des valeurs entre deux individus partageant des chemins de longueur 2 ou plus peuvent être ajoutées à la matrice de covariance en la modélisant selon l'équation (3.2) avec une valeur de $K > 1$.

En interprétant les résultats obtenus, on s'aperçoit que la corrélation entre deux individus liés dans le réseau $cor(Y_i, Y_j) = \frac{\beta_1}{\beta_0} = \frac{-4.964}{22.859} = -0.217$ est négative. Cela signifie, qu'en général, si l'individu i a un nombre d'amis élevé, l'individu j a tendance à avoir un nombre d'amis faible et vice versa. Les différentes corrélations entre les individus se retrouvent dans le tableau 3.2. La variance du modèle est donnée par le coefficient β_0, 22.859.

Tableau 3.2 – Matrice des corrélations

	1	2	3	4	5	6	7	8
1	1	−0.217	0	−0.217	0	0	0	0
2	−0.217	1	0	−0.217	0	0	0	0
3	0	0	1	0	−0.217	0	0	0
4	−0.217	−0.217	0	1	0	0	0	0
5	0	0	−0.217	0	1	−0.217	0	0
6	0	0	0	0	−0.217	1	−0.217	0
7	0	0	0	0	0	−0.217	1	0
8	0	0	0	0	0	0	0	1

Les paramètres ont été estimés avec l'algorithme des moindres carrées décrit à l'équation (3.5). Comme il a été présenté précédemment, il existe d'autres algorithmes d'estimation des paramètres et les résultats sont présentés au tableau 3.3. À première vue, les estimations des paramètres semblent être cohérentes entre les différents algorithmes. Les signes des trois paramètres sont toujours les mêmes. Le ratio $\hat{\tau}$ représente la corrélation entre deux individus. Ayant un signe négatif, cela signifie que la corrélation est négative entre deux individus liés dans le réseau selon les quatre méthodes d'estimation. Au niveau des valeurs des paramètres, elles semblent également être assez constantes d'un algorithme à l'autre sauf pour l'algorithme de maximum de vraisemblance de Lan et al., 2018 qui obtient des valeurs plus élevées. L'algorithme des moindres carrés est celui qui a les plus petites variances pour les paramètres β_0 et β_1. Cependant, cet exemple n'est pas conçu pour comparer les propriétés des différents estimateurs. Cette étude est faite dans la prochaine sous-section.

Tableau 3.3 – Estimations des paramètres selon chacun des algorithmes

	EMV Lan	EMV	Moindres carrés	Assortativié
$\hat{\beta}_0$	32.4373	19.2608	22.859	26.125
$\hat{\sigma}_{\beta_0}$	18.4015	10.0147	6.580	X
$\hat{\beta}_1$	-15.8243	-4.2316	-4.9640	-7.4464
$\hat{\sigma}_{\beta_1}$	9.3761	7.3573	5.372	X
$\hat{\tau}$	-0.4878	-0.2197	-0.2171	-0.2850

En conclusion, l'approche proposée par Lan et al., 2018 est très intéressante pour être en mesure d'estimer une matrice de variance covariance à partir d'un réseau et d'une variable réponse. La méthodologie permet de réduire un problème d'estimation de grande dimensionnalité en un problème d'estimation à faible dimensionnalité. Si la matrice d'adjacence A est observée, il suffit d'estimer uniquement K paramètres β selon le degré du modèle choisi. L'interprétation des résultats obtenus permet aussi de mesurer la corrélation entre deux individus selon les différents liens qui peuvent les unir. Une des limites de l'approche est qu'elle est conçue pour

traiter une variable réponse continue. Elle ne peut donc pas être utilisée si la variable d'intérêt est discrète ou catégorielle.

3.2 Propriétés échantillonnales des différents estimateurs

Pour comparer les méthodes d'estimation des paramètres, on utilise une étude Monte-Carlo (Mooney, 1997). On initialise une matrice d'adjacence A, puis, pour chaque itération, un vecteur de valeurs pour la variable aléatoire Y est simulé. Par la suite, les différents algorithmes d'estimation des paramètres sont employés et les résultats sont compilés. Ainsi, il est possible de comparer les résultats et voir les avantages et les inconvénients des différents algorithmes.

Application

L'objectif ici est de simuler un réseau Bernoulli pour tester les quatre estimateurs des paramètres présentés à la sous-section 3.1.5. Le réseau est formé de 1000 individus et la probabilité qu'un lien existe entre deux personnes est de 0.005. Il n'existe aucun lien qui part d'un nœud et qui revient sur celui-ci. La diagonale de la matrice d'adjacence est donc composée de 0 uniquement.

Le nombre de liens dans le réseau simulé est de 2534 et la moyenne des degrés des nœuds est de 5.068. La figure 3.2 permet de voir la distribution des degrés du réseau. Comme il a été vu à la sous-section 1.2.2, cette distribution suit une loi de Poisson dans un réseau Bernoulli.

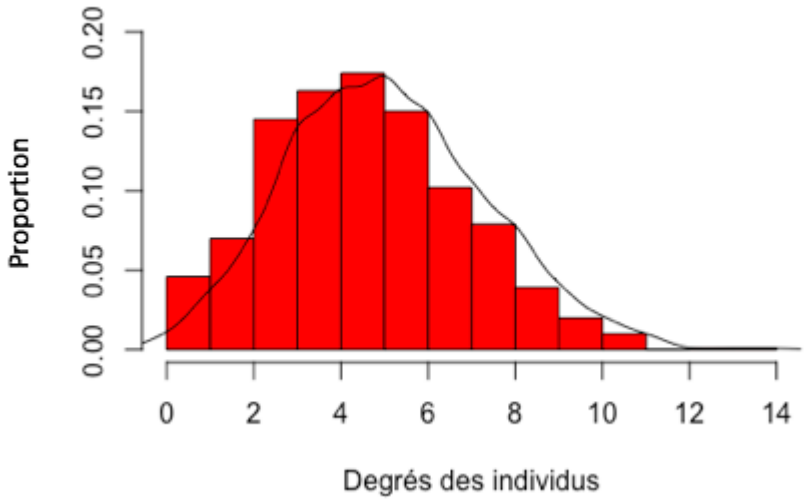


FIGURE 3.2 – Distribution des degrés des individus dans le réseau utilisé dans les simulations

Le réseau simulé contient également une composante géante qui regroupe 99.10% des nœuds,

ce qui signifie que parmi les 1000 individus formant le réseau, uniquement 9 d'entre eux ne sont pas liés à la composante géante. Ce type de composante est présent dans le réseau puisque la moyenne des degrés est supérieure à 1, comme il a été discuté à la sous-section 1.2.4.

Le tableau 3.4 présente certaines statistiques descriptives pour le réseau utilisé dans l'étude de Monte-Carlo.

Tableau 3.4 – Statistiques sur les composantes du réseau de l'étude de Monte-Carlo

Nb compo	Transitivité	Taille minimale des composantes	Taille maximale des composantes	Taille moyenne des composantes	Longueur moy des chemins	Diamètre
10	0.005	1	991	100	4.42	10

La composante géante regroupe 99.10% des individus et les 9 autres sont tous orphelins. Dans cette composante, on peut passer d'un nœud à un autre en passant par 4.42 individus en moyenne. La valeur de la transitivité est de 0.005. La structure triangulaire est donc assez rare dans le réseau puisque si deux nœuds partagent un voisin commun, ceux-ci sont également connectés que 5 fois sur 1000.

À partir de la matrice d'adjacence, il est possible de calculer une matrice de covariance à l'aide de l'équation (3.1). Pour cela, il faut fixer des valeurs pour les paramètres β_0 et β_1. Comme la matrice de covariance doit être définie positive, on ne peut pas prendre n'importe quelles valeurs pour β_0 et β_1. Il faut que pour toutes les valeurs propres λ de la matrice A, l'équation suivante soit respectée,

$$\beta_0 + \beta_1 \lambda_i > 0 \quad \forall \lambda_i \in \lambda_A. \tag{3.11}$$

où λ_A représente l'ensemble des valeurs propres de la matrice A. La plus faible valeur propre de A est de -4.87. En initialisant les paramètres β_0 et β_1 aux valeurs respectives de 12 et 2 l'équation (3.11) est donc respectée. La matrice de covariance simulée suit donc la formule suivante,

$$\Sigma = Cov(\mathbb{Y}|\mathbb{A}) = 12I + 2A. \tag{3.12}$$

Pour être en mesure de comparer les différentes méthodes d'estimation des paramètres, il faut également simuler le vecteur $\mathbb{Y}$ de façon répétitive ainsi,

$$\mathbb{Y} = \Sigma^{1/2} Z, \tag{3.13}$$

où Z est un vecteur de variables aléatoires normales centrées réduites indépendantes et $\Sigma^{1/2}$ est la racine carrée de la matrice de covariance obtenue à l'équation (3.12). Le nombre d'itérations faites est de 10 000 et les résultats sont présentés au tableau 3.5.

Tableau 3.5 – Propriétés échantillonnales des quatre estimateurs des paramètres β

	EMV Lan	EMV	Moindres carrés	Assortativié
Biais $\hat{\beta}_0$	-0.0227	-0.0228	-0.0392	-0.0200
Écart-type $\hat{\beta}_0$	0.5617	0.5616	0.5700	0.5706
$\mathbb{E}[\hat{\sigma}_{\beta_0}]$	0.5637	0.5637	0.5714	X
Biais $\hat{\beta}_1$	-0.0074	-0.0066	-0.0314	-0.0317
Écart-type $\hat{\beta}_1$	0.1855	0.1855	0.2690	0.2608
$\mathbb{E}[\hat{\sigma}_{\beta_1}]$	0.1866	0.1866	0.2538	X
Biais $\hat{\tau}$	-0.0004	-0.0004	-0.0021	-0.0023
Écart-type $\hat{\tau}$	0.0112	0.0112	0.01867	0.01793
$\mathbb{E}[\hat{\sigma}_{\tau}]$	0.0113	0.0113	X	X
Nombre définie positive	10 000	10 000	9846	9881

La valeur τ représente le rapport $\frac{\beta_1}{\beta_0}$.

