Statistique bayésienne
Rapport final
31 janvier 2017

Analyse de *Social Networks and the Identification of Peer Effects*

Tom Duchemin
Mehdi Miah
Benoit Robaglia

Enseignants :
Anna Simoni
Rémi Bardenet

Résumé : Rédigé en 2013, *Social Networks dans the Identification of Peer Effets* traite de l'effet des amis dans l'évolution des notes dans un lycée américain. A travers trois modèles, les auteurs montrent comment l'amitié et des caractères cachés influencent les notes et les amitiés. Après une analyse et une critique de l'article, une modélisation sous `R` a été faite concernant les deux premiers modèles. Bien que les données soient simulées, les résultats obtenus sont en cohérence avec ceux des auteurs.
Mots clés : peer effects, network, endogénéité, Bayes, Metropolis-Hastings

Table des matières

1 Introduction

L'analyse des réseaux sociaux est aujourd'hui un domaine de recherche en pleine expansion notamment grâce aux opportunités ouvertes par le *Big Data*. L'article de Goldsmith-Pinkham et Imbens intitulé *Social Networks and the Identification of Peer Effets* et publié en 2013 dans le *Journal of Business & Economic Statistics* s'inscrit dans ce domaine de recherche et s'intéresse tout particulièrement aux effets de pair. Les effets de pair sont définis comme l'influence des proches sur les variables étudiées et le sujet étudié est ici l'impact des amis sur l'évolution des notes dans un réseau de lycéens américains.

Pour étudier cet effet, les auteurs utilisent plusieurs modèles économétriques qui décrivent directement l'effet de pair mais aussi la formation des réseaux et pointent les difficultés à travailler avec des données de réseau. Ces problèmes sont divers. Ils peuvent provenir de l'endogénéité des modèles (la formation des réseaux est un problème très complexe), de problèmes dans les données (comment définir clairement ce qu'est un lien d'amitié ?) ou bien même de problèmes de calculs (comment estimer nos paramètres avec autant de données ?).

Ce rapport présente cet article et tente de reproduire les méthodes utilisées et de discuter ses conclusions. Dans un premier temps, les différents modèles économétriques utilisés dans l'article, ainsi que les résultats principaux des auteurs, sont présentés. Ensuite, un discours critique sera formulé et, notamment, les différentes hypothèses des modèles seront discutées. Enfin, dans un dernier temps, les modèles présentés seront reproduits sur des données simulées et l'accent sera mis sur les méthodes bayésiennes utilisées pour l'estimation des paramètres.

2 Synthèse de l'article de Goldsmith-Pinkham et Imbens

2.1 Notations et généralités

Notations

La variable d'intérêt de cette étude est le *"grade point average"* (GPA) pour un étudiant présent dans la seconde période ; elle est notée Y_i. Le N-vecteur $\mathbf{Y}$ regroupe le GPA de tous les individus. Les covariables exogènes sont réunies dans la matrice de taille $N \times K$, $\mathbf{X}$ ayant pour ligne i les caractéristiques de l'individu i. Les auteurs ne considèrent uniquement comme covariable le GPA initial à la période 1. $\mathbf{X}$ est donc un vecteur colonne ($K = 1$).
De plus, la structure du réseau à la période 2 est donnée par la matrice adjacente et symétrique $\mathbf{D}$ où $D_{i,j} = 1$ si i et j sont amis et 0 sinon (il suffit que l'un des deux considère l'autre comme un ami pour que la liaison soit faite). De la même manière, on note la matrice des relations à la période 1 $\mathbf{D}_0$. De là, plusieurs variables décrivant le réseau sont créées :

- le vecteur $\mathbf{M}$, où $M_i = \sum_{j=1}^{N} D_{i,j}$, est le nombre d'amis de i dans la seconde période ;
- $\mathbf{G} = \text{diag}(\mathbf{M})^{-1}\mathbf{D}$, la matrice $\mathbf{D}$ normalisée ;
- la matrice $\mathbf{F}_0$ où $F_{0,i,j} = 1$ si i et j ont un ou plusieurs amis en commun et 0 sinon.

La question de l'identification

Dans un problème impliquant des effets de pairs, la question de l'identification est très délicate car il est difficile de déterminer quelles causes impliquent quels effets, en raison des effets endogènes importants présents dans le réseau. Pour résoudre ce problème, les auteurs "imposent de la structure" au réseau pour limiter les dépendances entre observations. Nous avons trouvé ce point un peu obscur notamment à cause du grand nombre de références citées et par le faible développement de cette théorie. Nous ne nous attarderons donc pas sur cette question.

2.2 Le modèle de Manski "linéaire en moyenne" : estimer les effets de pairs sous l'exogénéité du réseau

L'hypothèse effectuée pour ce premier modèle est que le réseau est exogène.
Le premier modèle implémenté par Goldsmith-Pinkhan et Imbens est le modèle linéaire-en-moyennes de Manski (1993) (pour une synthèse graphique du modèle, voir la Figure 1) :

$$Y_i = \beta_0 + \beta_x X_i + \beta_{\overline{y}} \overline{Y}_{(i)} + \beta_{\overline{x}} \overline{X}_{(i)} + \eta_i$$

où :

$$\overline{Y}_{(i)} = \frac{1}{M_i} \sum_{j=1}^{N} D_{i,j} Y_j = \sum_{j=1}^{N} G_{i,j} Y_j$$

$$\overline{X}_{(i)} = \frac{1}{M_i} \sum_{j=1}^{N} D_{i,j} X_j = \sum_{j=1}^{N} G_{i,j} X_j$$

ou sous forme matricielle :

$$\mathbf{Y} = \beta_0 \mathbf{t}_N + \beta_x \mathbf{X} + \beta_{\overline{y}} \mathbf{G}\mathbf{Y} + \beta_{\overline{x}} \mathbf{G}\mathbf{X} + \eta$$

avec $\mathbf{t}_N$ le vecteur où tous les termes sont égaux à 1. Les effets de pairs sont représentés par les coefficients $\beta_{\overline{y}}$ et $\beta_{\overline{x}}$. Les auteurs interprètent le premier coefficient comme l'effet causal de fournir de l'aide scolaire à un ami sur son propre GPA et le deuxième comme comme l'effet de modifier les covariables de ses amis.

Deux hypothèses sont posées pour estimer et identifier le modèle :

1. (Exogénéité) $\eta \perp \mathbf{X}, \mathbf{D}$
2. (Normalité) $\eta | \mathbf{X}, \mathbf{D} \sim \mathcal{N}(0, \sigma^2 I_N)$

Ainsi, dans la configuration de Bramouillé, Djebbari et Fortin (2009) où le réseau ne forme pas une partition de la population ($\mathbf{GG} \neq \mathbf{G}$), alors le modèle peut être estimé et identifié.

La technique d'estimation choisie par les auteurs est une approche bayésienne qui présente de nombreux avantages. Tout d'abord, elle permet de contourner le problème de maximisation de vraisemblance car la loi a posteriori est calculée par des simulations MCMC. De plus, elles permettent d'incorporer des informations supplémentaires et subjectives (grâce à un prior) aux données. Enfin, la souplesse de ces méthodes permettent de les étendre relativement facilement à d'autres modèles plus complexes comme c'est le cas dans cet article.

