Université de Montréal

Interactions Sociales et Réseaux Sociaux en Économique : Modélisation et Estimation

par
Vincent Boucher

Département de Sciences Économiques
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en sciences économiques

mars, 2013

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

Interactions Sociales et Réseaux Sociaux en Économique : Modélisation et Estimation

présentée par:

Vincent Boucher

a été évaluée par un jury composé des personnes suivantes:

Yves Sprumont	président-rapporteur
Marc Henry	directeur de recherche
Onur Özgür	codirecteur
Yann Bramoullé	membre du jury
Brian Krauth	examinateur externe
Éric Lacourse	Faculté des arts et sciences

RÉSUMÉ

Cette thèse comporte trois essais sur les interactions sociales en sciences économiques. Ces essais s'intéressent à la fois au côté théorique qu'empirique des interactions sociales. Le premier essai (chapitre 2) se concentre sur l'étude (théorique et empirique) de la formation de réseaux sociaux au sein de petites économies lorsque les individus ont des préférences homophilique et une contrainte de temps. Le deuxième essai (chapitre 3) se concentre sur l'étude (principalement empirique) de la formation de réseau sociaux au sein de larges économies où les comportement d'individus très distants sont aproximativement indépendants. Le dernier essai (chapitre 4) est une étude empirique des effets de pairs en éducation au sein des écoles secondaires du Québec. La méthode structurelle utilisée permet l'identification et l'estimation de l'effet de pairs endogène et des effets de pairs exogènes, tout en contrôlant pour la présence de chocs communs.

Mots Clés : Interactions Sociales, Réseaux Sociaux, Formation de Réseaux, Effets de Pairs

ABSTRACT

This thesis includes three essays on social interactions in economics, both from a theoretical and applied perspective. The first essay (chapter 2) focusses on the (theoretical and empirical) study of a network formation process in small economies characterized by the fact that individuals have homophilic preferences and a time constraint. The second essay (chapter 3) is focussed on the study (mostly empirical) of a network formation process in large economies characterized by the fact that distant individuals have approximately independent behaviours. The last essay (chapter 4) is an empirical study of peer effects in education for Quebec high-school teenagers. The structural method used allows for the identification and estimation of the endogenous peer effect and the exogenous peer effects, while controlling for the presence of common shocks.

Keywords: Social Interactions, Social Networks, Network Formation, Peer Effects

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

What you do in this world is a matter of no consequence.
The question is what can you make people believe you have done.

- Arthur Conan Doyle, A Study in Scarlet

REMERCIEMENTS

"Not everything that can be counted counts,
and not everything that counts can be counted."
- Albert Einstein

Il m'est impossible de présenter ici une liste exhaustive des gens qui m'ont supporté durant ces dernières années, ni d'énumérer l'ensemble des raisons pour lesquelles je leur serai éternellement reconaissant. De même, je sais que ces remerciements sont loin de compenser l'étendue de leur support tout au long de mes études doctorales.

Tout d'abord, je voudrais remercier mes directeurs de recherche (en ordre alphabétique) Yann Bramoullé, Marc Henry et Onur Özgür. Merci de m'avoir aidé à y voir clair. Merci de m'avoir permis de prendre du recul sur mes recherches, ce qui a beaucoup contribué à diminuer mon niveau de stress ! Merci surtout pour votre temps et votre confiance.

Je voudrais aussi remercier tout spécialement, Lars Ehlers pour l'appui qu'il m'a apporté. Merci beaucoup pour ces nombreuses oppportunités ! !

Merci à mes coauteurs pour les articles présents dans cette thèse : Ismael Y. Mourifié (chapitre 3), ainsi que Yann Bramoullé, Habiba Djebbari et Bernard Fortin (chapitre 4) . Merci pour votre implication dans ces projets. J'ai appris beaucoup.

Merci aussi aux professeurs du département de sciences économiques de l'université de Montréal. Merci pour vos enseignements, conseils, et les nombreux : "Est-ce que tu as juste 5 minutes pour une question rapide ?" Merci aussi à tout le personnel non-enseignant du département pour votre apport indispensable.

Merci aussi à mes collègues étudiants à l'université de Montréal et ailleurs pour les nombreuses discussions (pertinentes ou non), notamment, Louis-Philippe Beland et Youcef Msaïd, ainsi que mes collègues de bureau Maxime Agbo, Catherine Gendron-Saulnier, Anabelle Maher et Ismael Y. Mourifié.

Merci à Sanjeev Goyal ainsi qu'au personnel de la Faculté d'économique de l'Université de Cambridge et au Cambridge-INET pour leur accueil durant ma

dernière année doctorale.

Finalement, la vie professionnelle n'est rien sans une vie personnelle équilibrée. Merci beaucoup à vous, parents et amis, qui avez été là pour me faire prendre le recul nécessaire. Un merci particulier à Isabelle pour avoir été à mes cotés ces dernières années.

Je n'aurais jamais pu mettre l'effort nécessaire à la réalisation de cette thèse sans l'appui financier des organismes suivants : le Fond de Recherche Société et Culture (FRSC), le Conseil de Recherche en Sciences Humaines du Canada (CRSH), le Centre Interuniversitaire de Recherche en Économie Quantative (CIREQ) et le Cambridge-INET.

À vous tous et à ceux que j'ai oublié, un immense merci !

Specific Acknowledgements

Structural Homophily : I would like to thank Paolo Pin for his helpful comments. I would also like to thank Ismael Mourifie, Louis-Philippe Béland, Yousef Msaid and David Karp, as well as the participants at various conferences, including those of the Canadian Economic Association (2011), Coalition Theory Network (2010), Societe Canadienne de Science Economique (2010), Econcon (2011), and Groupe de Recherche International (2011) for their questions, comments and suggestions. Finally, I gratefully acknowledge financial support from the CIREQ, the FQRSC and the SSHRC.

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

My Friend Far Far Away : A significant part of this work has been done while Ismael Mourifié was visiting the economic department at PennState. This paper benefited from many highly interesting discussion. We would especially want to

thank Tim Conley, Sílvia Gonçalves, Bryan S. Graham, Paul Grieco, Joris Pinkse, Benoit Perron and Ilze Kalnina for their comments and suggestions. A special thanks also to Yann Bramoullé and Marc Henry for their helpful comments and suggestions. We would also like to thank the participants to many conferences and seminars, including : the SCSE2012, the economic department at PennState, the economic department at U. de Montreal, and the workshop CeMMAP-CIREQ-SciencesPo-X on "The Estimation of Complementaries in Matching and Social Networks". This paper uses data from Yahoo !. We would like to thank Yahoo staff, especially Kim Capps-Tanaka, for their help. Thanks also to Sinan Aral and Lev Muchnik for their help with the database.

Do Peers Affect Student Achievement ? : We thank Lung-fei Lee, seminar participants at University of Toronto, Northwestern University, Université de Paris 1, and the CIRANO-CIREQ Conference on the Econometrics of interactions for insightful discussions, and three anonymous referees and the co-editor Thierry Magnac for very helpful comments. We are also grateful to the Québec Ministry of Education, Recreation and Sports (MERS) for providing the data, in particular Raymond Ouellette and Jeannette Ratté for their assistance in obtaining and interpreting the data used in this study. The views expressed in this paper are solely our own and do not necessarily reflect the opinions of the MERS. We receive excellent research assistance from Steeve Marchand. Support for this work has been provided by the Canada Chair of Research in Economics of Social Policies and Human Resources, and le Fonds Québécois de Recherche sur la Socité et la Culture and le Centre Interuniversitaire sur le Risque, les Politiques économiques et l'Emploi.

AVANT-PROPOS

"The temptation to form premature theories upon insufficient data is the bane of our profession."
-Sherlock Holmes

Le manque de données de qualité est un des principaux problèmes de l'étude des intéractions sociales. La difficulté d'interpréter correctement les données disponibles en est un autre. Les chapitres qui suivent tentent surmonter certains défis imposés par ces problèmes. Il reste néanmoins beaucoup de travail à faire. Je reste persuadé qu'une meilleure compréhension des phénomènes économiques sous-jacents est la clé vers la résolution de nombreux problèmes d'origine statistique.

CHAPITRE 1

INTRODUCTION

Ces dernières années, la litérature théorique et empirique sur les interactions sociales et les réseaux sociaux en économique a pris beaucoup d'importance. Dans cette thèse, je m'intéresse à deux branches de cette litérature : l'estimation des effets de pairs, et l'estimation des processus de formation de réseaux sociaux.

Afin de faciliter la discussion, j'introduis les trois objets suivants : $(\mathbf{y}, \mathbf{X}, \mathbf{G})$, pour une population composée de n individus. Le vecteur $\mathbf{y}$, de taille $n \times 1$ représente une variable endogène potentiellement sujette à des effets de pairs (e.g. habitudes alimentaires, résultats scolaires...). La matrice $\mathbf{X}$, de taille $n \times k$ représente une série de k variables exogènes (e.g. age, genre, revenu, groupe racial...). Finalement, la matrice $\mathbf{G}$, de taille $n \times n$ est une matrice (symétrique) d'adjacence donnant la structure d'un réseau social. Si deux individus i et j sont liés, alors $\mathbf{G}_{ij} = 1$. Dans le cas contraire, $\mathbf{G}_{ij} = 0$. L'estimation des effets de pairs se concentre principalement sur l'étude et l'estimation de $\mathbb{P}(\mathbf{y}|\mathbf{X}, \mathbf{G})$ alors que l'estimation des processus de formation de réseaux sociaux s'intéresse principalement à $\mathbb{P}(\mathbf{G}|\mathbf{X})$.[1]

Effets de Pairs

Une des problématiques avec l'estimation de $\mathbb{P}(\mathbf{y}|\mathbf{X}, \mathbf{G})$ est d'identifier les effets de pairs des potentiels chocs communs. Supposons que la population de taille n peut être réparée en différents groupes indépendants et considérons la forme structurelle suivante (issue de Bramoullé et al., 2009) pour le groupe r :

$$\mathbf{y}_r = \alpha_r + \mathbf{X}_r\beta + \tilde{\mathbf{G}}_r\mathbf{X}_r\gamma + \lambda\tilde{\mathbf{G}}_r\mathbf{y}_r + \varepsilon_r$$

où $\tilde{\mathbf{G}}_r$ est construit à partir de $\mathbf{G}_r$ en normalisant la somme de chaque ligne à 1. Une méthode classique de contrôler pour l'effet fixe α_r et de contourner le problème

[1] où $\mathbb{P}$ représente la densité.

causé par la présence du paramètre fortuit α_r ("incidental parameter problem", Newman et Scott, 1948) est de réécrire le modèle en déviations.

$$(\mathbf{I}-\mathbf{M}_r)\mathbf{y}_r = (\mathbf{I}-\mathbf{M}_r)\mathbf{X}_r\beta+\lambda(\mathbf{I}-\mathbf{M}_r)\tilde{\mathbf{G}}_r\mathbf{y}_r+(\mathbf{I}-\mathbf{M}_r)\tilde{\mathbf{G}}_r\mathbf{X}_r\gamma+(\mathbf{I}-\mathbf{M}_r)\varepsilon_r \quad (1.1)$$

où $\mathbf{M}_r$ est une matrice carrée symétrique telle que la somme de chaque ligne est normalisée à 1.

Dans ce cas, l'identification des paramètres β, λ, γ est loin d'être triviale. De plus, l'estimation est compliquée par la présence de la variable endogène $\tilde{\mathbf{G}}_r\mathbf{y}_r$ ("reflection problem", Manski, 2003). Pour un réseau arbitraire, Bramoullé et al. (2009) montent que l'identification est possible si les matrices $\mathbf{I}, \tilde{\mathbf{G}}, \tilde{\mathbf{G}}^2$ sont linéairement indépendantes. Dans le cas spécifique où le réseau représente des interactions en groupes (le réseau est complètement connecté pour chaque groupe r), Lee (2007) développe un estimateur par maximum de vraisemblance basé sur la forme réduite de (1.1). C'est cet estimateur qui est utilisé dans le chapitre 4.

Formation du Réseau

La problématique principale en ce qui a trait à l'étude de $\mathbb{P}(\mathbf{G}|\mathbf{X})$ est l'important degré de dépendance entre les observations. Par exemple, la probabilité que deux individus créent un lien d'amitié peut dépendre des liens que chacuns ont. Donc, en général, $\mathbb{P}(\mathbf{G}|\mathbf{X}) \neq \Pi_{ij:i<j}\mathbb{P}(\mathbf{G}_{ij}|\mathbf{X})$. Dans les chapitres suivants, je présente deux approches pour contourner ce problème. La première approche (chapitre 2) est basée sur l'observation de plusieurs petites sous-économies indépendantes. La deuxième (chapitre 3) est basée sur l'observation d'une large économie et sur un argument d'indépendence asymptotique.

Une autre problématique est de pouvoir modéliser $\mathbb{P}(\mathbf{G}|\mathbf{X})$ afin de répliquer les faits stylisés observés. J'utilise l'une des principales caractéristiques des réseaux sociaux : l'homophilie. L'homophilie caractérise le fait que deux individus liés sont en moyenne davantage similaires que deux individus non liés. Dans le chapitre 2, j'utilise une version forte de l'homophilie (l'homophilie structurelle) alors que dans

le chapitre 3, j'utilise une forme très faible de l'homophilie (l'homophilie asymptotique).

Trois Essais...

Cette thèse comprend trois essais (chapitres 2,3 et 4). Le premier essai intitulé *Structural Homophily* s'intéresse à la formation de réseaux d'amitié dans des écoles secondaires de taille relativement faibles, aux États-Unis. Une des caractéristiques de ces petites écoles est que les individus connaissent l'ensemble de leurs amis potentiels. Chaque individu a une contrainte de temps (inobservée) qui introduit une limite quant au nombre possible d'amis qu'il peut avoir. Lorsque les individus préfèrent créer des liens avec des individus qui leurs sont similaires, l'unique réseau d'équilibre possède une structure extrêmement particulière : l'homophilie structurelle. Cette structure particulière permet de définir une méthode d'estimation originale. Je développe un estimateur permettant d'estimer les poids relatifs de la contribution de différentes caractéristiques socioéconomiques au processus de formation du réseau. Cet estimateur est basé sur la maximization de la probabilité que le réseau observé soit caractérisé par l'homophilie structurelle. J'utilise une base de données américaine (AddHealth) contenant de l'information sur des adolescents de niveau secondaire ("high-school"). Pour chaque adolescent, j'observe son réseau social ainsi que son groupe racial. Je trouve que le niveau de ségrégation varie considérablement selon le groupe racial considéré, les noirs étant davantage ségrégués que les autres groupes raciaux.

Le deuxième essai est intitulé *My Friend Far Far Away : Asymptotic Properties of Pairwise Stable Networks*.[2] Comme pour l'essai précédent, nous nous intéressons à $\mathbb{P}(\mathbf{G}|\mathbf{X})$, mais maintenant, dans le cas où on observe l'unique réseau d'équilibre d'une très vaste économie. L'estimateur est basé sur la maximisation de la probabilité que le réseau d'équilibre soit Pairwise Stable (Jackson et Wolinsky, 1996). En utilisant une généralisation des modèles spatiaux de dépendence limitée, nous montrons que l'estimateur est convergent et asymptotiquement normalement distribué.

[2]Cooécrit avec Ismael Mourifié.

Un des grands avantages de cet estimateur est sa simplicité. L'estimation peut se faire à l'aide de commandes préprogrammées pour la majorité des logiciels statistiques. Nous présentons une application en utilisant une base de données fournie par Yahoo ! contenant de l'information sur l'utilisation de leur service de messagerie instantanée pour plus de 22 millions d'utilisateurs. Nous trouvons que la probabilité que deux individus interragissent par le service de messagerie instantanée est fonction de la densité locale du réseau, du degré d'utilisation d'Internet par les individus, ainsi que de la différence entre leurs caractéristiques socioéconomiques et le contenu des pages qu'ils ont visitées.

Le troisième essai est intitulé *Do Peers Affect Student Achievement ? Evidence from Canada Using Group Size Variation.*[3] Cet essai est différent des deux essais précédent puisqu'il utilise la forme du réseau (ici un réseau complètement connecté) afin d'estimer les effets de pairs, i.e. $\mathbb{P}(\mathbf{y}|\mathbf{G}, \mathbf{X})$. À l'aide d'une base de donnnée fournie par le Ministère de l'éducation et des loisirs et des sports du Québec, et de l'approche structurelle de Lee (2007), nous estimons l'effet de pair endogène et les effets de pairs exogènes, tout en contrôlant pour la présence de chocs communs pour quatre matières : Français, Histoire, Sciences et Mathématiques. Les effets endogènes sont positifs lorsqu'ils sont significatifs et de magnitude comparable avec ceux trouvés dans la litérature. Nous trouvons aussi quelques effets exogènes, dont un effet négatif de l'âge des pairs, qui est interprété comme un effet négatif de la présence de doubleurs au sein d'un groupe.

[3]Cooécrit avec Yann Bramoullé, Habiba Djebbari et Bernard Fortin.

CHAPITRE 2

STRUCTURAL HOMOPHILY

2.1 Introduction

The fact that similar individuals tend to interact with each other is a prominent feature of social networks. The phenomenon, referred to as *homophily*, is increasingly being studied by economists.[1] Indeed, the structure of the social networks in which individuals interact has been shown to significantly influence many social outcomes such as segregation,[2] information transmission and learning,[3] and employment and wages.[4] Being able to understand, identify, and measure how the social characteristics of an individual influence network formation is therefore of central importance. However, most studies to date overlook the equilibrium implications of homophily, and disregard key factors such as the impact of time constraints.

In this paper, I develop an empirically realistic model of strategic network formation incorporating homophilic preferences and capacity constraints on the number of links. My analysis uncovers novel structural predictions generated by the equilibrium interplay between the individuals' homophilic preferences and capacity constraints. Building on the explicit structure of homophily obtained in equilibrium, I develop a new estimation technique that allows one to recover underlying preferences parameters. I show as an illustration that the formation of friendship networks among American teenagers is strongly influenced by racial considerations. I also show that this preference bias toward individuals of the same race varies considerably with respect to the racial group considered.

The emphasis on the equilibrium implications of homophilic preferences is new

[1] See for example Currarini et al. (2009), Bramoullé et al. (2012), and Rivas (2009).
[2] Echenique and Fryer (2007), Watts (2007), and Mele (2010).
[3] Golub and Jackson (2010a,2010b).
[4] van der Leij et al. (2009) and Patacchini and Zenou (2012).

to the literature. The equilibrium network resulting from the theoretical model exhibits more structure than the known stylized facts regarding homophilic patterns in social networks.[5] The equilibrium network architecture allows for an original empirical methodology using a maximum likelihood approach. A key feature of the estimation strategy is that it recovers explicit preference parameters characterizing homophily in social networks. In other words, the estimation strategy allows for the identification of *preference interactions* from *constraint interactions*.[6]

The theoretical framework produces sharp predictions. There exists a generically unique, empirically realistic equilibrium network. A key assumption is that the homophilic preferences of individuals can be represented by a distance function on the set of characteristics of the individuals. This idea is implicitly or explicitly exploited by many papers looking at homophily in social networks.[7] This assumption allows me to introduce enough heterogeneity in the model to generate empirically realistic equilibrium networks. I also assume that individuals have link-separable utilities, and an explicit resource constraint, such as time. For example, while a teenager may prefer to be friends with other teenagers who have similar characteristics, he must take into account the fact that he has limited time to spend with the friends he chooses to have. Hence, the resource constraint explicitly introduces an upper bound on the number of bilateral relationships an individual can sustain.[8] The specific notion of homophily emerging in equilibrium results from the tension between the individuals' homophilic preferences and the individuals' resource constraint. These two premises imply a novel theoretical prediction on the shape of homophily in equilibrium. I call this specific network architecture *structural homophily*.

Structural homophily describes an explicit relationship between individuals' socioeconomic characteristics and the network architecture. An individual is charac-

[5]See Bramoullé et al. (2012), and Currarini et al (2009).

[6]Manski (2000) distinguishes between three sources of social interactions : Preference interactions, Constraint interactions, and Expectations interactions.

[7]See for instance, Johnson and Gilles (2000), Marmaros and Sacerdote (2006), Iijima and Kamada (2010), Mele (2010) and Christakis et al. (2010).

[8]It relates to the sociological and psychological observation referred to as the Dunbar's number.

terized by a "social neighborhood" on the space of individual characteristics.[9] This neighborhood explicitly determines the set of acceptable bilateral relationships. In a network characterized by structural homophily, two individuals are linked if and only if they belong to the intersection of their neighborhoods. These neighborhoods are not directly observable, but are implied by equilibrium predictions of the theoretical model for a given a distance function. This novel theoretical prediction has empirical power.

I use structural homophily to develop an original estimation strategy. This strategy is based on the duality between the equilibrium network structure, and structural homophily. Any equilibrium network exhibits structural homophily, and any observed network that exhibits structural homophily is an equilibrium network. I develop a maximum likelihood approach, defined over a population of distinct social networks. The empirical method allows for the identification and estimation of prominent socioeconomic characteristics affecting the equilibrium network structure. This is relevant for policy making since it allows the policy maker to *target* relevant socioeconomic characteristics. As an illustration, I use data on the friendship networks of American teenagers provided by the Add Health database.[10] I focus the analysis on race-based choices and show that the same-race preference bias substantially varies with respect to racial group. Blacks have a stronger bias than Asians, while Whites have the smallest bias. The estimated coefficients are preference parameters, and hence do not depend on the distribution of the racial groups in the population, nor do they depend on the individuals' resource constraints.

This paper contributes to the theoretical and the empirical literature on network formation. Most theoretical models of network formation produce relatively structured equilibrium networks such as stars, circles or chains.[11] These models, although highly relevant from a theoretical perspective, are not well suited for em-

[9]It relates to the sociological notion of a "social niche" ; see for instance McPherson et al. (2001)

[10]Carolina Population Center, University of North Carolina at Chapel Hill ; see http ://www.cpc.unc.edu/projects/addhealth.

[11]Bala and Goyal (2000), Jackson (2008, chapter 6), Jackson and Rogers (1997), Jackson and Wolinsky (1996), and Johnson and Gilles (2000).

pirical purposes. Indeed, the resulting set of equilibrium networks is both too large (many equilibrium networks) and too constraining (stars, chains, circles, etc.) to represent actual, observable, social networks. Most theoretical models assume that payoffs depend on detailed features of the network structure, but neglect the capacity constraints on the number of links an individual can make.[12] I show that the introduction of this constraint, combined with explicit ex-ante homophilic and link-separable utilities, implies the existence of a unique, empirically realistic equilibrium network.[13]

Two alternative explanations of homophily have been proposed. The first is through correlations in the meeting process :[14] individuals have no preference bias, but individuals with similar characteristics have a higher probability of meeting. The second is through preference biases :[15] individuals prefer to link with similar individuals. In this paper, I assume that individuals have homophilic preferences, but evolve in a deterministic world. I analyze the equilibrium implication of these preferences in a fully strategic, non-cooperative setting.

The empirical literature on network formation is still in an early stage. The few existing papers clearly identify homophily as a driving factor of the network formation process.[16] This paper contributes to the literature on strategic network formation by providing an estimation strategy based on the equilibrium structure of homophilic preferences. Equilibrium considerations are important, as they imply a departure from link-level estimation techniques. The model defines a precise dependence structure which allows for the definition of an explicit maximum likelihood estimator.[17]

The remainder of the paper is organized as follows. In section 2.2, I present the

[12]Exceptions include Bloch and Dutta (2009) and Rubí-Barceló (2010).

[13]I concentrate on strategic models of network formation. There exists a large literature on random network formation, which is not directly concerned with the current setting. The interested reader can see for instance Jackson (2008, chapters 4 and 5) and the references therein.

[14]See for instance Bramoullé et al. (2012)

[15]See Currarini et al. (2009), and Mele (2010)

[16]See for instance Christakis et al. (2010), Mele (2010), Currarini et al. (2010), and Franz et al. (2008)

[17]As opposed to the simulated Bayesian approaches in Christakis et al. (2010), and Mele (2010).

theoretical model and key definitions. In section 2.3, I find and characterize the (unique) equilibrium network. In section 2.4, I describe the empirical methodology and explore its properties using Monte Carlo simulations. In section 2.5, I present an application of race-based homophily in friendship networks using the Add Health database. I conclude in section 2.6.

2.2 The Theoretical Model

In this section, I present a non-cooperative model of network formation that characterizes the equilibrium effects of homophily. The model generically produces a unique equilibrium. I first provide a formal definition of Structural Homophily. Next, I outline the theoretical framework, and finally, I briefly present the main definitions and equilibrium concepts.

2.2.1 Structural Homophily

In order to introduce this new notion of homophily, we need some preliminary assumptions. There is a finite set of individuals N. Individuals may be linked together through a network. Let $g_i \subseteq N$ be the set of individuals linked to individual i for all $i \in N$. Each individual $i \in N$ is characterized by a type $\theta_i \in \Theta$, where Θ is the type space. An individual's type could represent, for instance, a series of socioeconomic characteristics. I consider a distance d on Θ. For notational simplicity, let $d_{ij} \equiv d(\theta_i, \theta_j)$ for any $i, j \in N$. Then, *structural homophily* is defined as follows.

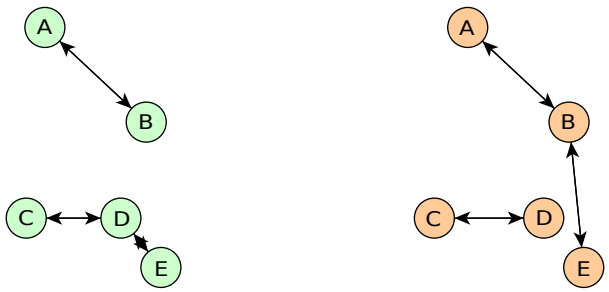
Definition 1 *A network g exhibits **structural homophily** with respect to a distance $d(.,.)$ if whenever two individuals, i and j, are not linked, either $d_{ij} \geq \max_{k \in g_i}\{d_{ik}\}$ or $d_{ij} \geq \max_{k \in g_j}\{d_{jk}\}$.*

This definition formalizes the fact that two individuals that are "close" should be linked. Intuitively, if two individuals are not linked, it is because, from the point of view of one of the individuals, the other is located relatively too far. Notice that this definition only makes sense when the creation of a link requires mutual

consent. Figure 2.1 shows two examples of networks for $\Theta = \mathbb{R}^2$. The first network exhibits Structural Homophily, but the second does not. In Figure 2.1b, the closest individuals (i.e. D and E) are not linked, which is in contradiction with structural homophily since D is linked to C, and E is linked to B.

Figure 2.1 – Structural Homophily

(a) Respected (b) Violated



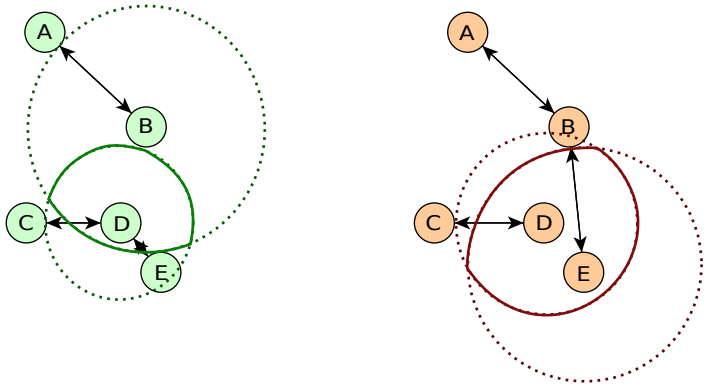
More insight can be obtained by drawing the equivalence (or indifference) curves corresponding to the *farthest link* for each individuals considered (i.e. for B and D in Figure 2.2a, and for D and E in Figure 2.2b). These equivalence curves define neighborhoods ; every individual inside the neighborhood of i is at a distance smaller the distance between i and his farthest link. If both individuals belong to the intersection of the two neighborhoods generated by the equivalence curves (as in Figure 2.2b), then Structural Homophily is violated.[18]

Structural homophily has an interpretation in terms of revealed preferences. Suppose that individuals have preferences over links with other individuals, and that such preferences are a function of the distance between the individuals. Suppose also that we observe the network (i.e. the individuals and their links), and the types of the individuals in the network (i.e. a series of individual characteristics). Then, under mutual consent, we should not observe networks such as the one depicted in Figure 2.2b. That is, structural homophily should hold.

[18]This closely relates to the *cutoff rule* of Iijima and Kamada (2010).

Figure 2.2 – Structural Homophily : Equivalence curves

(a) Respected (b) Violated



It it interesting to note that small-worlds networks respect structural homophily for a specific type space.[19] In a small world model, individuals are located on *islands*. In that setting, structural homophily implies that individuals are linked first with individuals of the same island. Hence, if there is a link between two islands, those islands have to be fully connected. I now present a social networking game, which produces Structural Homophily at equilibrium.

2.2.2 The Game

There are n individuals, each of whom is endowed with a fixed amount of resources $\bar{x}_i = \kappa_i \xi$, where $\xi \in \mathbb{R}_+$ and $\kappa_i \in \mathbb{N}$. We will see that, in equilibrium, κ_i is interpreted as the maximum number of links that an individual i can have. A strategy for an individual i is a vector $x_i = (x_i^1, ..., x_i^n) \in X_i$, where $X_i = \{x_i \in \mathbb{R}_+^n | x_i^j \leq \xi, \ and \ \sum_{j \in N} x_i^j \leq \kappa_i \xi\}$. Then, ξ plays the role of a link-level constraint. The introduction of the link-level constraint is motivated by the empirical fact that the number of links varies across individuals. Let $X = \times_{i \in N} X_i$. We say that there is a link between an individual i and an individual j iff $x_i^j > 0$ and $x_j^i > 0$. Let $g_i = \{j \in N | i \text{ and } j \text{ are linked}\}$, so $j \in g_i$ iff $i \in g_j$. That is, a link exists iff both individuals invest a strictly positive amount of resources in it. Notice that individual i can be linked to himself.

[19] See for instance Jackson and Rogers (2005) and Galeotti et al. (2006).

The utility of an individual is given by the function $u_i : X \to \mathbb{R}$. It is additive in the different links he has, and it is represented by :

$$u_i(x) = \sum_{j \in N \setminus \{i\}} v_i(x_i^j, x_j^i, d_{ij}) \cdot \mathbb{I}_{\{j \in g_i\}} + w_i(x_i^i) \cdot \mathbb{I}_{\{i \in g_i\}}$$

where $\mathbb{I}_{\{P\}}$ is an indicator function that takes value 1 if P is true, and 0 otherwise. The function $v_i(x, y, d)$ gives the value of any link for i. It is assumed to be twice continuously differentiable with $v_x(x, y, d) > 0$ if $y > 0$, $v_y(x, y, d) > 0$ if $x > 0$, and $v_d(x, y, d) < 0$ if $x, y > 0$. The function $w_i(x_i^i)$ represents the payoff received from the private investment of i.[20] It is also twice continuously differentiable with $w'(x) > 0$. I also allow for the presence of fixed costs, i.e. $v_i(0, 0, d) \leq 0$ and $w_i(0) \leq 0$. Notice that an individual benefits from a link only if both individuals invest in the link. The model induces a game Γ between the n individuals. Formally, we have $\Gamma = (N, \{X_i\}_{i \in N}, \{u_i\}_{i \in N})$.