Le calcul des valeurs du tableau 3.5 a été fait avec les formules suivantes :

— Biais $\beta_0 = \mathbb{E}\left[\hat{\beta}_0\right] - \beta_0$

— Écart-type $\beta_0 = \sqrt{\frac{\sum_{i=1}^{10000}(\hat{\beta}_0 i - \mathbb{E}[\hat{\beta}_0])^2}{n-1}}$ où $\mathbb{E}\left[\hat{\beta}_0\right]$ représente la moyenne empirique des $\hat{\beta}_0$

L'espérance de $\hat{\sigma}_\beta$ est quant à elle obtenue à partir de la matrice hessienne fournie par la fonction `optim` de la librairie Stats (R Core Team, 2021). Cette fonction a été utilisée pour obtenir les estimations des paramètres pour les deux estimateurs de vraisemblance. Elle fournit une matrice hessienne qui permet d'avoir la variance des estimateurs lorsqu'elle est inversée (Rice, 2006). Par la suite, il suffit de prendre la racine carrée des variances des estimateurs pour obtenir leur écart-type. L'espérance de $\hat{\sigma}_\beta$ constitue la moyenne des 10 000 écarts-types des estimateurs.

À première vue, il ne semble pas avoir de différences majeures dans les estimations entre les quatre algorithmes. Particulièrement entre l'algorithme du maximum de vraisemblance présenté par Lan et al., 2018 et le maximum de vraisemblance que nous avons développé. Pour rappel, Lan et al., 2018 estiment la moyenne de la variable Y par la moyenne des valeurs du vecteur $\mathbb{Y}$ alors que nous l'estimons par une moyenne pondérée par l'inverse de la variance des valeurs de $\mathbb{Y}$. Cette différence entre les deux approches ne semble donc pas affecter les résultats de façon significative. Quant à l'algorithme des moindres carrés, ses résultats sont très similaires à l'approche développée avec le coefficient d'assortativité. Ces deux méthodes offrent des estimations assez précises ayant elles aussi des biais très faibles.

La dernière ligne du tableau 3.5 présente également un résultat intéressant. Comme l'objectif

est d'estimer une matrice de covariance, il est souhaitable que l'algorithme utilisé permette d'obtenir une matrice définie positive. Pour cela, l'équation (3.11) doit être respectée comme il a été mentionné précédemment. On voit alors qu'il y a une différence entre les deux algorithmes du maximum de vraisemblance et ceux des moindres carrés et d'assortativité. Pour ce qui est des deux premiers, lors des 10 000 simulations, la matrice de covariance obtenue a toujours été définie positive. Alors que pour l'algorithme des moindres carrés et d'assortativité, respectivement 154 et 119 des simulations ont engendré une matrice de covariance non définie positive. Cela peut être expliqué par les écarts-types des paramètres qui sont plus grands pour les méthodes des moindres carrés et de l'assortativité, particulièrement pour l'estimation de β_1. Effectivement, la théorie mentionne que les estimateurs du maximum de vraisemblance sont ceux qui offrent la plus petite variance (Bélisle, 2018). Cette plus grande variance des deux autres estimateurs affecte donc la précision des estimations et les valeurs des β peuvent sortir de l'espace des paramètres à respecter. Au vu des propriétés échantillonnales qui sont très similaires pour les quatre algorithmes, le fait que les algorithmes du maximum de vraisemblance obtiennent toujours une matrice de covariance définie positive leur donne un avantage sur les deux autres. De plus, les algorithmes du maximum de vraisemblance permettent également d'estimer l'écart-type des paramètres estimés, ce qui n'est pas le cas avec l'approche avec l'assortativité. Pour ce qui est de l'algorithme des moindres carrés, il n'est pas en mesure de fournir l'estimation de l'écart-type pour le paramètre τ.

3.2.1 Application aux réseaux réels

Après avoir étudié l'approche de Lan et al., 2018 sur un réseau Bernoulli, regardons s'il est possible de l'appliquer sur certains des réseaux de la compagnie d'assurance. En se référant au tableau 1.17, il est judicieux de sélectionner une variable qui a une assortativité non nulle dans un réseau. Si celle-ci est trop faible, l'estimation des paramètres sera peu intéressante puisque le rapport β_1/β_0 tendra vers 0 et de ce fait la valeur de β_1 sera nulle. Deux réseaux ont donc été sélectionnés. Le réseau 2 pour lequel le nombre d'années d'ancienneté des individus représente la variable Y ainsi que le réseau 5 avec le nombre d'interactions entre les individus comme variable Y.

Comme il est présenté au tableau 1.11, ces deux réseaux font partie des plus grands en termes de nœuds, particulièrement le deuxième. Utiliser la totalité des individus qui les constituent requiert beaucoup d'espace en mémoire. La matrice d'adjacence des réseaux 2 et 5 contiendraient plus de 310 milliards et 3 milliards de valeurs, respectivement. Un échantillon de chaque réseau est donc nécessaire pour faciliter les calculs dus à la complexité computationnelle. Comme tout échantillon, il est important que celui-ci garde des caractéristiques similaires à la population de départ. Avec les réseaux bipartis, une des façons de procéder est en échantillonnant les composantes. Comme celles-ci sont indépendantes les unes des autres, en choisissant un certain nombre de composantes, aucun lien n'est perdu. Cela n'est pas garanti si l'échantillonnage se

fait selon les individus ou les groupes. Par exemple, si l'individu a est connecté au b et qu'un seul des deux est échantillonné, alors ce lien est perdu.

Le tableau 1.16 présente certaines statistiques en lien avec les composantes des différents réseaux. De plus, les tableaux 1.12 et 1.13 permettent de remarquer que la majorité des groupes sont composés de 2 individus et que ceux-ci font partie la plupart du temps que d'un seul groupe. On peut donc affirmer qu'une majorité des composantes sont de taille 2. Pour avoir un échantillon représentatif du réseau, les composantes sont séparées en deux types. Celles qui sont formées de 2 nœuds et celles qui sont formées de plus de 2 nœuds. Le premier groupe constitue 75% des composantes de l'échantillon et le second 25%. On s'assure ainsi que les composantes de l'échantillon ressemblent à celles du réseau initial et qu'aucun lien n'est perdu. Le tableau 3.6 présente certaines statistiques en lien avec l'échantillonnage fait sur les réseaux 2 et 5.

Tableau 3.6 – Statistiques sur les échantillons des réseaux 2 et 5

	Nb compo	Nb individus	Taille maximale	Taille moyenne	Longueur moy des chemins	Diamètre
Réseau 2	4000	10 000	52	2.64	1.79	9
Réseau 5	20 000	10 000	6	2.07	1.03	3

Pour être en mesure d'échantillonner 10 000 individus par réseau, 4000 composantes ont été nécessaires pour le réseau 2 et 20 000 pour le réseau 5. L'échantillon de ce dernier est évidemment encore une fois composé de composantes de petites tailles avec un maximum de 6. Pour l'échantillon du réseau 2, le nombre de nœuds maximum formant une composante est de 52. Le diamètre des deux échantillons est également différent avec des valeurs de 9 pour le réseau 2 et de 3 pour le réseau 5.

Pour être en mesure d'obtenir une matrice de covariance définie positive, il faut respecter l'équation (3.4). La valeur propre la plus élevée en valeur absolue est de -2.4105 pour l'échantillon du réseau 2 et de -1 pour celui du réseau 5. Le système d'équations à respecter lors de l'estimation des paramètres est donc défini ainsi pour le réseau 2,

1. $\beta_0 - 2.4105\beta_1 > 0$
2. $\beta_0 + 2.4105\beta_1 > 0$

et ainsi pour le réseau 5,

1. $\beta_0 - \beta_1 > 0$
2. $\beta_0 + \beta_1 > 0$.

Les quatre algorithmes ont encore une fois été utilisés et les résultats sont présentés au tableau 3.7 pour le réseau 2 et au tableau 3.8 pour le réseau 5.

Tableau 3.7 – Résultats des estimations des paramètres selon chaque algorithme pour le réseau 2

	EMV Lan	EMV	Moindres carrés	Assortativié
$\hat{\beta_0}$	19.5443	19.5377	19.5361	19.5379
$\hat{\sigma_{\beta_0}}$	0.2692	0.2690	0.2858	X
$\hat{\beta_1}$	0.4601	0.4599	0.4062	0.4320
$\hat{\sigma_{\beta_1}}$	0.2187	0.2187	0.2099	X
$\hat{\tau}$	0.0233	0.0233	0.0207	0.0221
$\hat{\sigma_{\tau}}$	0.0111	0.0111	X	X
Définie positive	Oui	Oui	Oui	Oui

Encore une fois, les estimations des paramètres sont très similaires d'un algorithme à l'autre. L'échantillonnage effectué sur le deuxième réseau à fait diminuer l'assortativité de 0.06 à 0.0221 pour la variable du nombre d'années d'ancienneté. La plus grande différence au niveau des paramètres estimés se situe à l'estimation de β_1 par l'algorithme des moindres carrés. Sa valeur est un peu plus faible d'environ 5 centièmes par rapport aux deux algorithmes de maximum de vraisemblance. Les quatre méthodes d'estimation permettent d'obtenir une matrice de covariance définie positive. La corrélation du modèle étant positive, cela signifie que deux individus liés dans le réseau ont tendance à avoir une ancienneté avec la compagnie d'assurance similaire.