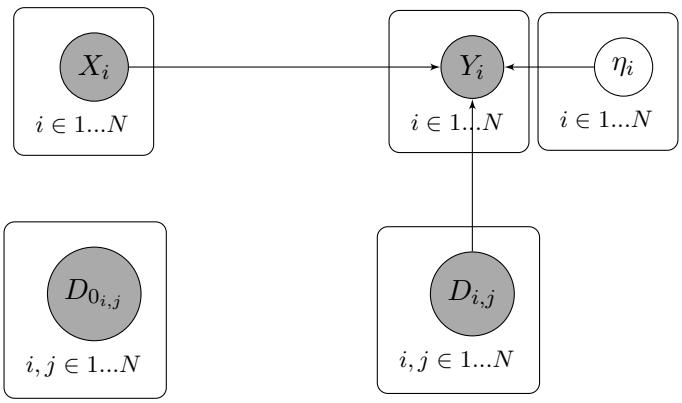

FIGURE 1 – Graphe orienté pour le premier modèle

2.3 Un modèle de formation de réseau exogène

Dans un second temps, le réseau est créé à travers un modèle de formation. Cette variante indique que les relations d'amitié à la seconde période sont impactées par les notes obtenues lors de la première période. En effet, le premier modèle ne prend pas en compte l'effet des notes $\mathbf{X}$ sur $\mathbf{D}$.

Les auteurs de l'article utilisent la notion d'utilité pour modéliser la formation du réseau d'amitié : deux individus créeront un lien d'utilité si les deux y trouvent un intérêt. Pour les individus i et j, cet intérêt dépendra :

— de la différence des notes obtenues dans la période 1 : $|X_i - X_j|$;
— de l'existence d'une relation d'amitié à la période précédente : $D_{0,i,j}$;
— de l'existence d'amis en commun : $F_{0,i,j}$

Mathématiquement, cela se modélise à travers l'équation d'utilité suivante (pour une synthèse graphique du modèle, voir la Figure 2) :

$$U_i(j) = \alpha_0 + \alpha_x|X_i - X_j| + \alpha_d D_{0,i,j} + \alpha_f F_{0,i,j} + \epsilon_{i,j},$$

où $\epsilon_{i,j}$ suit une loi logistique telle que $\epsilon_{i,j}$ soit indépendant de $\epsilon_{j,i}$.

Ainsi, une liaison sera faite entre les individus i et j dés lors que chaque utilité est positive :

$$D_{i,j} = \mathbb{1}_{U_i(j)>0}\mathbb{1}_{U_j(i)>0}$$

Ensuite, $\mathbf{D}$ étant une matrice d'adjacence, chaque élément $D_{i,j}$ suit une loi de Bernoulli de paramètre :

$$\mathbb{P}(D_{i,j} = 1|\mathbf{D}_0, \mathbf{X}) = p_{i,j} \times p_{j,i} \text{ avec}$$

$$p_{i,j} = p_{j,i} = \frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_d D_{0,i,j} + \alpha_f F_{0,i,j})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_d D_{0,i,j} + \alpha_f F_{0,i,j})}$$

Les paramètres sont estimées par des méthodes bayésiennes et, lors de l'estimation de ces α, les priors choisis seront des gaussiennes centrées réduites indépendantes. Les résultats sont peu sensibles à ces choix-ci.

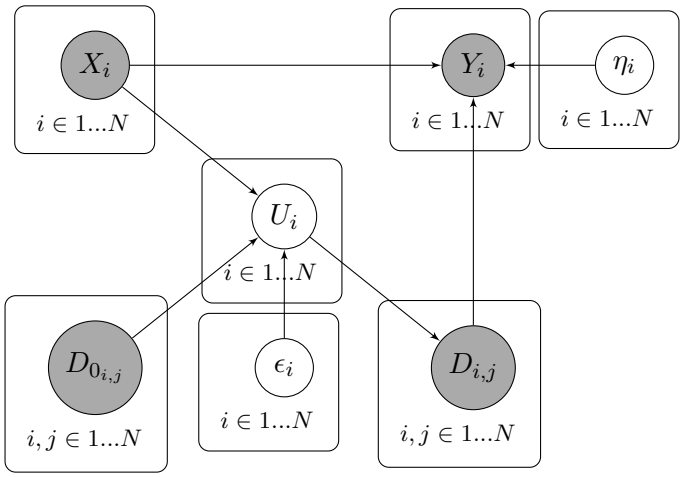

FIGURE 2 – Graphe orienté pour le second modèle

Quant aux résultats, l'article montre que l'existence d'une amitié passée ou d'un ami en commun dans le passé favorise une amitié future et qu'un écart fort au niveau des notes pénalise cette fonction d'utilité.

2.4 Un modèle de formation de réseau endogène

Les paramètres estimés par le modèle précédent ne peuvent cependant pas être interprétés de manière causale si les groupes de pair eux-même sont endogènes. Cette endogénéité s'entend par le fait que les individus d'un même groupe peuvent être similaires en termes de caractéristiques non-observées et qui peuvent influencer la variable à expliquer. Il s'agit d'un problème de variable omise : le bruit de l'équation définissant $\mathbf{Y}$ n'est plus indépendant de $\mathbf{D}$ et de $\mathbf{X}$. Par exemple, dans notre cas, le milieu social pourrait être une caractéristique non-observée qui influencerait les amitiés mais aussi les notes des élèves.

Afin de pallier ce problème d'endogénéité, un nouveau modèle est proposé. Introduisons ξ_i une variable décrivant les caractéristiques inobservées de l'individu i. On introduit aussi η_i telle que $\mathbf{E}[\eta_i|C_i = c] = \delta_c$ avec C_i la variable qui assigne chaque individu à son groupe de pairs. Nous reformulons l'équation du modèle MLIM de la manière suivante :

$$Y_i = \beta_0 + \beta_x X_i + \beta_{\bar{y}} \bar{Y}_{(i)} + \beta_{\bar{x}} \bar{X}_{(i)} + \beta_\xi \xi_i + \eta_i$$

De même, l'équation d'utilité définissant la formation du réseau peut-être redéfinie :

$$U_i(j) = \alpha_0 + \alpha_x |X_i - X_j| + \alpha_\xi |\xi_i - \xi_j| + \alpha_d D_{0ij} + \alpha_F F_{0,ij} + \epsilon_{ij}$$

Une synthèse graphique du modèle peut être lue en Figure 3.

Manski étudie un cas où les caractéristiques inobservées ξ_i correspondent à l'indicatrice de groupe de pairs C_i. Dans ce cas, on s'attend à un α_ξ négatif et très grand en valeur absolue, ce qui signifierait que les individus de groupes de pairs différents ont une utilité très faible à former un lien.

Les auteurs ont estimé ce modèle en effectuant les hypothèses suivantes :
— $\eta|\xi \sim \mathbb{N}(0, \sigma^2 I_N)$
— $\epsilon_{i,j}$ suit une loi logistique
— $p(\xi_i = 1|X, D_0) = 1 - p(\xi_i = 0|X, D_0) = p = 1/2$

Les résultats de l'estimation, effectuée par des méthodes bayésiennes avec des priors gaussiens sur la plupart des paramètres, montre que le paramètre α_ξ impacte significativement le modèle de formation de réseau alors que le paramètre β_ξ a peu d'impact sur la variable observée.

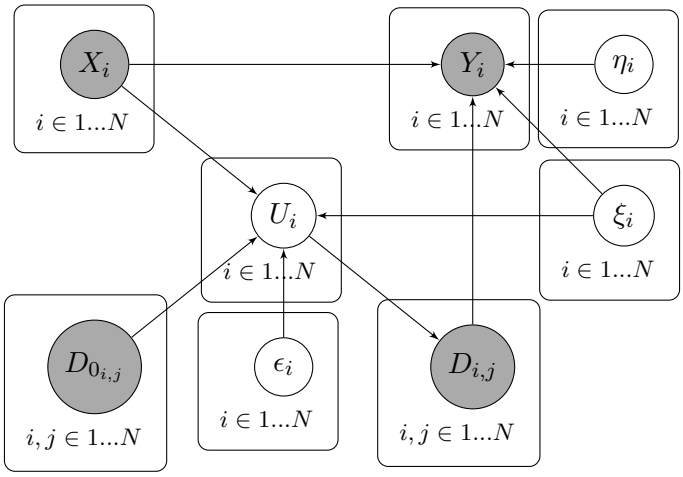

FIGURE 3 – Graphe orienté pour le troisième modèle

2.5 Extensions et limites du modèle soulevées par les auteurs

Le modèle MLIM présenté dans l'article ainsi que les données utilisées soulèvent quelques problèmes relevés par l'auteur. Premièrement, le modèle MLIM assume que l'effet de tous les

pairs est similaire. Ensuite les données sont construites sur des déclarations des élèves et il se peut qu'il y ait donc des "erreurs de mesure" dans la base.