The model has two important features. First, the initial endowment creates scarcity and induces a feasibility constraint. This effect is typical of any matching model. If some individual i invests resources in a link with individual j, he will have less available resources to create a link with another individual. That is, the feasibility constraint implies a tradeoff between the distance between two individuals, and the level of investment they put in the link. This is what Manski (2000) refers to as "constraint interactions". Second, the preferences are affected by the presence of direct externalities. The amount of resources invested by some individual in a given link directly affects the utility of the individuals he links to. That is, in Manski's terms, "preference interactions". Those two features will play an important role in equilibrium.

This completes the description of the game. I now present the main definitions.

[20] The function w_i can also be interpreted as the private value of the resource x for i

2.2.3 Definitions

Before turning to the analysis of the model, I introduce some definitions. The collection of links between individuals generates a *network* $g = (N, E)$. A network is characterized by a set of individuals (here, N), and a set E of links, which are (unordered) pairs of individuals. The set of all possible networks is denoted by $\mathbb{G}$. Any network g can be represented by a $n \times n$ *adjacency matrix* A that takes values $a_{ij} = 1$ if $j \in g_i$, and 0 otherwise, for all $i, j \in N$. The *degree* $\delta_i(g)$ of an individual i is the number of links attached to i, i.e. $\delta_i(g) = |g_i|$.

I am interested in the following solution concepts :

Definition 2 *A Nash Equilibrium (NE) is a profile $x^* \in X$ such that $u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*)$ for all $x_i \in X_i$, and for all $i \in N$.*

The set of Nash equilibria is very large. Since an individual benefits only from a collaborative link when both individuals invest in the link, it will never be profitable to unilaterally start a new link. For this reason, I will focus on the following solution concept, introduced by Goyal and Vega-Redondo (2007).

Definition 3 *A Bilateral Equilibrium (BE) is a profile $x^* \in X$ such that :*
(1) x^* is a Nash Equilibrium
(2) There exists no $i, j \in N$, such that $u_i(x_i, x_j, x_{-i-j}^*) > u_i(x^*)$ and $u_j(x_i, x_j, x_{-i-j}^*) \geq u_j(x^*)$ for some $x_i \in X_i$ and $x_j \in X_j$.

This solution concept allows for bilateral deviations. This is a natural extension of individual rationality, since individuals can benefit from the creation of links. For certain economies, however, the BE concept will be too constraining. Accordingly, I also introduce the following weakened equilibrium concept.

Definition 4 *A Weak Bilateral Equilibrium (WBE) is a profile $x^* \in X$ such that :*
(1) x^* is a Nash Equilibrium
(2) There exists no $i, j \in N$, such that $u_i(x_i, x_j, x_{-i-j}^*) > u_i(x^*)$ and $u_j(x_i, x_j, x_{-i-j}^*) > u_j(x^*)$ for some $x_i \in X_i$ and $x_j \in X_j$.

In a WBE, a deviation must strictly increase the payoff of both individuals involved. Notice that $BE \subseteq WBE \subseteq NE$. I discuss the distinction between these concepts in section 2.3.1 (lemma 2.3.1 and proposition 2.3.5).

2.3 Equilibrium Characterization

I first show the existence of an equilibrium. Since the payoff functions are not continuous, we cannot directly use the standard fixed-point arguments. The existence of a NE is straightforward. Let $x_i^j = 0$ for all $j \neq i$. Then, for every individual, the maximization problem becomes : $\max_{x_i \in X_i} w(x_i^i) \cdot \mathbb{I}_{\{i \in g_i\}}$. The allocation $x^* \in X$ that maximizes this problem for all $i \in N$ is obviously a NE. In order to show the existence of a WBE (or a BE), I will need to introduce additional assumptions. The next result provides an intuition on the additional restrictions imposed by the bilateral stability on the solution set. It states that if a deviation is jointly profitable, but not unilaterally profitable, the deviating individuals have to invest more in their collaborative link. All proofs can be found in appendix I.1.

Lemma 2.3.1 *If $x^* \in X$ is a NE, but not a WBE, given any deviating pair (i, j), with profitable deviations $x_i \in X_i$ and $x_j \in X_j$, we have $x_i^j > x_i^{j*}$ and $x_j^i > x_j^{i*}$.*

Since x^* is a NE, it is individually rational. Also, since the utility functions are additive in the different links, the action of individual j on individual i only affects i through the link between i and j. If x^* is not jointly rational for i and j, the incentive to deviate must come from the link i and j have together.

Throughout this section, I consider two alternative assumptions :

Assumption 1 (Finiteness) *For all $i, j \in N$, $x_i^j \in \{0, \xi\}$*

Assumption 2 (Convexity) *For all $i \in N$, $\frac{\partial^2 v_i}{\partial x^2}(x, y, d) \geq 0$, $\frac{\partial^2 w_i}{\partial x^2}(x) \geq 0$*

The finiteness assumption is extensively used in the literature.[21] Convexity is often assumed when the network formation process involves continuous strategies.

[21] See for instance Jackson (2008) chapters 6 and 11.

For example, Bloch and Dutta (2009) define the strength of a link between individuals i and j as the sum of a (strictly) convex function of the individuals' investment, i.e. $s_{ij} = f(x_i^j) + f(x_j^i)$, with $f' > 0$ and $f'' > 0$. Rubi-Barceló (2010) uses a linear (hence convex) function to represent the payoff from scientific collaboration between two researchers.[22] I provide existence results and show that those two assumptions imply that the equilibrium network exhibits structural homophily.

The next results are based on an algorithm referred to as the *assignment algorithm*, and formally defined in Appendix I.2. The assignment algorithm uses as inputs : (1) the list of preferences $\{u_i(x)\}_{i\in N}$, (2) the individual characteristics $\{\theta_i\}_{i\in N}$, (3) the resource constraints $\{\kappa_i\}_{i\in N}$, and (4) the distance function d on Θ. It produces at least one allocation $x \in X$, and any allocation produced is such that $x_i^j \in \{0, \xi\}$ for all $i, j \in N$. When $x_i^j \in \{0, \xi\}$, the payoff that an individual receives from the links can be ranked using the distance function (a small distance implies a big payoff). Accordingly, the assignment algorithm proceeds first by linking the pairs of individuals with the smallest distances (provided that the link is profitable for both individuals, and leads to a higher payoff than the private investment). The following results show that any allocation constructed in that fashion is a WBE, and induces a network that exhibits structural homophily.

Let's start with the finite case. Under Finiteness, the involvement of an individual in some link does not affect the amount of resources he invests in his other (existing) links. The value of a link between two arbitrary individuals is then independent of the other (potential) links. Consequently, we have the following :

Theorem 2.3.2 (Finite Strategy Space) *Under Finiteness, an allocation is a WBE iff it is produced by the assignment algorithm.*

Under convexity, for a given link, it is also rational for both individuals to invest resources until the link-level constraint ξ is met, provided that it leads to a positive payoff. We then have the following :

[22] The value of a scientific collaboration as defined by Rubi-Barceló (2008, p.7) is interpreted as a distance in my model.

Theorem 2.3.3 (Existence) *Under Convexity, any allocation produced by the assignment algorithm is a WBE.*

Proposition 2.3.4 gives sufficient conditions so that any individual *has* to invest up to the link-level constraint, in any WBE.

Proposition 2.3.4 (Uniqueness) *Suppose that the inequalities in Assumption 2 are strict, then any WBE can be produced by the assignment algorithm.*

Then, under Finiteness or Strict Convexity, any equilibrium can be constructed through the assignment algorithm. It is worth noting that under Finiteness, $x_i^j \in \{0, \xi\}$ by assumption, while under Strict Convexity it must hold only in equilibrium.

The above results show the existence of a WBE, but not of a BE. The intuition is the following. Suppose that Finiteness holds, and that the economy contains only 3 individuals : i, j, k. Suppose also that $d_{ij} = d_{ik} < d_{jk}$, and that $\bar{x}_i = \bar{x}_j = \bar{x}_k = \xi$. Finally, suppose that $v_i(\xi, \xi, d_{ij}) = v_j(\xi, \xi, d_{ij}) = v_k(\xi, \xi, d_{ik}) > 0$, while any other link has a negative value. Then, in this example, there is no BE, but there are two WBE (see Figure 2.3). The reason is that i is indifferent between a link with j or a link with k. So, if i is linked with j, but receives a proposition from k, he will be indifferent between keeping his link with j and replacing it with a link with k (while k would be strictly better off with such a deviation).

In many contexts, however, individuals have many characteristics, and the likelihood of such a circumstance is small. In the absence of such a circumstance, we can show the existence of a BE. Formally,

Proposition 2.3.5 *Suppose that $d_{ij} \neq d_{kl}$ for any $i \neq j$ and $k \neq l$, then any WBE produced by the assignment algorithm is a BE. Moreover, if d is such that $v_i(\xi, \xi, d_{ij}) \neq 0$ and $v_i(\xi, \xi, d_{ij}) \neq w_i(\xi)$ for all $i, j \in N$, this equilibrium is unique.*

This implies that if for all $i \in N$, the types $\theta_i \in \Theta$ are drawn from a distribution with a dense support on Θ, then there exists a unique WBE, which is also a BE, $[a.s.]$

Figure 2.3 – WBE and BE

(a) The First WBE (b) The Second WBE

Let's now turn to the characterization of the equilibrium network. Since the level of investment of an individual in a potential link does not depend on the number of links he has, the payoffs are only influenced by the distance. Suppose i and j are linked. Then, the creation of a new link between j and k has no spillover effects on i. This produces important consequences on the shape of the equilibrium network. The next proposition characterizes the allocations produced by the assignment algorithm.

Proposition 2.3.6 (Characterization) *Let g^* be the network generated by some allocation produced by the assignment algorithm, then*
(1) For all $i \in N$, $\delta_i(g^*) \leq \kappa_i$.
(2) The network g^* exhibits Structural Homophily.

The proof is immediate from the construction through the assignment algorithm. Since investments are maximal in every link, the number of links an individual can have is bounded by the resource constraint κ_i. Also, since the assignment algorithm creates links starting from the ones associated with the smallest distances, the induced network exhibits structural homophily. In essence, under Finiteness or (strict) Convexity, any equilibrium network can be constructed through the assignment algorithm, hence satisfying structural homophily.

Let's now turn to efficiency issues. There are many ways to define efficiency. The first one would be to consider the Pareto criterion. Given Finiteness or Convexity, any BE is Pareto efficient. In fact, we have an even stronger result, which is the

fact that any BE is a Strong Nash equilibrium (Aumann, 1959).

Proposition 2.3.7 *Under Finiteness or Strict Convexity, any BE is a Strong Nash equilibrium.*

Since the utility functions are additive, bilateral stability implies stability in the sense of a Strong Nash equilibrium. However, since the utility functions are non-continuous (and utilities are not transferable), Pareto efficiency does not imply efficiency in the sense of the utilitarian criterion. Consider the following social welfare function :

$$W(x) \quad = \quad \sum_{i \in N} u_i(x)$$

In this case, efficiency is not guaranteed. In particular, one can find examples of economies where the unique BE is efficient (in the sense of the utilitarian and the Pareto criterion), as well as examples of economies where the unique BE is inefficient (in the sense of the utilitarian criterion). This inefficiency comes from two principle sources.

First, under the Finiteness assumption, any efficient allocation $z \in X$ is such that $z_i^j \in \{0, \xi\}$ for all $i, j \in N$ (by assumption). Since an individual values only his own payoff, while the social planner (SP) cares about all individuals, a collaborative link is more valuable for the SP than it is for an individual. (It enters the utility function of both the individuals involved.) The tradeoff between the individual and the collaborative links is then different for an individual than for the SP.

Second, under the (strict) Convexity assumption, another issue arises. Since the SP is willing to trade off the utilities of the individuals, an efficient allocation $z \in X$ need not be such that $z_i^j \in \{0, \xi\}$. For example, suppose that there are no fixed costs, then any network g^* such that $\delta_i(g^*) < \kappa_i$ for some $i \in N$ is inefficient. The reason is that if $\delta_i^* < \kappa_i$ for some $i \in N$, the creation of a link with some agent j (who is willing to invest a small amount ϵ) leads to $v_i(\xi, \epsilon, d_{ij})$ for i. If ϵ is small enough, the loss for j is compensated by the discrete jump in the utility of

i. Hence, g^* is inefficient. However, it is possible that such a network g^* is induced by a BE.

This concludes the analysis of the theoretical model. In section 4, I develop an estimation technique derived from structural homophily, and present Monte Carlo simulations.

2.4 The Econometric Model

In this section, I present the econometric model. I use Structural Homophily in order to estimate the weights of the distance function.[23] I would like to emphasize that the method and results of this section are self-contained. If one was willing to *assume* structural homophily (instead of viewing it as the equilibrium outcome of the non-cooperative game presented in the last section), all the results of this section would apply.

In order to present the econometric model, I introduce the following definition :

Definition 5 *An observation q is*
1) a network $g = (N_q, E_q)$, and
2) for each individual $i \in N_q$, a vector of R individual socioeconomic characteristics, i.e. $\{\theta^i\}_{i \in N}$, where θ^i is a $1 \times R$ vector.

For a given observation $q \in 1, ..., Q$, I note (g_q, θ_q), where θ_q is $n_q \times R$. Definition 5 implies that an econometrician does not observe the specific level of investment in a link (i.e the link-level constraint), nor does he observe the resource constraint κ_i.[24] Accordingly, given a set of observations $(g_q, \theta_q)_{q=1}^{Q}$, we do not possess enough information to construct the equilibrium network through the assignment algorithm, even assuming some structural form for the utility functions. Specifically, a standard econometric model would be the following. Given a parametric form

[23] Čopič et al. (2009) also exploit homophily, although in a very different setting, in order to develop their estimation technique.

[24] Notice that while κ_i is an upper bound to $\delta_i(g)$, they are not necessarily equals. See proposition 2.3.6.

for the payoff functions (i.e. $\{v_i(x,y,d), w_i(x)\}_{i\in N_q}$), and the distance function (i.e. $d(i,j)$), one would assume that the data is generated by :

$$g_q = \Lambda(\theta_q, \kappa_q, \xi_q, \varepsilon_q; \beta) \tag{2.1}$$

where Λ is the assignment algorithm, κ_q is the $n_q \times 1$ vector of individual resource constraints, ξ_q is the link-level resource constraint, ε_q is the error term, and β is the vector of parameters to be estimated. Provided that one observes $\theta_q, \kappa_q, \xi_q$, one could, in principle, estimate β. Since κ_q and ξ are typically unobserved in existing datasets, I use a different approach.[25] From section 2.3's results, I have established that any allocation produced by the assignment algorithm respects structural homophily.[26] My approach will then be to maximize *the likelihood that the observed network exhibits structural homophily.* Accordingly, the distance function will play a central role. I assume the following structural form for the distance function :

$$\ln(d_{ij}) = \sum_{l=1}^{L} \beta_l \rho_l(\theta_i, \theta_j) + \varepsilon_{ij} \tag{2.2}$$

where $\varepsilon \sim_{iid} N(0,1)$, and $\rho_l(.,.)$ is a dimension-wise distance function.[27] The vector $(\beta_1, ..., \beta_L) \in \Xi \subset \mathbb{R}^L$ are the weights of the distance function. Equation (2.2) highlights two important features of the model.

First, instead of trying to specifically identify the parameters of the utility function, I limit myself to the estimation of the relative importance of the social characteristics in the network formation process. That is, I only seek to estimate the parameters of the distance function, and not the parameters of the utility functions (for instance, I do not estimate the value of the resource for the individuals). This

[25] There are also severe computational and identification issues using the specification in (2.1).

[26] Also, by observing a network that exhibits structural homophily, one can always find some $v_i(x,y,d)$, κ_i and ξ such that it is produced by the assignment algorithm.

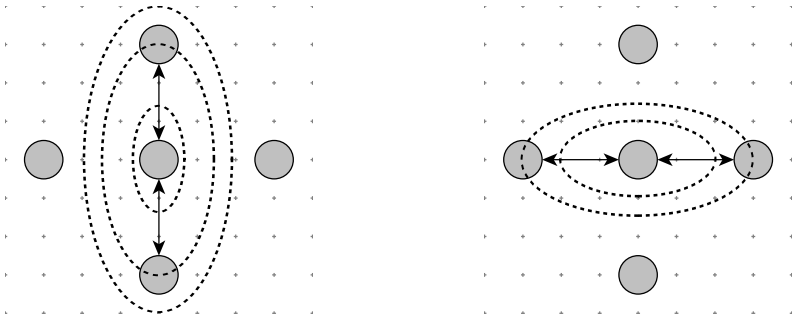
[27] For instance, if $\Theta \in \mathbb{R}^2$, one could choose $\rho_l(\theta_i, \theta_j) = |\theta_i^l - \theta_j^l|$. The proposed structural form is by no means the only possibility. Any positive and symmetric function could be used. I prefer to use the specification in 2.2 to simplify the exposition.

is illustrated in Figure 2.4. In Figure 2.4a, the individuals place more value on the characteristic on the *horizontal* axis. Then, the "closest" individuals for the central node are the ones on the top and bottom. Symmetrically, in Figure 2.4b, the individuals place more value on the characteristic on the *vertical* axis. Then, the "closest" individuals for the central node are the ones on the left and right. My aim is to estimate the relative weights placed on each characteristics.[28]

Figure 2.4 – Changing the Weight of the Distance Function

(a) Relative Importance on the Horizontal Characteristic

(b) Relative Importance on the Vertical Characteristic



Second, I assume that the distance function is observed with noise. That is, there exists a set of variables, observed by the individuals within the model, but unobserved by an econometrician, that affects the distance function.[29] This assumption is not standard and deserves a discussion.

A typical method to introduce unobserved heterogeneity into this type of models would be to assume that the value of a link depends on some unobserved set of characteristics, i.e. $v_i(x, y, d) + \varepsilon_{ij}$. However, this cannot be identified from a model where the distance is observed with noise, since we can always define a symmetric

[28]Centered ellipses like those depicted in Figure 4 are implied by the additive form we assumed in (2.2). The generalization to more general class of distance functions such as in Henry and Mourifie (2011) is straightforward.

[29]For instance, ε_{ij} can be interpreted as a measurement error.

function $\tilde{d} : \Theta^2 \to \mathbb{R}$ such that $v_i(\xi, \xi, \tilde{d_{ij}}) = v_i(\xi, \xi, d_{ij}) + \varepsilon_{ij}$ for all $i \neq j$.[30]

Now, given (2.2), we can compute the probability (conditional on an observation) that a network exhibits structural homophily. Let $\Psi = 1 - \Phi$, where Φ is the c.d.f. of the standard normal distribution, and let $\gamma = \left(\beta_1/\sqrt{2}, ..., \beta_L/\sqrt{2}\right)$. The probability that a network g (given a set of characteristics θ) exhibit Structural Homophily is (algebraic manipulations can be found in appendix I.3) :

$$\begin{aligned} \mathbb{P}(sh|g, \theta, \gamma) &= \Pi_{ij \notin g} \left\{ \Pi_{k \in g_i} \Psi \left[(s_{ik} - s_{ij})\gamma' \right] + \Pi_{k \in g_j} \Psi \left[(s_{jk} - s_{ij})\gamma' \right] \right. \\ &\quad \left. - \Pi_{k \in g_i} \Psi \left[(s_{ik} - s_{ij})\gamma' \right] \Pi_{k \in g_j} \Psi \left[(s_{jk} - s_{ij})\gamma' \right] \right\} \end{aligned} \tag{2.3}$$

where s_{ij} is the $1 \times L$ vector of dimension-wise distance, i.e. $s_{ij}^l = \rho_l(\theta_i, \theta_j)$.[31]

Then, given that there are Q observations, I propose the following maximum likelihood estimator :

$$\ell(\beta|\theta) = \frac{1}{Q} \sum_{q=1}^{Q} \ln[\mathbb{P}(sh|g_q, \theta_q, \gamma)] \tag{2.4}$$

Provided that there exists a unique $\gamma^0 \in \Xi$ which maximizes (2.4), the maximum likelihood estimator is well-behaved, and γ can be consistently estimated.[32]

The identification's strategy is based on a link-deference approach. A link exists if no individual refused it. There are two reasons for an individual to refuse a link : (1) because he has no resources left (constraint interactions), or (2) because the other individual is too distant (preference interactions). I want to identify the preference effect, given that the resource constraint is unobserved. The estimation strategy can be viewed as to minimize the probability that structural homophily is violated.

[30] If v_i is log quasi-linear in the distance, i.e. $v(x, y, d) = f(x, y) - ln(d)$, the two models are equivalent.

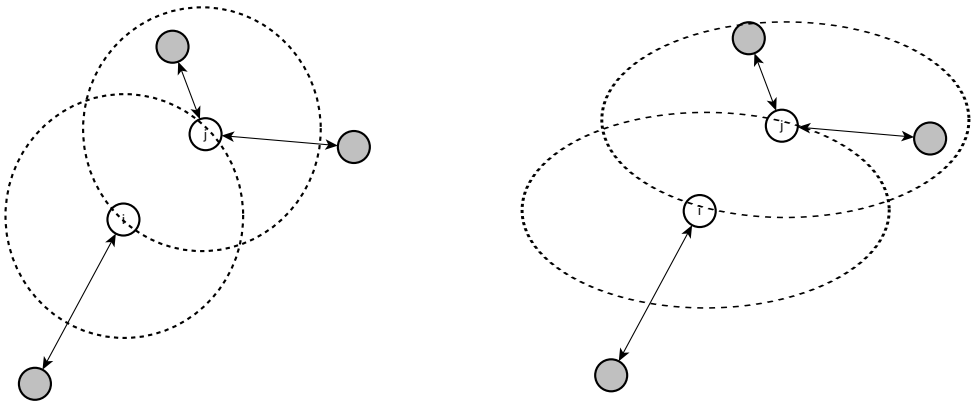
[31] Equation 2.3 assumes that there is no isolated individual (i.e. no individual i is such that $g_i \in \{\emptyset, \{i\}\}$). This is done without loss of generality since for any pair of individuals in which one of the individual is isolated, the condition imposed by structural homophily is trivially respected.

[32] Although the function in (2.3) looks peculiar, the MLE setting is standard and the estimation of (2.4) requires only usual the usual set of assumptions. See for instance Cameron and Trivedi (2005, p. 142-143) for the asymptotic properties of the maximum of likelihood estimator.

Lets consider two alternative parameters β and β'. Suppose that we observe two individuals, i and j, not linked together, as in Figure 2.5. According to β and β', i is linked to an individual, farther from him than j. This means that i would have been willing to create a link with j, but that j refused. This implies that j cannot be linked to individuals farther from him than i. If he does, structural homophily is violated. Thus, if j is linked to farther individuals than i under β, but not under β', then β' is chosen over β to represent individuals' preferences.

Figure 2.5 – Admissible Parameters, $\Theta = \mathbb{R}^2$

(a) Distance Weights according to β (b) Distance Weights according to β'



This shows why isolated individuals (i.e. individuals that have no link) provide no information : whatever the parameters' values, they never contradict structural homophily. In other words, for isolated individuals, we cannot identify whether they are isolated because they have limited resources, or because they have strong homophilic preferences. From a revealed preference approach, we gain information about an individual' preferences by observing his choices. If an individual is not connected, he does not "consume" any resource. We therefore cannot say anything about his preferences.

I now explore the properties of this method through Monte Carlo simulations.

2.4.1 Monte Carlo Simulations

I now present some Monte Carlo simulations. One of the advantage of section 2.3 is that it provides a simple algorithm allowing for the construction of the equilibrium network. Using the assignment algorithm, I will explore the finite sample properties of the estimator defined in the previous section. For simplicity and because of computational limitations, I assume that $\Theta = \mathbb{R}^2$ (this could represent, for example, the geographic position of the individuals), and $\rho_l(\theta_i, \theta_j) = |\theta_i^l - \theta_j^l|$. For all $i \in N$, I assume that $\theta_i \sim_{iid} N(\mathbf{0}, \sigma^2\mathbf{I})$. Thus, σ^2 controls for the dispersion of the individuals on the plane. As assumed, I let $\varepsilon_{ij} \sim N(0,1)$. I run 1000 replications of an economy composed of 150 independent populations (networks), each of which has 20 individuals, and I vary κ_i and σ^2 (I assume that κ_i is drawn from a uniform distribution).

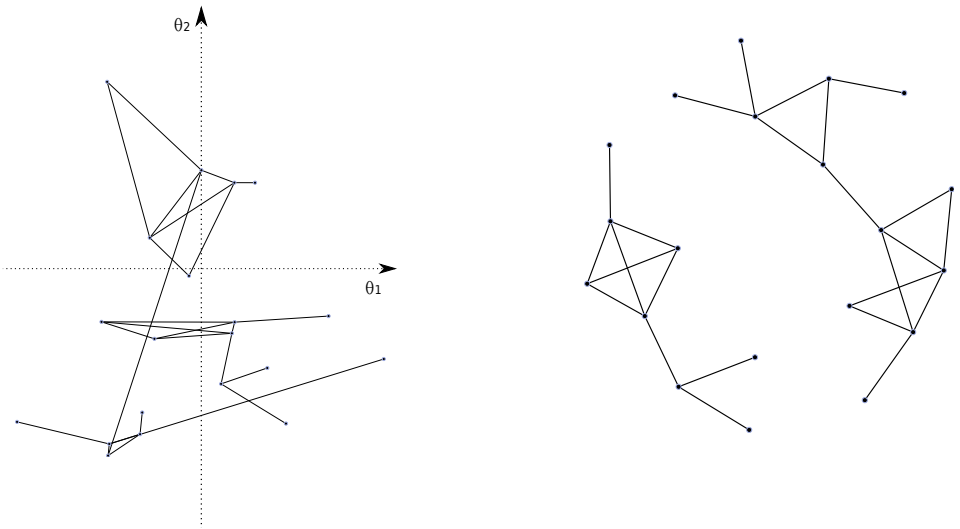
The simulated networks are generated using the assignment algorithm, assuming that $v_i(\xi, \xi, d_{ij}) > 0$ for all $i, j \in N$ and that $w_i(\xi) < 0$ for all $i \in N$. I assume that the weights are $\beta = (2, 6)$, so the distance is $d(\theta_i, \theta_i) = 2|\theta_i^1 - \theta_j^1| + 6|\theta_i^2 - \theta_j^2|$. Figure 2.6 displays a typical equilibrium network for this economy. Figure 2.6a shows the simulated network on the plane while Figure 2.6b rearranges the individuals in order to see clearly the network structure. Notice that the individuals value the vertical characteristic more than the horizontal one.

The small size of each observation (i.e. 20 individuals in every network) has an impact on the precision of the estimator. Take the following limiting case. Suppose that, as in the simulation framework, every link is profitable. Then, if the resource constraint is large enough, the equilibrium network is the complete network, and Structural Homophily is not binding. As a result, the model in (2.4) is not identified. I now explore the precision of the estimator when individuals have a relatively large resource constraint, compared to the size of the population. I find that the estimator performs better when the maximal number of links is small compared to the size

[33] Using the Kamada-Kawai algorithm is a standard way of drawing networks on the plane.

Figure 2.6 – Typical network, with $\beta = [2\ 6]$, and $\kappa_i \sim U[1, 4]$

(a) In the type space (b) K.K. representation[33]



of the population, and that the precision of the estimator can be improved by increasing the dispersion of the population on the type space.

Table 2.1 and Figure I.1 to I,4 (Appendix I.4) show the simulation results. Since the parameters are only scale-identified, I report only the relative estimates. Simulations show that as the number of links increases (relative to the size of the population), the precision of the estimator is increased, but the estimates can be slightly biased upward. However, this problem vanishes as the distribution of the population over the type space increases.

I now turn to the implementation of the estimation technique. In the next section, I use the Add Health database to address the role of race in the formation of friendship networks.

2.5 Empirical Application : High-School Friendship Networks

I wish to estimate the weights of the distance function that leads to the formation of the friendship networks of American teenagers. I am particularly inter-

Tableau 2.I – Monte Carlo Simulations

	Standard Deviation (σ)			
κ_i	10	12	14	16
$\{1,2\}$	3.031	3.024	3.02	3.01
	(0.026)	(0.028)	(0.02)	(0.02)
$\{3,4\}$	3.077	3.045	3.03	3.02
	(0.027)	(0.028)	(0.03)	(0.02)
$\{5,6\}$	3.089	3.050	3.03	3.03
	(0.029)	(0.029)	(0.03)	(0.03)
$\{7,8\}$	3.104	3.069	3.05	3.03
	(0.032)	(0.030)	(0.03)	(0.03)
$\{9,10\}$	3.107	3.081	3.05	3.04
	(0.033)	(0.030)	(0.03)	(0.03)
$\{11,12\}$	3.112	3.082	3.05	3.04
	(0.034)	(0.033)	(0.03)	(0.03)
$\{13,14\}$	3.117	3.082	3.05	3.04
	(0.044)	(0.039)	(0.04)	(0.04)
$\{15,16\}$	3.122	3.090	3.06	3.05
	(0.047)	(0.071)	(0.06)	(0.06)

ested in the role of race, as previous studies have suggested there is a significant race-based preference bias in the choice of friendship relations among teenagers. Currarini et al. (2010) use a search model in order to estimate the preference bias for Asians, Blacks, Hispanics and Whites. They show that Asians have the largest preference bias, followed by Whites, Hispanics and Blacks. Using a different approach, Mele (2011) estimates the role that homophilic preferences toward race plays in the formation of friendship networks. He shows that all racial groups have strong homophilic preferences, although he does not capture any strong differences between groups. Interestingly, I find strong evidence that the racial preference bias varies across racial groups, although I find that Blacks have the strongest bias, followed by Asians and Whites.

As in the two papers mentioned, I use the Add Health database as it is particularly well suited for my model. Recall that the model presented in sections 2.2 and 2.3 assumes that the individuals of the same population meet with probability one. A convincing empirical implementation then requires that the observed populations

are small enough. To that effect, the Add Health database provides information on students' high-schools, which are quite small entities.[34] Specifically, the sample includes the race, and the friendship networks of 5,466 teenagers, coming from 98 high schools in the U.S. The variable of interest is race. I assume that a student's type is his or her race. Thus the type space has 4 dimensions : White, Black, Asian, Native. Formally, $\Theta = \{0,1\}^4$, so a student who considers himself as Black-Asian would be of type $\theta = (0,1,1,0)$. I assume the following distance function :

$$\ln d(x_i, x_j) = \sum_{r=1}^{4} \beta_r \mathbb{I}_{\{x_i^r \neq x_j^r\}} + \varepsilon_{ij} \tag{2.5}$$

where $\mathbb{I}_{\{P\}}$ is an indicator function that takes value 1 if P is true, and 0 otherwise. For instance, the distance between a teenager i who is White, and a teenager j who is Black, is $d(x_i, x_j) = \beta_{white} + \beta_{black}$. The β's measure the relative strength of the preference bias toward individuals of a particular racial group, e.g. being Black, v.s. being non-Black.