Tableau 3.8 – Résultats des estimations des paramètres selon chaque algorithme pour le réseau 5

	EMV Lan	EMV	Moindres carrés	Assortativié
$\hat{\beta_0}$	528.8526	529.0231	529.172	529.226
$\hat{\sigma}_{\beta_0}$	7.6199	7.6232	7.941	X
$\hat{\beta_1}$	-56.6692	-56.3333	-62.651	-60.2882
$\hat{\sigma}_{\beta_0}$	6.8276	6.8342	7.651	X
$\hat{\tau}$	-0.1071	-0.1054	-0.1183	-0.1139
$\hat{\sigma}_{\tau_0}$	0.0126	0.0126	X	X
Définie positive	Oui	Oui	Oui	Oui

Les conclusions pour l'échantillon du réseau 5 sont similaires à celles pour le réseau 2. Les estimations sont très proches pour les trois paramètres et encore une fois la plus grande différence est observée par l'estimation du paramètre β_1 par l'algorithme des moindres carrés. Cette fois-ci l'échantillonnage ne semble pas affecter l'assortativité qui reste d'environ -0.11. Une fois de plus, les quatre matrices de covariance obtenues sont définies positives. Cette fois-ci la corrélation du modèle est négative. Donc, si deux individus sont en relation, un aura tendance à avoir un nombre d'interactions élevé avec la compagnie d'assurance alors que l'autre interagira moins.

Chapitre 4

Approche alternative

Les précédents chapitres ont présenté plusieurs concepts en lien avec un réseau. Plusieurs notions permettant de mieux comprendre sa structure ont été abordées. Des tests statistiques ont également permis de voir l'influence qu'il peut avoir sur une variable aléatoire externe. Puis un modèle statistique permettant d'extraire une matrice de covariance a également été présenté. Dans ce chapitre, une approche similaire est exposée. De plus, comme ce mémoire fait partie d'un projet plus grand visant l'introduction des réseaux dans des modèles conjoints, une petite introduction à ceux-ci est également présente.

4.1 Modèle alternatif pour estimer une matrice de covariance

Reprenons la variable Y représentant le nombre d'amis des individus d'un réseau. Pour chaque i allant de 1 à n, $Y_i \sim N(\mu_y, \sigma_y^2)$. Si deux individus i et j sont en relation dans le réseau, leurs nombres d'amis Y_i et Y_j risquent d'être corrélés également. Cela peut être validé en calculant leur covariance $cov(Y_i, Y_j) \neq 0$. Cela suggère que la covariance de $\mathbb{Y} = (Y_1, ..., Y_n)^\top$ est dépendante de la matrice d'adjacence A. On a alors $cov(\mathbb{Y}|A) = \Sigma(A)$.

Comme mentionné au début du chapitre 3, Y_i peut être décomposée en deux parties. Une première est expliquée par les relations de l'individu i dans le réseau. Supposons qu'il partage un lien avec l'individu j et posons s_{ij} une variable aléatoire représentant l'information engendrée par ce lien. Nous avons alors $s_{ij} = s_{ji}$, $\mathrm{E}(s_{ij}) = 0$ et $var(s_{ij}) = \beta_1$. La deuxième partie, qui représente l'information provenant de l'extérieur du réseau, est dénotée par e_i. La variable réponse Y_i est alors modélisée selon l'équation suivante,

$$Y_i = \sum_{j \neq i} a_{ij} s_{ij} + e_i, \tag{4.1}$$

où s_{ij} et e_i sont des variables aléatoires indépendantes.

À partir de l'équation (4.1) il devient possible de calculer la covariance de $\mathbb{Y}$ et elle s'écrit ainsi,

$$Cov\left(Y_i, Y_j\right) = Cov\left(\sum_{l=1}^{N} a_{il}s_{il} + e_i, \sum_{m=1}^{N} a_{jm}s_{jm} + e_j\right). \tag{4.2}$$

Posons, les hypothèses suivantes :

1. $Cov\left(s_{il}, e_j\right) = 0$
2. $Cov\left(e_i, e_j\right) = \begin{cases} \beta_0 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$
3. $Cov\left(s_{il}, s_{jm}\right) = \begin{cases} \beta_1 & \text{si } i{=}j \text{ et } l{=}m \\ 0 & \text{sinon} \end{cases}$
4. $s_{ij} = s_{ji} \quad \forall i, j$

Développons l'équation (4.2),

$$Cov\left(Y_i, Y_j\right) = Cov\left(e_i, e_j\right) + Cov\left(e_i, \sum_{m=1}^{N} a_{jm}s_{jm}\right) + Cov\left(\sum_{l=1}^{N} a_{il}s_{il}, e_j\right) + Cov\left(\sum_{l=1}^{N} a_{il}s_{il}, \sum_{m=1}^{N} a_{jm}s_{jm}\right),$$

$$= Cov\left(e_i, e_j\right) + \sum_{m=1}^{N} a_{jm}Cov\left(e_i, s_{jm}\right) + \sum_{l=1}^{N} a_{il}Cov\left(s_{il}, e_j\right) + \sum_{l=1}^{N} a_{il} \sum_{m=1}^{N} a_{jm}Cov\left(s_{il}, s_{jm}\right).$$

À l'aide de l'hypothèse 1, les deux covariances du milieu sont nulles et on obtient l'équation suivante,

$$Cov\left(Y_i, Y_j\right) = Cov\left(e_i, e_j\right) + \sum_{l=1}^{N} a_{il} \sum_{m=1}^{N} a_{jm}Cov\left(s_{il}, s_{jm}\right).$$

Étudions le cas où $i = j$,

$$Cov\left(Y_i, Y_i\right) = Cov\left(e_i, e_i\right) + \sum_{l=1}^{N} a_{il} \sum_{m=1}^{N} a_{im}Cov\left(s_{il}, s_{im}\right),$$

$$= Cov\left(e_i, e_i\right) + \sum_{l=1}^{N} a_{il}a_{il}Cov\left(s_{il}, s_{il}\right) + \sum_{l=1}^{N} \sum_{\substack{m=1 \\ m \neq j}}^{N} a_{il}a_{im}Cov\left(s_{il}, s_{im}\right),$$

$$= \beta_0 + \beta_1 \left(\sum_{l=1}^{N} a_{il} \right).$$

Sur la deuxième ligne, le premier terme est égal à β_0 grâce à l'hypothèse 2. La covariance du deuxième terme est égale à β_1 par l'hypothèse 3 et la même hypothèse permet d'annuler la covariance du troisième terme. Sur la troisième ligne, la sommation représente le degré du nœud i qui sera noté d_i pour la suite.

Étudions le cas où $i \neq j$,

$$Cov\left(Y_i, Y_j\right) = Cov\left(e_i, e_j\right) + \sum_{l=1}^{N} a_{il} \sum_{m=1}^{N} a_{jm} Cov\left(s_{il}, s_{jm}\right),$$

$$= a_{ij} a_{ji} Cov\left(s_{ij}, s_{ij}\right) + \sum_{\substack{l=1 \\ l \neq j}}^{N} \sum_{\substack{m=1 \\ m \neq i}}^{N} a_{il} a_{jm} Cov\left(s_{il}, s_{mj}\right),$$

$$= \beta_1 a_{ij}.$$

Sur la première ligne, l'hypothèse 2 permet d'annuler le premier terme. Puis, pour le deuxième terme, l'hypothèse 4 permet de réécrire s_{jm} par s_{mj} dans la covariance. Ainsi, nous pouvons sortir le cas où $m = i$ et $l = j$. Celui-ci représente le premier terme de la seconde ligne et la covariance est égale à β_1 grâce à l'hypothèse 3. De plus, nous avons $a_{ij}a_{ji} = a_{ij}a_{ij} = a_{ij}$ ce qui permet de simplifier la notion des éléments de la matrice d'adjacence A. Cette même hypothèse permet d'annuler la covariance du second terme.

En conclusion nous obtenons,

$$Cov\left(Y_i, Y_j\right) = \begin{cases} \beta_0 + d_i \beta_1 & \text{si } i = j \\ \beta_1 a_{ij} & \text{si } i \neq j. \end{cases} \tag{4.3}$$

Lorsque l'on compare les équations (4.3) et (3.1) on se rend compte qu'il y a certaines similitudes, plus particulièrement dans le cas où $i \neq j$ où les deux équations sont équivalentes. Effectivement en prenant l'équation (3.1), si $i \neq j$ tous les éléments de la matrice identité I_p sont alors nul. Ainsi le premier terme comprenant β_0 disparaît et on retrouve l'équation (4.3).

Il existe toutefois une différence entre les deux équations dans le second cas, où $i = j$. Le premier terme comprenant β_0 est similaire. Par contre, dans le deuxième terme composé de β_1 il n'est pas multiplié par la même constante. Dans l'équation (3.1), β_1 est multiplié par la valeur de la matrice d'adjacence a_{ij}. Comme c'est une matrice 0, 1, le paramètre β_1 est

multiplié par 1 si i et j sont liés dans le réseau. La constante multiplicative est différente pour l'équation (4.3). Le paramètre β_1 est plutôt multiplié par le degré de l'individu i.

4.1.1 Matrice définie positive

Comme il a été mentionné plusieurs fois, la matrice résultante doit être définie positive pour être par la suite utilisée dans un modèle prédictif. Réécrivons l'équation (4.3) sous la forme suivante,

$$\Sigma(A) = \beta_0 I_n + (A + D)\beta_1, \tag{4.4}$$

où I est une matrice identité, A est la matrice d'adjacence décrivant le réseau et dont la diagonale est nulle et D est une matrice diagonale dont les éléments sont les degrés des individus.