Un modèle MLIM fondé sur deux réseaux plutôt qu'un seul est alors proposé. On note A le premier réseau et B le second réseau. L'équation du modèle devient :

$$Y_i = \beta_0 + \beta_X X_i + \beta_{\bar{y},A}\bar{Y}_{A,(i)} + \beta_{\bar{x},A}\bar{X}_{A,(i)} + \beta_{\bar{y},B}\bar{Y}_{B,(i)} + \beta_{\bar{x},B}\bar{X}_{B,(i)} + \eta_i$$

On doit supposer $\eta \perp X, D_A, D_B$ afin de pouvoir effectuer l'estimation correctement.

Trois limites du modèle sont étudiés par les auteurs :

— La première limite étudiée est la possibilité d'erreur de mesure. Pour cela, les deux groupes suivants sont construits : $D_{A,ij} = D_{ij}$ et $D_{B,ij} = (1 - D_{ij})D_{0,ij}$. Les individus du groupe B (c'est-à-dire les individus déclarés amis en première période mais pas en seconde) devraient avoir un effet de pair beaucoup plus faible que les individus du groupe A. L'estimation de ce modèle montre au contraire que l'effet de pair du groupe A ($\beta_{\bar{y},A}$) est très proche de l'effet de pair du groupe B ($\beta_{\bar{y},B}$). Ce résultat pose problème puisque le résultats des individus qui étaient déclarés amis en première mais ne le sont plus ensuite ont le même impact que les individus déclarés amis en seconde période. Cela montre les limites des données construites sur la déclaration des individus.

— La seconde limite étudiée est que le modèle de Manski suppose toutes les relations d'amitiés bilatérales or il se peut que les amitiés unilatérales aient un impact plus faible sur la variable étudiée. Pour cela, deux groupes sont construits : $D_{A,ij} = D_{ij}$ et $D_{B,ij} = D_{M,ij}$ avec $D_{M,ij}$ l'indicatrice qui prend 1 si l'amitié entre les individus i et j est déclarée par les deux individus i et j et 0 sinon. Les résultats estimés par les auteurs ne montrent pas de résultats significatifs pour le groupe B et il ne semble donc pas y avoir d'effets liés aux relations unilatérales.

— La troisième limite étudiée est que les effets des "amis de nos amis" n'est pas pris en compte. Deux groupes sont construits : $D_{A,ij} = D_{ij}$ et $D_{B,ij} = F_{M,ij}(1 - D_{ij})$ avec $F_{M,ij}$ l'indicatrice que les individus i et j ont au moins un ami en commun. Des résultats significatifs sont montrés pour l'effet des "amis de nos amis" sur les résultats d'intérêt.

Ces résultats complémentaires montrent donc que certains paramètres ne sont pas pris en compte dans le modèle MLIM. Cependant, les paramètres $\beta_{\bar{y},A}$ et $\beta_{\bar{x},A}$ estimés par ces trois modèles différents présentent des valeurs très proches des paramètres $\beta_{\bar{y}}$ et $\beta_{\bar{x}}$ estimés dans les modèles précédents. Ainsi, il ne semble pas que ces omissions aient un véritable impact sur l'estimation des paramètres d'intérêt.

3 Discussion

3.1 L'apport de cet article à la littérature scientifique et les politiques

L'article de Goldsmith-Pinkham a été très remarqué au sein de la communauté scientifique grâce aux apports méthodologiques qu'ils proposent à l'étude et à la modélisation des effets de pair comme en témoignent les nombreuses réponses et commentaires sur cet article.

Tout d'abord, comme le souligne Jackson (2013), les auteurs se sont confrontés aux grands problèmes posés par l'analyse des effets de pair à savoir :

- l'identification : il est en effet difficile d'identifier les moyens par lesquels les pairs influencent le comportement d'un individu ;
- l'endogénéité du réseau et les effets d'homophilie : ne pas prendre en compte les caractéristiques inobservées peut biaiser l'estimateur des effets de pair ;
- la complexité algorithmique : les problèmes impliquant des réseaux ont généralement une complexité en exponentielle de la taille du réseau (Chandrasekhar et Jackson 2012), d'où

le choix des méthode bayésiennes dans l'article qui permettent une complexité polynomiale ;
- l'erreur de mesure : les relations sont difficiles à observer et à mesurer. En effet, de nombreuses relations peuvent avoir été omises par les sujets et l'intensité des relations n'est pas prise en compte (toutes les relations n'ont pas le même poids).

Plus particulièrement, le cœur de leur étude est l'incorporation de l'endogénéité du réseau dans la modélisation de la variable dépendante. En effet, l'homophilie, le fait de partager des caractéristiques communes, observées ou inobservées, augmente fortement la probabilité de former un lien (Currarini, Jackson, and Pin 2009, 2010). Elaborer un modèle avec des caractéristiques inobservées est donc un apport considérable à la compréhension des effets de pairs. Ils ont par ailleurs testé les conséquences possibles d'une non prise en compte de l'endogénéité sur le biais de l'estimateur des effets de pair. Ils trouvent en effet que les caractéristiques inobservées jouent un rôle significatif dans la formation du réseau mais est quasiment nul dans l'explication sur les notes (le GPA). Ce résultat est surprenant car il est presque certain que des éléments inobservés soient corrélés au GPA. Nous reviendrons sur ce point ultérieurement.

Un second apport que nous pouvons souligner concerne les politiques publiques et plus particulièrement les interprétations que peuvent exploiter les politiques sur les estimations du modèle linéaire en moyenne. Pour comprendre ce point, nous allons exposer les travaux de Kline et Tamer (2013) qui fournissent une interprétation intéressante de ces coefficients. Ils énoncent pour cela le théorème suivant :

Théorème : considérons le modèle suivant sous hypothèse d'exogénéité ($\mathbb{E}(\eta|\mathbf{X},\mathbf{G}) = 0$) :

$$\mathbf{Y} = \beta_0 \mathbf{t}_N + \beta_x \mathbf{X} + \beta_{\bar{y}} \mathbf{G}\mathbf{Y} + \beta_{\bar{x}} \mathbf{G}\mathbf{X} + \eta$$

Et supposons que $\Gamma = \mathbf{I} - \mathbf{G}\beta_{\bar{Y}}$ est inversible.
Alors les effets marginaux de $\mathbf{X}$ sur $\mathbb{E}(y_i|\mathbf{X},\mathbf{G})$ sont donnés par :

$$\beta_x e_i \Gamma^{-1} + \beta_{\bar{x}} e_i \Gamma^{-1} \mathbf{G}$$

Avec $\mathbf{e} = (e_1, ...e_N)$ la base canonique. Ainsi, l'effet marginal de la k-ème covariable de l'individu j est donné par le coefficient (k,j) de cette matrice.