The Add Health questionnaire asks each teenager to identify their best friends (up to 10, and a maximum 5 males and 5 females). I assume that two individuals are friends only if they attend the same school. This assumption is standard in the literature using Add Health data. This allows each school (the set of teenagers and the network) to be treated as an observation. Thus, the database contains 98 observations (i.e. 98 schools). Table 2.II summarizes the data :

Tableau 2.II – Descriptive Statistics

Variable	Mean	Standard Deviation
White	0.733	0.442
Black	0.150	0.357
Asian/Pacific	0.031	0.174
Native	0.062	0.242
Degree	2.064	1.284

[34] For that reason, and for computational reasons, I limit myself to schools for which I observe less than 300 students, which is about 68% of the schools in the database. I also remove the isolated individuals, as they provide no relevant information (see p.18, last paragraph).

I estimate the model (2.4), using the distance function in (2.5). The estimated weights $(\hat{\beta}_1, ..., \hat{\beta}_4)$ and the corresponding standard errors are shown in Table 2.III. Since the weights are only scale-identified, I report the relative effects. The estimation shows that the weight associated with the Blacks' dimension is the greatest (2.270 times greater than the Whites', and 1.796 times greater than the Asians'). The Asians' dimension is the second in magnitude (1.264 times greater than the Whites'). I find no statistically significant relative weight for the Natives' dimension. Notice that this is independent of the relative proportion of each racial group in the population, and the (unobserved) individuals' time constraints.

Tableau 2.III – Relative Estimated Weights (White normalized to 1)[†]

	Black	Asian	Native
Estimate[††]	2.270**	1.264**	-0.199
SE	(0.244)	(0.157)	(0,150)
Robust SE[†††]	[0.304]	[0.294]	[0.171]

[†] S.E computed using the delta method.
[††] ** for 1% significance level.
[†††] Robust SE using the (sandwich) variance-covariance matrix for pseudo-m.l.e.

Turning back to the distance functions, one can reconstruct the distance between the different racial groups from the estimates in Table 2.III. Recall that, for instance, the distance between a Black and a White is $d(black, white) = \beta_{black} + \beta_{white}$. Then, according to Table 2.III, the distance between Blacks and Asians is the greatest ($d = 3.534$), followed by the distance between Blacks and Whites ($d = 3.270$) and the one between Whites and Asians ($d = 2.264$). This shows that, in order to correctly specify the impact of homophilic preferences on the creation of links, one has to take in to account the impact of the preference biases of both individuals involved. Structural homophily allows to identify those preference biases.

I now discuss the limitations of my approach and suggest some potential generalizations.

2.6 Going Further

I have shown that structural homophily can be obtained by a non-cooperative game of network formation. Under Finiteness or (strict) Convexity, any Bilateral Equilibrium of the game features structural homophily. I also have shown that structural homophily has empirical implications. I develop an estimation technique that can be used to estimate some parameters of the model, namely the weights of the distance function. I can then identify which social characteristics significantly influence the network formation process. Being able to estimate the magnitude of these relevant characteristics is an important step in the process of designing efficient policies, as it allows the policy makers to target relevant characteristics. To illustrate this method, I estimated the weights of the distance function in the context of friendship networks for teenagers. I found significant differences in the homophilic preference bias between racial groups.

The model developed in this paper is a first step toward a better understanding of network formation processes under time constraints. However, there are still many unanswered questions. For instance, the results in section 3 are based on the Finiteness or Convexity assumption. Those are arguably strong assumptions as they imply that individuals invest as much as they can in their existing links. This may not be true in general. However, the study of the model under a concavity assumption faces difficult existence issues. One could address this issue by considering weaker solution concepts such as Pairwise Stability (Jackson and Wolinsky, 1996) which potentially exhibit less structured equilibrium networks.

Another potential extension would be to introduce probabilities of meeting between individuals. Without meeting probabilities, the set of potential friends is the same for every individuals, i.e. the whole population. In general, in large population, some individuals may not know themselves, which would obviously prevent them from creating a link. A simple way to introduce meeting probabilities would be to assume that the set of potential friends is limited to individuals that have "met". Hence, individuals can only invest resources in links with individuals in a

subset of the population. In that case, the (ex-post) strategy space would not be the same for every individual, but structural homophily would still hold in equilibrium (provided that the set of potential friends is known). More elaborate models could however assume that meeting friends is a costly process. The individuals would then be allowed to endogenously choose the amount of resource they spend searching for friends.[35] As the estimation technique does not require the observation of the time constraints, structural homophily is likely to hold in equilibrium. However, in both extensions, the estimated parameters may not be interpreted in terms of preferences. If homophily affects the preferences *and* the random meeting process, it is unclear how those two effects can be identified.

[35] A nice example of a search model with homophilic preferences is Currarini et al. (2009).

CHAPITRE 3

MY FRIEND FAR FAR AWAY : ASYMPTOTIC PROPERTIES OF PAIRWISE STABLE NETWORKS[1]

3.1 Introduction

How do social networks form ? Specifically, how can we measure the influence of an individual's socioeconomic characteristics on the identity of his peers ? We know that many social networks exhibit strong racial or religious segregation (see for instance Echenique and Fryer 2007, Watts 2006, and Mele 2007). This observation raises many interesting questions regarding the cause of this segregation. For instance, we would like to be able to distinguish the impact of the individuals' characteristics (e.g. race), and the impact of the individuals' positions in the networks (e.g. popularity). The shape of the existing social networks also have measurable effects on individuals' choices. Many studies show a strong influence of an individual's peers on his actions, ranging from unhealthy consumption choices (e.g. Fortin and Yazbeck 2011 and the references therein) to labor force participation (e.g. van der Leij et al. 2009, and Patacchini and Zenou 2012). However, since most social networks are endogenously formed, the estimated influence of peers is likely to be biased.[2] Understanding how the networks are formed could then allow us to control for this endogeneity and suggest policy instruments that would help influence network formation processes.

In this paper, we provide a simple Maximum Likelihood estimator which allows us to recover underlying preference parameters for pairwise stable networks (Jackson and Wolinsky, 1996). The approach is compelling as it only requires the observation of a single network. We show that the estimator is consistent and asymptotically normally distributed provided that individuals' preferences exhibit

[1] This chapter is a joint work with Ismael Mourifié.

[2] The literature on peer effects have only recently considered explicitly the endogeneity of social networks. See for instance Goldsmith-Pinkham and Imbens (2011), and Blume et al. (2011).

a weak version of *homophily*. Homophily is one of the most robust empirical fact about social networks. It formalizes the observation that similar individuals are more likely to interact with each other. As homophily is featured by both theoretical (e.g. Bramoullé et al. 2012, and Currarini et al. 2009), and empirical (e.g. Mele 2007, and Christakis et al. 2010) models of network formation, our methodology is applicable to many existing models of network formation. We apply this new methodology to the formation of communication networks, using a database on the Yahoo!'s Instant Messaging service. We find that the probability that a link is created is strongly influenced by the local density of the network, by their general Internet usage, and by their socio-economic and Internet behavior differences.

A fundamental challenge in estimating a network formation process is the highly dependent nature of most socio-economic relationships. Consider for instance the case of friendship networks. The probability that Adam and Beth are friends depends on their individual characteristics. However, it may also depends on the fact that Beth is friend with Charlotte (who maybe does not like Adam). The probability that Adam and Beth are friends may then depend on Charlotte's individual characteristics. Hence, the observation "Adam and Beth are friends" depends on Charlotte's characteristics. However, if individuals have homophilic preferences, the probability that Adam and Beth are friends should be primarily influenced by individuals similar to them. If Adam and Beth are high-school teenagers for instance, the probability that they become friends increases if they go the the same school, or if they attend the same classes. Accordingly, if Beth and Charlotte are friends, there is a greater probability that they go to the same school, or at least that they live in the same country. Then, Donald, a elderly man, living in a different country (hence having individual characteristics quite different from those of Adam, Beth and Charlotte) probably does not influence much the probability that Adam and Beth become friends. We generalize this argument and show that homophily implies a generalization of the ϕ-mixing property used in time-series and spatial econometric models. This fact allows us to define a consistent estimation strategy based on a Quasi Maximum Likelihood estimator.

This paper contributes to the empirical literature on strategic network formation. Two main approaches have been proposed. The first approach is specifically interested in estimating homophilic preferences (see for instance Boucher 2012, and Currarini et al. 2010) and uses standard frequentist approaches, i.e. standard Maximum Likelihood estimators. As these papers assume ex-ante homophily, they are limited in their scope of applications. Also, the maximum likelihood methods proposed require the observation of many (mostly independent) social networks, which is not always available in existing databases.

The second approach requires the observation of only one network, at one point in time. As the observations are highly dependent, standard maximum likelihood methods are not consistent. Accordingly, most papers use a Bayesian approach, and as the likelihood function cannot usually be written explicitly, most papers rely on simulation methods such as Markov Chain Monte Carlo (in particular Christakis et al. 2010, Mele 2010, and Goldsmith-Pinkham and Imbens 2011). If they put less restrictions on the individuals' preferences, those methods are however quite complex to implement in practice, and the computing time needed makes them unsuitable for large database.

We contribute to this literature by providing a explicit Quasi Maximum Likelihood Estimator (QMLE) when we observe only one social network, at one point in time. We introduce a weakened notion of homophily, and show that it implies that our QMLE is consistent and asymptotically normally distributed. In order to do so, we use Large Laws of Numbers and Central Limit Theorems due to Jenish and Prucha (2009), as well as estimators for the variance-covariance matrices due to Conley (1999) and Bester et al. (2011).

The remaining of the paper is organized as follows. In section 3.2.1, we present the economy. In section 3.2.2, we propose an estimator of the equilibrium social network which allows to recover the underlying individuals' preferences. In section 3.3, we derive the asymptotic distribution of our estimator, and in section 3.4, we define a class of network formation models suited to our econometric framework. In section 3.5, we provide an application using the formation of online communication

network, and we discuss policy-making implications and potential avenues for future research in section 3.6.

3.2 The basic framework

3.2.1 The Economy

Let $N = \{1, ..., n\}$ be the set of individuals. Each individual is characterized by a random vector of $T \geq 1$ characteristics $x_i = (x_i^1, ..., x_i^T) \in \mathcal{X}$. We assume that $\mathcal{X} \subset \mathbb{R}^T$ and we define the distance between two individuals as $d(i, j) = d(x_i, x_j)$, where d is a distance on $\mathbb{R}^T$. Finally, we note $x = (x_1, ..., x_n) \in \mathcal{X}^n$ the matrix of individual characteristics. Is it worth noting that the choice of the distance function d is arbitrary. In general, the choice of this distance function will be context-dependent. In particular, the distance can represent spatial preferences of the individuals.[3] We provide an example in section 3.5.

Let $m = \frac{n(n-1)}{2}$ be the number of possible pairs of individuals (i, j) for $i \neq j$ in the economy. We assume that individuals interact in a network $g_m = (N, \mathbf{W})$, where $\mathbf{W}$ is a $n \times n$ symmetric matrix that takes values $w_{ij} = 1$ if $i \in N$ and $j \in N$ are linked by a socio-economic relationship (e.g. friendship), and $w_{ij} = 0$ otherwise. For a given set of individuals N, the set of all possible networks is noted $\mathbb{G}_m$. For a given network $g_m \in \mathbb{G}_m$, we will note $ij \in g_m$ if $w_{ij} = 1$. We will also denote by $g - ij$, the network g_m from which we removed the link between i and j. If $ij \notin g_m$, then $g_m - ij = g_m$. We define $g_m + ij$ similarly.

The set of links an individual has is noted $N_i(g) = \{j \in N : ij \in g_m\}$. The cardinality of that set is the *degree* of an individual, formally $n_i(g_m) = |N_i(g_m)|$. The *geodesic distance* (or shortest path) between i and j in the network g_m equals the minimal number of existing links in g_m such that j can be reached from i. Let $\rho_{ij}(g_m)$ be the geodesic distance between i and j in the network g_m. We say that i and j are *connected* in g_m if $\rho_{ij}(g_m) < \infty$. If i and j are not connected, we let $\rho_{ij}(g_m) = \infty$. Let $R_{ij}^{g_m} = \{k \in N | \min(\rho_{ik}(g_m), \rho_{jk}(g_m)) < \infty\}$ be the set of

[3] See in particular Henry and Mourifié (2011) for spatial preferences on the euclidean space.

individuals connected either to i or to j. For $V \subset N$, we note $g_{m|V}$ the network restricted to individuals in V, i.e. for all $i, j \in V$, we have $(w_{|V})_{ij} = w_{ij}$, while we have $(w_{|V})_{ij} = 0$ if $i \in N \setminus V$ or $j \in N \setminus V$. Let also $x_V \in \mathcal{X}^{|V|}$ be the matrix of individual characteristics of individuals in V.

We assume that the network $g_m = (N, \mathbf{W})$ is endogenous and determined as a function of the individuals' (stochastic) utilities. An individual has preferences over the set of characteristics and the network structure in the economy, i.e. $u_i : \mathbb{G}_m \times \mathcal{X}^n \to \mathbb{R}$. Specifically, we write $u_i(g_m, x; \theta, \varepsilon_i)$ where $\theta \in (\theta^1, ..., \theta^K) \in \Theta$ is the set of parameters to be estimated, and the vector $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{in})$ is the unobserved component of the utility function. It will be convenient to use the following representation of the utility function.

Definition 6 *Given g_m and x, the value for $i \in N$ of a link with $j \in N \setminus \{i\}$ is given by*

$$H_i^j(g_m, x; \theta, \varepsilon_i) = u_i(g_m, x; \theta, \varepsilon_i) - u_i(g_m - ij, x; \theta, \varepsilon_i)$$

Given $H_i^j(g_m, x; \theta, \varepsilon_i)$ for all $i, j \in N$, we want to know what information can be retrieved from the observation of a single network $g_m \in \mathbb{G}_m$, and a set of individual characteristics $x \in \mathcal{X}^n$. We concentrate on the properties of the network g_m and not on the specific dynamic process by which the network is created. For instance, we do not require the links to be added in a specific order to the network. We rather assume that the observed network g_m is *stable*. We are interested in a particular notion of stability, introduced by Jackson and Wolinsky (1996).

Definition 7 *A network g_m is **Pairwise Stable** if, for all $i, j \in N$, the two following conditions hold simultaneously :*
1) if $w_{ij} = 1$ then [$H_i^j(g_m, x; \theta, \varepsilon_i) \geq 0$ and $H_j^i(g_m, x; \theta, \varepsilon_j) \geq 0$]
2) if $w_{ij} = 0$ then [$H_i^j(g_m + ij, x; \theta, \varepsilon_i) > 0$ implies $H_j^i(g_m + ij, x; \theta, \varepsilon_j) < 0$]

Then, a link is created iff it is profitable for both individuals involved.[4] Let $PSN \subseteq \mathbb{G}_m$ be the set of pairwise stable networks. The existence and multiplicity

[4]Notice that conditions 1 and 2 of definition 7 are mutually exclusives as $w_{ij} \in \{0, 1\}$.

of equilibria are discussed in section 3.4.3. For now, assume that there exists a unique pairwise stable network. Pairwise stability is extensively used in the literature on strategic network formation.[5] Any potential deviation from a pairwise stable network results from a *single* pair of individuals changing the status of *its* link. That is, any admissible deviation is such that $g_m \in \mathbb{G}_m$ goes from g_m to $g_m + ij$ for some $i, j \in N$, or from g_m to $g_m - ij$ for some $i, j \in N$. Pairwise stability can then be viewed as the weakest bilateral extension from the set of individually rational networks.[6] We study the asymptotic properties of pairwise stable networks. In the next section, we present the econometric framework.

3.2.2 The Econometric Framework

We want to know what information can be retrieved from the observation of a single pairwise stable network. Specifically, suppose that we observe a set of m pairs of individuals. The set of pairs is noted S_m, with typical elements s and r. Any two individuals i and j necessarily belong to some pair s, where $s = (s_1, s_2) = (i, j)$. For each pair, we observe the *status* (linked or not) of the pair and the socio-economic characteristics of the individuals in the pair (age, gender, income...). We formally define the position of a pair $s \in S_m$ in $\mathcal{X}$ as the average point between s_1 and s_2, i.e. $x_s \in \mathcal{X}$ such that $x_s = \frac{x_{s_1} + x_{s_2}}{2}$.[7] Accordingly, the distance between two pairs r and s is equal to $d(s, r) = d(x_r, x_s) = d(\frac{s_1+s_2}{2}, \frac{r_1+r_2}{2})$.

In this section, we show that pairwise stability allows to express the probability of a link's status in terms of the observable socio-economic characteristics. We present our first assumption.

Assumption 3 (Preferences) *For all $i, j \in N$,*
(3.1) $H_i^j(g_m, x; \theta, \varepsilon_i) = h_i^j(g_m, x; \theta) + \varepsilon_{ij}$, *with* $\varepsilon_{ij} | g_m, x \sim N(0, 1)$.

[5] See for instance Jackson (2008, chapter 6).

[6] For comparisons between stability concepts on networks, see for instance Bloch and Jackson (2006) and Chakrabarti and Gilles (2007).

[7] This is done without loss of generality. The method is robust to other definitions of a pair's position in $\mathcal{X}$, as long as x_s is located in a given neighbourhood of x_{s_1} and x_{s_2}.

(3.2) $h_i^j(g_m, x; \theta)$ *is three times continuously differentiable in* θ*.*
(3.3) Θ *is a compact subset of* $\mathbb{R}^K$*, for* $K \geq 1$*.*

Assumption 3.2 and 3.3 are standard technical requirements. Assumption 3.1 deserves more attention. The error term ε_{ij} is interpreted as a random shock on the value of the pair, hence $\varepsilon_{ij} = \varepsilon_{ji}$. The separability of the error term is quite standard (see Additive Random Utility Models, following McFadden, 1981). Also, as our endogenous variable (i.e. the status of a pair) is discrete, only scale-identification can be achieved. There is then no loss of generality in normalizing the variance of the error term. We assume that ε_{ij} follows a normal distribution for convenience (for instance, it allows to present our estimator as a standard Probit, see below). In general, our method can be adapted to many distributional assumptions. In particular, all our results are valid for any distribution for which the left-tail of the cdf distribution is exponentially bounded. Notice that while the ε_{ij} are identically distributed, they are not necessarily independent.

We want to estimate $\theta \in \Theta$, given the fact that the observed network g_m is pairwise stable. Given definition 2, a link ij is created (i.e. $w_{ij} = 1$) if and only if $H_i^j(g_m, x; \theta, \varepsilon_i) \geq 0$ and $H_j^i(g_m, x; \theta, \varepsilon_j) \geq 0$. Then, under assumption 3.1, the probability that $w_{ij} = 1$ for $i, j \in N$ is equal to $\Phi(\min\{h_i^j(g_m, x; \theta), h_j^i(g_m, x; \theta)\})$, where Φ is the c.d.f. for the standardized normal distribution. We then propose the following QMLE.

$$\begin{aligned} \mathcal{L}_m(\theta) &= \frac{1}{m} \sum_{ij:i<j} w_{ij} \ln[\Phi(\min\{h_i^j(g_m, x; \theta), h_j^i(g_m, x; \theta)\})] \\ &+ (1 - w_{ij}) \ln[1 - \Phi(\min\{h_i^j(g_m + ij, x; \theta), h_j^i(g_m + ij, x; \theta)\})] \quad (3.1) \end{aligned}$$

This is actually a standard probit model.[8] However, the estimator $\hat{\theta} = argmax_{\theta \in \Theta} \mathcal{L}_m(\theta)$ is not necessarily consistent (as $m \to \infty$) since the observations can be dependent. For instance, $h_i^j(g_m, x; \theta)$ may depends on the number of

[8]Notice that $P(w_{ij} = 0) + P(w_{ij} = 1) = 1$ since the two conditions in definition 7 are mutually exclusives.

links i and j have in the network g_m. In the next section, we find sufficient conditions for the consistency and asymptotic normality of $\hat{\theta} = argmax_{\theta \in \Theta} \mathcal{L}_m(\theta)$ when the number of pairs m goes to infinity.

3.3 Limited Dependence Theorems

In this section, we present two theorems for dependent observations. We show that under ϕ-mixing, $\theta \in \Theta$ can be consistently estimated using the model in (3.1). Those theorems are useful since, as we show in section 4, there exist simple conditions on h_i^j which imply ϕ-mixing.[9]

We start by introducing the following random variable, for all pairs $s \in S_m$:

$$Z_{s,m} = \begin{cases} 1 \text{ if } H_{s_1}^{s_2}(g_m, x; \theta, \varepsilon_{s_1}) \geq 0 \text{ and } H_{s_2}^{s_1}(g_m, x; \theta, \varepsilon_{s_2}) \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

The random field $\{Z_{s,m}; s \in S_m, m \in \mathbb{N}\}$ is defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \{0,1\}^m$, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure on Ω. To clarify the exposition, we use the following simplifying notation :

$$\begin{aligned} q_{s,m}(z_{s,m}|x, g_m, \theta) &= w_s \ln[\Phi(\min\{h_{s_1}^{s_2}(g_m, x; \theta), h_{s_2}^{s_1}(g_m, x; \theta)\})] \\ &+ (1 - w_s) \ln[1 - \Phi(\min\{h_{s_1}^{s_2}(g_m + s, x; \theta), h_{s_2}^{s_1}(g_m + s, x; \theta)\})] \end{aligned}$$

so (3.1) can be written as :

$$\mathcal{L}_m(\theta) = \frac{1}{m} \sum_{s \in S_m} q_{s,m}(z_{s,m}|x, g_m, \theta) \tag{3.2}$$

We also use $q_{s,m}(\theta) = q_{s,m}(z_{s,m}|x, g_m, \theta)$ when there is no ambiguity.

We now turn to the dependence structure of (3.2). For any two events $A \in \mathcal{A}$ and $B \in \mathcal{B}$, where $\mathcal{A}, \mathcal{B}$ are sub-σ-algebras of $\mathcal{F}$, the ϕ-mixing coefficient is given

[9] Our results can easily be adapted to other mixing definitions such as α-mixing.

by

$$\phi(\mathcal{A},\mathcal{B}) = \sup\{|\mathbb{P}(A|B) - \mathbb{P}(A)|, A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(B) > 0\}$$

This is analog to standard time-series models. In a time dependent model, the estimation is consistent if $\lim_{r\to\infty} \sup_t \phi(\mathcal{F}^t_{-\infty}, \mathcal{F}^\infty_{t+r}) = 0$, where $\mathcal{F}^{t_2}_{t_1}$ is the σ-algebra for the realizations from time t_1 to time t_2.[10] We want to apply the same basic approach when the dependence between A and B goes through $\mathcal{X}$. Then, instead of characterizing an observation by its position in time, we define it by its position in $\mathcal{X}$. Since the dependence in $\mathcal{X}$ is more complex than time-dependence, the asymptotic convergence of the ϕ-mixing coefficient is not sufficient. In order to show the consistency and asymptotic normality of $\hat{\theta} = argmax_\theta \mathcal{L}_m(\theta)$, we use Large Laws of Numbers and Central Limit theorems for dependent observations on random fields developped by Jenish and Prucha (2009, Theorems 1,2 and 3). Lets introduce the following definition.

Definition 8 *Let $A, B \subset \Omega$, with corresponding σ-algebra $\mathcal{A}_m$ and $\mathcal{B}_m$. Let also $|A|$ and $|B|$ denote the number of pairs of individuals in A and B. We define the following function :*

$$\bar{\phi}_{k,l}(r) = \sup_m \sup_{A,B} (\phi(\mathcal{A}_m, \mathcal{B}_m), |A| \leq k, |B| \leq l, d(A,B) \geq r)$$

where $d(A,B)$ is the Hausdorff distance on $\mathcal{X}$ for the set of pairs in A and B.

We will show that a sufficient condition for the consistency and the asymptotic normality of $\hat{\theta} = argmax_\theta \mathcal{L}_m(\theta)$ is the following :

Assumption 4 (ϕ-mixing)
(4.1) $\sum_{r=1}^\infty r^{T-1} \bar{\phi}^{1/2}_{1,1}(r) < \infty$
(4.2) $\sum_{r=1}^\infty r^{T-1} \bar{\phi}_{k,l}(r) < \infty$, *for* $k + l \leq 4$
(4.3) $\bar{\phi}_{1,\infty}(r) = O(r^{-T-\epsilon})$ *for some* $\epsilon > 0$.

Recall that $T \geq 1$ is the dimension of $\mathcal{X}$. In words, not only $\bar{\phi}_{k,l}(r)$ has to converge to 0, but this convergence has to be *fast enough*. In section 4, we give

[10] See for instance White (2001).

sufficient conditions under which assumption 4 holds. For the moment, we show the validity of the estimation technique given that ϕ-mixing is respected. The first theorem concerns the consistency of $\hat{\theta} = arg\max_{\theta\in\Theta}\mathcal{L}_m(\theta)$. First, we need some regularity conditions.

Assumption 5 (Regularity I)

(5.1) There exists a unique $\theta_0 \in int\ \Theta$ *maximizing* $\lim_{m\to\infty}\mathbb{E}[\mathcal{L}_m(\theta)]$.

(5.2) For all $s_1, s_2 \in N$, $d(s_1, s_2) \geq d_0$ *for some* $d_0 > 0$.

(5.3) $\sup_m \sup_s \mathbb{E}[\sup_{\theta\in\Theta} |q_{s,m}(\theta)|^{(1+\eta)}] < \infty$ *for some* $\eta > 0$.

(5.4) $\sup_m \sup_s \mathbb{E}[\sup_{\theta\in\Theta} |\frac{\partial q_{m,s}(\theta)}{\partial\theta}|] < \infty$.

Assumption 5.1 is the identification condition. Assumption 5.2 is the *increasing domain* assumption. It ensures that the distance goes to infinity as the number of individuals goes to infinity. Given the existence of a minimal distance d_0, the sub-space of $\mathcal{X}$ which contains all the individuals has to expand (with respect to d) as the number of individuals increases. This assumption describes how the space of individual characteristics $\mathcal{X}$ is filled as the number of pairs m goes to infinity. Finally, assumption 5.3 and 5.4 require standard moment conditions on the payoff function. We have the following.

Theorem 3.3.1 (Consistency) *Suppose that assumptions 3 and 5 hold, and that assumption (4.2) is respected for* $k = l = 1$. *Then, the estimator* $\hat{\theta} = arg\max_{\theta\in\Theta}\mathcal{L}_m(\theta)$ *is consistent as* $m \to \infty$.

We still need to derive the asymptotic distribution of $\hat{\theta}$. We define the following matrices :

$$\begin{aligned} D_0(\theta_0) &= lim_{m\to\infty}\mathbb{E}[\frac{\partial^2\mathcal{L}_m(\theta_0)}{\partial\theta\partial\theta'}] \\ B_0(\theta_0) &= lim_{m\to\infty} m\mathbb{E}[\frac{\partial\mathcal{L}_m(\theta_0)}{\partial\theta}\left(\frac{\partial\mathcal{L}_m(\theta_0)}{\partial\theta}\right)'] \end{aligned}$$

Now, since the asymptotic normality of the estimator requires more structure than the one needed for consistency, we need assumptions 4.1-4.3, as well as the following additional regularity conditions.

Assumption 6 (Regularity II)

(6.1) $B_0(\theta_0) > 0$.

(6.2) $D_0(\theta_0)$ *is invertible.*

(6.3) $\sup_m \sup_s \mathbb{E}[\sup_{\theta\in\Theta} \|D_{m,s}(\theta)\|^{1+\eta}] < \infty$ *for some* $\eta > 0$.

(6.4) $\sup_m \sup_s \mathbb{E}[\sup_{\theta\in\Theta} \|\frac{\partial D_{m,s}(\theta)}{\partial\theta}\|] < \infty$.

(6.5) $\sup_m \sup_s \mathbb{E}[\sup_{\theta\in\Theta} |\frac{\partial q_{s,m}(\theta)}{\partial\theta}|^2] < \infty$

where $D_{m,s}(\theta) = \frac{\partial^2 q_{s,m}(\theta)}{\partial\theta\partial\theta'}$. Those assumptions are quite standard and are sufficient to show the asymptotic normality of our estimator.[11]

Theorem 3.3.2 (Asymptotic Normality) *Let* $m \to \infty$. *Under assumptions 3, 4, 5 and 6, the estimator* $\hat{\theta} = arg\max_{\theta\in\Theta} \mathcal{L}_m(\theta)$ *is normally distributed with variance-covariance matrix given by* $D_0^{-1}B_0D_0^{-1}/m$.

The Variance-Covariance Matrix is the equivalent for our setting of the Heteroskedasticity and Autocorrelation Consistent (HAC) variance-covariance matrix. The estimation of those variances is not straightforward. The estimation of $D_0(\theta_0)$ follows from theorems 3.3.1 and 3.3.2 since $D_0(\theta)$ has the same dependence structure as $\lim_{m\to\infty} \mathbb{E}\mathcal{L}_m(\theta)$. A consistent estimator is then $D_m(\hat{\theta}) = \frac{1}{m}\sum_{s=1}^{m} D_{s,m}(\hat{\theta})$. Defining a consistent estimator for $B_0(\theta_0)$ is more challenging. We suggest two approaches to estimate $B_0(\theta_0)$. The first one is based on a generalization of standard HAC estimators and is due to Conley (1999). The estimator $B_m(\theta)$ is formally described in the appendix II. Under mixing conditions, $B_m(\theta)$ is a consistent estimator for $B_0(\theta_0)$. Although valid, this estimator can be very computationally intensive when the number of dimensions of $\mathcal{X}$ increases (say, $T \geq 4$). An alternative approach have been suggest by Bester et al. (2011), where they propose to use the well known Variance Cluster (VC) estimator (also formally described in appendix II). Although the estimator is not consistent under weak dependence, they show that the estimator converges to a well defined random variable and that the standard t-test are still valid. In other words, under mixing conditions, inference using the

[11] Formally, the proof of theorem 3.3.2 derives the limit distribution for $\sqrt{m}(\hat{\theta} - \theta_0)$. We report the asymptotic distribution of $\hat{\theta}$ for presentation purposes.

VC estimator is valid, even if the estimator itself is not consistent. This estimator has the advantage of requiring little computational time and to be simple to implement.

In this section, we have shown that under ϕ-mixing and some regularity conditions, $\theta \in \Theta$ can be recovered using (3.1). In the next section, we show that an asymptotic version of the homophily principle is a sufficient condition for ϕ-mixing, as defined in assumption 4.