Les matrices I et D ayant uniquement des éléments positifs sur leurs diagonales sont définies positives (Meyer, 2000). La matrice résultante de l'addition $A+D$ est également définie positive et la preuve est disponible en annexe A.2. Une matrice définie positive conserve cette propriété lorsqu'elle est multipliée par un élément supérieur à 0 (Horn and Johnson, 2012). Il faut donc que les coefficients β_0 et β_1 soient strictement supérieurs à 0 pour que les matrices résultantes des opérations $\beta_0 I_n$ et $(A + D)\beta_1$ restent également définies positives. Leur addition donne aussi une matrice définie positive.

Donc, l'espace des paramètres à respecter pour le modèle de l'équation (4.4) est,

$$\Theta = \{\beta_0, \beta_1 \in [0, \infty)\}. \tag{4.5}$$

Le lien entre l'assortativité et le rapport β_1/β_0 a été abordé à la sous-section 3.1.5. Considérant l'espace des paramètres (4.5), l'utilisation du modèle (4.4) devient très intéressante lorsque la mesure d'assortativité est positive puisque la matrice de covariance résultante est définie positive par définition du modèle. Aucun algorithme n'est nécessaire pour estimer les paramètres β puisque β_0 représente la variance de la variable aléatoire Y et β_1 est obtenue avec le produit assortativité $\times\ \beta_0$.

Évidement, une des limites du modèle (4.4) est qu'il ne permet pas d'obtenir une matrice définie positive lorsque l'assortativité en lien avec la variable Y est négative.

En théorie, cette méthode semble être meilleure que celle de Lan et al., 2018 lorsque l'assortativité est positive. Dans ce cas, la matrice résultante est toujours définie positive sans contrainte sur les valeurs des paramètres β. De plus, le calcul de la matrice $\Sigma(A)$ ne requiert aucun algorithme. Une étude de simulation similaire à ce qui a été présenté à la section 3.2

est toutefois nécessaire pour comparer les deux approches et déterminer laquelle offrent de meilleurs résultats. Cependant, dans le cas où au moins un des paramètres β est négatif, l'approche de Lan et al., 2018 est sans aucun doute celle à adopter puisque c'est la seule qui peut obtenir une matrice de covariance définie positive.

4.2 Les modèles

Comme il est mentionné précédemment, ce mémoire n'est qu'une partie d'un projet plus grand ayant comme objectif d'intégrer l'information pouvant être extraite d'un réseau à un modèle prédictif. La modélisation du temps restant avant l'abandon de la police d'assurance d'un client est un problème de survie. La disponibilité de données longitudinales sur les clients rend donc l'utilisation d'un modèle conjoint possible. C'est donc ce type de modèle qui est introduit dans cette section. Un petit rappel essentiel est aussi fait sur le modèle linéaire généralisé mixte et le modèle de Cox.

4.2.1 Modèle linéaire généralisé mixte

Le modèle linéaire généralisé mixte permet de modéliser une variable réponse en relation avec certaines variables explicatives. La relation qui les unit est spécifiée par la fonction de lien g. Lorsque la variable réponse est continue, la fonction de lien utilisée est la fonction identité et nous retrouvons le modèle linéaire mixte. Lorsque la variable à modéliser est catégorique, deux scénarios sont possibles. Si celle-ci est dichotomique, la fonction de lien choisie sera généralement la fonction logit. Ce type de modèle est appelé régression logistique. Si, au contraire, la variable endogène peut prendre plus de deux valeurs possibles, le lien logarithmique sera employé. Le nom associé à ce type de modèle est régression de Poisson.

Le modèle linéaire généralisé mixte permet également l'ajout d'un terme aléatoire qui explique une partie de la variabilité du jeu de données. En général, le terme aléatoire est soit associé aux sujets soit à des groupes de sujets ayant des caractéristiques similaires qui ne sont pas associées aux variables explicatives fixes du modèle (Gelman and Hill, 2006).

Considérons une étude menée sur N sujets où des mesures sont prises sur le sujet i à J_i occasions. La variable d'intérêt y_{ij} représente la $j^{ième}$ mesure du $i^{ième}$ sujet. Les p variables explicatives fixes sont représentées par le vecteur x_{ij} de dimension $(p+1) \times 1$. Quant aux q variables explicatives aléatoires, elles sont représentées par le vecteur z_{ij} de dimension $q \times 1$.

Posons un variable aléatoire y_{ij} suivant une certaine loi de probabilité. La moyenne de cette variable satisfait alors l'équation suivante,

$$m_{ij} = g\left\{E\left(y_{ij}\right)|b_i\right\} = x_{ij}^{\top}\beta + z_{ij}^{\top}b_i, \tag{4.6}$$

où

— g : la fonction de lien

— β : vecteur de dimension $(p+1) \times 1$ des coefficients à estimer liés aux variables explicatives

— b_i : vecteur de dimension $q \times 1$ des effets aléatoires associés au sujet i. $b_i \sim \mathcal{N}_q(0, D)$ où D est une matrice variances covariances de dimension $q \times q$ pour les effets aléatoires

4.2.2 Modèle de survie de Cox

Le modèle de Cox permet de modéliser une durée de vie. Il cherche à comprendre ou à prédire le temps restant avant la survenue d'un évènement à l'aide de variables explicatives. Plus précisément, la modélisation porte sur le taux d'évènements conditionnel. Celui-ci correspond à la probabilité de subir l'évènement au temps t pour un sujet à risque. Ce type de modèle est dit à taux proportionnel, puisque le rapport des taux d'évènement de deux sujets est constant dans le temps (Kleinbaum and Klein, 1996).

Pour bien comprendre cela, supposons une étude qui tente d'évaluer l'efficacité d'un traitement W sur le temps de vie restant à une personne ayant reçue un diagnostic positif à une certaine maladie. Le groupe A reçoit donc le traitement $(W = 1)$ et le groupe B est le groupe placebo $(W = 0)$. Le taux d'évènement pour le groupe ayant reçu le traitement W s'écrit : $h(t|W) = h_0(t) \exp(W\gamma)$.

Le rapport des taux des deux groupes est donc (Talbot, 2021),

$$RT = \frac{h(t|W=1)}{h(t|W=0)} = \frac{h_0(t)\exp(1\times\gamma)}{h_0(t)\exp(0\times\gamma)} = \frac{\exp(\gamma)}{\exp(0)} = \exp(\gamma). \tag{4.7}$$

Nous remarquons donc que le modèle de Cox stipule que peut importe le temps t le taux d'évènement du groupe A est $\exp(\gamma)$ fois celui du groupe B.

Reprenons la même étude menée sur N sujets où nous désirons estimer la survie d'une personne atteinte d'une certaine maladie. Le vecteur w_i de dimension $r \times 1$ représente maintenant plusieurs caractéristiques d'un sujet. Il peut contenir le traitement reçu et d'autres variables comme l'âge, le sexe et le poids. Ce modèle fait également intervenir un indicateur de censure C qui prend la valeur 0 lorsque l'information récoltée concernant le temps de survie d'un sujet est incomplète et 1 sinon (Talbot, 2021). Le type de censure le plus fréquent se nomme censure à droite. Dans ce cas, le temps de survie du sujet est supérieur au temps de suivi, mais sa valeur exacte est inconnue. Au contraire, si le temps de survie est égal au temps de suivi, il n'y a pas de censure, car le sujet a vécu l'évènement d'intérêt. Le modèle de Cox avec r variables explicatives correspond à un taux conditionnel d'évènement défini ainsi,

$$h\left(t|w\right) = h_0\left(t\right) \exp\left(w^\top \gamma\right). \tag{4.8}$$

où

— $h_0\left(t\right)$: taux de base lorsque toutes les variables w_i sont nulles

— γ : vecteur de dimension $r \times 1$ des coefficients à estimer

4.2.3 Modèle conjoint

Le modèle conjoint combine les deux modèles décrits précédemment. Il est utilisé lorsque l'on veut modéliser une durée de vie qui est en relation avec une variable longitudinale. Comme chaque sujet peut avoir un nombre de mesures différentes pour la variable longitudinale et que plusieurs variables longitudinales peuvent être en relation avec celle de la durée de vie, le modèle de Cox devient rapidement inefficace. Le modèle conjoint est donc utilisé pour ce type de situation.

Considérons une étude menée sur N sujets. Posons K le nombre de variables longitudinales. Des mesures x_{ijk} et w_i sont prises pour le sujet i avec $i = 1, ..., N$. Comme dans l'équation (4.6), les vecteurs x_{ijk} de dimension $(p+1) \times 1$ permettent d'expliquer le comportement de la variable longitudinale k. Les vecteurs w_i de dimension $(r \times 1)$ sont quant à eux en relation avec la durée de vie comme le montre l'équation (4.8).