Qualitativement, les effets $\mathbf{X}$ sur $\mathbf{Y}$ peuvent se décomposer en 3 :
- *Effet d'interaction non-sociale exogène* : toutes choses égales par ailleurs une modification de la covariable de i modifie son GPA à travers le coefficient β_x
- *Effet d'interaction sociale exogène* : transformer la covariable d'un individu $j \neq i$ affectent, toutes choses égales par ailleurs, les GPA de tous ses amis à travers $\beta_{\bar{x}}$
- *Effet d'interaction sociale endogène* : les variables dépendantes y_i sont déterminées simultanément selon le paramètre $\beta_{\bar{y}}$

Toutefois, malgré l'apport méthodologique proposé par Goldsmith-Pinkham et Imbens, des critiques peuvent être émises concernant les hypothèses formulées dans leur article.

3.2 Discussions sur les hypothèses formulées par les auteurs

Dans cette partie, nous allons nous interroger sur la pertinence des hypothèses émises par les auteurs.

Tout d'abord, les auteurs soulèvent eux même diverses questions quant aux hypothèses qu'ils ont émises dans la section 7. En premier lieu, l'hypothèse d'homogénéité des effets de pair, à savoir que tous les liens possèdent la même influence semble étrange. En effet, on peut imaginer par exemple qu'une amitié extra-scolaire serait de nature différente que celle avec un camarade

de classe ou encore qu'une relation avec un garçon ou une fille affecterait le type de lien entre les individus. Les auteurs ont tenté de résoudre ce problème d'hétérogénéité en décomposant le réseau en réseaux multiples. Leurs conclusions sont assez mitigées et cela remet donc en question leur méthodologie pour traiter l'hétérogénéité inobservée. En revanche, ils n'émettent pas de doute quant à la symétrie de la fonction d'utilité de posséder un lien ce qui nous a paru étrange. Dans l'équation (5.1), on remarque que l'utilité pour un individu i de former un lien avec un individu j est totalement symétrique. Or il existe probablement des effets individuels propres à chacun qui déséquilibreraient la valeur que 2 personnes attribueraient à un même lien d'amitié. Par exemple, un élève en difficulté peut donner une grande importance à se lier d'amitié avec un élève au GPA élevé (pour de l'aide scolaire par exemple) alors que ce dernier n'aurait pas le même intérêt pour établir ce lien. De plus, la notion de centralité et de hiérarchie, pourtant très présente en théorie des réseaux sociaux n'a pas été évoquée et prise en compte dans les modèles. On distingue en effet dans un réseau d'école entre adolescents des différences de "popularité" entre les étudiants, certains ayant beaucoup de relations et d'autres très peu (Ladd, 1983). Si nous supposons l'utilité comme fonction (décroissante ?) du nombre d'amis alors l'hypothèse de symétrie de la fonction d'utilité est donc discutable.

En outre, les hypothèses concernant les résidus ont également attiré notre attention. Dans un premier temps, dans l'équation (5.1) modélisant l'utilité d'un individu, les auteurs supposent que le terme inobservé $\epsilon_{i,j}$ suit une loi logistique. Bramouillé, en réponse à cet article, écrit que cette hypothèse impose trop de restrictions entre les interactions entre variables observées et variables inobservées et qu'il faudrait favoriser plutôt des effets de substitution grâce à une plus grande gamme de probabilités. Quant à la composante inobservée ξ, il n'est pas expliqué pourquoi la distribution choisie est une loi de Bernouilli de paramètre 1/2. Enfin, nous avons trouvé que les hypothèses bayésiennes posées par Goldsmith-Pinkham et Imbens n'étaient pas assez justifiées. En effet, leur spécification des lois a priori semblent non justifié lorsque l'on sait que ce choix a un impact sur l'inférence (en particulier pour les modèles à faible observation comme c'est le cas avec le premier). De plus, ils ne précisent pas si leur décision de prior gaussien est subjective (distribution informative) auquel cas une analyse d'un expert aurait du être fournie, ou objective (ditribution non-informative) qui nécessiterait l'explication de la méthode utilisée (Méthode de Jeyffreys ou analyse de référence de Bernardo par exemple). Une explication que l'on pourrait formuler pour l'utilisation d'une telle hypothèse est la suivante : sachant que le nombre de données est important pour les modèles avec exogénéité et endogénéité, sous les bonne hypothèses, le théorème de Bernstein-Von Mises affirme qu'asymptotiquement la distribution a posteriori "efface" la loi a priori. Comme par ailleurs sa loi asymptotique est gaussienne, le choix d'une loi conjuguée s'explique.

3.3 Critiques de la méthode résolvant l'endogénéité dans le réseau

Enfin, de nombreuses critiques peuvent être formulées quant à leur méthode destinée à résoudre le problème d'endogénéité. Tout d'abord, il est surprenant que dans leurs résultats, les caractéristiques inobservées ξ contribuent significativement à la formation du réseau et de manière quasiment nulle et non significative à l'estimation la variable dépendante. Or il est difficile de croire que le GPA ne soit pas corrélé à des caractéristiques comme les habitudes de travail, la participation aux activités extra-scolaires, le milieu-socio-économique etc. La source de cette erreur provient peut-être de la mauvaise spécification du terme inobservé ξ. En effet, la simplification de ξ en une variable binaire est peut-être trop restrictive pour observer et encore moins quantifier un potentiel impact sur le GPA. Leur hypothèse est toutefois utile pour déterminer l'importance des caractéristiques inobservées à la formation d'un lien.

En outre, il est naturel de penser que les caractéristiques observées et inobservées sont de même nature. En effet, par homophilie, si 2 personnes sont amies mais que leurs caractéristiques observées sont différentes et éloignées alors il est probable que leurs caractéristiques inobservées soient assez proches et il en va de même avec les "non-amis". Il faudrait donc, comme pour les

variables observées, rajouter un terme de contextualisation des variables inobservées $\alpha_{\bar{\xi}\bar{\xi}_i}$. Par ailleurs, il est assez légitime de penser que le nombre d'amis en communs entre 2 personnes à une date t influe positivement l'amitié entre ces 2 personnes à cette même date (Comme le dit le proverbe : *"l'ami de mon ami est mon ami"*). Il faudrait donc rajouter dans l'équation d'utilité (Équation 6.2), un terme $\alpha'_f F_{i,j}$ pour capter le fait d'avoir 1 ou plusieurs amis en communs dans le réseau actuel.

4 Application

4.1 Les données

L'article de Goldsmith-Pinkham utilise, pour illustrer les modèles présentées, des données de la base *Add Health* qui est une base de données construites à partir d'un sondage effectué auprès d'étudiants américains en 1994 et 1995. Des observations provenant de 534 élèves d'un même lycée ont été utilisées dans l'article.

Cependant, pour des raisons d'anonymat, les données ne sont pas immédiatement disponibles et nous n'avons pas pu les utiliser dans ce rapport. D'autres méthodes ont dû êtres utilisées.

Pour pouvoir implémenter et tester les modèles présentés dans l'article, quatre paramètres ont dû être simulés : les notes moyennes des élèves pendant la première période (X) et pendant la seconde période (Y, la variable que l'on souhaite expliquer) et le réseau d'amitié entre les élèves pour la première période (D_0) et la seconde période (D). Différents processus de simulation ont été mis en place pour chacune de ces variables et vont être présentés :

- D_0 est simulée à partir d'un algorithme Watts-Strogatz. Cet algorithme permet de simuler un réseau qui satisfait les propriétés des *petits mondes*, c'est-à-dire que les amis sont plus susceptibles d'être amis avec des amis de leurs amis plutôt qu'avec des individus sans lien avec leur cercle d'amis. Ainsi, pour présenter de manière simplifier l'algorithme il faut, pour simuler le réseau, choisir le nombre d'individus du réseau, le nombre d'amis par cercle d'amis mais aussi la probabilité d'être ami avec des amis à l'extérieur du cercle. La fonction *sample_smallworld* du package *igraph* de R a été utilisée.