3.4 Models of network formation

3.4.1 A First Example

We now turn to economic models of network formation. We want to find sufficient conditions on $h_i^j(g_m, x; \theta)$ such that assumption 4 holds. To clarify the presentation, we start with a simple example. Assume for the moment that

$$h_i^j = h_i^j[N_i(g_m), N_j(g_m), d(i,j)]. \tag{3.3}$$

That is, the value of a link depends only on the (direct) links the individuals have, and the distance between them. Given this specific dependence structure, we will show that a weak version of the homophily principle is sufficient to achieve ϕ-mixing.

Homophily is a prominent feature of social networks. It characterizes the empirical fact that similar individuals have a higher probability of being linked.[12] We assume the following :

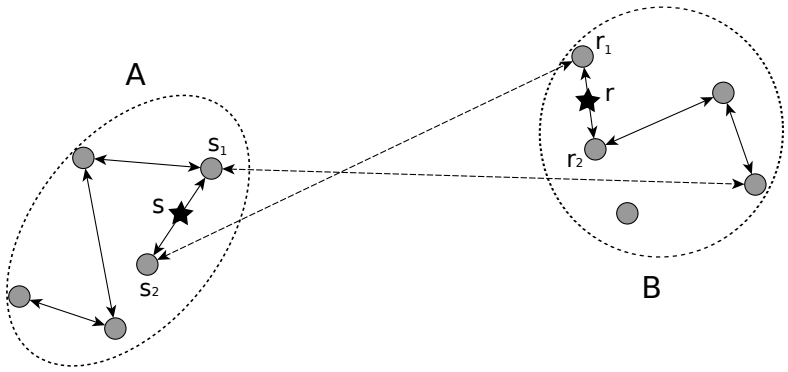
Assumption 7 (Asymptotic Homophily) *For all* $i, j \in N$,
(7.1) $h_i^j(g_m, x; \theta) \to -\infty$ *as* $d(i,j) \to \infty$.
(7.2) $lim_{d\to\infty} \exp\left\{-\frac{h_i^j(g_m,x;\theta)^2}{2d}\right\} \in [0,1)$.

[12] Many definitions of homophily exist in the economic literature, see for instance Currarini et al. (2009) and Bramoullé et al. (2012). In particular, some papers explicitly define homophily using a distance function on the space of individual characteristics : for instance, Johnson and Gilles (2000), Marmaros and Sacerdote (2006), and Iijima and Kamada (2010).

Assumption (7.1) simply says that if the distance between two individuals is infinite, the probability that they form a link is equal to 0. Condition (7.2) limits the asymptotic concavity of h_i^j in d. For example, suppose that $h_i^j(d) = O(d^\eta)$ for some η. Then, assumption 7.2 holds if $\eta > \frac{1}{2}$, but not if $\eta \leq \frac{1}{2}$. Notice that assumption 7 only requires that homophily holds asymptotically hence allowing for a wide range of non-homophilic preferences. We provide an example in section 5.

We show that, under the specification in (3.3), Asymptotic Homophily is sufficient for ϕ-mixing. Before we present the formal result, we provide a graphical intuition. Consider Figure 3.1, where we assumed that $\mathcal{X} = \mathbb{R}^2$. Individuals are represented as circles, and pairs as stars.

Figure 3.1 – ϕ-mixing on Networks (I)



The ϕ-mixing condition says that, as the distance between A and B tends to infinity, the realizations on A and B (i.e. the status of the pairs within those subsets) are independent. Consider pairs s and r. As the distance between r and s increases, the distance between the individuals within those pairs (i.e. s_1, s_2 and r_1, r_2) increases as well. Under assumption 7, as the distance between, s_2 and r_1 goes to infinity, the probability that they form a link goes to zero. Since, under the specification in (3.3), payoffs only depends on direct links, the status of s will therefore be independent of the status of r. The argument holds for any pairs in A and B.

Before presenting the formal statement, we need to add one more regularity assumption. Recall that a necessary condition for theorems 3.3.1 and 3.3.2 was

the existence of a minimal distance d_0. However, in order to show that asymptotic homophily is sufficient for ϕ-mixing, we need to be more specific about the how the space of individual characteristics is filled as the number of individuals goes to infinity. Specifically, we assume :

Assumption 8 $\lim_{m\to\infty} m d_m^{T+\epsilon}\eta^{d_m} < \infty$ *for all* $\eta \in [0,1)$ *and for some* $\epsilon > 0$.

where d_m represent the fact that the distance increases as $m \to \infty$ (increasing domain).[13] This is in essence a distributional assumption for the individuals in $\mathcal{X}$. It requires that the tails of the distributions are large enough. If the distribution of individuals on the type is too concentrated, the mixing coefficient $\bar{\phi}_{1,\infty}(r)$ will decrease as m increases, but not enough for assumption 4 to hold. Given this last regularity assumption, we have the following :

Proposition 3.4.1 *Let* $m \to \infty$. *Suppose that the payoff function is given by (3.3) for all* $i, j \in N$. *Then, assumptions 3, 7 and 8 imply assumption 4.*

When the payoffs are only dependent through direct links, it is sufficient to show that the probability of a link between an individual in a pair in A and an individual in a pair in B goes to zero fast enough. Since we assumed (assumption 3) that the error term is normally distributed, this probability decreases at exponential rate, which is sufficiently *fast* in the sense of assumption 4.

Assumption 7 is quite natural, and allows to adapt many known theoretical models to our setting. Consider for instance the "Local Spillover" model from Goyal and Joshi (2006) :[14]

$$h_i^j(g_m, x) = \psi(n_i(g_m) - 1, n_j(g_m) - 1) - c_{ij}$$

where $\psi : \mathbb{N}^2 \to \mathbb{R}$, and c_{ij} is some positive constant. In this example, the value of a link between i and j is equal to a function of the number of links they have, minus

[13] Specifically assumption 8 must be satisfied for any sequence d_m.

[14] Formally, we are assuming the homogeneity of the function ψ, compared to their original model.

a link-dependent cost. One could adapt their model, and introduce the observed heterogeneity by letting $c_{ij} = d(i,j)$, i.e. a cost equal to the distance between the two individuals in $\mathcal{X}$. Doing so would guarantee the Asymptotic Homophily assumption. We now turn to more general network formation processes.

3.4.2 More General Models

Proposition 3.4.1 provides a first encouraging result for the estimation of preferences on networks. However, the specification in (3.3) excludes many interesting models of network formation. For instance, one could be interested in the following model. Let $\mathbf{C}(g_m, \lambda) = (\mathbf{I} - \lambda \mathbf{W})^{-\mathbf{1}}\mathbf{W1}$ be the $n \times 1$ vector of Bonacich centrality in the network g, represented by the adjacency matrix $\mathbf{W}$, for some $\lambda \in (0,1)$. The Bonacich centrality accounts for the total number of links (direct and indirect) an individual has, and can be interpreted as a measure of popularity.[15]

Now, define the payoffs as : $h_i^j = h(c_i(g_m, \lambda), c_j(g_m, \lambda), d(i,j))$. This payoff function does not respect the conditions of proposition 3.4.1 since it depends on indirect links. We will see that we can nonetheless use the same argument to allow for such models. First, we provide some intuition on the class of models which do not respect the ϕ-mixing condition. Suppose that the payoff function is of the following form.[16]

$$h_i^j(g_m, x) = \psi(n_i(g_m), n_j(g_m), L(g_{m,-i-j})) - c_{ij}$$

where $L(g_{m,-i-j}) = \sum_{k \neq i,j} n_k(g_{m,-i-j})$ is the total number of links in the network $g_{m,-i-j}$, obtained from g_m by removing all links individuals i and j have in g_m.[17] In that case, the value of a link depends on the whole network, irrespective of the individuals' characteristics. This model does not have the property that the dependence vanishes as the distance between individuals increases, and hence ϕ-mixing is not respected. In order to achieve ϕ-mixing, we have to limit the dependence to the network structure. Specifically :

[15] See for instance Mihaly (2009).

[16] This is a loose adaptation of the "Playing the Field" model from Goyal and Joshi (2006)

[17] Specifically, $g_{m,-i-j} = g_m - i1 - ... - in - j1 - ... - jn$.

Assumption 9 (Component Dependence) *For all* $i, j \in N$, $h_i^j(g_m, x; \theta) = h_i^j(g_{m|R_{ij}^g}, x_{R_{ij}^g}; \theta)$

This condition states that the dependence through the network is limited to (finitely) connected individuals. Suppose that the number of individuals in the population is finite. Then, the probability that i and j form a link depends only on the characteristics of the individuals in the same component as i or j.[18] When however, the number of individuals (hence the number of pairs) goes to infinity, we may have two individuals connected through an infinite path. Assumption 9 states that, in that case, those individuals can be treated as disconnected. In other words individuals are unaffected by infinitely distant (in the network) neighbors. Most models of network formation respect this condition as they assume some decay factor.[19] Notice that the previous example where $h_i^j(g_m, x) = \psi(n_i(g_m), n_j(g_m), L(g_{m,-i-j})) - c_{ij}$ does not respect assumption 9. Since h_i^j depends on $L(g_{m,-i-j})$, the payoff function may depend on links between individuals not connected to i nor to j.

Now, by analogy to the specification in (3.3), we see that it is sufficient for assumption 4 to hold to show that the probability that any two individuals, say s_2 and r_1 are connected through some path goes to zero, i.e. $P(s_2 \leftrightarrow r_1) \to 0$. However, this probability does not only depend on the individuals in pairs in A and B, but also on the individuals in pairs "between" the sets. Figure 3.2 illustrates.

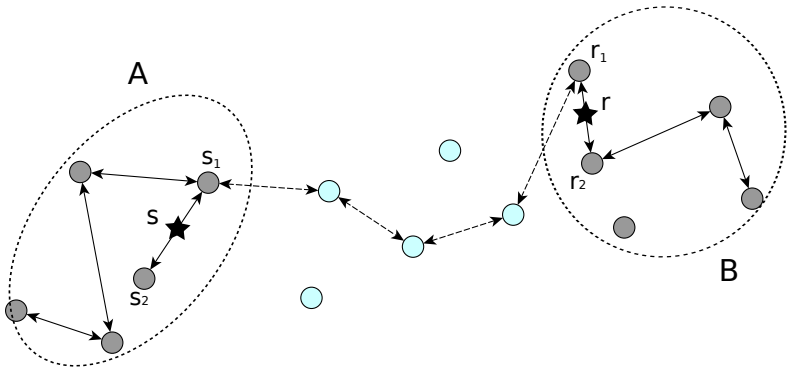
When the number of pairs m (hence the number of individuals n) goes to infinity, there may exists a path of individuals, each of them separated by a finite distance, so $P(A \leftrightarrow B)$ may well be strictly positive. However, since the distance between A and B goes to infinity, this path has to be infinite (i.e. contains an infinite number of individuals). Hence, under assumption 9, the realizations over A and B are independent. Formally,

Proposition 3.4.2 *Assumptions 3, 7, 8 and 9 imply assumption 4 as* $m \to \infty$.

[18] A component is a maximally connected subnetwork.

[19] Links of degree 1 have more influence than links of degree 2, which have more influence than links of degree 3... and so on. Examples include generalizations the Connection Model from Jackson and Wolinsky (1996), and models based on the Bonacich centrality.

Figure 3.2 – ϕ-mixing on Networks (II)



Proposition 3.4.2 shows that the class of models that can be estimated using (3.1) is quite large. It also provide easy to check conditions for applied researchers wanting to estimate some arbitrary model of network formation. In practice, provided that the choosen structural form for $h_i^j(g, x; \theta)$ respects Asymptotic Homophily and Component Dependence, one can estimate $\theta \in \Theta$ using the ML estimator defined in (3.1).

In the next section, we discuss the existence and potential multiplicity of pairwise stable networks.

3.4.3 Existence and Multiplicity

In the previous sections, we implicitly assumed that the set of pairwise stable networks was non-empty, and unique. In general, this may not be true. General conditions for the existence of a pairwise stable network are well known.[20] One result that is particularly adapted to our setting is the fact that monotone preferences imply the existence of at least one pairwise stable network. Formally :

Definition 9 (Monotonicity) *A payoff function is* ***monotone*** *if for any* $g_m, g'_m \in \mathbb{G}_m$ *such that* $g_m \subseteq g'_m$*, we have that* $h_i^j(g_m, x, \theta) \leq h_i^j(g'_m, x, \theta)$ *for all* $i, j \in N$*.*

[20] For general existence results for pairwise stable networks, see Jackson and Watts (2001) and Chakrabarti and Gilles (2007).

Monotone payoff functions have the convenient property that the set of pairwise stable networks is non-empty, irrespective of the value of the unobserved term ε_{ij}. To see why, consider the following simple algorithm. Starting from the empty network, we add links sequentially if $H_i^j(g_m + ij, x; \theta, \varepsilon_i) \geq 0$ and $H_j^i(g_m + ij, x; \theta, \varepsilon_j) \geq 0$. The link creation process stops when there exists no such profitable link creation. Since the payoff function is monotone, the creation of a link increases the value of the existing links so $H_i^j(g + ij, x; \theta, \varepsilon_i) \geq 0$ implies that $H_i^j(g + ij + kl, x; \theta, \varepsilon_i) \geq 0$ for any link kl. The network generated by this sequential creation of links is then pairwise stable.

Another issue that has not been addressed is the potential existence of multiple equilibria.[21] A specific feature of pairwise stable networks is the complexity of the equilibrium set. In general, one cannot explicitly find the set of pairwise stable networks, as showing existence is already challenging. Also, recall that we assumed that we observe only one equilibrium of the game, and not the other (potential) equilibria. Then, in the presence of multiple equilibria, our estimator should not be interpreted as a QMLE, but remains a well defined a extremum estimator, where the objective function is a specific feature of the model : *the probability that the observed network is pairwise stable*. However, the validity of the estimation procedure under the potential presence of other potential equilibria is unclear. Formally understanding the properties of the estimator under multiple equilibria goes far beyond the scope of this paper and is left for future research.

In the next section, we provides an empirical application of our method using communication networks.

3.5 Instant Messaging Networks

In this section, we apply the methodology developed in the previous sections to estimate a model of network formation using a database provided by Yahoo!. The Instant Messaging (IM) database is particularly well suited to our estimation

[21] See Bisin et al. (2011), Galichon and Henry (2011), and Tamer (2003).

strategy as the variance of the observed characteristics is quite large (see Tables 3.II and 3.III).

We use the Yahoo ! IM database which includes data on the communications among users of their IM service. We assume that there exists a link between two individuals if we recorded at least one communication between them. The database includes data on a little more than 20 million individuals. Each individual is characterized by his age, gender, reported country and Internet usage.[22] The precise description of the variables used can be found in Table 3.I.

We use the following structural model, which assumes that the probability that two individuals form a link is explained by the local density of the network, the individuals' general Internet usage (measured by the number of Yahoo's pages viewed) and by the social distance between the individuals. Here, "social distance" means gender, age, geographical distances, and differences in the topic of the Internet pages visited (sports, finance, news...). Specifically, we define :

$$\begin{aligned} h_i^j(g,x;\theta) &= \theta_1(n_i(g)+n_j(g)) + \theta_2(PV_i+PV_j) + \theta_3\Delta(Gender_{ij}) \\ &\quad +\theta_4\Delta(Age_{ij}) + \theta_5\Delta^*(country_{ij}) + \theta_6\Delta(PV_{Weather,ij}) \\ &\quad +\theta_7\Delta(PV_{News,ij}) + \theta_8\Delta(PV_{Finance,ij}) + \theta_9\Delta(PV_{Sports,ij}) \\ &\quad +\theta_{10}\Delta(PV_{Flickr,ij}) + \theta_{11} \end{aligned} \tag{3.4}$$

where $\theta_1 > 0$, and θ_{11} represents the intrinsic value of a link. The restriction $\theta_1 > 0$ is needed to ensure that preferences are monotonic, which implies the existence of a Pairwise Stable network (see section 3.4.3). It's easy to show that under the specification in (3.4), the estimator in (3.1) is globally concave.

[22]Data was collected in October 2007 using a snowball procedure. For a more detailed description of the database, see Sinan et al. (2009a) and Sinan et al. (2009b).

Tableau 3.I – Variables Description

Variable	Variable Name	Description
$n_i(g) + n_j(g)$	Local Density	The total number of links i and j have in the network g.
$PV_i + PV_j$	Total Internet Usage	Total number of Internet pages viewed by i and j.
$\Delta(Gender)$	Gender Distance	Takes value 1 if i and j are of the same gender, and 0 otherwise.
$\Delta(Age)$	Age Distance	Distance (in years) between the ages of i and j.
$\Delta^*(Country)$	Geographic Distance	Geographic distance (in 1000Km) between i and j.[23]
$\Delta(PV_{Weather})$	Weather's Pages Distance	Distance (absolute value) between the percentage of weather pages viewed by i and j.
$\Delta(PV_{News})$	News' Pages Distance	Distance (absolute value) between the percentage of news pages viewed by i and j.
$\Delta(PV_{Finance})$	Finance's Pages Distance	Distance (absolute value) between the percentage of finance pages viewed by i and j.
$\Delta(PV_{Sports})$	Sports' Pages Distance	Distance (absolute value) between the percentage of sports pages viewed by i and j.
$\Delta(PV_{Flickr})$	Flickr's Pages Distance	Distance (absolute value) between the percentage of Flickr pages viewed by i and j.[24]

Now, a particular issue with our database is that it is far too big to be used in its totality.[25] We then use a random sample of the database. The sub-sampling procedure is as follows.

First, we randomly select a subset of pairs. Then, for every individuals in every pairs of the subset, we compute $n_i(g)$ and $n_j(g)$ over the whole sample. Using this procedure, our sub-sample includes the total number of links the individuals have, including links with individuals that are excluded from our sub-sample. The final sub-sample has 74229 pairs of individuals, including 173781 individuals. Table 3.II gives summary statistics for individuals, and Table 3.III gives summary statistics for pairs.

Tableau 3.II – Descriptive Statistics for the individuals

Variable	**Mean**	**Std. Dev.**
$Gender$ $(Female = 1)$	0.598	0.490
Age	29.782	11.494
PV	794.595	1402.652
$PV_{Weather}$	0.241	2.718
PV_{News}	3.807	44.494
$PV_{Finance}$	4.028	115.700
PV_{Sports}	13.139	170.008
PV_{Flickr}	6.590	179.007
$n_i(g)$	1.062	2.400
nb countries	228	
nb indiv.	173 781	

Following Bester et al. (2011), we estimate the specification in (3.4) using Cluster-Robust standard errors.[26] Marginal effects are reported in Table 3.IV. Notice that while we assumed $\theta_1 > 0$ in (3.4), we did not used that restriction for the estimation.

[25] Recall that the number of pairs is $m = n(n-1)/2 \approx (20\ 000\ 000)^2/2 = 2 \times 10^{14}$.

[26] The estimation procedure is simple enough. One can simply use the preprogrammed probit command available in most statistical packages.

Tableau 3.III – Descriptive Statistics for the pairs

Variable	Mean	Std. Dev.
$Links$	0.008	0.090
$n_i(g) + n_j(g)$	4.070	3.333
$PV_i + PV_j$	1.583	1.988
$\Delta(gender)$	0.479	0.500
$\Delta(age)$	10.921	12.058
$\Delta^*(country)(1000Km)$	6.656	5.829
$\Delta(PV_{Weather})$	0.001	0.008
$\Delta(PV_{News})$	0.005	0.028
$\Delta(PV_{Finance})$	0.006	0.046
$\Delta(PV_{Sports})$	0.014	0.085
$\Delta(PV_{Flickr})$	0.007	0.052
nb. pairs	74 229	

3.5.1 Results

The estimation (Table 3.IV) shows that the probability that a link is created is influenced by the local density of the networks (i.e. $n_i(g) + n_j(g)$), as well as by the general Internet usage. Non-surprisingly, frequent Internet users have a higher probability of interacting through the IM service. Interestingly, however, the connectivity of the individuals in the IM network seems to have an additional positive effect on the probability of creating a link. This could reflect the fact that frequent users of the IM service have higher probability of interacting together. The estimation also shows strong effects of the distance on the probability of a link. The positive effect of the variable $\Delta(Gender)$ seems to indicate that the IM service is highly used by heterosexual couples. The effect of the age distance is negative, which is coherent with homophily on with respect to the age of the individuals. The geographic distance seems to have a negative impact, although it is not significantly different from zero. The proximity in the topic of the Internet pages viewed by the users also seems to have a negative effect, however the difference in the percentage of sports pages viewed his captured significantly.

As this application shows, the approach used in this paper is promising as it has the advantage of being intuitive, flexible, and simple to implement.

Tableau 3.IV – Estimation Results (Marginal Effects)

Variable	**Coefficient**	**(Std. Err.)**
$n_i(g) + n_(g)$	0.00049**	(0.00008)
$PV_i + PV_j$	0.00024*	(0.00011)
$\Delta(Gender)$	0.00105*	(0.00052)
$\Delta(Age)$	-0.00006*	(0.00003)
$\Delta^*(Country)$	-0.00004	(0.00005)
$\Delta(PV_{Weather})$	-0.03598	(0.05562)
$\Delta(PV_{News})$	-0.00045	(0.01775)
$\Delta(PV_{Finance})$	-0.00412	(0.00627)
$\Delta(PV_{Sports})$	-0.00918*	(0.00446)
$\Delta(PV_{Flickr})$	0.00124	(0.00457)

Significance levels : $*$: 5% $**$: 1%

3.6 Conclusion and Discussions

In this paper, we have developed a micro-founded econometric model of network formation which requires the observation of only one social network. We have shown that an asymptotic version of homophily is sufficient for ϕ-mixing, which implies that the estimation of the underlying preference parameters can be achieved using a simple Maximum Likelihood estimator. The methodology is appealing as it is simple, and allows to estimate many theoretical models of network formation. We have provided an empirical application using Yahoo ! Instant Messaging database. We have shown that the probability that a link is created is strongly influenced by the local density of the network, by general internet usage, and by the social distance between individuals.

CHAPITRE 4

DO PEERS AFFECT STUDENT ACHIEVEMENT ? EVIDENCE FROM CANADA USING GROUP SIZE VARIATION[1]

4.1 Introduction

Evaluating peer effects in academic achievement is important for parents, teachers and schools. These effects also play a prominent role in policy debates concerning ability tracking, racial integration and school vouchers (for a recent survey, see Epple and Romano 2011). However, despite a growing literature on the subject, the evidence regarding the magnitude of peer effects on student achievement is mixed (*e.g.,* Sacerdote 2001, Hanushek *et al.* 2003, Stinebrickner and Stinebrickner 2006, Ammermueller and Pischke 2009). This lack of consensus partly reflects various econometric issues that any empirical study on peer effects must address. Identifying and estimating peer effects raises three basic challenges. First, the relevant peer groups must be determined. Who interacts with whom ? Second, peer effects must be identified from confounding factors. Especially, spurious correlation between students' outcomes may arise from self-selection into groups and from common unobserved shocks. Third, identifying the precise type of peer effect at work may be hard. Simultaneity, also called the *reflection problem* by Manski (1993), may prevent separating contextual effects, *i.e.*, the influence of peers' characteristics, from the endogenous effect, *i.e.*, the influence of peers' outcome. This issue is important since only the endogenous effect is the source of a *social multiplier*. Researchers have adopted various approaches to solve these three issues ; we discuss the methods and results of previous studies in more detail in the next section. As will be clear, however, there is no simple methodological answer to these three challenges.

[1]This chapter is a joint work with Yann Bramoullé, Habiba Djebbari and Bernard Fortin, and published in The Journal of Applied Econometrics

In this paper, we provide, to our knowledge, the first application of a novel approach developed by Lee (2007) for identifying and estimating peer effects. In principle the approach is promising, as it allows to solve the problem of correlated effects and the reflection problem with standard observational (non-experimental) data. Moreover the exclusion restrictions imposed by the model are explicitly derived from its structural specification and provide natural instruments. The econometric model does rely on a number of crucial assumptions, however, which makes its confrontation to real data particularly important. We empirically assess the approach using original administrative data on test scores at the end of secondary school in the Canadian province of Québec. We investigate the presence of peer effects in student achievement in Mathematics, Science, French, and History. In the process, we also provide new economic insights regarding the sources of identification in the model. This matters in particular to assess its robustness to alternative (non-linear) approaches.

The econometric model relies on three key assumptions. First, individuals interact in groups known to the modeler. This means that the population of students is partitioned in groups (*e.g.*, classes, grade levels) and that students are affected by all their peers in their groups but by none outside of it. This assumption is typical in studies of academic achievement but clearly arises from data constraints. Second, each individual's peer group is everyone in his group *excluding* himself. While this assumption seems innocuous and has been used in most empirical studies, it is a key source of identification in the model, as it will become clear below. In fact, it is a main source of difference between Manski's (1993) and Lee's models. Manski's approach can be interpreted as one in which each individual's peer group *includes* himself.[2] Third, individual outcome is determined by a linear-in-means model with group fixed effects. Thus, the test score of a student is affected by his characteristics and by the average test score and characteristics in his peer group. In addition, it

[2]More precisely, Manski studies a social interactions model which, in terms of identification, has the same properties as a model where individuals interact in groups and each individual is included in his peer group (see Bramoullé *et al.* 2009).

may be affected by any kind of correlated group-level unobservable.

Lee (2007) shows that peer effects are identified in such a framework when there are sufficient groups of different sizes. One important contribution of our paper is to clarify the economic intuition behind identification. Regarding the estimation of parameters, one potentially important limitation of the method, however, is that convergence in distribution of the peer effect estimates may occur at low rates when the average group size is large relative to the number of groups in the sample (Lee 2007). This is also intuitive : excluding the individual or not from his peer group does not change much when its size is relatively large.

Here two remarks are in order. First, these results are to be distinguished from the idea that the group size is a factor in a school's production function (*e.g.*, Krueger 2003). In Lee's model, the effects of group sizes which are separable from the peer effects are controlled for by fixed effects in the structural model. Second, Lee's identification method differs from the *variance contrast* approach developed by Graham (2008). The basic idea in this approach is that peer effects will induce intra-group dependencies in behavior that introduce variance restrictions on the error terms. These restrictions are used to identify the composite (endogenous + contextual) social interaction effects under the assumption that the variance matrix parameters are independent of the reference group size.

We use administrative data on academic achievement for a large sample of secondary schools in the Province of Québec obtained from the Ministry of Education, Recreation and Sports (MERS). Our dependent variables are individual scores on four standardized tests taken in June 2005 (Math, Sciences, French and History) by fourth and fifth grade secondary school students. All 4th and 5th grade students in the province must pass these tests to graduate. One advantage of these data is that all candidates in the province take the same exams, no matter their school and location. This feature effectively allows us to consider test scores as draws from a common underlying distribution. Another advantage is that our sample is

representative and quite large. We have the scores of all students for a 75% random sample of Québec schools which, over the four subjects, yields 194,553 test scores for 116,534 students. In terms of interaction patterns, the structure of the data leads us to make the following natural assumption. We assume that the peer group of a student contains all other students in the same school qualified to take the same test in June 2005. In practice, a small number of students postpone test-taking to August 2005. We extend Lee's methodology in the empirical modeling to address this issue. However, since the difference between observed group sizes and actual group sizes is small, the correction has little effect on the results. Following Lee (2007), we estimate the model in two ways : through generalized instrumental variables (IV) and, under stronger parametric conditions, through conditional maximum likelihood robust to non-normal disturbances (pseudo CML).

Our results are mixed though consistent with the model. We do provide evidence of some endogenous and contextual peer effects. Based on pseudo CML estimates, we find that the endogenous peer effect is positive, significant and quite high in Math (0.83). Moreover it is within the range of previous estimates (see Sacerdote 2011 for a recent survey). However, the effect is smaller and non significant in History (0.64), French (0.30), and Science (-0.23).[3] Endogenous peer effects estimates obtained from IV methods are highly imprecise with our data even in Math. The higher precision of our pseudo CML estimates is consistent with results in Lee (2007) showing that CML estimators are asymptotically more efficient than IV estimators. As regards contextual peer effects, we find evidence that some of them matter, based on both pseudo CML and IV estimators. For instance, results from pseudo CML indicate that interacting with older students (a proxy for repeaters) has a negative effect on own test score in all subjects except Math (not significant).

It is remarkable that even with large average group size relative to the number of groups, we are able to identify some peer effects. However there is also much

[3]The effect of individual characteristics, such as gender, age, and socioeconomic background, on test scores are precisely estimated by either method, and these estimates generally conform to expectations.

dispersion in group sizes within our samples. We suspect that this helps identification. We study this issue systematically through Monte-Carlo simulations. We find that indeed increasing group size dispersion has a positive impact on the precision of estimates.

The remainder of the paper is organized as follows. We discuss past research in section 4.2 and present our econometric model and the estimation methods in section 4.3. We describe our dataset in section 4.4. We present our empirical results in section 4.5 and run Monte Carlo experiments in section 4.6. We conclude in section 4.7.

4.2 Previous research

In this section, we give a brief overview of the recent literature on student achievement and peer effects, and we explain how our study complements and enhances current knowledge on peer interactions in academic outcomes.[4]

As discussed above, measuring peer effects is complex as it raises three basic interrelated problems : the determination of reference groups, the problem of correlated effects and the reflection problem. The choice of reference groups is often severely constrained by the availability of data. In particular, there are still few databases providing information on the students' social networks ; the Add Health dataset is an exception, see *e.g.* Calvo-Armengol *et al.* (2009) and Lin (2010).[5] For this reason, many studies focus on the grade-within-school level (*e.g.*, Hanushek *et al.* 2003, Angrist and Lang 2004). Other studies analyze peer effects at the classroom level (*e.g.*, Kang 2007, Ammermueller and Pischke 2009). The administrative data we use in this study do not provide information on classes or teachers. Therefore, we assume that for each subject the relevant reference group for a student

[4]For two recent comprehensive surveys on peer effects in education, see Sacerdote (2011) and Epple and Romano (2011).

[5]Bramoullé *et al.* (2009) determine conditions under which endogenous and contextual peer effects are identified when students interact through a social network known by the modeler and when correlated effects are fixed within subnetworks. See also section 3.4.2. in this paper.

taking the test contains all other students in the same school who have completed all courses in the subject matter by June 2005. Thus, given that the reference group is likely to include students from other classes, one should probably expect peer effects to be smaller than at the classroom level.[6]

Two main strategies have been used to handle the problem of correlated effects. A first strategy has been to exploit data where students are randomly or quasi-randomly assigned within their groups (*e.g.*, Sacerdote 2001, Zimmerman 2003, Kang 2007). Results on the impact of contextual effects using randomly assigned roommates as peers are usually low though significant. However, Stinebrickner and Stinebrickner (2006) have argued that these studies tend to underestimate true peer effects as the true influence of roommates is unclear. A second strategy uses observational data to estimate peer effects. This approach is usually based on two assumptions. First, fixed effects allow to take correlated effects into account. With cross section data, these effects are usually defined at a level higher than peer groups. Otherwise, peer effects are absorbed in these effects and cannot therefore be identified. For instance, Ammermueller and Pischke (2009) introduce *school* fixed effects to estimate peer effects at the *class* level for fourth grader in six European countries. Contrary to this approach, our model allows to include fixed effects at the *peer group* level even with cross-section data. This is so because each student within a group has his own reference group (since he is excluded from it). The second assumption is that one observes exogenous shocks to peer group composition which allow to identify a composite (endogenous + contextual) peer effect. The strategy uses either cross-section or panel data. With cross-section data, demographic variations across grades but within schools are usually exploited (see Bifulco *et al.* 2010). With panel data, demographic variations across cohorts but within school-grades are usually exploited (see Hanushek *et al.* 2003).