Le modèle conjoint s'écrit dans sa forme la plus simple de la façon suivante,

$$h\left(t|M_i\left(t\right)\right) = h_0\left(t\right) \exp\left(w_i^\top \gamma + \sum_{k=1}^{K} \alpha_k m_{ik}\left(t\right)\right), \tag{4.9}$$

où

— $h\left(t|M_i\left(t\right)\right)$: taux conditionnel d'évènement au temps t sachant l'historique des variables longitudinales

— $h_0\left(t\right)$: taux de base

— γ : vecteur de dimension $r \times 1$ des coefficients à estimer

— α_k : poids de la $k^{ième}$ variable longitudinale sur le risque de l'évènement

— $m_{ik}\left(t\right)$: prédiction de la moyenne transformée par la fonction de lien pour la $k^{ième}$ variable longitudinale du $i^{ième}$ sujet au temps t. La valeur est obtenue avec l'équation (4.6)

4.3 Mise en pratique

4.3.1 Jeu de données

Lorsqu'on utilise un modèle conjoint, les données doivent contenir au moins une variable longitudinale et une durée de vie. Le tableau 4.1 montre les observations prises sur les trois premiers sujets du jeu de données *aids* venant de la librairie JM (Rizopoulos, 2010). Il a été collecté lors d'une étude clinique comparant l'efficacité de deux traitements menée auprès de 467 sujets atteints du sida ayant échoués ou étant intolérants au traitement par la zidovudine (AZT). Le point d'origine du temps d'un sujet est le moment où il se joint à l'étude. La variable *patient* identifie les sujets. Elle permet de construire le vecteur z_{ij} de l'équation (4.6). *Time* correspond au temps de suivi pour chaque sujet. *Death* informe si le sujet est décédé au cours de l'étude (1 si oui, 0 sinon). Si le patient est toujours en vie lors du dernier suivi, il y a censure à droite. $CD4$ est la valeur de la variable longitudinale qui correspond ici à un nombre de cellules. La variable *obstime* indique le temps où la mesure de $CD4$ a été prise. Cette variable ainsi que son interaction avec la variable *drug* qui indique le traitement reçu pour chaque sujet forme le vecteur x_{ij} de l'équation (4.6). De plus, *drug* est aussi la seule variable explicative dans le vecteur w_i de l'équation (4.9). Au temps t, la moyenne estimée de la variable longitudinale $CD4$ pour un individu i représentera l'élément $m_{ik}(t)$ de l'équation (4.9). Finalement, le taux de base à une valeur constante de 1. L'équation (4.9) du modèle conjoint peut donc être réécrite comme suit,

$$h(t|M_i(t)) = 1 \times \exp\{\gamma \times drug + \alpha(\beta_0 + \beta_1 \times obstime + \beta_2 \times obstime \times drug + patient_i)\}.$$

Dans cette formule, *patient* est un terme aléatoire d'où l'indice i qui lui est attribué.

Tableau 4.1 – Mesures prises sur les trois premiers individus du jeu de données aids

patient	Time	death	CD4	obstime	drug
1	16.97	0	10.68	0	ddC
1	16.97	0	8.43	6	ddC
1	16.97	0	9.43	12	ddC
2	19.00	0	6.32	0	ddI
2	19.00	0	8.12	6	ddI
2	19.00	0	4.58	12	ddI
2	19.00	0	5.00	18	ddI
3	18.53	1	3.46	0	ddI
3	18.53	1	3.61	2	ddI
3	18.53	1	6.16	6	ddI

4.3.2 Analyse

La fonction `jointModel` de la librairie JM du logiciel R a été utilisée pour traiter ce jeu de données. Son architecture est assez simple. Il suffit de spécifier un modèle linéaire mixte

avec la fonction `lme` de la librairie `nlme` (Pinheiro et al., 2021) ainsi qu'un modèle de survie avec la fonction `coxph` de la librairie `Survival` (Therneau, 2021). Dans la fonction `lme`, l'effet aléatoire est spécifié dans l'argument *random*. Dans notre cas, chaque sujet a un effet aléatoire pour l'ordonnée à l'origine ainsi que pour la pente liée à la variable *obstime*. De plus, l'effet aléatoire est indépendant entre les différents sujets. Pour la fonction `coxph`, le jeu de données contient seulement une observation par individu. C'est pour cela que l'argument *data* prend comme valeur *aids.id*. Ce jeu de données est le même que *aids*, mais il n'utilise que la première observation de chaque sujet. Nous indiquons aussi la variable représentant le temps pour le modèle conjoint. Par la suite, plusieurs autres hyperparamètres sont disponibles notamment *method* qui permet d'indiquer la technique avec laquelle nous voulons estimer les paramètres du modèle en particulier le taux conditionnel d'évènement de base $h_0(t)$. Voici le code qui a été implémenté,

```
lmeFit <- lme(CD4 ~ obstime + obstime:drug,
           random = ~ obstime | patient, data = aids)
coxFit <- coxph(Surv(Time, death) ~ drug, data = aids.id, x = TRUE)
jointFit <- jointModel(lmeFit, coxFit, timeVar = "obstime").
```

4.3.3 Résultats

Les résultats sont présentés ci-dessous dans le tableau 4.2. Le coefficient lié au traitement *ddI* a une valeur de 0,3424. Quant à l'interaction entre ce même traitement et le temps d'observation, le coefficient est positif avec une valeur de 0,0119. Il semblerait donc que le traitement *ddI* soit moins performant que le traitement *ddC* puisqu'il contribue à augmenter le taux conditionnel d'évènement. De plus, la valeur *alpha* représentant le poids entre la variable longitudinale $CD4$ et le taux conditionnel d'évènement est de -0.2802. Cela signifie que plus la valeur de la variable $CD4$ représentant le nombre de cellules est grande, plus le taux conditionnel d'évènement diminue.

Tableau 4.2 – Estimation des coefficients du modèle obtenu avec la fonction `jointModel`

Partie longitudinale				
	Valeur	Écart-type	Valeur z	Valeur p
ordonnée à l'origine	7.2080	0.2221	32.4513	<0.0001
obstime	-0.1877	0.0216	-8.7078	<0.0001
obstime*drugddI	0.0119	0.0301	0.3961	0.6920
Partie de survie				
	Valeur	Écart-type	Valeur z	Valeur p
ordonnée à l'origine	-3.0640	0.3039	-10.0828	<0.0001
drugddI	0.3424	0.1567	2.1856	0.0288
association	-0.2802	0.0356	-7.8650	<0.0001

4.4 Simulation de données

Dans certaines situations, il est très pratique de simuler un jeu de données pour tester certains modèles. En sachant les paramètres exacts qui ont été utilisés pour produire les données, nous pouvons être en mesure de voir si le modèle offre de bonnes estimations. Il est possible d'étudier la stabilité des estimations à l'aide de technique de Monte-Carlo. Nous pouvons, avec plusieurs simulations, tester certaines propriétés des estimateurs comme le biais ainsi que leur variance. Simulons un jeu de données assez simple venant d'un modèle conjoint. Une variable explicative X est en relation avec la durée de vie ainsi qu'une variable longitudinale. Le modèle conjoint s'écrit donc ainsi,

$$h_i\left(t|M_i\left(t\right)\right) = h_0\left(t\right)\exp\left(\gamma_0 + \gamma_1 X_i + \alpha m_i\left(t\right)\right), \tag{4.10}$$

avec,

$$m_i\left(t\right) = \beta_0 + \beta_1 t + \beta_2 X_i + b_{0i} + e_i\left(t\right). \tag{4.11}$$

Faisons la simulation pour 1 individu. Posons $\beta_0 = 1$, $\beta_1 = 4$ et $\beta_2 = -0.1$ les coefficients des variables fixes de l'équation (4.11). Pour la partie de survie, supposons $h_0\left(t\right) = 1$ $\gamma_0 = -4.4$ et $\gamma_1 = 0.1$. La valeur pour *alpha* est de 0.5 et l'effet aléatoire b_{0i} ainsi que les résidus $e_i\left(t\right)$ proviennent tous les deux d'une loi normale de moyenne 0 et de variance 1 et 4 respectivement. La variable X, jouant le rôle du traitement dans la section 4.3, provient d'une loi de Bernoulli avec une probabilité de 0.5. Pour simuler le jeu de données, il faut aussi avoir une variable pour le temps de censure qui provient d'une loi uniforme entre 1.25 et 3.25. Il faut aussi simuler le temps de suivi de l'individu. La formule (4.12) tirée de l'article de Furgal et al., 2019 est

utilisée pour calculer le temps de survie. Les détails des calculs sont disponibles à l'annexe A.1,

$$Time = \frac{1}{\alpha\beta_1} \ln\left(\frac{-\ln(U)(\alpha\beta_1)}{\exp(\gamma_0 + \alpha\beta_0 + \alpha b_0 + (\gamma_1 + \alpha\beta_2) X_1)} + 1\right). \tag{4.12}$$

La variable $Time$ provient d'une distribution de Gumbel puisque le logarithme d'une loi Weibull est utilisé dans le calcul (Pham, 2006). Dans l'équation (4.12), U suit une loi uniforme entre 0 et 1. Posons les temps d'observations possibles $t = (0, 0.5, 1, 1.5, 2, 2.5, 3)$. Pour chaque individu, il suffit donc de simuler une observation chaque fois que le temps de suivi (minimum entre $Time$ et le temps de censure) est plus petit qu'un des temps d'observation. La valeur de la variable longitudinale est calculée selon l'équation (4.11). L'individu observe l'évènement d'intérêt lorsque la valeur de $Time$ est plus faible que le temps de la censure ainsi que du plus grand temps d'observation possible qui est de 3. Dans ce cas, la variable $Event$ prend la valeur 1 sinon, elle sera égale à 0. Toutes les mesures de temps sont calculées en année. Le tableau 4.3 montre la simulation du jeu de données faite pour un individu.

Tableau 4.3 – Simulation d'un jeu de données ayant un sujet

patient	Time	X1	var_long	t	Event
1	1.38	TRUE	2.52	0.00	0.00
1	1.38	TRUE	0.66	0.50	0.00
1	1.38	TRUE	5.98	1.00	0.00

Le patient numéro 1 a donc un temps de suivi de 1.38 année. Puisque nous sommes en présence de censure à droite, nous savons que l'évènement d'intérêt représentant ici le décès s'est produit après au moins 1.38 année sans pour autant connaître le moment exact. Le premier sujet fait partie du groupe ayant reçu le traitement, ses valeurs pour la variable longitudinale sont de 2.52, 0.66 et 5.98 lors des trois premiers temps d'observation.

