RIETI Discussion Paper Series 17-E-023

# Accounting for Heterogeneity in Network Formation Behavior: An application to Vietnamese SMEs

**HOSHINO Tadao**
Tokyo University of Science

**SHIMAMOTO Daichi**
Waseda University

**TODO Yasuyuki**
RIETI

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 17-E-023

**March 2017**

# Accounting for Heterogeneity in Network Formation Behavior: An application to Vietnamese SMEs[1]

HOSHINO Tadao

Tokyo University of Science

SHIMAMOTO Daichi

Waseda University

TODO Yasuyuki

Research Institute of Economy, Trade and Industry and Waseda University

Abstract

Network formation is often characterized by homophily—the tendency that agents connect with others who have similar attributes. However, while most agents are homophilous, others may be heterophilous, namely, aiming to create ties with dissimilar agents. This study finds evidence supporting this hypothesis for the first time in the literature by applying random coefficient models to information-sharing network data for Vietnamese small and medium-sized enterprises (SMEs). One possible interpretation for this heterophily is that firms can obtain more useful and performance-improving information from those dissimilar—as opposed to similar—to themselves, as suggested by certain social network studies.

*Keywords*: Network formation, Homophily, Heterophily, Random coefficient model.

*JEL classification*: L14, D85, Z13



[1] This research was conducted as part of a project entitled 'Firms' Domestic and International Networks', undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The authors would like to thank JSPS Kakenhi Grant (No. 25101003 and 26245037) and Waseda University for financial support. The opinions expressed and the arguments employed in this paper are the sole responsibility of the authors and do not necessarily reflect those of RIETI, Tokyo University of Science, Waseda University, or any institution with which the authors are affiliated.

## 1. Introduction

Over the past two decades, it has become increasingly evident that social interaction plays an important role in a variety of economic activities and economic outcomes by facilitating the diffusion of information, knowledge, and technologies. Accordingly, recent studies have focused on how social networks are formed and evolve, and they have typically found homophily as a major driving force of network formation; in other words, agents are likely to be connected to others who are close socially, economically or geographically to themselves (e.g., McPherson et al., 2001; Fafchamps and Gubert, 2007; Currarini et al., 2009; Jackson, 2010, Section 3.2.6.). For example, Fafchamps and Gubert (2007) find that farmers in rural Philippines seeking mutual help tend to form ties with neighboring farmers rather than with distant farmers or non-farmers.

In the presence of homophily in network formation, connected agents are similar in their attributes. Therefore, connected agents share not only the same advantages but also the same disadvantages, which can be a shortcoming of homophilous networks (Beugelsdijk and Smulders, 2004; McFadyen and Cannella, 2004; Perry-Smith, 2006; Todo et al., 2016). For example, Fafchamps and Gubert (2007) argue that homophilous networks among farmers within regions can be ineffective when farmers request help from nearby farmers because neighboring farmers necessarily share the same weather shocks. Similarly, a number of existing studies find that the knowledge of agents in homogeneous networks is often redundant, whereas heterogeneous links can be a conduit for new information and knowledge. A seminal paper by Granovetter (1973) reveals that job seekers obtain more important information from people who they meet only infrequently than from their close friends, emphasizing the importance of "weak ties". In another seminal paper, Burt (1992) argues that the role of "structural holes" is to connect different groups by facilitating knowledge diffusion.

Despite the potential benefits from such heterophilous links, most theoretical and empirical works have assumed that network formation is mainly characterized by homophily. A possible reason for this observation is that agents have more opportunities to interact with others who have similar—as opposed to dissimilar—attributes (Jackson, 2010, Section 3.2.6). However, in reality, although most agents in many situations prefer homophilous partnerships for the reasons described above, it is likely that other agents may have heterophilous preferences in terms of their network formation with regard to obtaining new knowledge and technologies. Thus, to elucidate this heterogeneous nature of the mechanism of network formation, it is necessary to explicitly and simultaneously model both homophily and heterophily, which is the objective of this study.