The reflection problem is handled using two main strategies. In most papers,

[6]In fact, at the end of secondary level, classes and teachers are usually different depending on the subject matter taught.

no solution for this difficult problem is provided. Rather, researchers estimate a reduced-form linear-in-means model, and no attempt is made to separate the contextual and endogenous peer effects. Only composite parameters are estimated (Sacerdote 2001, Ammermueller and Pischke 2009). Note however that a number of these papers (often implicitly) assume that there are no contextual effects. In this case, the composite parameter(s) allow(s) to identify the endogenous peer effect. In a second strategy, one uses instruments to obtain consistent estimates of the endogenous peer effect (*e.g.*, Evans *et al.* 1992, Gaviria and Raphael 2001). The problem here is to choose suitable instruments. For instance, Rivkin (2001) argues that the use of metropolitan-wide aggregate variables as instruments in the Evans *et al.* (1992) study exacerbates the biases in peer effect estimates. In our paper, we provide some results based on instrumental methods. However, our instruments are naturally derived from the structure of the model.

In short, various strategies have been proposed to address the three basic issues that occur in the estimation of peer effects. But most rely on strong assumptions that are difficult to motivate and may not hold in practice. Some of them require panel data while others rely on experiments that randomly allocate students within their peer group. This makes the results in Lee (2007) particularly interesting, as they show that both endogenous and contextual peer effects may be fully identified even with observational data in cross-section.

4.3 Econometric model and estimation methods

4.3.1 Econometric model

We review and adapt the structural model suggested by Lee in the context of our application. Lee's model builds on and extends the standard linear-in-means model of peer effects (Moffitt 2001) to groups with various sizes. The set of students $\{i = 1, ...M\}$ is supposed to be partionned into groups of peers indexed by $r = 1, ..., R$. Let M_r be the r^{th} group of peers, of size m_r. All students in the same group have the same number of peers since they interact with all others in the

group. We assume that student i who belongs to group r is excluded from his own reference group. Let M_{ri} be student i's group of peers, of size $m_r - 1$. A peer is any fellow student whose academic performance and personal characteristics may affect i's performance. Let y_{ri} be the test score obtained by student i. Let $\mathbf{x}_{ri}$ be a $1 \times K$ vector of characteristics of i and $\mathbf{X}_r$ be the $m_r \times K$ matrix of individual characteristics. For expository purposes, the model is first presented with a unique characteristic ($K = 1$), defined by his family socio-economic background. Another departure from the linear-in-means model is the inclusion of a term α_r that captures all group invariant unobserved variables (*e.g.*, same learning environment, similar preferences of school or motivation towards education). The error term ϵ_{ri} reflects other unobservable characteristics associated with i.

We do not change any other assumption of the linear-in-means model. In particular, we assume that a student's performance to the standardized test may be affected by the average performance in his group of reference, by his family socioeconomic background, and by the average socioeconomic background in his group. Formally, the basic structural equation is given by :

$$y_{ri} = \alpha_r + \beta \frac{\sum_{j \in M_{ri}} y_{rj}}{m_r - 1} + \gamma x_{ri} + \delta \frac{\sum_{j \in M_{ri}} x_{rj}}{m_r - 1} + \epsilon_{ri}, \qquad \mathbb{E}(\epsilon_{ri} |\, \mathbf{X_r}, m_r, \alpha_r) = 0, \quad (4.1)$$

where β captures the endogenous effect, γ the individual effect and δ the contextual effect. Observe that eq. (4.1) can be derived from the first-order conditions of a choice-theoretic non-cooperative (Nash) model where each student's performance is obtained from the maximisation of his quadratic utility function which depends on his individual characteristics, his performance and his reference group's mean performance and mean characteristics.

Importantly, we assume strict exogeneity of m_r and $\{x_{ri} : i = 1, ..., m_r\}$ *conditional* on the unobserved effect α_r, *i.e.*, $\mathbb{E}(\epsilon_{ri} |\, \mathbf{X_r}, m_r, \alpha_r) = 0$. This exogeneity assumption can notably accommodate situations where peer group size is endogenous.

Suppose that, everything else equal, brighter students attend smaller schools, *i.e.*, schools where the cohort of students eligible to take the province-wide test in the subject matter (our peer groups) is small. In this case, peer group size m_r may well depend on unobserved common characteristics of the student's group, α_r : $\mathbb{E}(\alpha_r|\ \mathbf{X_r}, m_r) \neq 0$. Our model allows for this type of correlation. However, conditional on these common characteristics, peer group size m_r is assumed to be independent of the student's idiosyncratic unobserved characteristics : $\mathbb{E}(\epsilon_{ri}|\ \mathbf{X_r}, m_r, \alpha_r) = 0$. We maintain this assumption throughout our analysis.

To eliminate group-invariant correlated effects, we next apply a *within* transformation to eq. (4.1). In particular, as we noted above, when the effect of group size is separable from peer and individual effects, it is captured by α_r. The model can address the problem of selection or endogenous peer group formation. For instance, school choice may depend on some unobserved factors specific to a school (*e.g.*, reputation, unobserved quality) and determine the type of students who are attracted by these schools. The advantage of the within transformation is that we compare students of the same type. This transformation also allows to control for common environment effects. Resources available at the school level (*e.g.*, teaching, physical infrastructure) may affect the performance of all the students. Again, by comparing students within the same school, we can abstract from these effects. The within reduced form equation for students in the r^{th} group can be written as :

$$y_{ri} - \bar{y}_r = \frac{\gamma - \frac{\delta}{m_r - 1}}{1 + \frac{\beta}{m_r - 1}}(x_{ri} - \bar{x}_r) + \frac{1}{1 + \frac{\beta}{m_r - 1}}(\varepsilon_{ri} - \bar{\varepsilon}_r) \tag{4.2}$$

where means $\bar{y}_r$, $\bar{x}_r$ and $\bar{\epsilon}_r$ are computed over *all* students in the group. Now assume that $\gamma\beta + \delta \neq 0$. Only one composite parameter can be recovered from the reduced form for each group size m_r. At least three sizes are thus necessary to identify the three structural parameters β, γ and δ.[7]

[7]It is easy to show that when $\gamma\beta + \delta = 0$, only γ is identified.

4.3.2 Interpretation of identification

The fact that the parameters of the structural within eq.(4.2) may be fully identified is quite surprising, and deserves some elaboration. Indeed, under the alternative assumption that means are inclusive, that is, $i \in M_{ri}$, peers are the same for everyone in a group $M_{ri} = M_r$, and peer effects cannot be separated out from group fixed effects. So somehow assuming that the individual is excluded from his own peer group allows to solve two difficult identification problems : distinguishing true peer effects from correlated effects and further distinguishing endogenous from contextual peer effects. Intuitively, where does identification come from ?

Suppose first that the endogenous effect is absent $\beta = 0$. Note that each individual has different peers : $i \neq k$ implies that $M_{ri} \neq M_{rk}$. A first key observation is that, within a group, individual attributes x_i are perfectly negatively correlated with mean peer attributes $(\sum_{j \in M_{ri}} x_j)/(m_r - 1)$.[8] Thus, students with an ability above average necessarily have peers with a mean ability below average, and vice versa. If the individual and the contextual effects γ and δ are positive, this negative correlation tends to *reduce* the dispersion in outcomes. In such a group setting, peer effects lower the difference in achievement between high and low ability students.[9] Formally, the impact of the difference in attributes on the difference in outcomes changes from γ to $\gamma - \delta/(m_r - 1)$ when introducing peer effects [see eq. (4.2)]. So variations in group sizes can be used to identify contextual peer effects. The second key observation is that this reduction is stronger in smaller groups. The variance in mean peer attributes is simply higher in smaller groups, reflecting the relatively larger effect of excluding one individual from the mean. And as group size increases, mean peer attributes converge to the group mean, and peer effects have increasingly less bite on how differences in covariates affect differences in outcomes.

Next, consider the reflection problem. Observe that outcomes are subject to a

[8] To see this, observe that $\sum_{j \in M_{ri}} x_j = (\sum_{j \in M_r} x_j) - x_i$. So if $x_i < x_k$ then $\frac{1}{m_r - 1} \sum_{j \in M_{ri}} x_j > \frac{1}{m_r - 1} \sum_{j \in M_{rk}} x_j$.

[9] In contrast if $\gamma > 0$ and $\delta < 0$, this negative correlation helps *amplify* the dispersion in outcomes.

similar negative correlation : within a group, students with grades above average necessarily have peers with grades below average. So if $\beta > 0$, endogenous peer effects lead to a further reduction in outcome dispersion. However, simultaneity now implies that this decrease in impact is *non-linear* in the peer coefficient : from $\gamma - \delta/(m_r - 1)$ to $(\gamma - \delta/(m_r - 1))/(1 + \beta/(m_r - 1))$ [see eq. (4.2)]. The difference in the shapes of impact reduction can then be used to identify endogenous from contextual peer effects.

Finally, this understanding is useful to assess the robustness of the identification strategy to changes in the econometric model. In particular, it is easy to see that if $x_i < x_k$ then the distribution of attributes in i's peer group M_{ri} first-order stochastically dominates the distribution in M_{rk}. So identification is likely to hold, in general, if we replaced the mean in equation (1) by the median, the variance, or many other moments of the distribution.[10]

4.3.3 Treatment of missing values

One problem we face in our sample is that we do not always observe the scores of all students within a group. For instance, some students may postpone test-taking to the next session due to illness. We next use a correction first developed by Davezies et al. (2009) to allow for this possibility. Our setting is one where the total number of students (including those who postpone test-taking) in each group is known, but we only observe the test scores of subsamples N_r of size n_r of each group M_r, with $n_r \leq m_r$ and $\sum_{r=1}^{R} n_r = N$. We assume that a student's decision to postpone exam-taking is random or depends on the observable strictly exogenous variables, conditional on the fixed group effect. We show how to adapt Lee's analysis to this more general setting. Let L_r be the complement of N_r, *i.e.* ,

[10] Of course, one has to address a basic modeling question first, that is, whether the implied model is coherent. A model has this property when a specific nonlinear structure generates a unique solution for outcomes.

$L_r = M_r - N_r$.[11] The structural equation becomes :

$$y_{ri} = \widetilde{\alpha}_r + \beta \frac{\sum_{j \in N_{ri}} y_{rj}}{m_r - 1} + \gamma x_{ri} + \delta \frac{\sum_{j \in N_{ri}} x_{rj}}{m_r - 1} + \epsilon_{ri}, \qquad \mathbb{E}(\epsilon_{ri} |\, \mathbf{X_r}, m_r, \alpha_r) = 0, \quad (4.3)$$

where i now denotes an observed individual in the sample (but not any one in the rth group) and $\widetilde{\alpha}_r = \alpha_r + \beta \frac{\sum_{j \in L_r} y_{rj}}{m_r - 1} + \delta \frac{\sum_{j \in L_r} x_{rj}}{m_r - 1}$ is the new group fixed effect. Under our assumptions, estimators are consistent, even if we do not observe test scores for all students in each group. Moreover, effects stemming from unobserved individuals are the same for all the individuals observed in the sample from the rth group. They are therefore picked up by the group fixed effect. Using the within transformation, one obtains the same equation as (4.2) but where means $\bar{y}_r$, $\bar{x}_r$ and $\bar{\epsilon}_r$ are computed only over all *observed* students in the group.

4.3.4 Estimation methods

4.3.4.1 CML Estimator

We consider estimation under both pseudo Conditional Maximum Likelihood (or CML) and Instrumental Variables (or IV) identification conditions.

To present pseudo CML and IV estimators, it is easier to express eq. (4.3) in matrix notations. We now allow for any number of characteristics, so that γ is a $K \times 1$ vector of individual effects and δ a $K \times 1$ vector of contextual ones. Recall that in this setting, students are affected by all others in their group and by none outside of it. This means that the observed social interactions can be modelled as a $N \times N$ block-diagonal matrix $\mathbf{G} = Diag(\mathbf{G}_1, ..., \mathbf{G}_R)$, such that for all r, $\mathbf{G}_r$ is comprised of elements $g_{rij} = \frac{1}{m_r - 1}$ if $i \neq j$ and $g_{rii} = 0$. In other terms, $\mathbf{G}_r = \frac{1}{m_r - 1}(\iota_{n_r} \iota_{n_r}^{'} - \mathbf{I}_{n_r})$, where ι_{n_r} is a $n_r \times 1$ vector of ones and $\mathbf{I}_{n_r}$ the identity matrix of dimension n_r. Eq. (4.3) can be re-written in matrix form as follows :

$$\mathbf{y}_r = \iota_{n_r} \widetilde{\alpha}_r + \beta \mathbf{G}_r \mathbf{y}_r + \mathbf{X}_r \gamma + \mathbf{G}_r \mathbf{X}_r \delta + \epsilon_r, \quad (4.4)$$

[11] If N_{ri} denotes the group of peers of student i, we also have $L_r = M_{ri} - N_{ri}$.

where $\mathbb{E}(\epsilon_r \mid \mathbf{X}_r, \mathbf{G}_r, \widetilde{\alpha}_r)$=0.

Applying the operator matrix $\mathbf{J}_r = \mathbf{I}_{n_r} - \frac{1}{n_r}\iota_{n_r}\iota_{n_r}^{'}$ allows us to obtain deviations with respect to the mean for the observed group members. Pre-multiplying eq. (4.4) by $\mathbf{J}_r$ eliminates the group fixed effect and yields :

$$\mathbf{J}_r\mathbf{y}_r = \beta\mathbf{J}_r\mathbf{G}_r\mathbf{y}_r + \mathbf{J}_r\mathbf{X}_r\gamma + \mathbf{J}_r\mathbf{G}_r\mathbf{X}_r\delta + \mathbf{J}_r\epsilon_r \tag{4.5}$$

Elementary linear algebra tells us that $\mathbf{J}_r\mathbf{G}_r = -\frac{1}{m_r-1}\mathbf{J}_r$. Letting $\mathbf{J}_r\mathbf{A}_r = \mathbf{A}_r^*$, we obtain

$$\frac{m_r - 1 + \beta}{m_r - 1}\mathbf{y}_r^* = \mathbf{X}_r^*\frac{(m_r-1)\gamma - \delta}{m_r - 1} + \epsilon_r^*$$

which is equivalent to eq. (4.2).

To derive the pseudo CML estimator, we assume (possibly wrongly) that the ϵ_{ir}'s are i.i.d. $\mathbf{N}(0, \sigma^2)$. It follows that, given $\mathbf{X}_r$, m_r, and n_r, the pseudo density of $\mathbf{y}_r^*$ is a multivariate normal distribution with mean $\mathbf{X}_r^*\frac{(m_r-1)\gamma-\delta}{m_r-1+\beta}$ and variance $(\sigma\frac{m_r-1}{m_r-1+\beta})^2\mathbf{J}_r$.[12] The pseudo log likelihood function to be maximized can then be expressed as follows :

$$\begin{aligned}
\ln L \;=\; & c + \sum_{r=1}^{R}(n_r - 1)\ln(m_r - 1 + \beta) - \frac{N-R}{2}\ln\left(\sigma^2\right) \\
& -\frac{1}{2\sigma^2}\sum_{r=1}^{R}\left(\frac{m_r - 1 + \beta}{m_r - 1}\mathbf{y}_r^* - \mathbf{X}_r^*\frac{(m_r-1)\gamma - \delta}{m_r - 1}\right)' \times \\
& \left(\frac{m_r - 1 + \beta}{m_r - 1}\mathbf{y}_r^* - \mathbf{X}_r^*\frac{(m_r-1)\gamma - \delta}{m_r - 1}\right),
\end{aligned}$$

where c is a constant. This log likelihood function excludes any fixed effects. It is a *conditional* log likelihood function as it is conditional on the sufficient statistics $\overline{\mathbf{y}}_r$, (as well as on the $\mathbf{X}_r$'s, the m_r's, and the n_r's), for $r = 1, ...R$. Under the assumption that the ϵ_{ir}'s are correctly specified and i.i.d. $\mathbf{N}(0, \sigma^2)$, Lee (2007) shows that the CML estimators of β, γ, δ and σ are consistent and asymptotically efficient under regularity conditions and provided there is sufficient variation in group sizes.

[12]Note that only $n_r - 1$ elements of ϵ_r^* are linearly independent.

Even if the assumed density of $\mathbf{y}_r^*$ is misspecified, the pseudo CML estimator is consistent provided that the conditional mean of the $\mathbf{y}_r^*$'s is correctly specified. This is the case since the normal density belongs to the Linear Exponential Family (see Gourieroux *et al.* 1984). Of course, the estimator is no longer asymptotically efficient. Moreover, one has to compute the robust covariance matrix using the sandwich formula $J^{-1}IJ^{-1}$, where J is minus the expectation of the Hessian matrix and I the expectation of the outer-product-of-the-gradient matrix. A further advantage of this computation is that it allows us to see whether an apparent precision of CML estimators is driven by the normality assumption used in Lee (2007).

4.3.4.2 2SLS and Generalized 2SLS estimators

Alternatively, the structural equation (4.4) can be estimated by instrumental (IV) methods. To see how the methods work, define a $N \times N$ block-diagonal matrix $\mathbf{J} = Diag(\mathbf{J}_1, ..., \mathbf{J}_R)$. Concatenating eq. (4.5) over all groups yields :

$$\mathbf{Jy} = \beta\mathbf{JGy} + \mathbf{JX}\gamma + \mathbf{JGX}\delta + \mathbf{J}\epsilon. \tag{4.6}$$

where $\mathbf{y}$ (resp. $\mathbf{X}$) is obtained by stacking the vectors $\mathbf{y}_r$ (resp. the matrices $\mathbf{X}_r$), for $r = 1, ..., R$.

The reduced form of the model is :

$$\mathbf{Jy} = (\mathbf{I}-\beta\mathbf{G})^{-1}(\mathbf{JX}\gamma + \mathbf{JGX}\delta) + (\mathbf{I} - \beta\mathbf{G})^{-1}\mathbf{J}\epsilon. \tag{4.7}$$

Identification can be given a natural interpretation in terms of instrumental variables. If $i \notin M_{ri}$ and there are at least three different group sizes, $\mathbb{E}[\mathbf{JGy}|\mathbf{X}, \mathbf{G}]$ is not perfectly collinear to $(\mathbf{JX}, \mathbf{JGX})$ and the model is identified [see Bramoullé *et al.* (2009) for more details]. Moreover $\mathbf{JG^2X}$ can be used as a matrix of valid

instruments for $\mathbf{JGy}$.[13]

One advantage of an IV approach over pseudo CML is that it requires less structure. Specifically, we do not assume that the specified density function of the y_r's, potentially partially misspecified, is normal. Also we do not use the structure on the error terms for identification purpose. Thus, identification in this case is semi-parametric, or "distribution-free". Of course, this comes at a price : the IV estimator is asymptotically less efficient than the pseudo CML, since the latter imposes more structure on the distribution of error terms.

In addition, we can derive a Generalized IV estimator as proposed in Kelejian and Prucha (1998), and discussed in Lee (2007). Assuming homoskedasticity, it yields an asymptotically optimal (best) IV estimator and reduces to a two-step estimation method in our case. More precisely, our first step consists in estimating a 2SLS as described above, by using as instruments $\mathbf{S} = (\mathbf{JX}, \mathbf{JGX}, \mathbf{JG}^2)$. The second step consists in estimating a G2SLS estimator using as instruments $\widehat{\mathbf{Z}} = (\widehat{\mathbf{JGy}}, \mathbf{JX}, \mathbf{JGX})$, where $\widehat{\mathbf{JGy}}$ is computed from the reduced form (4.7) premultiplied by $\mathbf{G}$ and using the first-step estimates.

4.4 Data

We gathered for this analysis original data from the Québec Government MERS. These administrative data provide detailed information on individual scores on standardized tests taken in June 2005 on four subjects (Math, Sciences, French and History) by fourth and fifth grade secondary school students. They also include information on the age, gender, language spoken at home and socioeconomic status of students. Sampling has been done in two steps. The population of interest is the set of all fourth and fifth grade secondary school students who are candidates to the MERS examinations in June 2005. This population is comprised of 152,580 students

[13] In fact, $\mathbf{J}_r\mathbf{G}_r = -\frac{1}{m_r-1}\mathbf{J}_r$ and $\mathbf{J}_r\mathbf{G}_r^2 = \frac{1}{(m_r-1)^2}\mathbf{J}_r$, hence instruments are built here by premultiplying characteristics (in deviation) by group-dependent weights and by stacking them across groups.

in total. In the first step, a 75% random sample of secondary schools offering fourth and fifth grade classes in the 2004-2005 school year have been selected. In the second step, all fourth and fifth grade students in these schools have been included. Overall, we have 194,553 individual test scores for 116,534 students.[14]

There are many advantages to the use of our data. First, all 4th and 5th grade students must take tests on these four subjects to qualify for secondary school graduation. This means that our results do not pertain to a selected sample of schools. In particular, both public and private school students have to take these tests. Another advantage is that the tests are standardized, *i.e.*, designed and applied uniformly within the province of Québec. We use test results gathered by the MERS, so there is less scope for measurement error with these data than with survey data on grades. Finally, although survey data may have provided information on a larger set of covariates, sample sizes in our study are larger than in typical school surveys.

Given the lack of information on the structure of relevant social interactions, we assume that the peer group for a student taking a test is comprised of all other students in the same school who are qualified to take the test in June 2005. Two test sessions are offered for those who completed coursework in the Spring semester. We thus consider as belonging to the same group all those who belong to the same school and who take a subject test in one of the two consecutive sessions of June and August 2005. We know the number of students in each of these groups. But we only observe test scores for the set of students who took the test in June. Therefore we do not always observe the scores of all students within a group. We offered a correction for this problem in our discussion of the econometric model, and our empirical results below incorporate this correction. In any case an overwhelming majority of the students do take the tests in June, so the correction has little effect on the results.

[14]There are more individual test scores than students as some students take test in more than one subject matter.

We use for this study French, History, Science and Math test results as reported in the MERS administrative data. Students in a regular track take History and Science tests in Secondary 4. The French test is commonly taken in Secondary 5. Finally, we focus on students who take the Math test in Secondary 5 (Math 514). This completes their mathematical training for secondary school. Note that the MERS administers a unique test to all secondary school students in French, History and Science. In contrast, it administers different tests in Math, depending on academic options chosen early on by the students. We report here results for students following the regular mathematical training (Math 514). We focus on this test in our analysis.

We provide descriptive statistics in Table III.I.[15] For each subject, the dependent variable in our econometric model is the test score obtained in the provincial standardized test. The average score is between 70% and 75% in French, Science and History tests. It is lower and about 62% in Math. In samples for which the regular track for the test is Secondary 5 (resp. Secondary 4), the average age of students is close to 16 (resp. 15). Most students taking French and Math (98% and 96%) are enrolled in Secondary 5. Most of those taking Science and History are enrolled in Secondary 4 (92% and 96%). Between 52% and 55% of students are female, and between 11% and 13% of students speak a language at home which is different from the language of instruction (*Foreign* variable).[16] Between 30% and 34% of students come from a relatively high socioeconomic background and between 40% and 42% from a medium one. We use an index of socio-economic status provided by the MERS. This index is computed from data from the 2001 census. It uses information on the level of education of the mother (a weight of 2/3) and the job status of parents (weight of 1/3). Low socio-economic status corresponds to the three lowest deciles of the index (high socio-economic status to the three highest deciles).

[15] All Tables can be found in appendix III.

[16] The language of instruction is French in most schools, and English otherwise.

We observe test scores and characteristics of students taking the same test in June 2005. Sample sizes are 41, 778 for French, 54, 981 for Science, 15, 771 for Math, and 55, 057 for History. We also observe the number of students who completed coursework but postpone test-taking to August 2005. There are 118 students postponing French, 186 postponing History, 195 postponing Science, and 160 postponing Math. We observe between 314 and 382 peer groups depending on the subject matter considered. The average group size is between 50 (Math) and 146 (Science). The ratio between the number of groups and the average group size varies between 2.36 (French) and 7.23 (Math). These numbers are relatively small, which suggests that our estimates could be subject to weak identification problems. The group size standard deviation is quite large, however, varying between 50 (in Math) and about 135 (in Science and History). We expect such dispersion in group sizes to help identification. We analyze these issues in more details in Section 4.6.

4.5 Empirical Results

4.5.1 CML and pseudo CML estimates

Table III.II reports the results of maximum likelihood estimation with unrobust (CML) and robust (pseudo CML) standard errors. The model estimated is the linear-in-means model with group fixed effects, individual impacts, and endogenous and contextual peer effects. We find that the estimated endogenous peer effect lies between -0.24 and 0.83. Using unrobust standard errors (in brackets), the endogenous effect is significantly different from zero and positive for Math ($\widehat{\beta} = 0.82$), and History ($\widehat{\beta} = 0.65$). It is not significant for French ($\widehat{\beta} = 0.33$) and for Science ($\widehat{\beta} = -0.23$). Based on robust standard errors, it is no longer significant for History (p-value= 10,82%) but still significant for Math. One thus concludes that regarding this peer effect, inference appears to be driven by normality for one subject (History). In general standard errors are larger using pseudo CML than CML, but their differences are not so important.

Two reasons may explain why the endogenous peer effects in Math is significant

in our sample. First, the standard error of the estimates is smaller in Math than in other subjects. This is consistent with the fact that the average group size relative to the number of groups is close to three times smaller in Math than in other subjects. Second, our endogenous effect estimate is much larger in Math (0.82). How does this result compare with other studies ? Sacerdote (2011) has recently provided a survey of studies of endogenous peer effects in test scores for primary and secondary schools based on linear-in-means models (see his Table 4.2.). Interestingly, in most reported studies (5 over 6) which analyze achievement in both Math and Reading, the endogenous peer effect is larger in Math. In addition, this effect is often very high and exceeds the value we have estimated. Thus Hoxby (2000) reports a 1.7 to 6.8-point increase in own score in relation with a 1-point increase in mean score of peers in some specifications. Betts and Zau (2004) show a 1.9-point increase in association with a 1-point increase in mean math score of peers. On the other hand, Hanushek *et al.* (2003) obtain a Math peer effect of 0.4.[17] So our estimate lies on the average to high side of the range of previous estimates. Observe finally that our results in Math are larger than those usually obtained in studies based on randomized experiments (*e.g.*, Sacerdote 2001, Zimmerman 2003). One possible explanation is that peers used in these papers are often people from the same dorm. These individuals do not necessary represent those who exercise significant influence on students' scholar achievement.

The relatively large endogenous peer effect in Math may reflect the fact that mathematics provide more opportunities for interactions among students. And, probably more than in other subjects, it may also reflect general effects such as disruption. For instance, it is likely that success in Math requires much concentration in class from the average student. Now suppose that there is a student (with low grade in Math) in class who is characterized by his propensity to disrupt learning by bad behavior or asking poor questions. His behavior may have large negative effects on his peers' scholar achievement (*e.g.*, see Lazear 2001) and thus

[17]Kang (2007, p. 475) also provide a survey of endogenous peer effects in achievement in mathematics which is broadly consistent with results reported in Sacerdote (2011).

generates strong endogenous peer effects.

Regarding the individual characteristics, most of them have a significant effect on test scores, and the signs of these effects essentially conform to expectations. All test scores decrease significantly with age. Since older students have often repeated a grade, being younger is a natural proxy for ability. Test scores are significantly higher for female students than for male students, except for History where male students perform significantly better than female students. This is broadly consistent with results from previous studies. For instance, results from the 2000 Program for International Student Assessment (PISA) show that Québec female students perform better than males on reading literacy tests but that the differences in performance on mathematics and science tests are smaller and not significant (see Québec Government 2001). Similarly, in our analysis, the difference in performance is quantitatively large in French but much smaller in the other disciplines. The performance of foreign students is, non surprisingly, significantly lower than for non-foreign students on the French test, but higher for Science and History and not significantly different for Math. Secondary 5 students tend to perform significantly better on all tests than Secondary 4 students, which reflects the positive impact of an additional year of schooling on test scores. Finally, students from a higher socioeconomic category perform significantly better in all tests.

As far as contextual variables are concerned, a few of them have a significant impact on student performance. Average age of other students has a negative and significant effect on all test scores except Math where it is positive but not significant. These results also conform our expectations. When the number of repeaters rises (as reflected by an increase in mean age of our peers at a given grade level), this will tend to reduce own test score. Proportion of other students enrolled in Secondary 5 have a large positive and significant effect on own score in French. Peers' socioeconomic background has little effect on own schooling performance. The proportion of female students among peers has a positive and significant effect in Math. When significant, the magnitude of contextual effects is always larger

than the magnitude of individual effects. This is not surprising as it captures the effect of a unit change in the characteristic of *every* other student in the group.[18]

4.5.2 Reflection problem

One way of addressing the simultaneity problem without exploiting group size variations is to exclude at least one contextual variable from the outcome equation and to use it as an instrument for average test score. We estimate a model similar to the one presented in Table III.II but excluding contextual effects that are not individually significant in the pseudo CML specification (*i.e.*, for which the null that $\delta = 0$ is not rejected) ; see Table III.VI. Using likelihood ratio tests, we reject the null that these δ's are jointly equal to zero for French but not for the other subjects. This suggests that the exclusion restrictions may be valid for these latter samples. Therefore, the pseudo CML estimators provided in Table III.VI should be consistent and asymptotically more efficient than those provided in Table III.II for the Science, Math and History tests. Results however appear to be robust to these new specifications. Observe finally that we could not have known this *a priori* without an estimation of the full model.

Overall, this shows the interest of Lee's solution to the reflection problem. Estimating a model with both endogenous and contextual peer effects is needed to recover the different types of peer effects at work.