Conclusion

Tout au long de ce mémoire, plusieurs notions en lien avec les réseaux ont été abordées. Il a été vu comment passer d'un réseau biparti avec deux types de nœuds à un réseau ne mettant en relation que des nœuds d'individus. Plusieurs notions ont été présentées comme celle de degré, de densité, de composante, d'assortativité et de transitivité. Le tout dans le but de bien comprendre la structure du réseau qui est analysé. C'est cette compréhension qui guide la suite des analyses.

Au départ de l'étude, une compagnie d'assurance canadienne a mis à notre disposition cinq réseaux mettant en relation ses clients ainsi qu'un jeu de données longitudinales contenant des informations sur les clients et leur police d'assurance. Pour que l'analyse soit cohérente, il a fallu conserver seulement les individus ayant au moins une relation dans un des cinq réseaux et étant présents dans les données longitudinales. Puis, il a fallu déterminer si l'analyse des réseaux allait se faire de façon séparée ou bien en les regroupant pour n'en former qu'un. Le tableau 1.10 a montré que plus de 80% des individus ne sont présents que dans un seul réseau. De plus, les tableaux 1.12 et 1.13 ont permis de voir que même si une majorité des réseaux sont formés de petits groupes de 2 ou 3 individus, il existe une différence majeure sur les plus grandes valeurs. Par exemple, les réseaux 3 et 4 possèdent des groupes très grands qui vont mettre en relation de nombreux individus, ce qui n'est pas le cas pour le réseau 5. Le tableau 1.16 présente également des informations très importantes en lien avec les composantes. Nous avons vu, grâce à l'équation (1.8), que certains réseaux sont très connectés comme le troisième alors que c'est moins le cas pour le premier et le cinquième par exemple. De plus, la taille de la plus grande composante variait également énormément d'un réseau à l'autre ainsi que leur diamètre. En calculant l'assortativité selon plusieurs variables différentes au tableau 1.17, on a remarqué aussi que les relations de certains réseaux comme le deuxième et le cinquième pouvaient influencer les valeurs que prennent ces variables alors que ce n'étaient pas le cas pour les autres. En voyant toutes ces différences entre les différents réseaux, cela nous a donc poussés à continuer leur analyse de façon séparée.

Après avoir couvert les statistiques descriptives dans la première partie, nous avons par la suite fait certaines analyses pour étudier l'impact que pouvait avoir un réseau sur une variable aléatoire discrète décrivant l'abandon d'une police d'assurance par un client de la compagnie

d'assurance. Cette variable était disponible dans les données longitudinales. L'idée était également de voir si l'influence sur cette variable pouvait être différente d'un réseau à l'autre. Le tableau 2.1 a donné un premier élément de réponse. Il a permis de voir que le taux d'abandon de la police est assez stable d'un réseau à l'autre. En effet, il varie entre 27% et 31%. Cependant, le pourcentage de liens entre deux individus qui abandonnent leur police varient plus entre les différents réseaux. Il est même jusqu'à deux fois plus élevées entre les réseaux 1 et 2 avec des taux respectifs de 6.21% et 12.14%. On pourrait donc être porté à croire que si un individu abandonne sa police d'assurance dans le réseau 2, les individus en relation avec lui auront une probabilité plus élevée d'abandonner à leur tour leur police d'assurance. Toutefois, pour valider cette hypothèse, il faut être en mesure de l'évaluer à l'aide d'un test statistique. C'est ce qui a été fait à la section 2.1 avec un test de permutation. Comme nous l'avons expliqué, ce test permet de voir si le nombre de relations qui existe dans le réseau entre deux individus qui abandonnent leur police est différent de ce que le hasard prédit. Ce fut le cas pour les réseaux 1, 2 et 5 au seuil de 5%. Par la suite, à l'aide de la probabilité conditionnelle d'abandon, calculée avec l'équation (2.4), ou bien avec un graphique comme celui de la figure 2.1 on peut déterminer le type d'effet qu'ont les relations du réseau sur l'abandon de la police. Pour les réseaux 2 et 5, la probabilité conditionnelle étant plus élevée que la probabilité marginale, cela signifie que les individus liés à un nœud ayant abandonné leur police ont une probabilité plus grande de l'abandonner à leur tour. Pour le réseau 1, c'est le raisonnement contraire puisque la probabilité conditionnelle est plus faible que la marginale. Donc, dans ce réseau, les individus liés à une personne ayant abandonné sa police ont une probabilité plus faible de l'abandonner à leur tour.

Après avoir vu que les relations d'un individu pouvaient influencer sa probabilité d'abandon de la police d'assurance, on a voulu explorer la façon de quantifier ces relations. Les tests statistiques non paramétriques performés visaient à déterminer s'il était préférable de mesurer la force d'un lien avec l'équation (1.2) développée au chapitre 1 ou bien avec la méthode plus traditionnelle de lui attribuer la valeur 1 si celui-ci existe et 0 sinon. Nous avons commencé par diviser les liens en trois types. Ceux où les deux individus ont conservé leur police, ceux où un des deux individus a abandonné sa police et ceux où les deux individus ont abandonné leur police. L'objectif était ainsi de voir si le nombre d'abandons influençait la force d'un lien. Le premier test effectué était un test non paramétrique de Kruskal-Wallis qui permet de déterminer si les trois types de lien proviennent de la même distribution. Selon les résultats obtenus au tableau 2.4, le réseau 3 était le seul à ne pas obtenir un résultat significatif au seuil de 5%. Dans ce cas, cela signifie que les forces de liens ne varient pas en fonction du nombre d'abandons que contient celui-ci. Pour les quatre autres réseaux, l'analyse a été poussée un peu plus loin avec un test de Wilcoxon. Celui-ci permet de comparer les types de lien deux à deux. Maintenant qu'on sait qu'au moins une paire parmi les trois possibles provient de différentes distributions, on veut être en mesure de la ou les identifier. Comme il a été mentionné, toutes les paires de liens possibles ont obtenu des résultats significatifs pour le test de Wilcoxon dans

les quatre réseaux. Les moyennes des forces de lien disponibles au tableau 2.4 peuvent nous informer à savoir si les forces de lien augmentent ou diminuent en fonction du type de lien. L'information est également résumée au tableau 4.4.

Tableau 4.4 – Impact du nombre d'abandons dans un lien sur la moyenne des forces de liens

	$0 \to 1$	$0 \to 2$	$1 \to 2$
Réseau 1	↑	↑	↑
Réseau 2	↓	↓	↓
Réseau 4	↓	↓	↓
Réseau 5	↑	↑	↑

Dans ce tableau, la première colonne représente le fait de passer des liens avec 0 abandon à 1, la deuxième colonne le fait de passer de 0 à 2 abandons et la troisième de 1 à 2 abandons. La flèche vers le haut signifie que la moyenne des forces de liens augmente et celle vers le bas signifie que la moyenne des forces de liens diminue. On voit que pour les deuxième et quatrième réseaux, plus le nombre d'abandons dans un lien augmente, plus sa force diminue. Alors que cette corrélation est inverse dans les premier et le cinquième réseaux. Au vu des résultats, la force de lien pourrait être une bonne façon de représenter un lien pour les réseaux $1, 2, 4$ et 5 puisqu'elle est influencée par le nombre d'abandons.

Les tests non paramétriques ont permis d'analyser l'effet que pouvait avoir le nombre d'abandons dans un lien sur la force de celui-ci. On a également voulu tester l'effet inverse à l'aide de la régression logistique. Dans cette étude la force du lien était la variable explicative et le nombre d'abandons était la variable réponse. Dans le jeu de données analysé, chaque lien était divisé en deux lignes, une pour chaque individu le formant. On associait alors la force du lien avec la variable d'abandon correspondante à l'individu. La variable réponse était donc dichotomique avec la valeur 1 si l'individu a abandonné sa police d'assurance et 0 sinon. Le jeu de données avait alors $2m$ lignes avec m représentant le nombre de liens dans le réseau. Encore une fois, comme nous l'avons vu au tableau 2.6, le coefficient en lien avec la force de lien étant non nul pour tous les réseaux sauf le troisième. À première vue les résultats semblent donc similaires à ceux obtenus avec les tests non paramétriques. Cependant, il y a des différences, pour les quatre autres réseaux, lorsqu'on regarde le sens de la corrélation entre l'abandon et la force du lien. Les réseaux 1, 4 et 5 ont des coefficients positifs, ce qui signifie que plus la force d'un lien augmente, plus la probabilité d'avoir un abandon augmente également. Pour le réseau 2, la corrélation est négative entre les deux variables. Ces résultats sont légèrement différents à ceux observés au tableau 4.4 pour le test de Wilcoxon, où la corrélation entre la force de lien et l'abandon était négative pour le réseau 4.

Ensuite, nous avons exploré la méthode de Lan et al., 2018 qui permet d'estimer une matrice de covariance à partir de la matrice d'adjacence d'un réseau. Un des atouts principaux de cette méthode est qu'elle permet de réduire considérablement le nombre de paramètres à estimer.