- Les notes de la première période X sont, dans l'article, situées entre 0 et 4. X est donc simulé selon une loi normale tronquée sur l'intervalle [0,4] et de paramètres choisis grâce aux statistiques descriptives de l'article (moyenne de 2.6 et variance de 0.8). Pour que les données simulées aient un sens, les valeurs sont arrondies au dixième le plus proche.

- D est simulée grâce à une simulation basée sur le modèle de formation de réseau exogène présentée dans l'article (voir partie 2.4).Pour rappel, un lien se forme si l'utilité de former un lien pour chacun des deux partis est strictement positive. L'équation d'utilité suivante a été estimée par les auteurs de l'articles :

$$\hat{U}_i(j) = -2.26 - 0.21|X_i - X_j| - 1.06|\xi_i - \xi_j| + 2.63D_{0ij} + 1.22F_{0,ij} + \epsilon_{ij}$$

  Les ξ_i ont été simulés indépendamment selon une loi de $Bernoulli(1/2)$, comme dans l'article. De plus, comme les résidus ϵ_i de cette équation suivent une loi logisitique, la probabilité de former un lien a pu être calculée de la façon suivante :

$$p(U_i(j) > 0, U_j(i) > 0|X, D) = p_{ij}p_{ji} = p_{ij}^2 \text{ par symétrie et où : } p_{ij} = \frac{\exp(\hat{U}_i(j))}{1 + \exp(\hat{U}_i(j))}$$

  Ensuite, grâce à cette probabilité, un lien d'amitié entre chaque individu de la base a été tiré selon une loi de Bernoulli de paramètre p_{ij}^2.

	Moyenne	**Ecart-type**	**Min.**	**Max.**
Note période 1 (X)	2.53	0.75	0.3	4
Note période 2 (Y)	2.41	0.85	-0.34	4.61
Nombre d'amis période 1	6.01	0.47	10	14
Nombre d'amis période 2	4.52	2.31	2	17

TABLE 1 – Descriptif de la population simulée

		Période 1	
		Pas ami	**Ami**
Période 2	**Pas ami**	135749	1522
	Ami	733	1652

TABLE 2 – Dynamique de la population simulée

- Y a aussi été simulé selon les résultats de l'article pour le modèle endogène (voir partie 2.4). L'équation estimée utilisée est la suivante :

$$\hat{Y}_i = -0.10(I-0.15G)^{-1}+(I-0.15G)^{-1}0.73X_i+(I-0.15G)^{-1}0.11\bar{X}_{(i)}+(I-0.15G)^{-1}(-0.01)\xi_i+\epsilon_i$$

$$\text{avec } \epsilon_i \sim \mathbb{N}(0, 0.61)$$

Nous remarquons que les notes de $\mathbf{Y}$ peuvent être en dehors de l'intervalle $[0, 4]$. Afin de ne pas biaiser les résultats, les notes $\mathbf{Y}$ seront laissées telles quelles. De plus, le réseau d'amitié ainsi construit a une plus grande tendance à créer des liens d'amitié. Toutefois, ces différences par rapport aux données de la base *Add Health* ne sont pas rédhibitoires : les données simulées restent exploitables et la reproduction des résultats sera une preuve de la bonne méthodologie de l'article.

4.2 Reproduction des résultats

Dans cette partie, les deux modèles de l'article de Goldsmith-Pinkham et Imbens vont être reproduits sur les données simulées et présentées précédemment.

4.2.1 Le modèle de Manski "linéaire en moyenne"

L'objectif est d'estimer par une méthode bayésienne les différents paramètres du modèle de Manski. Rappelons tout d'abord l'équation définissant le modèle de Manski présenté par les auteurs :

$$\mathbf{Y} = \beta_0 \mathbf{t}_N + \beta_x \mathbf{X} + \beta_{\overline{y}} \mathbf{GY} + \beta_{\overline{x}} \mathbf{GX} + \eta$$

On peut réécrire cette équation de la façon suivante :

$$\mathbf{Y} = \beta \mathbf{V} + \eta \text{ avec } \beta = (\beta_0, \beta_x, \beta_{\overline{y}}, \beta_{\overline{x}})' \text{ et } \mathbf{V} = (\mathbf{t}_N, \mathbf{X}, \mathbf{GY}, \mathbf{GX})$$

Comme on sait que $\eta \sim \mathbb{N}(0, \sigma^2 I_N)$, on peut en déduire que $Y|\mathbf{V}, \beta, \sigma^2 \sim \mathbb{N}(\beta\mathbf{V}, \sigma^2 I_N)$.

Pour pouvoir estimer les paramètres du modèle, des lois a priori ont dû être choisies. Ainsi, nous avons supposé que $\beta|\sigma^2 \sim \mathbb{N}(\mu_0, \sigma^2\Sigma_0^{-1})$ et que σ^2 suit une loi du χ^2 inverse. Pour les paramètres des lois a priori des β, nous choisissons les paramètres estimés pour les données de l'article (voir Table 3). Aucune information n'est donnée à propos de l'estimation de σ^2, nous choisissons donc une loi du χ^2 inverse de degré de liberté $v = 4$ (comme le nombre de variables à estimer, sans le σ^2). On a donc les lois a priori suivantes :

$$\pi(\beta|\sigma^2, \mathbf{V}, \mathbf{Y}) \propto (\sigma^2)^{-4/2} \exp(-\frac{1}{2\sigma^2}(\beta - \mu_0)'\Sigma_0(\beta - \mu_0))$$

	Estimation de l'article		**Estimation selon nos données**	
	Moyenne	Ecart-type	Moyenne	Ecart-Type
β_0	-0.13	0.12	-0.13	0.02
β_x	0.74	0.04	0.78	0.06
$\beta_{\overline{y}}$	0.16	0.05	0.06	0.08
$\beta_{\overline{x}}$	0.11	0.07	0.04	0.04

TABLE 3 – Estimation bayésienne du premier modèle

$$\text{et } \pi(\sigma^2|\mathbf{V},\mathbf{Y}) \propto (\sigma^2)^{\frac{v}{2}-1}\exp(-\frac{1}{2x})$$

L'objectif est maintenant de déterminer les lois a posteriori de nos deux paramètres, on a :

$$\begin{aligned}
\pi(\beta,\sigma^2|\mathbf{Y},\mathbf{V}) &\propto \pi(\mathbf{Y}|\mathbf{V},\beta,\sigma^2)\pi(\beta|\sigma^2)\pi(\sigma^2)\\
&\propto \sigma^{-n/2}\exp(-\frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{V}\beta)'(\mathbf{Y}-\mathbf{V}\beta))(\sigma^2)^{-4/2}\\
&\qquad \exp(-\frac{1}{2\sigma^2}(\beta-\mu_0)'\Sigma_0(\beta-\mu_0))(\sigma^2)^{\frac{-v}{2}-1}\exp(-\frac{1}{-1\sigma^2})\\
\pi(\beta,\sigma^2|\mathbf{Y},\mathbf{V}) &\propto (\sigma^2)^{-4/2}\exp(-\frac{1}{2\sigma^2}(\beta-\mu_n)'(\mathbf{V'V}+\Sigma_0)(\beta-\mu_n))\\
&\qquad (\sigma^2)^{-\frac{v+n}{2}-1}\exp(-\frac{1+\mathbf{Y'Y}-\mu_n'(\mathbf{V'V}+\Sigma_0)\mu_n+\mu_0'\Sigma_0\mu_0}{2\sigma^2})
\end{aligned}$$

avec $\mu_n = (\mathbf{V'V}+\Sigma_0)^{-1}(\mathbf{V'Y}+\Sigma_0\mu_0)$.