4.5.3 2SLS and G2SLS estimates

Tables III.III, and Table III.VII provide the 2SLS and G2SLS estimation results of the linear-in-means model of peer effects with group fixed effects, individual impacts, and endogenous and contextual peer effects. In contrast to the CML and pseudo CML estimates of Table III.II, none of the endogenous effects is statistically significant. This is consistent with Lee's (2007, p. 345) result that the asymptotic

[18] We have also estimate a second-order pseudo CML in which restrictions are directly incorporated in the variance term and estimated. Results are quite similar with those presented in Table III.III.

efficiency of IV estimators is smaller than that of the CML. Estimated individual effects are quite similar to the corresponding CML estimates. Some contextual effects are similar while others are different. For instance, the proportion of other students in Secondary 5 still has a large and positive effect on own French score as well as no significant effects for the other subjects. In contrast, average age among peers now has a positive and significant effect on own score for most subjects, rather than a negative one. This could be explained by differences in small sample properties of both methods, possibly aggravated by the imprecision in the estimation of the endogenous peer effect.

Table III.III also reports two standard test results giving information on instrumental variables properties. We first look at Sargan tests on the validity of instruments and the over-identification restrictions of the model. We do not reject the null for Science, Math and History, but we reject it for French. While this may indicate a problem of model specification in this last case, one must be cautious in interpreting the test given the likely low convergence of peer effects IV estimates. We then compute Stock and Yogo test statistics on weak identification. Based on the definition that a group of instruments is weak when the bias of the IV estimator relative to the bias of ordinary least squares exceeds a certain threshold b, say 5%, one rejects the null that the instruments are weak for all subject matters. Finally, Hausman tests have been performed to test the equality of pseudo CML and G2SLS estimators. Under the null, both of these estimators are consistent, but pseudo CML estimators are asymptotically more efficient ; under the alternative, G2SLS estimators are consistent whereas pseudo CML estimators are not. For each subject, we could not reject the null. This suggests the absence of specification errors in the model.

4.6 Monte Carlo simulations

In this section, we study through simulations the effect of group sizes and their distribution on the precision and bias of our estimates. Lee (2007) shows that the

CML and IV estimators may converge in distribution at low rates when the ratio between the the number of groups and the average group size is small. Since this ratio varies between 2.36 and 7.23 in our samples, a problem of weak identification could in principle emerge. However, the standard deviation of the distribution of group sizes is also relatively large (see Table III.I), and we suspect that this may help identification. To study these issues, we realize two simulation exercises. First, we vary group sizes in a systematic manner and study how this affects the bias and precision of estimators. To focus on the approach which provides the most reasonable findings in our empirical analysis, we report results on the model using CML.[19] We look at uniform distributions, vary the size of the distribution's support and partly calibrate simulation parameters on our data. Second, we look at bias and precision of estimates for *fully* calibrated simulations, when group sizes are exactly the same as in the data. Overall, while our analysis confirms Lee's earlier results, we also find a strong positive impact of the dispersion in group sizes on the strength of identification. Especially, conditional maximum likelihood performs well on fully calibrated simulations. This suggests that the bias due to small sample issues is likely low in the results presented in Table III.II.

For each simulation exercise, we keep the number of observations fixed around $42,000$, and run $1,000$ replications. We first consider average sizes of 10, 20, 40, 80 and 120. We pick group sizes from the following intervals with decreasing length :

- Average size of 10 : $[3, 17]$, $[5, 15]$, $[7, 13]$ and $[9, 11]$,
- Average size of 20 : $[3, 37]$, $[8, 32]$, $[13, 27]$ and $[18, 22]$,
- Average size of 40 : $[3, 77]$, $[12, 68]$, $[21, 59]$, $[30, 50]$ and $[39, 41]$,
- Average size of 80 : $[3, 157]$, $[18, 142]$, $[33, 127]$, $[48, 112]$ and $[63, 97]$,
- Average size of 120 : $[3, 237]$, $[28, 212]$, $[53, 187]$, $[78, 162]$ and $[103, 137]$.

[19] In an earlier version of the paper, we also provided results for IV estimates. Basically, the results are qualitatively the same for IV as those for CML but, as expected, the magnitude of the bias and the loss in precision are always larger for IV than for CML.

For each of the intervals described above, we proceed in the following manner :
- pick a group size from a uniform distribution for which the support is defined by the minimum and maximum value of the interval ;
- truncate this value by eliminating its decimal portion ;
- repeat step 1 and 2 as long as the total number of observations is below or equal to $42,000$.

To reduce computing time, we assume that students have the same characteristics except for age and gender. We assume that age follows a normal distribution and gender follows a Bernoulli distribution. We calibrate the moments of these distributions on the sample of students taking the French test : average age is 16, variance of age is 0.25, and proportion of girls is 0.55. Values of the structural parameters β, γ and δ are set close to the estimated coefficients for the French test : $\beta = 0.35, \gamma_{age} = -8, \gamma_{gender} = 3.8, \delta_{age} = -40, \delta_{gender} = -25$.

We assume that the values of ϵ in the structural equation are drawn randomly from a normal distribution with mean zero and variance $\sigma^2 = 1$. We generate the endogenous variable y from the reduced-form equation in deviation form.

Looking at Table III.IV, we first compare simulation results across average group sizes and then we examine how estimators perform for a given average group size as dispersion in group size decreases. Separate horizontal panels in Table 4 pertain to different values of average group size. We report the average estimated coefficient and standard error for the endogenous effect (first vertical panel), the contextual effect associated with age (second vertical panel) and the contextual effect associated with gender (third vertical panel). We find that even for the largest average group size (*i.e.*, 120), CML may perform well in terms of bias and precision (first line in the last horizontal panel of Table 4). The biases of CML get in general larger as average group size increases. The CML estimate of the endogenous effect attains a plateau at the value 1. This is consistent with the fact that the CML estimator tends towards the naive OLS estimator as group sizes become larger. In

general, peer effects are also less precisely estimated in large groups than in small groups.

Our main new result concerns the effect of group size dispersion. When we fix the value of the average group size and reduce the length of the interval from which group sizes are picked, we find that the bias of CML typically increases while the precision typically decreases. In Table III.IV, this amounts to looking at each horizontal panel separately. Observe however that since we roughly pick group sizes from a uniform distribution holding average group size fixed, reducing the interval's length affects the two parameters of the size distribution (*i.e.*, the minimum and maximum value of its support) and a number of its moments. In particular, this leads to a reduction in variance and to an increase in the size of the smallest groups. In general, both the variance and the size of smallest groups may matter and the strength of identification may depend on the size distribution in complex ways. We leave a deeper investigation of this issue to future research.

We next fully calibrate the simulations' parameters on the data. We use observed group sizes in the French sample, calibrate the model parameters $\{\beta, \gamma_{age}, \gamma_{gender}, \delta_{age}, \delta_{gender}\}$ and moments of the explanatory variables as previously, and set the variance of the error term in the structural equation equal to the estimated variance in the French sample ($\hat{\sigma}^2 = 154.7$). Simulation results which now report both CML and IV estimates are reported in Table III.V. The CML estimator has small bias and standard error, while the IV estimator is not precisely estimated and the bias is large. These results confirm for CML what we obtained from picking group sizes at random ; they show that dispersion in group sizes help identification. Besides, this suggests that small sample bias may be relatively high in the IV estimates of Tables III.III, and of Table III.VII but relatively low for the CML estimates of Table III.II.

4.7 Conclusion

This paper provides an analysis of social interactions in scholar achievement when students interact through groups. Based on a linear-in-means approach with group fixed effects (Lee 2007), we make two main contributions regarding the identification and estimation of peer effects. First, we provide a new intuition for identification. We show that full identification of the model relies on three key properties : (1) Since the individual is excluded from his peer group, above average students have below average peers (with respect to any attribute). Therefore, when individual and peer effects are positive, peer effects then tend to reduce the dispersion in outcomes. (2) This reduction is stronger in smaller groups, reflecting the larger effect of excluding one individual from the mean. (3) Contextual and endogenous peer effects generate reductions of different shapes, which allow to identify both of them.

Second, as regards the estimation of peer effects, the model is applied to original administrative data providing individual scores on standardized tests taken in June 2005 in four subjects by fourth and fifth grade secondary school students in the Province of Québec (Canada). Based on a pseudo conditional maximum likelihood approach, our results indicate that students significantly benefit from their peers' higher test scores in Math but not in other subjects such as Science, History and French. Two reasons may explain these results. First, this is likely to reflect the fact that Math provides more opportunities for interactions among students. Second, in our sample, the average group size (relative to the number of groups) is close to three times smaller in Math that in other subjects. As suggested by Lee (2007), accurate estimation of peer effects requires relatively small groups. This is also confirmed by our Monte Carlo simulations. These results should be warning applied researchers in the future against using data in which the size of groups is too large. Besides, our simulations indicate that, for a given average group size, increasing group size dispersion improves the precision of peer effects estimates. In fact, our results suggest that, conditional on estimating on the whole sample,

even data on larger groups may provide useful information for estimation purposes. The basic intuition is that data on very large groups can be used to provide more precise individual effects estimators. In turn, this indirectly provides more efficient estimates of the peer effects from data on smaller groups. So, future estimations of Lee's model may benefit from data with relatively small average group size but relatively large group size dispersion, including both small and large groups.

In terms of public policy, the fact that the endogenous peer effects appear to be very large in Math suggests that a reform that improves the amount and quality of Math learning is likely to yield very high returns in terms of scholar achievement. This is so since such a reform will not only have direct effects on student performance in Math but also strong indirect effects through the additional external benefits generated by the social multiplier. Remarkably, our analysis also shows that the indirect peer effects of the reform will reduce performance inequalities in Math across students. This is the case because low-ability students have better peers (since their peers exclude them) and high-ability students have worse peers (for the same reason). Moreover, the strong negative effects of the average age of peers on scholar achievement (except in Math) suggest that resources invested by the government to reduce the number of repeaters may have an important indirect positive impact on student performance. One limitation of Lee 's linear-in-means approach is that it imposes that average test score over all schools are not influenced by a reallocation of students across schools (see Sacerdote 2011). Therefore, the model does not have much to say about issues such as optimal school composition by race or ability.

Our research could be extended in many directions. It would be interesting to evaluate the validity of this approach by using data where group membership is experimentally manipulated *and* group sizes are heterogenous (as in Sacerdote 2001). One could also analyze how group size variations may help to identify peer effects when the outcome is a discrete variable (*e.g.,* pass or fail). Brock and Durlauf (2007) have studied peer effects identification with discrete outcomes but they

ignore group size variations. A third potentially fruitful direction of research would be to analyze a nonlinear version of Lee's approach. Thus, student achievement could depend on the mean and standard deviation of peers attributes. Overall, we think that this first empirical application confirmed the interest of the method. Many more applications in different settings are needed, however, in order to gain a thorough understanding of the method's advantages, limitations, and applicability for public policy.

CHAPITRE 5

CONCLUSION

Comme discuté dans l'introduction au chapitre 1, les essais de cette thèse s'intéressent à la fois à l'analyse de $\mathbb{P}(\mathbf{G}|\mathbf{X})$ et de $\mathbb{P}(\mathbf{y}|\mathbf{G},\mathbf{X})$. Je prend ici le temps de discuter quelques limites de ces approches et certaines avenues potentielles pour la recherche future.

Une des faiblesses de la litérature sur la formation de réseaux est leur limitation en terme d'implications sur les politiques publiques. Bien sûr, être capable d'identifier quelles variables socioéconomiques influencent le processus de formation d'un réseau est un premier pas important. Par contre, le fait est que les causes pour lesquelles ces variables sont importantes demeurent en grande partie inconnues. La raison est que $\mathbb{P}(\mathbf{G}|\mathbf{X})$ représente une forme réduite de plusieurs phénomènes, i.e. $\mathbb{P}(\mathbf{G}|\mathbf{y}(\mathbf{X}),\mathbf{X})$. Par exemple, supposons que $\mathbf{y}$ représente la consommation de cigarettes d'adolescents et que cette consomation soit influencée par le niveau d'éducation des parents de ces adolescents. Supposons aussi que la consommation de cigarettes soit une variable importante quant à la formation de liens d'amitié. L'étude de $\mathbb{P}(\mathbf{G}|\mathbf{X})$ identifiera le niveau d'éducation des parents comme une variable importante en ce qui à trait à la formation de liens d'amitié chez les adolescents. Si cela est vrai, malheureusement, l'approche par forme réduite ne permet pas d'identifier la raison pour laquelle $\mathbf{X}$ influence $\mathbf{G}$. L'étude de $\mathbb{P}(\mathbf{G}|\mathbf{y},\mathbf{X})$ quant à elle est loin d'être triviale en raison de l'endogénéité de $\mathbf{y}$. L'identification de l'impact de $\mathbf{y}$ et $\mathbf{X}$ sur $\mathbf{G}$ est l'un des défis importants à surmonter.

Du côté de la littérature sur les effets de pairs, le problème inverse se pose. L'étude de $\mathbb{P}(\mathbf{y}|\mathbf{G},\mathbf{X})$ permet de bien identifier les effets importants pour la création de politique publiques. Par contre, l'endogénéité potentielle de $\mathbf{G}$ peut être problématique en pratique. Dans le cas ou les interactions se font en groupe (comme au chapitre 4), ce problème est moins important. Par contre, lorsque les effets de pairs passent, par exemple, par un réseau d'amitiés, il se peut très bien d'une va-

riable inobservée affecte la formation du réseau, ce qui entraîne naturellement un biais dans l'estimation de $\mathbf{y}$.

Une avenue prometteuse est l'étude de la probabilité jointe $\mathbb{P}(\mathbf{y}, \mathbf{G}|\mathbf{X})$. C'est entre autres cette avenue qui est empruntée par Goldsmith-Pinkham et Imbens (2012), où les auteurs étudient $\mathbb{P}(\mathbf{y}, \mathbf{G}|\mathbf{X}) = \mathbb{P}(\mathbf{y}|\mathbf{G}, \mathbf{X}) \cdot \mathbb{P}(\mathbf{G}|\mathbf{X})$. Encore ici, par contre, $\mathbb{P}(\mathbf{G}|\mathbf{X})$ représente une forme réduite comprenant simultanément plusieurs effets. De plus, on comprend encore mal les fondement microéconomiques de ce genre de models. C'est-à-dire, supposons que les individus choisissent en premier lieu le réseau $\mathbf{G}$, et ensuite leur action $\mathbf{y}$ sur ce réseau. Quelles sont les conditions sur $\mathbb{P}(\mathbf{y}|\mathbf{G}, \mathbf{X})$ et $\mathbb{P}(\mathbf{G}|\mathbf{X})$ tel que $(\mathbf{y}, \mathbf{G})$ soit un équilibre parfait en sous-jeu ?

La littérature empirique sur les réseaux sociaux est donc encore jeune et il existe encore beaucoup plus de questions que de réponses. À la lumière des phénomènes identifiés dans cette thèse je reste convaincu que beaucoup de réponses passeront par une meilleure compréhension des incitatifs des agents, donc par la création de modèles microéconomiques orientés vers une application empirique.

CHAPITRE 6

RÉFÉRENCES

Ammermueller, A. and Pischke, J-S. (2009) : "Peer Effects in European Primary Schools : Evidence from the Progress in International Reading Literacy Study", *Journal of Labor Economics,* 27(3), 315-348.

Angrist, J. D. and Lang, K (2004) : "Does School Integration Generate Peer Effects ? Evidence from Boston's Metco Program", *American Economic Review,* 94(5), 1613-1634.

Aumann, R. J. (1959) "Acceptable Points in General Cooperative n-person Games" In Contribution to the theory of game IV, Annals of Mathematical Study 40, 287-324

Bester A.C., Conley T.G. and Hansen C.B. (2011), "Inference with Dependent Data Using Cluster Covariance Estimators", *Journal of Econometric*, 165(2) p.137-151

Betts, J.R. and Zau, A. (2004) : "Peer groups and academic achievement : Panel evidence from administrative data", Unpublished manuscript.

Bifulco, R., Fletcher, J.M. and Ross, S.L. (2011) : "The Effect of Classmate Characteristics on Post-Secondary Outcomes : Evidence from the Add Health", *American Economic Journal : Economic Policy,*3(1), 25-53.

Bisin A., Moro A and Topa G. (2011) "The Empirical Content of Models with Multiple Equilibria in Economies with Social Interactions", *Working Paper*

Bloch F. and Dutta B. (2009) "Communication Networks with Endogenous Link Strength", Games and Economic Behavior, 66(1), 39-56

Blume L. Brock W., Durlauf S. and Ioannides Y. (2011) "Identification of Social Interactions", Chapter 18 in Social Economics, Benhabib, Bison and Jackson, North-Holland.

Boucher V. (2012) "Structural Homophily", *Working Paper*

Bramoullé Y., Currarini, S., Jackson, M.O., Pin P. and Rogers B. (2012) "Homophily and Long-Run Integration in Social Networks", Journal of Economic Theory, 147, p.1754-1786

Bramoullé Y., Djebbari H., and Fortin B. (2009) "Identification of peer effects through social networks", Journal of Econometrics, 150, 41-55

Brock, W. and Durlauf, S. (2007) : "Identification of Binary Choice Models with Social Interactions". *Journal of Econometrics,* 140, 57-75.

Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009) : "Peer Effects and Social Networks in Education", *Review of Economic Studies*, Vol. 76(4), 1239-1267.

Cameron A.C. and Trivedi P.K. (2005) "Microeconometrics, Methods and Applications", Cambridge University Press

Chakrabarti S. and Gilles R.P. (2007) "Network Potentials", The Review of Economic Design, 11, 13-52

Christakis N., Fowler J., Imbens G.W. and Kalyanaraman K. (2010) "An Empirical Model for Strategic Network Formation", *Working Paper*

Conley T. (1999) "GMM Estimation with Cross Sectional Dependence", Journal of Econometrics, 92, 1-45

Čopič J., Jackson M.O. and Kirman A. (2009) "Identifying Community Structures from Network Data via Maximum Likelihood Methods", B.E. Press Journal of Theoretical Economics 9 : Iss. 1 (Contributions), Article 30.

Currarini S., Jackson M. O., Pin P. (2009) "An Economic Model of Friendship : Homophily, Minorities, and Segregation", Econometrica, 77, 1003-1045

Currarini S., Jackson M. O., Pin P. (2010) "Identifying the roles of race-based choice and chance in high school friendship network formation," Proceedings of the National Academy of Sciences of the United States of America, 107(11) : 4857-4861

Davezies, L., d'Haultfoeuille, X. and Fougère, D. (2009) : "Identification of Peer Effects Using Group Size Variation", *Econometrics Journal*, Vol. 12, 397-413.

Echenique F. and Fryer R.G. (2007) "A Mesure of Segregation Based on Social Interactions" The Quarterly Journal of Economics, 122(2), 441-485

Epple, D. and Romano, R.E. (2011), "Peer Effects in Education : A Survey of the Theory and Evidence", in *Handbook of Social Economics,* edited by J.Benhabib, A. Bisin, and M. O. Jackson, Amsterdam : North Holland, pp. 1053-1163.

Evans, W., Oates, W. and Schwab, R. (1992) : "Measuring Peer Group Effects : a Study of Teenage Behavior", *Journal of Political Economy,* Vol. 100(5), 966-991.

Fortin B. and Yazbeck M.A. (2011), "Peer Effects, Fast Food Consumption and Adolescent Weight Gain", *Working Paper*

Franz S., Marsili M. and Pin P. (2008) "Observed Choices and Underlying Opportunities", *Working Paper*

Galeotti A., Goyal S. and Kamphorst J. (2006) "Network Formation with Heterogenous Players", Games and Economic Behavior, 54(2), 353-373

Galichon A. and Henry M. (201) "Set identification in models with multiple equilibria", Review of Economic Studies, 78(4), 1264-1298

Gallant R.A. and White H. (1988) "A unified theory of estimation and inference for nonlinear dynamic models", B. Blackwell, 155p.

Gaviria, A. and Raphael, S. (2001) : "School based Peer Effects and Juvenile Behavior", *Review of Economics and Statistics*. 83(2), 257-268.

Golub B. and Jackson M.O. (2010a) "Naive Learning in Social Networks : Convergence, Influence and the Wisdom of Crowds", the American Economic Journal : Microecnomics 2(1) : 112-149

Golub B. and Jackson M.O. (2010b) "Using selection bias to explain the observed structure of Internet diffusions", Proceedings of the National Academy of Sciences, 107(24) : 10833-10836

Goldsmith-Pinkham P. and Imbens G.W. (2011), "Social Networks and the Identification of Peer Effects", *Working Paper*

Gouriéroux, C, Montfort, A. and Trognon A. (2004) : "Pseudo Maximum Likelihood Methods : Theory", *Econometrica*, 52, 681-700.

Goyal S. and Joshi S. (2006) "Unequal Connections", International Journal of Game Theory, 34, 319-349

Goyal S. and Vega-Redondo F. (2007) "Structural Holes in Social Networks", Journal of Economic Theory, 137, 460-492

Graham, B.S. (2008) : "Identifying social interactions through conditional variance restrictions", *Econometrica,* 76 (3), 643-660.

Hanushek, E., Kain, J., Markman, J. and Rivkin, S. (2003) : "Does Peer Ability Affect Student Achievement ?", 18(5), *Journal of Applied Econometrics,* 527-544.

Henry M. and Mourifié I. (2011) "Euclidean Revealed Preferences : Testing the Spatial Voting Model", Journal of Applied Econometrics, *Forthcoming*

Hoxby, C. (2000) : "Peer Effects in the Classroom : Learning from Gender and Race Variation", WP no. 7867, National Bureau of Economic Research, Cambridge, MA.

Iijima R. and Kamada Y. (2010) "Social Distance and Network Structures", *Working Paper*

Irving R. W. (1985) "An efficient algorithm for the "stable roommates" problem", Journal of Algorithms, 6(4), 577-595

Jackson M. O. (2008) Social and Economic Networks, Princeton University Press

Jackson M.O. and Rogers B.W. (2005) "The Economics of Small Worlds", Journal of the European Economic Association, 3(2-3), 617-627

Jackson M.O. and Rogers B.W. (2007) "Meeting Strangers and Friends of Friends : How Random are Socially Generated Networks ?", American Economic Review, 97(3), 890-915

Jackson M. O. and Watts (2001) "The existence of Pairwise Stable Networks", Seoul Journal of Economics, 14(3), 299-321

Jackson M. O. and Wolinsky (1996) "A Strategic Model of Social and Economic Networks", Journal of Economic Theory, 71, 44-74

Jenish N. and Prucha I.R. (2009) "Central limit theorems and uniform laws of large numbers for arrays of random fields", Journal of Econometrics, 150, 86-98

Johnson C. and Gilles R. P. (2000) "Spatial Social Networks", Review of Economic Design, 5, 273-299

Kang, C. (2007) : "Classroom Peer Effects and Academic Achievement : Quasi-Randomization Evidence from South Korea", 61*, Journal of Urban Economics,* 61, 458-495.

Kelejian H.H. and Prucha I.R. (1998) : "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances", *Journal of Real Estate Finance and Economics,* Vol 17*,* 99-121.

Krueger, A.B. (2003) : "Economic Considerations and Class Size", *Economic Journal,* 113(485), F34-F63.

Lazear, E. P. (2001) : "Educational Production", *Quarterly Journal of Economics*, 116, 777-803.

Lee, L. F. (2007) : "Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects", *Journal of Econometrics,* 140(2)*,* 333-374.

van der Leij M., Rolfe M. and Toomet O. (2009) "On the Relationship Between Unexplained Wage Gap and Social Network Connections for Ethnical Groups", *Working Paper*

Lin, X. (2010) : "Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables", *Journal of Labor Economics*, 28(4), 825-860.

Manski, C. (1993) : "Identification of Endogenous Social Effects : The Reflection Problem", *Review of Economic Studies*, Vol. 60(3), 531-542.

Manski, C.F. (2000) "Economic Analysis of Social Interactions", The Journal of Economic Perspectives, 14(3), 115-136

Marmaros D. and Sacerdote B. (2006) "How Do Friendships Form ?", The Quarterly Journal of Economics, 121(1), 79-119

McFadden (1981) "Econometric Models of Probabilistic Choice", in C. Manski and D. McFadden (editors), Structural Analysis of Discrete Data with Econometric Applications, Cambridge, Mass. :M.I.T.Press.

McPherson M., Smith-Lovin L., and Cook M J. (2001) "Birds of a Feather : Homophily in Social Networks", Annual Review of Sociology, 27 :415-444

Mele A. (2010) "A Structural Model of Segregation in Social Networks", *Working Paper*

Moffitt, R. (2001) : "Policy Interventions, Low-Level Equilibria, and Social Interactions", in *Social Dynamics,* edited by Steven Durlauf and Peyton Young, MIT press.

Newman J. and E.L. Scott (1948) "Consistent Estimates Based on Partially Consistent Observations", Econometrica, 16(1), p.1263-1297

Patacchini E. and Zenou Y. (2012) "Ethnic Networks and Employment Outcomes", Science and Urban Economics, 42, p.938-949

Pinkse J. and Slade M.E. (1998), "Contracting in Space : An application of spatial statistics to discrete-choice models", Journal of Econometrics, 85, 125-154

Québec Government. (2001) : "PISA 2000. The Performance of Canadian Youth in Reading, Mathematics and Science. Results for Québec Students Aged 15", Ministry of Education, Recreation and Sports.

Rivas J. (2009) "Friendship Selection", International Journal of Game Theory, 38, 521-538

Rivkin, S. G. (2001) : "Tiebout Sorting, Aggregation and the Estimation of Peer Group Effects", *Economics of Education Review,* Vol. 20, 201-209.

Rubí-Barceló A. (2010) "Core/periphery scientific collaboration networks among very similar researchers", Theory and Decision, *Forthcoming.*

Sacerdote, B. (2001) : "Peer Effects with Random Assignment : Results for Darmouth Roommates", *Quarterly Journal of Economics*, Vol. 116(2), 681-704.

Sacerdote, B. (2011) : in *Handbook of the Economics of Education,* 3, edited by Hanushek E.A., Machin S., and Woessmann L., Amsterdam : North Holland.

Stinebrickner, R. and Stinebrickner, T. R. (2006) : "What Can be Learned About Peer Effects Using College Roommates ? Evidence from New Survey Data and Students from Disadvantaged Backgrounds", *Journal of Public Economics,* 90, 1435-1454.

Wang H., Iglesias E.M., and Wooldridge J.M. (2009) "Partial Maximum Likelihood Estimation of a Spatial Probit Model", *Working Paper*

Watts A. (2007) "Formation of Segregated and Integrated Groups", International Journal of Game Theory, 35 :505-519

White H. (2001) "Asymptotic Theory for Econometricians", Academic Press, 264p.

Yahoo ! Webscope dataset (2011) ydata-netuser-v2, [`http://research.yahoo.com/Academic_Relations`]

Zimmerman, D. (2003) : "Peer Effects in Academic Outcomes : Evidence from a Natural Experiment", *Review of Economics and Statistics,* 85(1), 9-23.

Annexe I

Appendix I

I.1 Appendix I.1

Proof of lemma 2.3.1

Let x^* be some NE, and suppose that (i, j) is a deviating pair in the sense of a WBE. Let $(\tilde{x}_i, \tilde{x}_j)$ be some joint deviation for (i, j). We need to show that $\tilde{x}_i^j > x_i^{j*}$ and $\tilde{x}_j^i > x_j^{i*}$.

Since $(\tilde{x}_i, \tilde{x}_j)$ is a profitable deviation (in the sense of a WBE), we have

$$\begin{aligned} u_i(\tilde{x}_i, \tilde{x}_j, x^*_{-i-j}) > u_i(x^*) \\ u_j(\tilde{x}_i, \tilde{x}_j, x^*_{-i-j}) > u_j(x^*) \end{aligned} \tag{I.1}$$

Since x^* is a NE, we have

$$\begin{aligned} u_i(x_i, x^*_{-i}) \leq u_i(x^*) \\ u_j(x_j, x^*_{-j}) \leq u_j(x^*) \end{aligned} \tag{I.2}$$

for all x_i, and x_j. In particular, condition (I.2) holds for $x_i = \tilde{x}_i$ and $x_j = \tilde{x}_j$. Putting conditions (I.1) and (I.2) together, we have : $u_i(\tilde{x}_i, \tilde{x}_j, x^*_{-i-j}) > u_i(\tilde{x}_i, x^*_{-i})$ and $u_j(\tilde{x}_i, \tilde{x}_j, x^*_{-i-j}) > u_j(\tilde{x}_j, x^*_{-j})$. Since the utility function is linear in the links, this is equivalent to $v_i(\tilde{x}_i^j, \tilde{x}_j^i, d_{ij}) > v_i(\tilde{x}_i^j, x_j^{i*}, d_{ij})$ and $v_j(\tilde{x}_i^j, \tilde{x}_j^i, d_{ij}) > v_j(\tilde{x}_j^i, x_i^{j*}, d_{ij})$. The production functions are strictly increasing in the second argument, so we must have $\tilde{x}_i^j > x_i^{j*}$ and $\tilde{x}_j^i > x_j^{i*}$. (If $x_i^{j*} = x_j^{i*} = 0$, we have $v_i(\tilde{x}_i^j, \tilde{x}_j^i, d_{ij}) > 0$ and $v_j(\tilde{x}_j^i, \tilde{x}_i^j, d_{ij}) > 0$, and the result is straightforward.) $\square$

Proof of theorem 2.3.2

First, we show that $\tilde{x}$ produced by the assignment algorithm (see appendix B) is a NE. By construction, we have $v_i(\xi, \xi, d_{ij}) \geq 0$, and $w_i(\xi) \geq 0$, hence removing a link is never profitable. Now, the only link that an individual can unilaterally create is the individual link. Suppose that it is profitable to do so for $i \in N$. Then either $[\delta_i < \kappa_i$ and $w_i(\xi) > 0]$, or $[\delta_i = \kappa_i$ and $w_i(\xi) > \min_{j \in g_i} v_i(\xi, \xi, d_{ij})]$. By construction, both are impossible.

Now, suppose that $\tilde{x}$ is a NE, but not a WBE. That is, there exists $i, j \in N$ such that $j \notin g_i$ (from lemma 2.3.1, since $x_i^j \in \{0, \xi\}$) who want to deviate, i.e. create a link between them. There are 2 cases :

1. $\delta_i = \kappa_i$. Then, i needs to remove a link in order to create a new link. (Since $\tilde{x}$ is a NE, he won't remove more than one link.) Then, this implies that there exists $k \in g_i$ such that $v_i(\xi, \xi, d_{ij}) > v_i(\xi, \xi, d_{ik}) \geq 0$. This implies that $d_{ij} < d_{ik}$.

   We now turn to j. If $\delta_j = \kappa_j$, the same argument applies for j, then $v_j(\xi, \xi, d_{ij}) > v_j(\xi, \xi, d_{jl})$ for some $l \in g_j$ (and $v_i(\xi, \xi, d_{ij}) > v_i(\xi, \xi, d_{ik})$). Since we have $d_{ij} < d_{ik}$ and $d_{ij} < d_{jl}$, this contradicts the fact that $\tilde{x}$ was created by the assignment algorithm.

   If $\delta_j < \kappa_j$, j has at least ξ to invest. Together with the fact that $d_{ij} < d_{ik}$, this contradicts the fact that $\tilde{x}$ is produced by the assignment algorithm.