Cependant, les estimations doivent respecter un espace des paramètres pour s'assurer que la matrice obtenue soit définie positive. Quatre algorithmes d'estimation des paramètres ont été présentés. Puis, à l'aide d'un réseau Bernoulli, une étude de simulation a permis de comparer les propriétés échantillonnales de ces quatre algorithmes. En général, tous les algorithmes avaient de faibles biais comme le montre le tableau 3.5. Les deux algorithmes du maximum de vraisemblance semblent toutefois légèrement plus précis. De plus, ils ont également une plus petite variance, comme c'est attendu en théorie. Cette plus grande instabilité sur les algorithmes des moindres carrés et d'assortativité a engendré certaines matrices de covariance non définies positives puisque les estimations ne respectaient plus l'espace des paramètres. Un autre avantage que possèdent les deux algorithmes du maximum de vraisemblance est qu'ils sont en mesure d'estimer la variance du paramètre τ représentant le rapport entre les deux paramètres du modèle β_1/β_0.

La méthode de Lan et al., 2018 a également été appliquée à deux des réseaux de la compagnie d'assurance. Les réseaux 2 et 5 ont été choisis avec comme variable réponse respective le nombre d'années d'ancienneté et le nombre d'interactions entre les individus et la compagnie. À cause de la trop grande taille des deux réseaux, un échantillon a été extrait pour faciliter les calculs. L'échantillonnage a été fait par rapport aux tailles des composantes des réseaux pour tenter de conserver une structure similaire à celle de départ. Environ 10 000 individus ont été sélectionnés dans les deux réseaux. Encore une fois les résultats obtenus par les quatre méthodes d'estimation ont été très proches les uns des autres. Le coefficient de corrélation τ était positif pour le réseau 2 et négatif pour le réseau 5. Comme il a été expliqué avec l'assortativité, cela signifie que deux individus liés dans le réseau 2 ont tendance à avoir un nombre d'années d'ancienneté similaire. Alors que pour le réseau 5, si un individu interagit beaucoup avec la compagnie d'assurance, ceux liés avec lui ont tendance à moins interagir.

Comme il a été expliqué, la méthode de Lan et al., 2018 est très intéressante pour estimer une matrice de covariance d'un réseau lorsque la matrice d'adjacence est disponible. Cependant, il existe aussi certaines limites. Prenons l'exemple de deux paires de nœuds reliés uniquement par un lien direct dans le réseau. Ils auront la même covariance et de ce fait également la même corrélation. Toutefois, dans le cas d'un réseau biparti, cela peut poser problème. Comme il a été mentionné précédemment, les liens entre les individus sont créés lorsqu'ils font partie d'un même groupe. Selon la taille de ces groupes, on peut supposer que certains liens sont plus forts que d'autres, comme il a été vu à l'aide de l'équation (1.2). Si l'une des deux paires forme un groupe à eux seul et que l'autre paire partage un groupe composé de plusieurs dizaines ou centaines d'individus, il est logique de croire que la corrélation entre les deux individus formant la première paire doit être plus grande. Malgré cette intuition, le modèle de Lan et al., 2018 ne fera pas de différence entre ces deux paires d'individus et leur covariance sera la même.

Le dernier chapitre a ouvert la voie à de futures explorations. L'équation (4.4) offre une alternative au modèle de Lan et al., 2018 pour estimer une matrice de covariance. Grâce à sa

simplicité, aucun algorithme n'est nécessaire pour estimer les paramètres β. Effectivement, il est possible à l'aide de la matrice d'adjacence A et d'un vecteur d'une variable aléatoire Y d'estimer la matrice de covariance. Ces deux éléments permettent de calculer l'assortativité pour estimer la valeur des paramètres β par la suite. La matrice d'adjacence permet également de trouver les degrés des individus comme il a été abordé à l'équation (1.3) pour former la matrice D. Certaines limites sont toutefois liées à ce modèle. Comme il a été mentionné, une assortativité négative ne permet pas d'obtenir une matrice définie positive. Un autre inconvénient comparativement à l'approche de Lan et al., 2018 est qu'il ne permet pas de considérer des liens dans le réseau de longueur plus grande que 1. Une étude de Monte-Carlo permettrait de mieux comprendre les propriétés de l'équation (4.4) ainsi que de tester sa robustesse.

Finalement, ce mémoire a permis d'explorer plusieurs techniques pour comprendre et extraire l'information que peut contenir un réseau. De futures recherches pourraient explorer la façon d'utiliser cette information dans des modèles prédictifs. L'utilisation d'un modèle de Cox ou d'un modèle conjoint de données de longitudinales et de survie permettrait de voir comment les relations présentent dans un réseau peuvent influencer le comportement des individus.

Annexe A

Preuves et calculs mathématiques en lien avec certaines équations

A.1 Calcul de l'équation (4.12) de la section 4.4

La fonction de hasard de la section 4.4 est décrite ainsi,

$$h_i\left(t|M_i\left(t\right)\right) = h_0\left(t\right)\exp\left(\gamma_0 + \gamma_1 X_i + \alpha m_i\left(t\right)\right), \tag{A.1}$$

avec,

$$Y_i\left(t\right) = m_i\left(t\right) = \beta_0 + \beta_1 t + \beta_2 X_i + b_{0i}. \tag{A.2}$$

Posons le taux de base égale à 1, la première équation peut donc se réécrire ainsi,

$$h_i\left(t|M_i\left(t\right)\right) = \exp\left(\gamma_0 + \gamma_1 X_i + \alpha\left(\beta_0 + \beta_1 t + \beta_2 X_i + b_{0i}\right)\right). \tag{A.3}$$

La fonction de hasard cumulative est,

$$\begin{aligned}
\Lambda\left(t\right) &= \int_0^t \lambda\left(y\right) dy, \\
&= \int_0^t \exp\left(\gamma_0 + \gamma_1 X + \alpha\left(\beta_0 + \beta_1 y + \beta_2 X + b_0\right)\right) dy, \\
&= \frac{\exp\left(\gamma_0 + \alpha\beta_0 + \alpha b_0 + \left(\gamma_1 + \alpha\beta_2\right) X\right)}{\alpha\beta_1}\left(\exp\left(\alpha\beta_1 t\right) - 1\right).
\end{aligned} \tag{A.4}$$

Posons $U \sim Unif\left(0,1\right)$ et $\Lambda\left(t\right) = -\ln\left(U\right)$. Alors,

$$-\ln\left(U\right) = \frac{\exp\left(\gamma_0 + \alpha\beta_0 + \alpha b_0 + \left(\gamma_1 + \alpha\beta_2\right) X\right)}{\alpha\beta_1}\left(\exp\left(\alpha\beta_1 t\right) - 1\right), \tag{A.5}$$

$$\exp\left(\alpha\beta_1 t\right) = \frac{-\ln\left(U\right)\left(\alpha\beta_1\right)}{\exp\left(\gamma_0 + \alpha\beta_0 + \alpha b_0 + \left(\gamma_1 + \alpha\beta_2\right) X\right)} + 1, \tag{A.6}$$

$$\alpha\beta_1 t = \ln\left(\frac{-\ln(U)(\alpha\beta_1)}{\exp(\gamma_0 + \alpha\beta_0 + \alpha b_0 + (\gamma_1 + \alpha\beta_2)X)} + 1\right), \tag{A.7}$$

$$t = \frac{1}{\alpha\beta_1}\ln\left(\frac{-\ln(U)(\alpha\beta_1)}{\exp(\gamma_0 + \alpha\beta_0 + \alpha b_0 + (\gamma_1 + \alpha\beta_2)X)} + 1\right). \tag{A.8}$$

A.2 Preuve que la matrice $A + D$ de l'équation (4.4) est définie positive

Démontrons que la matrice M est définie positive.

$$M = A + D \tag{A.9}$$

où

- A est un matrice d'adjacence contenant des 0 et des 1.
- D est une matrice diagonale des degrés. Les degrés sont des valeurs positives.

La démonstration est inspirée de l'essai de Sylvain-Morneau, 2021 ainsi que de l'article de Kelner, 2009. Posons un réseau biparti β ayant n noeuds. Sélectionnons au hasard une arête du graphe $\{u, v\} \in \beta$. Posons également $D_{\{u,v\}}$ et $A_{\{u,v\}}$ la matrice de degré et d'adjacence avec les mêmes n noeuds, mais contenant uniquement le lien entre u et v.

$D_{\{u,v\}}$ est donc de la forme,

$$D_{\{u,v\}} = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & 1 & \vdots & 0 & \vdots \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & 0 & \vdots & 1 & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix},$$

$A_{\{u,v\}}$ est donc de la forme,

$$A_{\{u,v\}} = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & 0 & \vdots & 1 & \vdots \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & 1 & \vdots & 0 & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}.$$

La formule A.9 pour l'arrête $\{u, v\}$ s'écrit alors,

$$M_{\{u,v\}} = D_{\{u,v\}} + A_{\{u,v\}}. \tag{A.10}$$

$M_{\{u,v\}}$ est donc de la forme,

$$M_{\{u,v\}} = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & 1 & \vdots & 1 & \vdots \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & 1 & \vdots & 1 & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}.$$

La matrice M peut donc être réécrite comme étant la somme des matrices $M_{\{u,v\}}$ de chaque lien existant dans le réseau. On obtient alors,

$$M = \sum_{\{u,v\}\in\beta} M_{\{u,v\}}. \tag{A.11}$$

À partir de cette expression, nous pouvons calculer directement pour un vecteur α non nul de longueur n l'expression suivante,

$$\alpha^\top M_{\{u,v\}}\alpha = \alpha_u^2 + 2\alpha_u\alpha_v + \alpha_v^2 = (\alpha_u + \alpha_v)^2. \tag{A.12}$$

Grâce aux deux dernières équations on obtient,

$$\alpha^\top M\alpha = \alpha^\top \sum_{\{u,v\}\in\beta} M_{\{u,v\}}\alpha = \sum_{\{u,v\}\in\beta} \alpha^\top M_{\{u,v\}}\alpha = \sum_{\{u,v\}\in\beta} (\alpha_u + \alpha_v)^2. \tag{A.13}$$

L'équation (A.13), montre que $\alpha^\top M\alpha$ ne peut pas être inférieur à 0. M est donc au moins une matrice semi-définie positive. Il faut donc ajouter à M une condition pour qu'elle soit définie positive. Si nous regardons la preuve dévelopée, l'équation (A.13) peut être égale à 0 pour tous les liens d'un réseau si pour chaque lien $\{u,v\}$ $\alpha_u = -\alpha_v$. Par contre, prenons le cas où un cycle de longueur 3 $\{u,v,w\}$ est présent dans le réseau. L'équation (A.13) sera alors,

$$(\alpha_u + \alpha_v)^2 + (\alpha_u + \alpha_w)^2 + (\alpha_v + \alpha_w)^2.$$

Dans ce cas de figure, il devient alors impossible de trouver une combinaison de $\alpha_u, \alpha_v, \alpha_w$ qui rendra la dernière équation égale à 0. On en conclut que la matrice M est toujours au minimum semi-définie positive et que si un cycle de longueur impair est présent dans le réseau M devient définie positive.