Des calculs ci-dessus, nous constatons que nous avons :

$$\pi(\beta,\sigma^2|\mathbf{Y},\mathbf{V}) \propto \pi(\beta|\sigma^2,\mathbf{Y},\mathbf{V})\pi(\sigma^2|\mathbf{Y},\mathbf{V})$$

$$\text{avec : } \pi(\beta|\sigma^2,\mathbf{Y},\mathbf{V}) \propto (\sigma^2)^{-4/2}\exp(-\frac{1}{2\sigma^2}(\beta-\mu_n)'(\mathbf{X'X}+\Sigma_0)(\beta-\mu_n))$$

$$\text{et : } \pi(\sigma^2|\mathbf{Y},\mathbf{V}) \propto (\sigma^2)^{-\frac{v+n}{2}-1}\exp(-\frac{1+\mathbf{Y'Y}-\mu_n'(\mathbf{X'X}+\Sigma_0)\mu_n+\mu_0'\Sigma_0\mu_0}{2\sigma^2})$$

Cette loi n'est pas usuelle et des simulations MCMC ont dû être effectuées pour simuler les paramètres. Nous avons utilisé l'algorithme de Metropolis-Hastings afin d'évaluer ces différents paramètres. Voici l'algorithme utilisé (un exemple d'algorithme de Metropolis-Hastings est présenté pour le modèle suivant).

Les résultats des estimations pour nos données sont lisibles en Table 3 (l'algorithme a été lancé pour 100 000 itérations dont les 10 000 premières ont été retirées pour les estimations pour des soucis de convergence). Nous observons de plus que les calculs de σ^2 donnent une moyenne empirique de 0.44 et écart-type de 0.03. La convergence des beta peut être observée en Figure 4.

On constate que les résultats sont très similaires entre les deux modèles pour les deux paramètres β_0 et β_x mais que les résultats diffèrent beaucoup plus pour $\beta_{\overline{y}}$ et $\beta_{\overline{x}}$. L'écart-type des estimations est d'ailleurs aussi basse dans les deux estimations.

La proximité des deux estimations pour les deux premiers paramètres montrent que les données ont bien été simulées mais interpréter la différence d'estimation entre les deux autres paramètres ? Rappelons-nous que les données ont été estimées avec une variable ξ qui est omise dans notre équation. Rappelons-nous de plus que cette variable ξ est aussi introduite dans la

construction du réseau. Ce problème de variable omise peut justifier la différence dans l'estimation et peut justifier le fait que cette différence n'apparaisse que dans les paramètres liés au réseau en période 2.

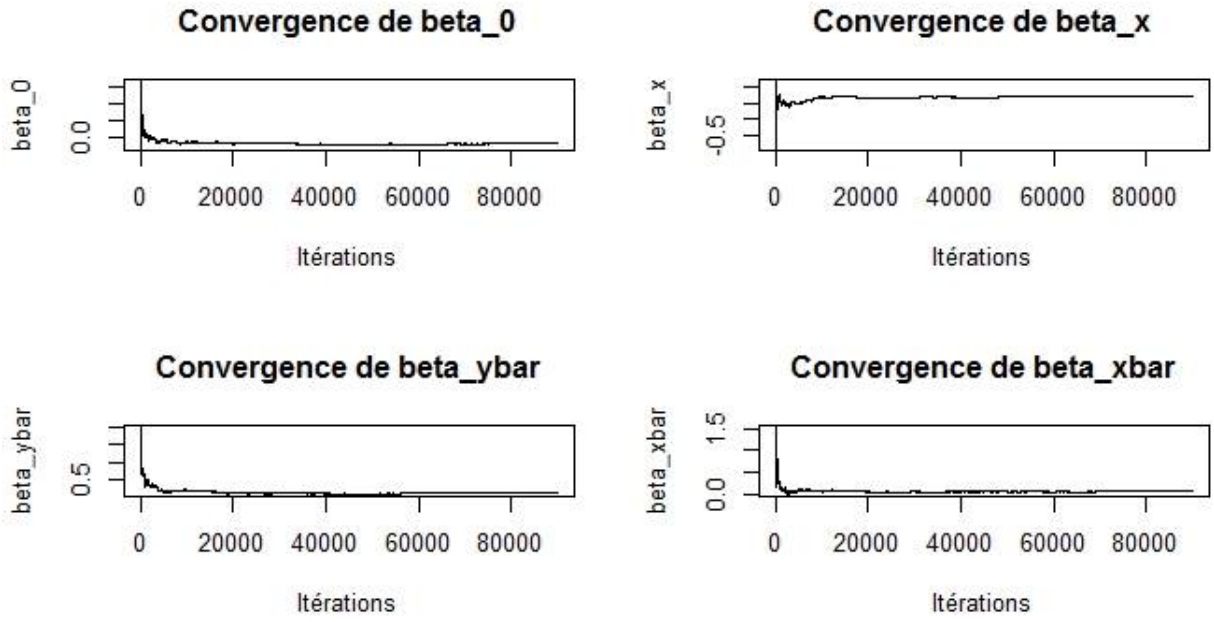

FIGURE 4 – Convergence de la moyenne des paramètres simulés

4.2.2 Modèle de formation de réseau exogène

Nous allons maintenant estimer les paramètres du second modèle, le modèle de formation de réseau exogène. Rappelons brièvement le modèle. Nous souhaitons expliquer la formation de lien entre les individus. Ce lien peut s'écrire de la manière suivante :

$$D_{ij} = \mathbf{1}_{U_i(j)>0}\mathbf{1}_{U_j(i)>0}$$

$U_i(j)$ est l'utilité pour l'individu i d'être ami avec j, ainsi un lien se crée si l'individu des deux individus est strictement positive. $U_i(j)$ prend la forme suivante :

$$U_i(j) = \alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij} + \epsilon_{ij}$$

On a de plus ϵ_{ij} iid et de loi logistique.

Notre objectif est d'estimer les paramètres α par des méthodes bayésiennes, il va donc falloir choisir une loi a priori pour ces paramètres pour leur trouver une loi a posteriori grâce à cette loi a priori et à la vraisemblance du modèle. Cherchons tout d'abord la vraisemblance du modèle. On a, tout d'abord, $\forall i, j$, $D_{ij} \sim Bernoulli(\mathbb{P}(U_i(j) > 0, U_j(i) > 0))$ et on a :

$$\begin{aligned}
\mathbb{P}(U_i(j) > 0) &= \mathbb{P}(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij} + \epsilon_{ij} > 0) \\
&= \mathbb{P}(\epsilon_{ij} > -(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})) \\
&= 1 - F(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij}) \text{ avec fonction de répartition d'une loi logistique} \\
&\qquad\qquad \text{car } \epsilon_{i,j} \text{ suit une loi logistique} \\
&= \frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}
\end{aligned}$$

Par indépendance des ϵ_{ij}, on constate de plus que $\mathbb{P}(U_i(j) > 0) = \mathbb{P}(U_j(i) > 0)$: On peut donc en déduire la loi d'un D_{ij} :

$$\pi(D_{ij}|X_i, X_j, F_0, D_{0,ij}) \propto \left(\frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}\right)^{2D_{ij}} \times \left(1 - (\frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})})^2\right)^{1-D_{ij}}$$

La vraisemblance du modèle est donc, par indépendance des ϵ_{ij} :

$$vraisemblance \propto \prod_{i \neq j} \left(\frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}\right)^{D_{ij}} \times \left(1 - \frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}\right)^{1-D_{ij}}$$