2. $\delta_i < k_i$ and $\delta_j < k_j$. This is impossible since, from the assignment algorithm, it implies that $v_i(\xi, \xi, d_{ij}) < 0$ or $v_j(\xi, \xi, d_{ij}) < 0$.

□

Proof of theorem 2.3.3

We need to show that the allocation $\tilde{x} \in X$, which is produced by the assignment algorithm (see appendix B), is a WBE of Γ.

We first show that $\tilde{x}$ is a NE. Suppose that it is not ; that is, there exists $i \in N$ such that $\tilde{x}_i$ is not individually rational. Since for any $i, j \in N$, we have $x_i^j \in \{0, \xi\}$. This means that i wants to create an additional link. (Unilaterally reducing the investment in a link necessarily lowers i's payoff.) The only link that i can create on his own is the individual link. There are two cases :

1. $\tilde{x}_i^i = 0$ and $\delta_i < \kappa_i$. Then, by construction from the assignment algorithm, this implies that $w_i(\xi) < 0$. So i has no individual profitable deviation, since $w_i(\tilde{x}_i^j) < w_i(\xi)$.

2. $\tilde{x}_i^i = 0$ and $\delta_i = \kappa_i$. Then, if i has a profitable deviation, there exists $J \subseteq g_i$ such that $w_i(\sum_{j\in J} \epsilon_j) > \sum_{j\in J}\{v_i(\xi, \xi, d_{ij}) - v_i(\xi - \epsilon_j, \xi, d_{ij})\}$. That is, i is reducing his investments in links in J in order to invest in his individual link. Let $d^* = \max_{j\in J} d_{ij}$, we have

$$\begin{aligned} w_i(\sum_{j\in J} \epsilon_j) &> \sum_{j\in J}\{v_i(\xi, \xi, d_{ij}) - v_i(\xi - \epsilon_j, \xi, d_{ij})\} \\ &\geq \sum_{j\in J}\{v_i(\xi, \xi, d^*) - v_i(\xi - \epsilon_j, \xi, d^*)\} \qquad \text{(I.3)} \\ &\geq v_i(\xi, \xi, d^*) - v_i(\xi - \sum_{j\in J} \epsilon_j, \xi, d^*) \qquad \text{(I.4)} \end{aligned}$$

   where (8) follows from $v_{xd}(x, \xi, d) \leq 0$, and (9) follows from $v_{xx}(x, \xi, d) \geq 0$. Now, since $v_{xx}(x, \xi, d) \geq 0$, if (8) is true for $\sum_{j\in J} \epsilon_j < \xi$, it is also true for $\sum_{j\in J} \epsilon_j = \xi$, hence $w_i(\xi) > v_i(\xi, \xi, d^*)$. This contradicts the fact that $\tilde{x}$ was created by the assignment algorithm.

We still need to show that $\tilde{x}$ is a WBE. Suppose that it's not, i.e. there exists (i, j) and (x_i, x_j) such that $u_i(x_i, x_j, \tilde{x}_{-i-j}) > u_i(\tilde{x})$ and $u_j(x_j, x_i, \tilde{x}_{-i-j}) > u_j(\tilde{x})$. From the construction of $\tilde{x}$, it must be the case that i, j are such that $\tilde{x}_i^j = \tilde{x}_j^i = 0$. Again, we have 2 cases :

1. $\delta_i < \kappa_i$ and $\delta_j < \kappa_j$. This is impossible since, from the assignment algorithm, it implies that $v_i(\xi, \xi, d_{ij}) < 0$.

2. $\delta_i = \kappa_i$. Then, if i has a profitable deviation, there exists $K \subseteq g_i$ such that $v_i(\sum_{k\in K} \epsilon_k, x^i_j, d_{ij}) > \sum_{k\in K}\{v_i(\xi, \xi, d_{ik}) - v_i(\xi - \epsilon_k, \xi, d_{ik})\}$. Let $d^*_i = \max_{k\in K} d_{ik}$, we have

$$\begin{aligned} v_i(\sum_{k\in K} \epsilon_k, x^i_j, d_{ij}) &> \sum_{k\in K}\{v_i(\xi, \xi, d_{ik}) - v_i(\xi - \epsilon_k, \xi, d_{ik})\} \\ &\geq \sum_{k\in K}\{v_i(\xi, \xi, d^*_i) - v_i(\xi - \epsilon_k, \xi, d^*_i)\} \qquad \text{(I.5)} \\ &\geq v_i(\xi, \xi, d^*_i) - v_i(\xi - \sum_{k\in K} \epsilon_j, \xi, d^*_i) \qquad \text{(I.6)} \end{aligned}$$

where (10) follows from $v_{xd}(x, \xi, d) \leq 0$, and (11) follows from $v_{xx}(x, \xi, d) \geq 0$. Now, since $v_{xx}(x, \xi, d) \geq 0$, if (11) is true for $\sum_{k\in K} \epsilon_k < \xi$, it is also true for $\sum_{k\in K} \epsilon_k = \xi$, hence $v_i(\xi, x^i_j, d_{ij}) > v_i(\xi, \xi, d^*_i)$.

We now turn to j. If $\delta_j = \kappa_j$, the same argument applies for j ; then $v_j(\xi, \xi, d_{ij}) > v_j(\xi, \xi, d^*_j)$ (and $v_i(\xi, \xi, d_{ij}) > v_i(\xi, \xi, d^*_i)$). Since we have $d_{ij} < d^*_i$ and $d_{ij} < d^*_j$, this contradicts the fact that $\tilde{x}$ was created by the assignment algorithm.

If $\delta_j < \kappa_j$, j has at least ξ to invest (and it is profitable to invest up to ξ since $v_x(x, y, d) > 0$), then together with the fact that $d_{ij} < d^*_i$, this contradicts the fact that $\tilde{x}$ is produced by the assignment algorithm.

□

Proof of proposition 2.3.4

From theorem 3.3, it is sufficient to show that for any $i, j \in N$, $x^j_i \in \{0, \xi\}$, at any NE.

Consider some $i, j \in N$, and suppose that $x^j_i \in (0, \xi)$. I show that this implies that there exists $k \in N$ such that $x^k_i \in (0, \xi)$. Suppose otherwise. Then, i still has resources available. Since $v_x(x, y, d) > 0$, i could increase x^j_i and be better off. Hence, x is not a NE, so it is not a WBE. Hence, there exists $k \in N \setminus \{i\}$ such that $x^k_i \in (0, \xi)$. There are 2 cases :

1. $[k = i]$. Since x is a NE, we must have the following.

   - If $x_i^i + x_i^j \geq \xi$, then

   $$w_i(x_i^i) + v_i(x_i^j, x_j^i, d_{ij}) \geq w_i(\xi) + v_i(x_i^j + x_i^i - \xi, x_j^i, d_{ij})$$
   $$w_i(x_i^i) + v_i(x_i^j, x_j^i, d_{ij}) \geq w_i(x_i^j + x_i^i - \xi) + v_i(\xi, x_j^i, d_{ij})$$

   Rewriting, we have

   $$w_i(\xi) - w_i(x_i^i) \leq v_i(x_i^j, x_j^i, d_{ij}) - v_i(x_i^j + x_i^i - \xi, x_j^i, d_{ij})$$
   $$w_i(x_i^i) - w_i(x_i^j + x_i^i - \xi) \geq v_i(\xi, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij})$$

   Since $v_{xx}(x, y, d) > 0$, we have $v_i(\xi, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij}) > v_i(x_i^j, x_j^i, d_{ij}) - v_i(x_i^j + x_i^i - \xi, x_j^i, d_{ij})$, and since $w''(x) > 0$, we have $w_i(\xi) - w_i(x_i^i) > w_i(x_i^i) - w_i(x_i^j + x_i^i - \xi)$. This is in contradiction with the above conditions, hence x is not a NE.

   - If $x_i^i + x_i^j < \xi$, then

   $$w_i(x_i^i) + v_i(x_i^j, x_j^i, d_{ij}) \geq w_i(x_i^i + x_i^j) + v_i(0, x_j^i, d_{ij})$$
   $$w_i(x_i^i) + v_i(x_i^j, x_j^i, d_{ij}) \geq w_i(0) + v_i(x_i^i + x_i^j, x_j^i, d_{ij})$$

   Rewriting, we have

   $$w_i(x_i^i + x_i^j) - w_i(x_i^i) \leq v_i(x_i^j, x_j^i, d_{ij}) - v_i(0, x_j^i, d_{ij})$$
   $$w_i(x_i^i) - w_i(0) \geq v_i(x_i^i + x_i^j, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij})$$

   Since $v_{xx}(x, y, d) > 0$, we have $v_i(x_i^j + x_i^i, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij}) > v_i(x_i^j, x_j^i, d_{ij}) - v_i(0, x_j^i, d_{ij})$, and since $w''(x) > 0$, we have $w_i(x_i^j + x_i^i) - w_i(x_i^i) > w_i(x_i^i) - w_i(0)$. Again, this is in contradiction with the above conditions, hence x is not a NE.

$i \neq k$ and $i \neq j$.

Since x is a NE, we must have the following :

- If $x_i^k + x_i^j \geq \xi$, then

$$v_i(x_i^k, x_k^i, d_{ik}) + v_i(x_i^j, x_j^i, d_{ij}) \geq v_i(\xi, x_k^i, d_{ik}) + v_i(x_i^j + x_i^k - \xi, x_j^i, d_{ij})$$
$$v_i(x_i^k, x_k^i, d_{ik}) + v_i(x_i^j, x_j^i, d_{ij}) \geq v_i(x_i^j + x_i^k - \xi, x_k^i, d_{ik}) + v_i(\xi, x_j^i, d_{ij})$$

Rewriting, we have

$$v_i(\xi, x_k^i, d_{ik}) - v_i(x_i^k, x_k^i, d_{ik}) \leq v_i(x_i^j, x_j^i, d_{ij}) - v_i(x_i^j + x_i^k - \xi, x_j^i, d_{ij})$$
$$v_i(x_i^k, x_k^i, d_{ik}) - v_i(x_i^j + x_i^k - \xi, x_k^i, d_{ik}) \geq v_i(\xi, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij})$$

Since $v_{xx}(x, y, d) > 0$, we have $v_i(\xi, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij}) > v_i(x_i^j, x_j^i, d_{ij}) - v_i(x_i^j + x_i^k - \xi, x_j^i, d_{ij})$, and $v_i(\xi, x_k^i, d_{ik}) - v_i(x_i^k, x_k^i, d_{ik}) > v_i(x_i^k, x_k^i, d_{ik}) - v_i(x_i^j + x_i^k - \xi, x_k^i, d_{ik})$. This is in contradiction with the above conditions, hence x is not a NE.

- If $x_i^i + x_i^j < \xi$, then

$$v_i(x_i^k, x_k^i, d_{ik}) + v_i(x_i^j, x_j^i, d_{ij}) \geq v_i(x_i^j + x_i^k, x_k^i, d_{ik}) + v_i(0, x_j^i, d_{ij})$$
$$v_i(x_i^k, x_k^i, d_{ik}) + v_i(x_i^j, x_j^i, d_{ij}) \geq v_i(0, x_k^i, d_{ik}) + v_i(x_i^j + x_i^k, x_j^i, d_{ij})$$

Rewriting, we have

$$v_i(x_i^j + x_i^k, x_k^i, d_{ik}) - v_i(x_i^k, x_k^i, d_{ik}) \leq v_i(x_i^j, x_j^i, d_{ij}) - v_i(0, x_j^i, d_{ij})$$
$$v_i(x_i^k, x_k^i, d_{ik}) - v_i(0, x_k^i, d_{ik}) \geq v_i(x_i^j + x_i^k, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij})$$

Since $v_{xx}(x, y, d) > 0$, we have $v_i(x_i^j + x_i^k, x_j^i, d_{ij}) - v_i(x_i^j, x_j^i, d_{ij}) > v_i(x_i^j, x_j^i, d_{ij}) - v_i(0, x_j^i, d_{ij})$, and $v_i(x_i^j + x_i^k, x_k^i, d_{ik}) - v_i(x_i^k, x_k^i, d_{ik}) > v_i(x_i^k, x_k^i, d_{ik}) - v_i(0, x_k^i, d_{ik})$. This is in contradiction with the above conditions, hence x is not a NE.

□

Proof of proposition 2.3.5

The proof is obvious from the proof of theorem 2.3.2 and theorem 2.3.3. One only has to remark that for any $i, j, k \in N$, $v_i(\xi, \xi, d_{ij}) \geq v_i(\xi, \xi, d_{ik})$ implies that $v_i(\xi, \xi, d_{ij}) > v_i(\xi, \xi, d_{ik})$ if we assume that $d_{ij} \neq d_{kl}$. □

Proof of proposition 2.3.7

The fact that any Strong NE needs to be produced by the assignment algorithm follows from propositions 2.3.2 and 2.3.4. Suppose that $x^* \in X$ is a BE, but not a Strong NE. There exists $S \subset N$ and $x_S \in \times_{i \in S} X_i$ such that $u_i(x_S, x^*_{-S}) > u_i(x^*)$ for all $i \in S$. We will show that under Strict Convexity or Finiteness, this implies that there exists a bilateral deviation.

Under Finiteness, $x_i \in \{0, \xi\}^n$ for all $i \in S$. Using the same argument as the one used in lemma 2.3.1, there exist at least one project created under a deviation by coalition S. That is, $\exists i, j \in S$, such that $x_i^{j*} = x_j^{i*} = 0$ and $x_i^j = x_j^i = \xi$. Since the utility functions are additive, this implies that i, j have a profitable bilateral deviation. Since resources invested in the link (i, j) must have come either from unused resources or from the deleation of another link since $x_i^j \in \{0, \xi\}$ for all $i, j \in N$.

Under Convexity, if it is profitable to withdraw resources from one link and invest in two new links, it is even better to invest in only one of those links. (This is exactly the argument used in proposition 2.3.3). Specifically, suppose that there exists $i, j, k \in S$ such that $x_i^j, x_i^k > 0$, and $x_i^{j*} = x_i^{k*} = 0$. Then, either $x_i^j = \xi$ and $x_i^k = 0$ or $x_i^j = 0$ and $x_i^k = \xi$ is better for i. Then, i is willing to make a bilateral deviation with j (wlog). Since the utilities are linear, it is also profitable for k (since it is under a joint deviation in S). Hence, there exists a bilateral deviation between i and j. □

I.2 Appendix I.2

The Assignment Algorithm

I generate a network g (represented by the adjacency matrix A) in which every individual invests as much as possible in every active link (i.e. $x_i^j \in \{0, \xi_i\}$ for all $i, j \in N$).

Let $\eta_i^j = v_i(\xi, \xi, d_{ij})$ for all $i, j \in N$ such that $i \neq j$, and $\eta_i^i = w_i(\xi)$, for all $i \in N$. This function represents the value of a link between two individuals. Now, define the (not necessarily unique) ordered list L^0 as follows : $L^0 = (d_{ij})_{i,j \in N: i<j}$, such that $L_1^0 \leq L_2^0 \leq ... \leq L_m^0$. The list L^0 is an ordered list of distance values, for all pairs of individuals. The number of elements in L^0 is the number of possible pairings between individuals in N, i.e. $n(n-1)/2$. Let L_l^0 be the element of position l in the list L^0. I note $(L_l^0)^{-1} = (i, j)$ if $L_l^0 = d_{ij}$.

The algorithm computes g and takes $L^t = L^0$ as inputs. It operates in two steps.

1 *Take the first element of the list L^t, i.e. L_1^t. Let $L_1^t = d_{ij}$.*
If $a_{ii} = 0$ or $a_{jj} = 0$,

1. *If $\eta_i^i \geq \eta_i^j$ and $\eta_i^i \geq 0$, then $a_{ii} = 1$*
2. *If $\eta_j^j \geq \eta_j^i$ and $\eta_j^j \geq 0$, then $a_{jj} = 1$*

Otherwise,

1. *If $\eta_i^j \geq 0$ and $\eta_j^i \geq 0$, then set $a_{ij} = a_{ji} = 1$.*
2. *If $\eta_i^j < 0$, then generate $L^{*i} = L^t \setminus \{d_{ik}\}_{k \in N: d_{ik} \in L^t}$. (That is, remove all distances associates with i, since all the following distances will be greater than d_{ij}.)*
3. *If $\eta_j^i < 0$, then generate $L^{*i} = L^t \setminus \{d_{jk}\}_{k \in N: d_{jk} \in L^t}$, i.e. do the same for j as we did for i.*

Generate $L^{t+1} = \{(d \in L^{i*} \cap L^{j*}) \setminus d_{ij}\}$.

2 *Repeat (1) for* $t = 1, ...$ *until* $|L^t| = 0$ *or until* $\exists i \in N$ *such that* $\delta_i = \kappa_i$.

For all $i \in N$ *such that* $\delta_i = \kappa_i$, *generate* $L^{*i} = L^t \setminus \{d_{ik}\}_{k \in N : d_{ik} \in L^t}$. *(That is, remove all distances associated with* i, *since he has no resources left.) Then, generate* $L^{t+1} = \cap_{i \in N} L^{i*}$ *and repeat (1).*

After the algorithm stops, I generate the allocation $\tilde{x}$ as follows. For all $i, j \in N$, if $a_{ij} = 1$, $\tilde{x}_i^j = \xi$, otherwise $\tilde{x}_i^j = 0$. Notice that by definition $\tilde{x} \in X$.

I.3 Appendix I.3

The Likelihood Function

I assume that no individual is isolated. The definition of structural homophily is : For all $ij \notin g$, $d_{ij} \geq d_{ik}$ for all $k \in g_i$ or $d_{ij} \geq d_{jk}$ for all $k \in g_j$. Then, since the ε_{ij} are independents, and $ln(d) \geq ln(d')$ iff $d \geq d'$, the probability that g exhibits structural homophily is

$$\Pi_{ij \notin g} \left\{ \Pi_{k \in g_i} \mathbb{P}(d_{ij} \geq d_{ik}) + \Pi_{k \in g_j} \mathbb{P}(d_{ij} \geq d_{jk}) - \Pi_{k \in g_i} \mathbb{P}(d_{ij} \geq d_{ik}) \Pi_{k \in g_j} \mathbb{P}(d_{ij} \geq d_{jk}) \right\}$$

This gives :

$$\mathbb{P}(d_{ij} \geq d_{ik}) = \mathbb{P}(\sum_{r=1}^{R} \beta_r \rho_r(\theta_i, \theta_j) + \varepsilon_{ij} \geq \sum_{r=1}^{R} \beta_r \rho_r(\theta_i, \theta_k) + \varepsilon_{ik})$$

At this point, the normalization of ε is necessary for the identification of β. Simplifying the last expression, we have :

$$\begin{aligned} \mathbb{P}(d_{ij} \geq d_{ik}) &= \mathbb{P}(Z \geq \sum_{r=1}^{R} \beta_r [\rho_r(\theta_i, \theta_k) - \rho_r(\theta_i, \theta_j)]) \\ &= 1 - \Phi(\sum_{r=1}^{R} \beta_r [\rho_r(\theta_i, \theta_k) - \rho_r(\theta_i, \theta_j)]) \end{aligned}$$

I.4 Appendix I.4

Figure I.1 – Standard deviation : 10

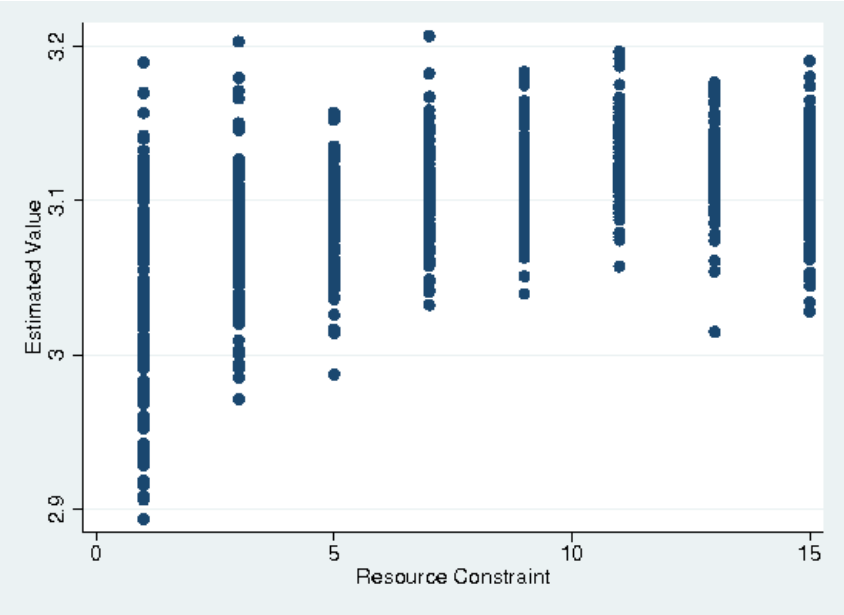


Figure I.2 – Standard deviation : 12

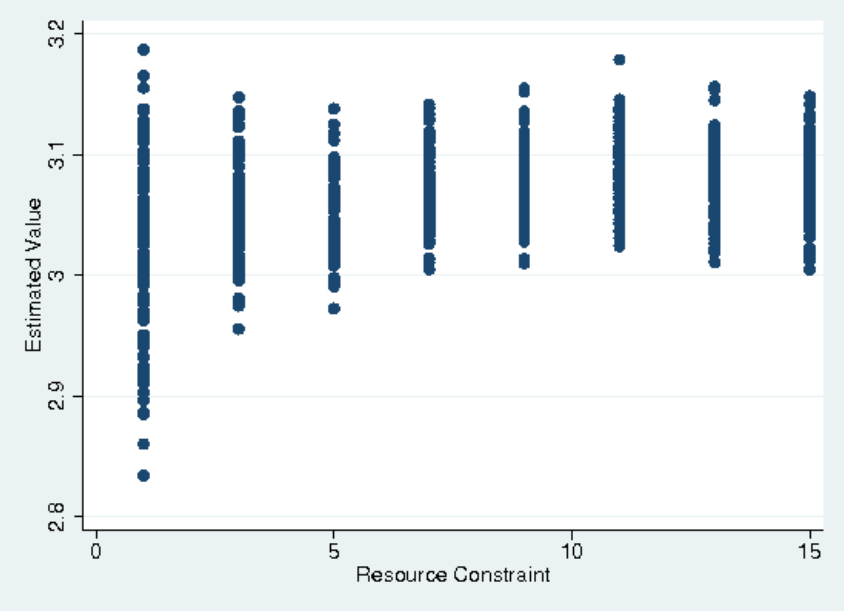


Figure I.3 – Standard deviation : 14

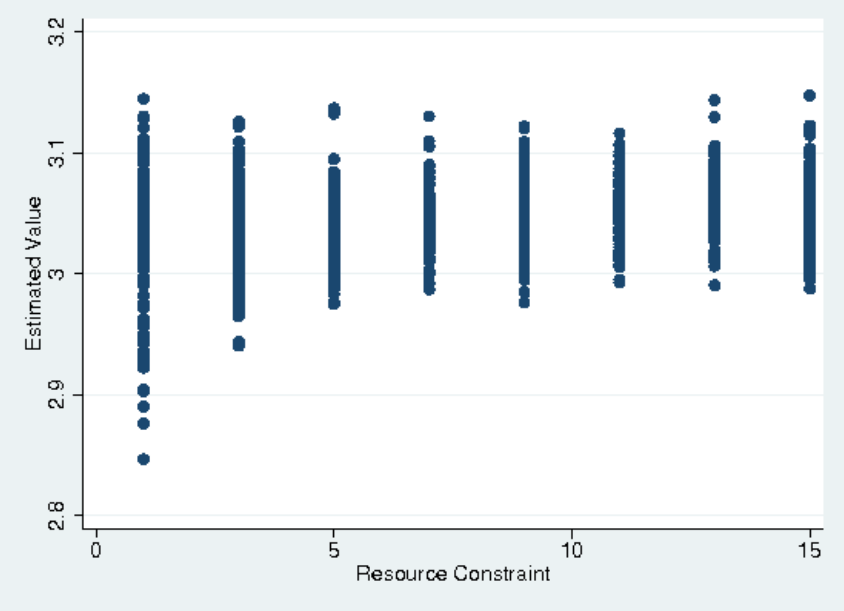
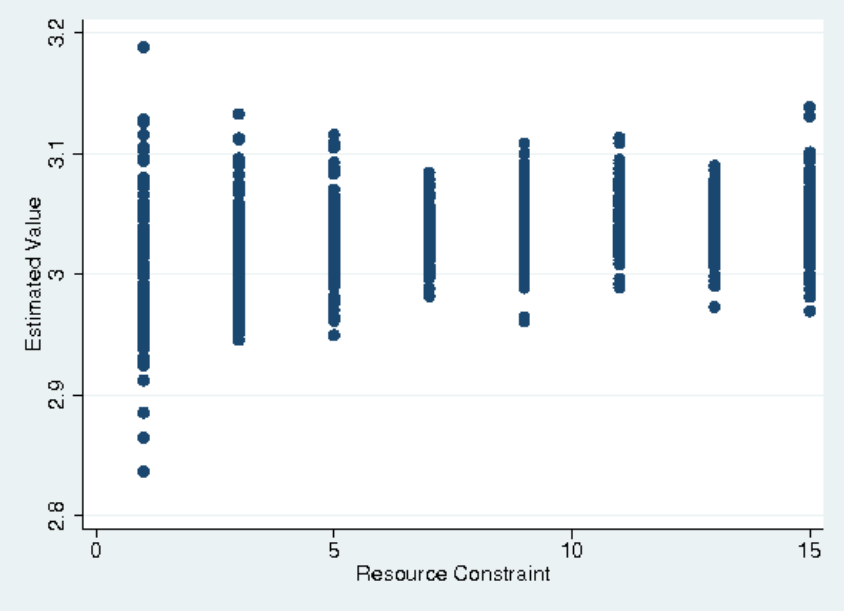


Figure I.4 – Standard deviation : 16



Annexe II

Appendix II

Proof of Theorem 3.3.1

Under assumption 5.1, it is sufficient to show that :[1]

$$\sup_{\theta\in\Theta} |\mathcal{L}_m(\theta) - \mathbb{E}(\mathcal{L}_m(\theta))| \to_{a.s.} 0, \text{ as } m \to \infty.$$

In order to show that this condition hold, it is sufficient to show that the conditions of theorem 2 and 3 from Jenish and Prucha (2009) hold. Specifically,

1. $d(r,s) > d_0 > 0$ for any $r, s \in S_m$
2. $(\Theta, \|.\|)$ is a totally bounded metric space.
3. Domination :

$$\lim \sup_{m\to\infty} \frac{1}{|S_m|} \sum_{s=1}^{m} \mathbb{E}(\bar{q}_{s,m}^p 1_{\{\bar{q}_{s,m}^p > k\}}) \to 0 \text{ as } k \to \infty,$$

   for some $p \geq 1$, and where $\bar{q}_{s,m} = \sup_\theta |q_{s,m}(z_{s,m}|x, g_m, \theta)|$.

4. Stochastic equicontinuity : For every $\epsilon > 0$,

$$limsup_m \frac{1}{|S_m|} \sum_{s=1}^{m} P(\sup_{\theta'\in\Theta} \sup_{\theta\in B(\theta',\delta)} |q_{s,m}(\theta) - q_{s,m}(\theta')| > \epsilon) \to 0 \text{ as } \delta \to 0,$$

   where $B(\theta', \delta)$ is the open ball $\{\theta \in \Theta : \|(\theta' - \theta)\| < \delta\}$.

5. $\sup_m \sup_{s\in S_m} \mathbb{E}[\sup_{\theta\in\Theta} |q_{s,m}(\theta)|^{(1+\eta)}] < \infty$ for some $\eta > 0$.
6. $\sum_{d=1}^{\infty} d^{T-1} \bar{\phi}_{1,1}(d) < \infty$.

[1] see for instance Gallant and White (1988), pp.18.

Condition 1 is implied by assumption 5.2. Condition 2 is verified by construction, and condition 5 and 6 are just assumption 5.3 and ϕ-mixing(2). Conditions 3 and 4 hold from the following : Under condition 5, $\sup_\theta |q_{s,m}(z_{s,m}|x, g_m, \theta)|$ is $L^{(1+\eta)}$ integrable which implies the uniform $L^{(1+\eta)}$ integrability of $|q_{s,m}(z_{s,m}|x, g_m, \theta)|$.

The next lemma shows that assumption 5.4 implies condition 4.

Lemma II.0.1 *Condition 4 is implied by assumption 5.4.*

Proof From the mean value theorem, we can write

$$q_{s,m}(\theta) \quad = \quad q_{s,m}(\theta') + \frac{\partial q_{s,m}(\tilde{\theta})}{\partial \theta}(\theta - \theta'),$$

Thus,

$$\begin{aligned} |q_{s,m}(\theta) - q_{s,m}(\theta')| &\leq |\frac{\partial q_{s,m}(\tilde{\theta})}{\partial \theta}|\|(\theta - \theta')\| \\ &\leq \sup_{\theta \in \Theta} |\frac{\partial q_{s,m}(\theta)}{\partial \theta}|\|(\theta - \theta')\|. \end{aligned}$$

According to Proposition 1 of Jenish and Prucha (2009), $q_{s,m}(\theta)$ is L_0 stochastically equicontinuous on Θ if the following $Cesàro$ sums is finite. i.e

$$limsup_m \frac{1}{|S_m|} \sum_{s=1}^{m} \mathbb{E}(\sup_{\theta \in \Theta} |\frac{\partial q_{s,m}(\theta)}{\partial \theta}|) < \infty.$$

However, under assumption 5.4, each term of the $Cesàro$ sums is finite, in the sense that $\sup_m \sup_{s \in S_m} \mathbb{E}[\sup_{\theta \in \Theta} |\frac{\partial q_{s,m}(\theta)}{\partial \theta}|] < \infty$. This fact completes the proof. □

From the previous lemma, conditions 1-6 are respected, hence theorem 2 and 3 from Jenish and Prucha (2009) apply. This completes the proof. □

Proof of Theorem 3.3.2

We want to show that $\sqrt{m}(\hat{\theta}_m - \theta_0) \Rightarrow N(0, D_0(\theta_0)^{-1}B_0(\theta_0)D_0(\theta_0)^{-1})$. From the mean value theorem, we have that

$$\begin{aligned}\frac{\partial \mathcal{L}_m(\hat{\theta}_m)}{\partial \theta} &= \frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta} + \frac{\partial^2 \mathcal{L}_m(\overline{\theta}_m)}{\partial \theta \partial \theta'} \\ 0 &= \frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta} + \frac{\partial^2 \mathcal{L}_m(\overline{\theta}_m)}{\partial \theta \partial \theta'}(\hat{\theta}_m - \theta_0).\end{aligned}$$

and

$$\begin{aligned}\sqrt{m}(\hat{\theta}_m - \theta_0) &= -\sqrt{m}[\frac{\partial^2 \mathcal{L}_m(\overline{\theta}_m)}{\partial \theta \partial \theta'}]^{-1}\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta} \\ &= -[\frac{\partial^2 \mathcal{L}_m(\overline{\theta}_m)}{\partial \theta \partial \theta'}]^{-1}[\frac{\sigma_m}{\sqrt{m}}][\sigma_m^{-1}Q_m],\end{aligned}$$

where $\sigma_m^2 = Var(Q_m)$ and $Q_m = \sum_{s=1}^{m} \frac{\partial q_{s,m}(\theta_0)}{\partial \theta}$.