A.3 Preuve que les degrés suivent une loi de Poisson dans un réseau Bernoulli

À partir de l'équation (1.16), évaluons le cas où le nombre de noeuds n est très grand.

Nous avons vu à la sous-section 1.1, qu'une grande majorité des réseaux sont clairsemés. Si l'on se fie à l'équation (1.6), cela signifie que la moyenne des degrés c augmente beaucoup moins rapidement que la taille du réseau n. En isolant p dans l'équation (1.15) on obtient $p = c/(n-1)$ et ce rapport va devenir extrêmement petit à mesure que n augmente. Rappelons également qu'en utilisant les séries de Taylor, nous pouvons développer $log(1-x)$ comme étant la série $-x - \frac{x^2}{2} - \frac{x^3}{3} + O(x^3)$.

On peut donc réécrire la 2^e partie de l'équation (1.16) en employant la nouvelle définition de p ainsi que les séries de Taylor pour développer le logarithme comme suit,

$$\ln\left[(1-p)^{n-1-d}\right] = (n-1-d)\ln\left(1-\frac{c}{n-1}\right) \simeq -(n-1-d)\frac{c}{n-1} \simeq -c. \tag{A.14}$$

Dans l'équation (A.14), uniquement le premier terme de la série de Taylor est conservé puisqu'on est dans le cas où n est très grand. Donc, comme discuté précédemment, la valeur de c va devenir très petite. Dans le développement de la série de Taylor les $2^e, 3^e, ...$ termes élèvent c à la puissance $2, 3, ...$ et donc la valeur du terme devient extrêmement faible, voire nulle. C'est pour cela qu'on ne conserve que le premier élément.

On peut maintenant réécrire $(1-p)^{n-1-d}$ comme étant égale à e^{-c} lorsque n est très grand. Avec la même supposition sur la taille du réseau, on peut également réécrire la combinaison de l'équation (1.16) comme étant,

$$\binom{n-1}{d} = \frac{(n-1)!}{(n-1-d)!d!} \simeq \frac{(n-1)^d}{d!}. \tag{A.15}$$

En rassemblant les deux dernières équations, on peut réécrire l'équation (1.16) ainsi,

$$p_d = \frac{(n-1)^d}{d!}p^d e^{-c} = \frac{(n-1)^d}{d!}\left(\frac{c}{n-1}\right)^d e^{-c} = e^{-c}\frac{c^d}{d!}. \tag{A.16}$$

On retrouve maintenant une distribution de Poisson avec paramètre c. Le réseau $G(n,p)$ a donc une distribution des degrés qui suit la loi de Poisson lorsque n est grand.

Annexe B

Utilisation de la librairie igraph

B.1 Example d'utilisation de la librairie igraph

B.1.1 Initialisation d'un réseau à partir d'un jeu de données

Reprenons les deux premières colonnes du tableau 1.3.

Tableau B.1 – Tableau des liens du réseau

ID Individu i	ID Individu j
1	2
1	4
2	4
3	5
5	6
6	7

À partir de ce tableau, on peut créer un jeu de données contenant la liste des noeuds ainsi que d'autres attributs des noeuds,

```
node_list = data.frame(node = unique(c(reseau$id_individu_i,reseau$id_individu_j))).
```

On est maintenant en mesure de créer un graphe `igraph`,

```
g = graph_from_data_frame(d=reseau, vertices=node_list, directed = FALSE),
```

où l'objet réseau est le tableau B.1. À partir de cet objet, il est maintenant possible de faire toute sorte d'opérations avec les fonctions de la librairie `igraph`. Les différentes statistiques du tableau 1.16 ont été obtenues avec les fonctions suivantes,

```
composantes = components(g),
Nb compo = composantes$no,
Taille minimale = min(composantes$csize),
Taille maximale = max(composantes$csize),
Taille moyenne = mean(composantes$csize),
Longueur moy des chemins = mean_distance(g),
Diamètre = diameter(g).
```

Il est également possible de calculer la densité, la transitivité ainsi que l'assortativité comme suit,

```
densité = edge_density(g),
transitivité = transitivity(g),
assortativité = assortativity(g, y),
```

où l'objet y représente les valeurs pour chaque noeud de la variable aléatoire y sur laquelle on veut calculer l'assortativité.

Il est également possible d'extraire à partir d'un graphe sa matrice d'adjacence,

```
matrice d'adjacence = as_adjacency_matrix(g).
```

Bibliographie

Arcagni, A., Grassi, R., Stefani, S., and Torriero, A. (2021). Extending assortativity : an application to weighted social networks. *Journal of Business Research*, 129 :774–783.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology : understanding the cell's functional organization. *Nature reviews genetics*, 5(2) :101–113.

Berry, K. J., Kvamme, K. L., Johnston, J. E., and Mielke Jr., P. W. (2021). *Permutation Statistical Methods with R*. Springer.

Bélisle, C. (2018). Statistique mathématique, stt-4000, université laval.

Chen, S., Kang, J., Xing, Y., Zhao, Y., and Milton, D. K. (2018). Estimating large covariance matrix with network topology for high-dimensional biomedical data. *Computational Statistics & Data Analysis*, 127 :82–95.

Cox, R. (1972). Regression models and life tables. *Journal of the Royal Statistic Society*, B(34) :187–202.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems :1695.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1) :269–271.

Duchesne, T. (2020). Théorie et applications des méthodes de régressions, stt-7125, université laval.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets : Reasoning about a highly connected world*. Cambridge university press.

Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1) :17–60.

Erdős, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae*, 6 :290–297.

Farine, D. (2016). *assortnet : Calculate the Assortativity Coefficient of Weighted and Binary Networks*. R package version 0.12.

Furgal, A. K., Sen, A., and Taylor, J. M. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review*, 87(2) :393–418.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6) :1360–1380.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Howell, D. C. (2007). *Statistical methods for psychology*. Cengage Learning.

Johnston, J. E., Berry, K. J., and Mielke Jr, P. W. (2007). Permutation tests : precision in estimating probability values. *Perceptual and Motor Skills*, 105(3) :915–920.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data : an introduction to cluster analysis*. John Wiley & Sons.

Kelner, J. (2009). An algorithmist's toolkit : lecture 2. https://ocw.mit.edu/courses/18-409-topics-in-theoretical-computer-science-an-algorithmists-toolkit-fall-2009/resources/mit18_409f09_scribe2/.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis : techniques for censored and truncated data*, volume 1230. Springer.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*. Springer.

Kleinbaum, D. G. and Klein, M. (1996). *Survival analysis a self-learning text*. Springer.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621.

Lan, W., Fang, Z., Wang, H., and Tsai, C.-L. (2018). Covariance matrix estimation via network structure. *Journal of Business & Economic Statistics*, 36(2) :359–369.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5) :603–621.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*, volume 71. Siam.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42.

Mooney, C. Z. (1997). *Monte carlo simulation.* Number 116. Sage.

Moreno, J. L. (1934). Who shall survive ? : A new approach to the problem of human interrelations. *Nervous and mental disease publishing co.*

Newman, M. (2010). Networks : An introduction : Oxford, uk : Oxford university press. 720 pp.

Newman, M. (2018). *Networks.* Oxford university press.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20) :208701.

Newman, M. E. (2003). Mixing patterns in networks. *Physical review E*, 67(2) :026126.

Pham, H. (2006). *Springer handbook of engineering statistics.* Springer Science & Business Media.

Pinheiro, J., Bates, D., et Deepayan Sarkar, S. D., and R Core Team (2021). *nlme : Linear and Nonlinear Mixed Effects Models.* R package version 3.1-152.

R Core Team (2021). *stats : The R Stats Package.* R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata.* STATA press.

Rice, J. A. (2006). *Mathematical statistics and data analysis.* Cengage Learning.

Rizopoulos, D. (2010). JM : An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9) :1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data : With applications in R.* CRC press.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Strang, G. (2006). *Linear algebra and its applications.* Belmont, CA : Thomson, Brooks/Cole.

Sylvain-Morneau, J. (2021). Essai : Méthodes statistiques pour données avec structure de réseau, université laval.

Talbot, D. (2021). Analyse de survie, epm-7028, université laval.

Taylor, B. (1715). *Direct and Indirect Methods of Incrementation*. J. Knapton.

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-11.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Journal of Psychology*, 21(1) :3–15.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Yuan, Y., Yan, J., and Zhang, P. (2021). Assortativity measures for weighted and directed networks. *Journal of Complex Networks*, 9(2) :cnab017.