Comme dans l'article de Goldsmith-Pinkham et Imbens, nous choisissons des a priori gaussiennes centrées réduites pour chacun des paramètres α, on en déduit donc la loi a posteriori suivante (on note $\alpha = (\alpha_0, \alpha_x, \alpha_f, \alpha_d)$) :

$$\begin{aligned}\pi(\alpha|\text{modèle}) &\propto \pi(\alpha) \times \text{ vraisemblance} \\ &\propto \prod_{k=(0,x,f,d)} \exp(-\frac{\alpha_k^2}{2}) \prod_{i \neq j} \left(\frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}\right)^{D_{ij}} \times \\ &\left(1 - \frac{\exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}{1 + \exp(\alpha_0 + \alpha_x|X_i - X_j| + \alpha_f F_{0,ij} + \alpha_d D_{0ij})}\right)^{1-D_{ij}}\end{aligned}$$

On ne peut déduire une distribution explicite de la forme de cette distribution a posteriori, mais les paramètres peuvent tout de même être estimés grâce à des méthodes MCMC. Nous avons utilisé l'algorithme de Metropolis-Hastings afin d'évaluer ces différents paramètres. Un code R a été écrit et voici l'algorithme construit :

1. On initialise les α_0 par les valeurs estimées par l'article (voir Table 4) : les valeurs des paramètres sont censées en être proches même si les réseaux sont différents entre l'article et cette application,
2. On réitère pour $j = 1, ..., n$ la procédure suivante :
   (a) On simule $\alpha_c \sim \mathbb{N}(\alpha_{j-1}, \mathbf{I}_4)$
   (b) On calcule le ratio $R = \frac{f(\alpha_c)g(\alpha_{j-1}|\alpha_c)}{f(\alpha_{j-1})g(\alpha_c|\alpha_{j-1})}$ avec f la densité de la loi a posteriori des α (à une constante près) et $g(.|\alpha)$ la densité d'une loi $\mathbb{N}(\alpha, \mathbf{I}_4)$.
   (c) On simule $B \sim Bernoulli(\min(R, 1))$ et si B=1, on choisit $\alpha_j = \alpha_c$. Sinon, on choisit $\alpha_j = \alpha_{j-1}$

L'algorithme a été effectué sur 5100 itérations, les 100 premières itérations ont été retirées afin de permettre la convergence. Les résultats peuvent être lus en Table 4.

On observe que les résultats sont relativement différents entre les deux estimations. Cela peut s'expliquer par le fait que l'algorithme n'ait pas totalement convergé dans notre cas (on peut le constater en Figure 5 : les temps de calculs sont très longs et plus d'itérations n'ont pas pu être effectuées. Cependant, les signes des paramètres sont cohérents et, outre un possible problème de convergence, la différence d'estimation peut venir d'un problème de variables omises puisque les données ont été simulées à partir d'une équation introduisant un paramètre ξ qui n'est pas dans prise en compte dans l'équation utilisée dans ce modèle.

	Estimation de l'article		**Estimation selon nos données**	
	Moyenne	Ecart-type	Moyenne	Ecart-Type
α_0	-2.56	0.04	-1.31	0.22
α_x	-0.20	0.03	-0.06	0.04
α_f	2.52	0.05	1.18	0.12
α_d	1.20	0.04	0.53	0.06

TABLE 4 – Estimation bayésienne du second modèle

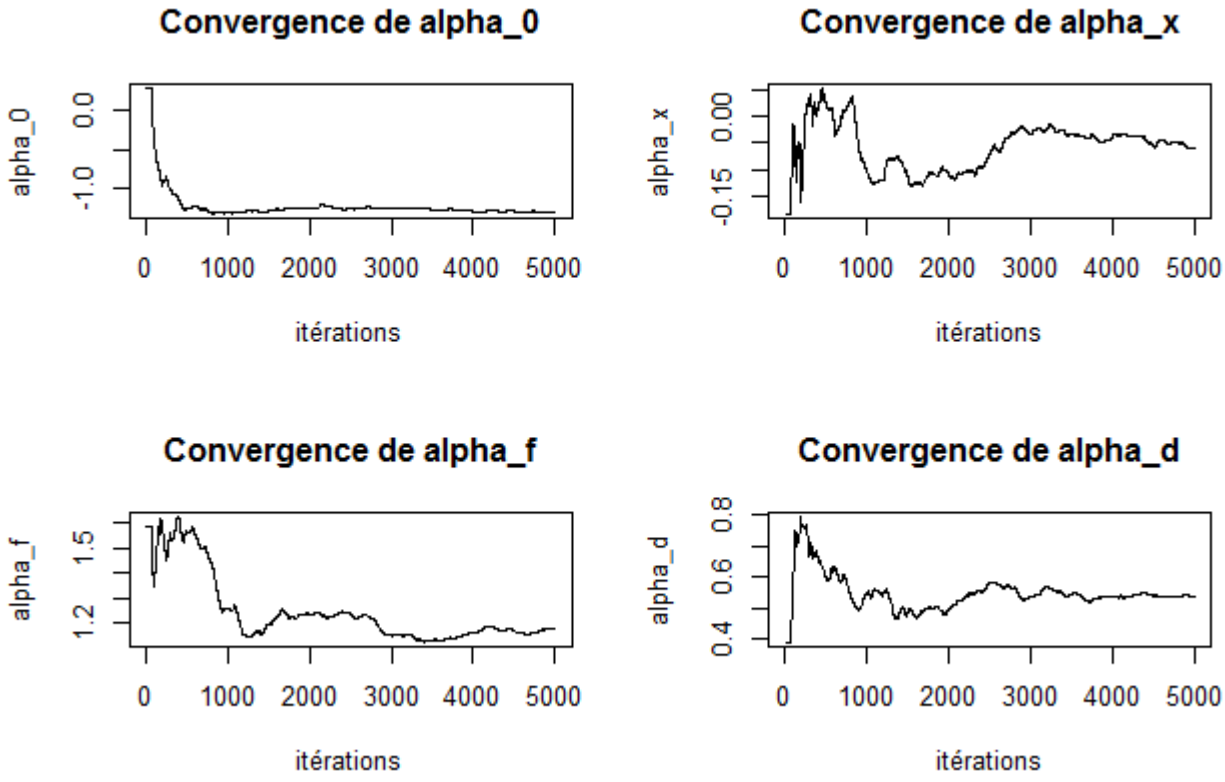

FIGURE 5 – Convergence de la moyenne des paramètres simulés

4.2.3 Modèle de formation de réseau endogène

Le troisième modèle est plus complexe puisqu'on suppose la présence d'endogénéité dans l'équation de formation du réseau et dans l'équation du modèle de Manski. Rappelons que les modèles réécrits pour prendre en compte l'endogénéité grâce à une variable $\xi_i \sim Bernoulli(\frac{1}{2})$ sont les suivants :

$$Y_i = \beta_0 + \beta_x X_i + \beta_{\bar{y}} \bar{Y}_{(i)} + \beta_{\bar{x}} \bar{X}_{(i)} + \beta_\xi \xi_i + \eta_i$$

$$U_i(j) = \alpha_0 + \alpha_x |X_i - X_j| + \alpha_\xi |\xi_i - \xi_j| + \alpha_d D_{0ij} + \alpha_F F_{0,ij} + \epsilon_{ij}$$

Bien que nous ayons simulé les données nous-même, il faut supposer les ξ_i inconnus pour comprendre la méthode d'estimation proposée par les auteurs. Les méthodes proposées sont à nouveau fondées sur des méthodes bayésiennes. Des distributions a priori sont proposées pour chacun des paramètres du modèle (les α, les β et σ^2) et les choix sont les mêmes que pour les deux modèles précédents. Le choix de lois gaussiennes permet de connaître explicitement la loi à

posteriori : la loi gaussienne est conjuguée à elle-même. Cependant, le calcul de la vraisemblance du modèle est plus complexe que précédemment mais le calcul est explicité dans l'article et le résultat se présente sous la forme du produit de la vraisemblance des deux modèles précédents, en introduisant le nouveau terme ξ_i :

$$L(\beta, \alpha, \sigma^2, p|Y, G, X, D_0, \xi) = L_{note}(\beta, \sigma^2|Y, D, X, \xi) L_{reseau}(\alpha|G, \xi, X, D_0)$$