Then, it is sufficient to show the following :

1. $\frac{\sigma_m^2}{m} \to B_0(\theta_0)$;
2. $\sigma_m^{-1}Q_m \Rightarrow N(0, I)$;
3. $[\frac{\partial^2 \mathcal{L}_m(\overline{\theta}_m)}{\partial \theta \partial \theta'}] \to_p D_0(\theta_0)$.

Again, we proceed in a series of lemmata.

Lemma II.0.2 *Under assumptions 5.1,* $\frac{\sigma_m^2}{m} \to B_0(\theta_0)$.

Proof

$$\begin{aligned}\frac{1}{m}\sigma_m^2 &= \frac{1}{m}Var(m\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}) \\ &= m\mathbb{E}[\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta'}] + m\mathbb{E}[\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}]\mathbb{E}[\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta'}] \\ &= m\mathbb{E}[\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}\left(\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}\right)'].\end{aligned}$$

where the last inequality holds since $\mathbb{E}[\frac{\partial \mathcal{L}_m(\theta_0)}{\partial \theta}] = 0$, as θ_0 maximizes $\mathbb{E}[\mathcal{L}_m(\theta)]$ (Assumption 5.1). Hence, $\frac{\sigma_m^2}{m} \to B_0(\theta_0)$. $\square$

Lemma II.0.3 *Under assumptions 3, and 6,* $\sigma_m^{-1} Q_m \Rightarrow N(0, I)$

Proof It is sufficient to show that the conditions for theorem 1 from Jenish and Prucha (2009) hold. Specifically,

1. $d(r, s) > d_0 > 0$ for any $r, s \in S_m$.
2. ϕ-mixing on Random Fields.
3. $\sup_m \sup_{s \in S_m} \mathbb{E}[\sup_{\theta \in \Theta} |\frac{\partial q_{s,m}(\theta)}{\partial \theta}|^2] < \infty$.
4. $\liminf_{m \to \infty} \frac{\sigma_m^2}{m} > 0$.

Condition 1 is implied by assumption 5.2. Condition 3 is just assumption 6.5, and condition 4 is implied by lemma II.0.2. $\square$

Lemma II.0.4 $\frac{\partial^2 \mathcal{L}_m(\bar{\theta}_m)}{\partial \theta \partial \theta'} \to_p D_0(\theta_0)$

Proof The proof is identical to the proof for the consistency of $\hat{\theta}$, replacing $q_{s,m}(\theta)$ by $D_{s,m}(\theta)$, and using assumptions 6.3 and 6.4 instead of assumptions 5.3 and 5.4. $\square$

Putting together lemmata 7.2, 7.3 and 7.4 completes the proof. $\square$

Proof of Proposition 3.4.1

Let $H_i^j = h_i^j[N_i(g), N_j(g), d(i, j)] + \varepsilon_{ij}$ where $\varepsilon \sim N(0, 1)$. We show that under assumption 7 and 8, ϕ-mixing is respected. Recall that $\phi(\mathcal{A}, \mathcal{B}) = \sup\{|P(A|B) - P(A)|, A \in \mathcal{A}, B \in \mathcal{B}, P(B) > 0\}$. Formally, A and B are subsets of pairs, i.e. $A, B \in S_m$. Let $i \in s \in A$ and $j \in s \in B$.

We have that $P(A) = P(A|\exists ij \in g)P(\exists ij \in g) + P(A|\nexists ij \in g)P(\nexists ij \in g)$ and $P(A|B) = P(A|B \cap \exists ij \in g)P(\exists ij \in g) + P(A|B \cap \nexists ij \in g)P(\nexists ij \in g)$. Since the

payoff function only depends on direct links, $P(A|B \cap \nexists ij \in g) = P(A|\nexists ij \in g)$. Hence, we can rewrite

$$\phi(\mathcal{A}, \mathcal{B}) = \phi(\mathcal{A}, \mathcal{B}|\exists ij \in g)P(\exists ij \in g)$$

Since, for any A, B, $\phi(\mathcal{A}, \mathcal{B}) \in [0, 1]$, we have that $\phi(\mathcal{A}, \mathcal{B}) \leq P(\exists ij \in g)$. Let $\bar{h}(d) = \sup_\theta \sup_g \sup_{ij} h_i^j(g, x, d; \theta)$ and $\underline{h}(d) = \inf_\theta \inf_g \inf_{ij} h_i^j(g, x, d; \theta)$. We then have that $\bar{\phi}_{k,l} \leq 4kl\Phi[\bar{h}(d)]$ since there can be a maximum of $2k$ individuals in A and $2l$ individuals in B. That is, the sum of the probabilities for each possible pairs between A and B, and for the maximal value for h_i^j. Notice that by the properties of the Hausdorf distance, if $d(i, j) \geq c$ for some $c > 0$ and all $i \in s \in A$ and $j \in r \in B$, then $d(A, B) \geq c$.

Now, we know that the Chernoff bound for Φ is such that $\Phi[\bar{h}(d)] \leq \frac{1}{2} \exp\{-\frac{1}{2}\bar{h}(d)^2\}$ for $\bar{h}(d) < 0$, which is true for d big enough from assumption (7.1). Then, a sufficient condition for assumption (4.1) and (4.2) is $\bar{\phi}_{k,l}(d) \leq 2kl \exp\{-\frac{1}{2}\bar{h}(d)^2\}$ for $k + l \leq 4$ or equivalently :

$$d^{T-1}\bar{\phi}_{k,l}(d) \leq 2kld^{T-1} \exp\{-\frac{1}{2}\bar{h}(d)^2\}$$

for all $d > \bar{d}$ for some $\bar{d} > 0$ and $k + l \leq 4$. Then, assumption (4.1) and (4.2) hold if $\sum_{d=1}^{\infty} d^{T-1} \exp\{-\frac{1}{2}\bar{h}(d)^2\}$ converges. According to Cauchy's rule, this last sum converges if $\lim_{d\to\infty} \exp\left\{-\frac{\bar{h}(d)^2}{2d}\right\} \in [0, 1)$. Which is true under assumption (7.2).

Now, ϕ-mixing (3) is different since $l = \infty$, so the upper bound goes to infinity. Specifically, condition (3) implies that there exists $C > 0$, $\overline{m}$, $\overline{d}$ such that $d^{T+\epsilon}\bar{\phi}_{1,m}(d) \leq C$, for some $\epsilon > 0$, for all $m > \overline{m}$ and $d > \overline{d}$. Using the Chernoff bound again we have that $d^{T+\epsilon}\bar{\phi}_{1,m}(d)$ is bounded when m goes to infinity if

$$\lim_{m\to\infty} md_m^{T+\epsilon} \exp\left\{-\frac{\bar{h}(d_m)^2}{2d_m}\right\} < \infty$$

assuming the increasing domain assumption 8, and the asymptotic homophily assumption (7.2). □

Proof of Proposition 3.4.2

Let $P(A \leftrightarrow B)$ be the probability that there exist a path between an individual in a site in A and and individual in a site in B. Using the same argument as in the proof of proposition 3.4.1, we have that $\phi(\mathcal{A}, \mathcal{B}) \leq P(A \leftrightarrow B)$. The probability $P(A-B)$ is however not trivial to compute. Instead, we use the fact that $P(A-B) = P(\exists k - B : ik \in g)$ for some i in a site in A. Since k is connected to B, there are two possibilities : (1) the distance between k and B is finite, or (2) the distance between k and B is infinite, and k is reached from B using an infinite number of links.

We start with the second possibility. In that case, from assumption 9, the realization on B does not depends on k, hence $P(A|B) = P(A)$.

Now, suppose that the distance between k and B is finite. Then, as in the proof of proposition 3.4.1, we can write

$$d^{T-1}\bar{\phi}_{k,l}(d) \leq 2kmd^{T-1}\exp\{-\frac{1}{2}\bar{h}(d)^2\}$$

for all $d > \bar{d}$ for some $\bar{d} > 0$. The remaining of the proof is omitted as it is identical to the proof of proposition 3.4.1. □

Conley's (1999) estimator

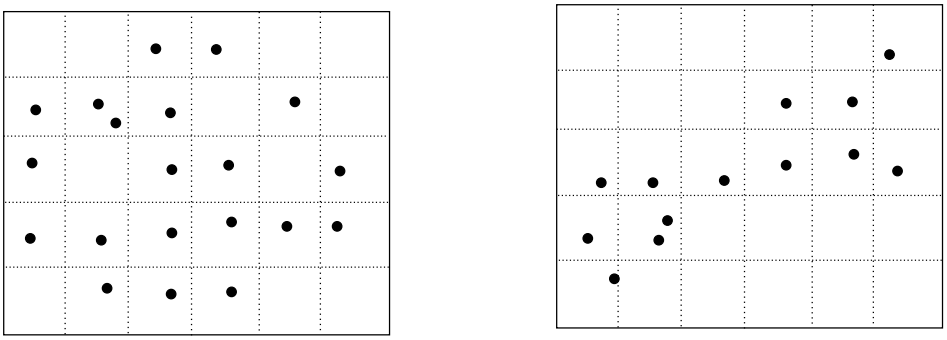
Conley (1999) provides an estimator when $\mathcal{X} \subset \mathbb{R}^2$ and $\{Z_{s,m}; s \in S_m, m \in \mathbb{N}\}$ is α-mixing and stationary. This approach has also been recently used by Wang et al. (2010) in the context of the estimation of a spatial probit. We propose to extend Conley's (1999) estimator for $\mathcal{X} \subset \mathbb{R}^T$, where $T \geq 1$.

We consider a compact subset of the space of individual characteristics, i.e. $Y \subset \mathcal{X}$. We define a random process Λ on a regular lattice on Y such that $\Lambda_y = 1$ if the location $y = (y_1, ..., y_T)$ is sampled, and $\Lambda_y = 0$ otherwise. We assume that Λ is independent of the underlying random field, has finite expectation, and is stationary. Intuitively, since the lattice is regular, it gives an idea of the dependence structure between the observations. Consider Figure 1 below, where $\mathcal{X} = \mathbb{R}^2$ for

presentation purposes. Sampled pairs are represented by the black circles.

Figure II.1 – Regular Lattice and Dependence Structure

(a) Uniform Dependence Structure (b) Directed Dependence Structure



In Figure 1a, sites are distributed more or less uniformly in Y. In Figure 1b however, the dependence structure seems to be more directed. Now, lets define $\bar{y} = (\bar{y}_1, ..., \bar{y}_T)$ to be the maximal location for Λ_y in every dimension. Notice that this quantity is well defined since Y is compact. For instance, for the lattice in Figure 1, $\bar{y} = (6, 5)$.

Now, let $\hat{q}_y(\theta) = \frac{1}{n(y)} \sum_{s \in y} q_{s,m}(\theta)$, where $s \in y$ is a sampled pairs s in location y, and $n(y)$ is the number of sampled pairs in location y. We define the following process, for any location y :

$$R_y(\theta) = \begin{cases} \frac{\partial \hat{q}_y}{\partial \theta}(\theta) & \text{if } \Lambda_y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let m^* be the number of sampled locations.[2] We can now present our proposed

[2] A simple way to compute m^* is to count the number of times $\Lambda_y = 1$.

estimator, based on a generalization of Conley (1999) :

$$\begin{aligned} B_m(\theta) &= \frac{1}{m^*}\sum_{y_1=0}^{\tilde{y}_1}...\sum_{y_T=0}^{\tilde{y}_T}\sum_{y_1'=y_1+1}^{\bar{y}_1}...\sum_{y_T'=y_T+1}^{\bar{y}_T}\Gamma_{\tilde{y}}(y)\left[R_{y'}(\theta)R'_{y'-y}(\theta)+R_{y'-y}(\theta)R'_{y'}(\theta)\right] \\ &- \frac{1}{m^*}\sum_{y_1=1}^{\bar{y}_1}...\sum_{y_T=1}^{\bar{y}_T}R_y(\theta)R'_y(\theta) \end{aligned} \tag{II.1}$$

Where $\tilde{y} < \bar{y}$, and $\Gamma_{\tilde{y}}(y)$ is a kernel function. For instance, Conley (1999) proposed to use $\tilde{y} = o(\bar{y}^{1/3})$, i.e. a bound of the same order as the cubic root of $\bar{y}$, and the following Bartlett window kernel :

$$\Gamma_{\tilde{y}}(y) = \begin{cases} (1-\frac{|y_1|}{\tilde{y}_1})...(1-\frac{|y_T|}{\tilde{y}_T}) & \text{for } |y_1|<\tilde{y}_1,...,|y_T|<\tilde{y}_T \\ 0 & \text{otherwise} \end{cases}$$

As in the estimation of HAC variances, the precise choice of $\tilde{y}$ and $\Gamma_{\tilde{y}}(y)$ will depend on the specific application. With that regard, we can easily show that the estimator in (II.1) when $T = 1$ is equivalent to a HAC estimator.

Lets rewrite the estimator for $T = 1$:

$$\begin{aligned} B_m(\theta) &= \frac{2}{m}\sum_{k=0}^{\tilde{y}_1}\sum_{y=k+1}^{\bar{y}_1}\Gamma_{\tilde{y}}(k)R_y(\theta)R'_{y-k}(\theta) - \frac{1}{m}\sum_{y=1}^{\bar{y}_1}R_y(\theta)R'_y(\theta) \\ &= \hat{\gamma}(0) + 2\sum_{k=1}^{\tilde{y}_1}\Gamma_{\tilde{y}}(k)\hat{\gamma}(k) \end{aligned}$$

where $\hat{\gamma}(0) = \frac{1}{m}\sum_{y=0}^{\bar{y}_1}R_y(\theta)R'_y(\theta)$ is the estimation of the variance of the process R_y, and $\hat{\gamma}(k) = \frac{1}{m}\sum_{y=k+1}^{\bar{y}_1}R_y(\theta)R'_{y-k}(\theta)$ the estimation of the autocovariance of the process R_y. Then, in one dimension our proposed estimator become exactly the HAC variance estimator for the covariance stationary process R_y, using the Bartlett kernel. In our case here, under some ϕ mixing conditions we may ensure that $\gamma(k) \to 0$ as $k \to \infty$.

Bester et al. (2012)

Let $\mathcal{X}$ be partitioned into groups, or clusters : $c = 1, ..., C$. Bester et al. (2012) propose to use the following CV estimator :

$$\hat{B}_m(\theta) = \frac{1}{m} \sum_{s \in S} \sum_{r \in S} \mathbb{I}(c_s = c_r) \frac{\partial q_{s,m}(\theta)}{\partial \theta} \left(\frac{\partial q_{s,m}(\theta)}{\partial \theta} \right)'$$

Where c_s is the group in which $s \in S$ is located. This is the usual Cluster-Variance estimator. It has the advantage of being easy and fast to implement. In practice, the constructions of those groups is not necessarily straightforward. Bester et al. (2012) recommend to use a relatively small number of large groups. An important requirement however is a boundary condition which states that most of the pairs in groups are located in the interior (i.e. not on the boundary) of those groups in $\mathcal{X}$. Specifically, let $\partial(c_s)$ be the boundary of the group c_s, and $\bar{c}_m$ be the average number of pairs in a group, then one should have $\partial(c_s) < \bar{c}_m^{(T-1)/T}$, where $T \geq 1$ is the dimension of $\mathcal{X}$.

Annexe III

Appendix III

Tableau III.I – Descriptive statistics

Course	Variable	Mean	S.D.
French	Score	72.647	14.086
(Sec. 5)	Age	16.142	0.488
	Socio-ec. Index	-	-
	Perc. High	0.328	0.469
	Perc. Med.	0.409	0.492
	Gender (Female=1)	0.549	0.500
	Foreign	0.111	0.310
	Secondary 5	0.985	0.120
	Number of observations	41778	
	Number of groups	314	
	Size of true groups	133.4	115.7
	Size of observed groups	133.1	115.4
Science	Score	74.689	17.671
(Sec. 4)	Age	15.255	0.610
	Socio-ec. Index	-	-
	Perc. High	0.338	0.470
	Perc. Med.	0.402	0.490
	Gender (Female=1)	0.527	0.499
	Foreign	0.127	0.333
	Secondary 5	0.077	0.267
	Number of observations	54981	
	Number of groups	378	
	Size of true groups	146.0	134.2
	Size of observed groups	145.5	133.7

Tableau III.I – Descriptive statistics (continued)

Course	Variable	Mean	S.D.
Math †	Score	62.088	15.83
(Sec. 5)	Age	16.272	0.574
	Socio-ec. Index	-	-
	Perc. High	0.303	0.460
	Perc. Med.	0.400	0.490
	Gender (Female=1)	0.540	0.498
	Foreign	0.111	0.314
	Secondary 5	0.957	0.202
	Number of observations	15771	
	Number of groups	361	
	Size of true groups	50.7	49.9
	Size of observed groups	49.9	49.7
History	Score	70.156	17.280
(Sec. 4)	Age	15.230	0.580
	Socio-ec. Index	-	-
	Perc. High	0.337	0.473
	Perc. Med.	0.403	0.491
	Gender (Female=1)	0.533	0.499
	Foreign	0.127	0.333
	Secondary 5	0.044	0.205
	Number of observations	55057	
	Number of groups	382	
	Size of true groups	144.6	134.8
	Size of observed groups	144.1	134.5

† Math refers to Math 514 (Secondary 5 regular course).

Tableau III.II – Peer Effects on Student Achievement[a]

Conditional Maximum Likelihood and Pseudo Conditional Maximum Likelihood

	French	Science	Math	History
Endogenous effect	0.296	-0.231	0.827**	0.641
	(0.605)	(0.414)	(0.319)	(0.399)
	[0.327]	[0.234]	[0.249]	[0.272]
Contextual effects				
Age	-39.435**	-19.493*	0.838	-31.607**
	(12.798)	(10.237)	(9.874)	(13.655)
	[10.987]	[8.893]	[7.382]	[9.471]
Socio-ec. Index (High)	16.613	8.941	29.310*	-6.367
	(15.096)	(21.637)	(15.422)	(17.505)
	[17.530]	[22.454]	[15.580]	[18.947]
Socio-ec. Index (Medium)	-4.765	22.156	18.246	-6.713
	(14.907)	(18.648)	(13.334)	(19.207)
	[16.870]	[17.783]	[13.726]	[18.565]
Gender (Female=1)	-24.870	14.852	15.558*	-11.837
	(15.927)	(13.425)	(9.006)	(12.633)
	[14.393]	[12.178]	[9.491]	[12.413]
Foreign	-26.699*	-8.844	-2.654	29.148*
	(14.828)	(13.737)	(12.802)	(15.304)
	[15.861]	[16.953]	[12.143]	[18.007]
Secondary 5	167.926**	-0.334	-6.080	24.041
	(54.842)	(25.048)	(39.168)	(24.027)
	[41.179]	[19.956]	[26.056]	[21.166]

(continued on the next page)

Tableau III.II – Peer Effects on Student Achievement (continued)[a]

Conditional Maximum Likelihood and Pseudo Conditional Maximum Likelihood

	French	Science	Math	History
Individual effects				
Age	-7.998**	-8.293**	-4.868**	-7.942**
	(0.239)	(0.269)	(0.330)	(0.253)
	[0.162]	[0.151]	[0.271]	[0.151]
Socio-ec. Index (High)	1.423**	1.609**	2.112**	2.019**
	(0.308)	(0.297)	(0.496)	(0.322)
	[0.245]	[0.268]	[0.500]	[0.261]
Socio-ec. Index (Medium)	0.670**	0.785**	1.189**	0.795**
	(0.266)	(0.260)	(0.464)	(0.272)
	[0.220]	[0.230]	[0.435]	[0.234]
Gender (Female=1)	3.807**	0.319	1.018**	-1.641**
	(0.196)	(0.200)	(0.325)	(0.207)
	[0.162]	[0.158]	[0.301]	[0.159]
Foreign	-2.596**	2.095**	-0.081	0.806**
	(0.314)	(0.380)	(0.513)	(0.384)
	[0.279]	[0.278]	[0.548]	[0.284]
Secondary 5	10.519**	1.653**	6.474**	3.126**
	(1.258)	(0.560)	(1.096)	(0.537)
	[0.676]	[0.328]	[0.767]	[0.399]
Log-likelihood	-162548.552	-226078.181	-62420.961	-226216.108

Notes :
CML unrobust standard errors in brackets. Pseudo CML robust standard errors in parentheses.
** indicates 5% significance level, based on robust s.e.
* indicates 10% significance level, based on robust s.e.
[a]The dependent variable is the score on June 2005 provincial secondary exams.

Tableau III.III – Peer Effects on Student Achievement[a]

2SLS Estimation with Group Fixed Effect[b]

	French	Sciences	Math	History
Endogenous effect	1.378	-0.509	-0.037	0.787
	(1.468)	(0.764)	(0.477)	(0.980)
Individual effects				
Age	-7.690**	-7.962**	-4.606**	-7.609**
	(0.197)	(0.167)	(0.228)	(0.163)
Socio-ec. Index (High)	1.373**	1.754**	1.836**	2.041**
	(0.242)	(0.250)	(0.423)	(0.248)
Socio-ec. Index (Medium)	0.661**	0.826**	1.069**	0.803**
	(0.221)	(0.219)	(0.365)	(0.221)
Gender (Female=1)	3.871**	0.333**	0.965**	-1.553**
	(0.164)	(0.159)	(0.265)	(0.157)
Foreign	-2.514**	2.128**	-0.005	0.716**
	(0.282)	(0.270)	(0.496)	(0.276)
Secondary 5	9.516**	1.415**	6.674**	2.910**
	(0.781)	(0.327)	(0.741)	(0.390)
Contextual effects				
Age	4.205	13.496**	6.713**	8.552**
	(4.845)	(3.050)	(1.712)	(4.036)
Socio-ec. Index (High)	7.364	30.997*	15.962**	-6.246
	(17.305)	(16.678)	(7.641)	(15.620)
Socio-ec. Index (Medium)	-7.103	26.344*	13.501*	-8.047
	(16.813)	(13.908)	(7.555)	(14.598)
Gender (Female=1)	-21.310*	15.637	13.237**	0.567
	(12.261)	(12.202)	(5.808)	(11.708)
Foreign	-15.732	-2.232	-0.065	19.385
	(12.571)	(11.449)	(7.189)	(12.903)
Secondary 5	40.184	-17.370	7.825	2.537
	(36.380)	(14.470)	(21.360)	(23.060)
Sargan Test	23.52	0.54	1.40	5.35
[*p-value*]	[0.00]	[1.00]	[0.97]	[0.50]
Stock and Yogo Test	706.84	1055.92	464.43	660.40
[*Critical Value for b=0.05 at sign. level of 5%*]	[18.37]	[18.37]	[18.37]	[18.37]

Notes :
Robust standard errors in parentheses
** indicates 5% significance level
* indicates 10% significance level
[a]The dependent variable is the score on June 2005 provincial secondary exams.

Tableau III.IV – Group Size Variation

Simulations using CML

Avg. Group size	Group sizes Range	**Endogenous effect** CML Avg. Coeff	CML SE	**Contextual effects - Age** CML Avg. Coeff	CML SE	**Contextual effects - Gender** CML Avg. Coeff	CML SE
10	[3 ;17]	0.35	0.00	-40.01	0.25	-25.01	0.33
10	[5 ;15]	0.35	0.00	-40.00	0.35	-24.99	0.74
10	[7 ;13]	0.35	0.02	-40.00	0.53	-25.01	1.50
10	[9 ;11]	0.57	0.38	-40.27	1.79	-26.97	5.78
20	[3 ;37]	0.35	0.00	-40.01	0.27	-25.03	0.44
20	[8 ;32]	0.35	0.02	-40.00	0.50	-25.02	1.10
20	[13 ;27]	0.35	0.09	-39.95	1.23	-25.04	2.11
20	[18 ;22]	0.94	1.56	-37.98	5.37	-28.55	8.47
40	[3 ;77]	0.35	0.00	-39.98	0.41	-25.03	0.65
40	[12 ;68]	0.36	0.07	-39.97	1.42	-25.05	1.66
40	[21 ;59]	0.39	0.20	-39.85	2.76	-25.14	2.67
40	[30 ;50]	0.72	0.85	-37.92	5.82	-26.93	5.30
40	[39 ;41]	1.00	155.98	-36.25	78.67	-26.28	69.22
80	[3-157]	0.35	0.01	-39.99	0.68	-25.05	0.98
80	[18-142]	0.43	0.19	-39.55	2.93	-25.40	2.49
80	[33-127]	0.57	0.46	-38.54	4.87	-25.96	3.66
80	[48-112]	0.87	1.20	-36.47	8.02	-27.10	5.75
80	[63-97]	1.00	5.27	-35.75	17.05	-27.74	11.97
120	[3-237]	0.36	0.01	-39.99	0.99	-25.10	1.50
120	[28-212]	0.64	0.51	-38.14	5.22	-26.34	3.79
120	[53-187]	0.89	1.30	-35.99	8.69	-27.25	5.85
120	[78-162]	1.00	3.85	-35.34	15.16	-27.65	9.94
120	[103-137]	1.00	25.15	-35.32	39.00	-28.20	25.29

Note :
True value of parameters : Endogenous effect : 0.35 ; Contextual effects - Age : -40 ; Contextual effects - Gender : -25.

Tableau III.V – Simulations Calibrated on French Sample

(1000 replications)

	CML	2SLS	G2SLS	OLS
Endogenous effect	0.391	-0.873	0.495	-33.571
	(0.101)	(0.852)	(167.702)	(3.688)
Individual effects				
Age	-8.002	-7.920	-8.006	-5.758
	(0.145)	(0.149)	(10.021)	(0.545)
Gender (Female=1)	3.798	3.822	3.828	4.480
	(0.147)	(0.139)	(1.693)	(0.554)
Contextual effects				
Age	-39.996	-38.085	-39.540	17.373
	(9.996)	(7.579)	(167.394)	(76.788)
Gender (Female=1)	-25.329	-16.703	-21.857	210.526
	(10.733)	(10.092)	(692.625)	(74.714)

Notes : Average standard errors are in parentheses. The group sizes are calibrated on our French sample. $\sigma^2 = \hat{\sigma}^2$ (calibrated)= 154.704. True value of parameters : Endogenous effect : 0.35 ; Individual effects - Age : -8 ; Individual effects - Gender : 3.8 ; Contextual effects - Age : -40 ; Contextual effects - Gender : -25.

Tableau III.VI – Peer Effects on Student Achievement[a]

Pseudo Conditional Maximum Likelihood (constrained)

	French	Science	Math	History
Endogenous effect	0.297	-0.31	0.801**	0.655
	(0.610)	(0.399)	(0.300)	(0.405)
Individual effects				
Age	-8.011**	-8.297**	-4.884**	-7.908**
	(0.238)	(0.266)	(0.267)	(0.247)
Socio-ec. Index (High)	1.314**	1.536**	1.845**	2.062**
	(0.264)	(0.255)	(0.487)	(0.290)
Socio-ec. Index (Medium)	0.698**	0.636**	0.818**	0.842**
	(0.217)	(0.220)	(0.374)	(0.231)
Gender (Female=1)	3.983**	0.221	1.026**	-1.563**
	(0.150)	(0.173)	(0.326)	(0.179)
Foreign	-2.537**	2.156**	0	0.812**
	(0.306)	(0.349)	(0.439)	(0.385)
Secondary 5	10.499**	1.657**	6.571**	2.944**
	(1.258)	(0.517)	(0.861)	(0.498)
Contextual effects				
Age	-41.056**	-20.456*	-	-27.305**
	(12.476)	(11.646)		(11.887)
Socio-ec. Index (High)	-	-	17.124	-
			(15.546)	
Socio-ec. Index (Medium)	-	-	-	-
Gender (Female=1)	-	-	16.215*	-
			(9.087)	
Foreign	-19.559	-	-	29.865**
	(11.969)			(15.082)
Secondary 5	165.537**	-	-	-
	(54.645)			
Log-likelihood	-162551.11	-226079.659	-62421.964	-226217.241
Likelihood Ratio[b]	5.122*	2.864	2.006	2.262

Notes :
Robust standard errors in parentheses
** indicates 5% significance level
* indicates 10% significance level
[a]The dependent variable is the score on June 2005 provincial secondary exams.
[b] The LR statistic is used to test the joint equality to zero of contextual effects that are not individually significant (at 10%) in the pseudo CML unconstrained version (see Table 2).

Tableau III.VII – Peer Effects on Student Achievement[a]

Generalized 2SLS Estimation[b]

	French	Sciences	Math	History
Endogenous effect	-2.104	-0.015	-0.162	-2.753
	(3.619)	(0.734)	(0.465)	(1.717)
Individual effects				
Age	-7.390**	-8.012**	-4.582**	-7.306**
	(0.348)	(0.165)	(0.227)	(0.203)
Socio-ec. Index (High)	1.542**	1.718**	1.844**	2.222**
	(0.293)	(0.251)	(0.421)	(0.250)
Socio-ec. Index (Medium)	0.867**	0.803**	1.080**	0.921**
	(0.293)	(0.220)	(0.362)	(0.219)
Gender (Female=1)	3.770**	0.318**	0.966**	-1.536**
	(0.186)	(0.161)	(0.264)	(0.150)
Foreign	-2.568**	2.144**	0.006	0.642**
	(0.283)	(0.273)	(0.494)	(0.268)
Secondary 5	9.797**	1.471**	6.701**	2.560**
	(0.799)	(0.327)	(0.739)	(0.393)
Contextual effects				
Age	15.211	11.514**	7.014**	22.639**
	(11.684)	(2.967)	(1.646)	(6.808)
Socio-ec. Index (High)	31.802	25.140	16.299**	25.748
	(30.455)	(16.904)	(7.431)	(18.608)
Socio-ec. Index (Medium)	21.574	23.091	14.010*	10.015
	(31.981)	(14.514)	(7.338)	(14.788)
Gender (Female=1)	-20.267*	13.639	13.265**	-1.936
	(11.040)	(12.713)	(5.676)	(9.533)
Foreign	-28.226	-1.320	0.354	9.226
	(18.593)	(12.377)	(7.041)	(12.411)
Secondary 5	79.885*	-12.062	9.953	-31.537
	(46.845)	(15.235)	(21.115)	(20.155)

Notes :
Robust standard errors in parentheses
** indicates 5% significance level
* indicates 10% significance level
[a]The dependent variable is the score on June 2005 provincial secondary exams.