Il faut aussi prendre en compte le fait que ξ est inconnu dans notre modèle et il faut donc intégrer la vraisemblance précédente sur ξ pour pouvoir procéder à l'estimation :

$$\begin{aligned} L(\beta, \alpha, \sigma^2, p|Y, G, X, D_0) &= \sum_{\xi} L(\beta, \alpha, \sigma^2, p|Y, G, X, D_0, \xi) \mathbb{P}(\xi|\beta, \alpha, \sigma^2, p, Y, G, X, D_0) \\ &= \sum_{\xi} L_{note}(\beta, \sigma^2|Y, D, X, \xi) L_{reseau}(\alpha|G, \xi, X, D_0) \times \\ &\quad \mathbb{P}(\xi|\beta, \alpha, \sigma^2, p, Y, G, X, D_0) \end{aligned}$$

Cette formule montre que la variable ξ intervient à la fois dans la création du réseau et dans l'évolution des notes. C'est pourquoi, lors de l'estimation des paramètres, les ξ devront être estimés dans un premier temps. Nous pouvons noter que ces variables latentes auraient pû être estimées à travers l'algorithme Expectation-Maximation.

Algorithm 1 Estimation des paramètres dans le modèle endogène

1. **Initialisation**
Un prior est utilisé pour les 11 variables :
— α_0, α_x, α_d et α_f ont un prior une loi $\mathcal{N}(0, 1)$;
— la variable α_ξ a un prior valant $\mathcal{N}(-1, 0.01)$;
— β_0, β_x, β_d et β_f ont un prior une loi $\mathcal{N}(0, 1)$;
— la variable β_ξ a un prior valant $\mathcal{N}(0, 0.01)$;
— la variable σ^2 suit une inverse chi-deux de 10 degrés de liberté
2. **Mise à jour des** ξ Pour une convergence plus rapide, les auteurs conseillent de mettre à jour ces variables binaires indivi par individu à partir des variables tirées précédemment. La loi a posteriori obtenue est $\mathbb{P}(\xi_i|\mathbf{X}, \mathbf{D}_0, \mathbf{Y}, \mathbf{D}) \propto Bernoulli(\frac{1}{2}) \times \mathcal{L}(\xi_i|\mathbf{X}, \mathbf{D}_0, \mathbf{Y}, \mathbf{D})$
3. **Mise à jour des paramètres** β avec l'algorithme de Metropolis-Hastings à travers une candidate gaussienne centrée sur la position actuelle ; la loi a posteriori obtenue est $\mathbb{P}(\beta_i|\mathbf{X}, \xi, \mathbf{Y}, \mathbf{D}) \propto \mathcal{N}(0, \text{diag}(1, 1, 1, 1, 0.01)) \times \mathcal{L}(\beta_i|\mathbf{X}, \mathbf{D}, \mathbf{Y}, \xi)$
4. **Mise à jour de** $\sigma 2$ avec l'algorithme de Gibbs sampler dont la loi a posteriori est $\mathbb{P}(\sigma^2|\mathbf{X}, \mathbf{D}, \mathbf{Y}, \xi) \propto \chi^2(10)^{-1} \times \mathcal{L}(\sigma^2|\mathbf{X}, \mathbf{D}, \mathbf{Y}, \xi)$
5. **Mise à jour des paramètres** α avec l'algorithme de Metropolis Hastings, avec la loi a posteriori $\mathbb{P}(\alpha|\mathbf{X}, \mathbf{D}_0, \mathbf{D}, \xi) \propto \mathcal{N}((0, 0, 0, 0, -1), \text{diag}(1, 1, 1, 1, 0.01)) \times \mathcal{L}(\alpha|\mathbf{X}, \mathbf{D}_0, \mathbf{D}, \xi)$
6. **Mise à jour des** ξ

L'esimation successive des paramètres dans le modèle endogène permet de déterminer en premier lieu les variables cachées. Celles-ci impactant à la fois la formation du réseau et les notes à la seconde période, il est nécessaire de les estimer avant les autres paramètres. Ensuite, ce sont les paramètres des notes qui sont déterminés puis celles du réseau. En dernier lieu, les variables cachées sont de nouveau ré-estimées (voir Algorithme 1). Malgré plusieurs tentatives, nous n'avons pas réussi à estimer les paramètres du modèle.

5 Conclusion

L'article de Goldsmith-Pinkham et Imbens présenté ici a donc permis de soulever une conclusion majeure, et de nombreuses interrogations : l'effet de pair au sein des réseaux sociaux est

un domaine d'étude très complexe et pour plusieurs raisons. La première raison est que les réseaux et les variables sur lesquelles on étudie l'effet de pair sont toujours intimement liés. Ainsi, dans notre cas, des variables omises déterminent ces deux dimensions simultanément et de l'endogénéité est ainsi présente. Les auteurs ont ainsi proposé un modèle pour prendre en compte ce phénomène. Ce modèle proposé a soulevé une deuxième raison pour la complexité des effets de pairs : l'estimation des paramètres n'est pas toujours aisée. Les modèles de cet article ne sont en effet pas des modèles classiques et des méthodes d'estimation bayésienne ont dû être mises en place. Ces méthodes bayésiennes ont dû être invoquées parce qu'elles sont souples et permettent de s'adapter à des modèles complexes comme ceux de l'article, en particulier le dernier qui incorporait un nouveau terme aléatoire, afin de capter l'endogénéité.
Cette présentation de l'article a été accompagnée d'une implémentation des différents modèles pour des données simulées selon les équations estimées par l'article. Cette implémentation a permis de présenter des algorithmes de Metropolis-Hastings pour estimer les différents paramètres des modèles (et la difficulté de construire ces algorithmes). Des résultats assez proches aux résultats de l'article ont été démontrés, notamment la difficulté à estimer l'effet de pair dû à la présence d'endogénéité.

Enfin, une méthode alternative permettant de comparer les résultats obtenus par l'approche bayésienne pourrait être proposé, l'utilisation des algorithmes d'approximation développés en théorie des graphes. On pourrait ainsi obtenir une approximation de la vraisemblance du modèle.

Bibliographie

Bramoulle, Y., Djebbari, H., and Fortin, B. (2009), "Identification of Peer Effects Through Social Networks," Journal of Econometrics, 150, 41–55.

Chandrasekhar, A., and Jackson, M. (2012), "Tractable and Consistent Random Graph Models," SSRN Working Paper.

Currarini, S., Jackson, M., and Pin, P. (2010), "Identifying the Roles of Choice and Chance in Network Formation : Racial Biases in High School Friendships," Proceedings of the National Academy of Sciences, 107, 4857– 4861.

Kline,B & Tamer,E (2013) Comment, Journal of Business & Economic Statistics, 31 :3, 276-279, DOI : 10.1080/07350015.2013.792264

Ladd,G. (1983) Social Networks of Popular, Average, and Rejected Children in School Settings Merrill-Palmer Quarterly, Vol. 29, No. 3, Invitational Issue : Popular, Rejected, and Neglected Children : Their Social Behavior and Social Reasoning , pp. 283-307

Matthew O. Jackson (2013) Comment, Journal of Business & Economic Statistics, 31 :3, 270-273, DOI : 10.1080/07350015.2013.794095

Manski, C. (1993), "Identification of Endogenous Social Effects : The Reflection Problem," Review of Economic Studies, 60, 531–542.