In the statistical literature, a growing number of studies are focusing on network

formation models. These studies can be classified into two types: those that attempt to incorporate the externalities on the realizing network structure endogenously affecting the network formation behavior itself (e.g., Christakis et al., 2010; Mele, 2013; Sheng, 2014; Leung, 2015) and those that simply ignore such endogeneity and emphasize modeling a sort of unobserved heterogeneity in each agent's behavior (e.g., Krivitsky et al., 2009; Graham, 2015; Jochmans, 2016).[2] In the former type, network formation is modeled as a game in which agents simultaneously form links, and the resulting estimator is typically computationally very demanding. Compared with the former, the latter type of model is more descriptive rather than structural but has great flexibility in its specification. In addition, following Graham (2015, Section 4), Graham (2016) and Goldsmith-Pinkham and Imbens (2013), if we can utilize network connections from the previous time period, we can explain some form of interdependency in link formation even within such a descriptive model. Thus, in terms of our research objectives, it is reasonable to adopt this type of modeling. Specifically, we extend a dyadic link-formation logit model to a random-coefficient model in which the effects of the given pairwise distance measures are allowed to distribute from negative (i.e., homophily) to positive (i.e., heterophily) values. We first consider a normal random-coefficient model in which the normal distribution of the random coefficients is assumed, then we relax the normality assumption by employing a Gaussian mixture sieve approach.

The proposed estimator is applied to the data on small and medium-sized enterprises (SMEs) in village industrial clusters in the apparel and textile industry in Vietnam. In particular, we focus on networks among firms to exchange business information, i.e., information-sharing partnerships, within the village industrial cluster. Village industry clusters are traditionally developed agglomerations of SMEs, including micro enterprises in a particular industry such as the apparel, wood furniture, and ceramic industries within village boundaries. Because the lack of access to information may be an obstacle to firms' activity, informal information-sharing partners play an important role in their economic activity.

In the results, we confirm the presence of homophily for all pairs of firms in terms of the age of firms and of their top managers. In other words, firms are likely to form information-sharing links with firms that are of similar age. By contrast, certain firm pairs show heterophilous behavioral patterns in terms of factors such as the size of firms, the type of clients, and the gender ratio, whereas more than half of the pairs remain homophilous in these aspects. In other words, certain firms prefer to link with dissimilar firms, although, on average, firms prefer to link with firms that are similar with respect to these attributes; this is consistent

[2] For a comprehensive summary of recent developments in econometric methods of network formation, see, e.g., De Paula (2015) and Chandrasekhar (2016).

with the previous findings in the literature. The presence of such heterophilous link-formation behavior would be explained by the arguments provided in the previous studies on social networks that show that agents can often benefit more from heterogeneous links than from homogeneous links. To our knowledge, the mixture of homophily / heterophily in link formation across agents is a topic that has not been analyzed clearly in the empirical literature and is thus our major contribution.

The remainder of this paper is organized as follows. Section 2 describes the data set used in the empirical analysis, whereas Section 3 demonstrates our network formation model and its estimation procedure. Section 4 presents the estimation results and a discussion regarding the obtained results. Section 5 concludes.

## 2. Data

### *2.1 Vietnamese SMEs in the apparel and textile industry*

This study focuses on the village industrial clusters of SMEs in the apparel and textile industry located in the Red River Delta region, Vietnam. To identify these village clusters, we utilized data collected by the Vietnam Enterprise Survey (VES) in 2010. The VES is conducted on an annual basis by the General Statistical Office, and it covers all the foreign-owned firms and randomly selected domestic private firms in Vietnam. We examined these data and extracted and isolated the villages and communes (the smallest administrative unit in Vietnam) with more than five registered firms in the textile and apparel industry in the Red River Delta region, which resulted in 16 apparel / textile village clusters. Then, for each of the 16 clusters, we obtained the full list of registered firms from the municipal government. There were a total of 354 SME firms operating in the apparel and textile industry in these 16 clusters.

We chose Vietnam because village industrial clusters have been traditionally developed in Vietnam so that relatively dense ties between firms within the cluster are observed. At the same time, such firm ties are often removed and newly created, as we will see later, because firms have to deal with recent external shocks in the industry. For example, lowering trade barriers encourages exports, while it also results in more competition with Chinese imports. Therefore, firms may have to look for new information exchange partners. These situations in Vietnam provides us a suitable situation for dynamic analysis of network formation.

In December, 2014 and January, 2015, we conducted the first round of face-to-face interviews with the owners, managing directors, or highly ranked managers of the 354 SME firms and obtained responses from 296 of them (the response rate was 84 percent). The second round of interviews was conducted in August, 2015, and we received responses from 284 of

the 296 firms who participated in the first round. The interview contained questions covering standard firm characteristics, such as sales, the number of workers, the number of subcontractors, main products, international trade activities, and ownership. In addition, we requested the interviewees to name partners with whom they exchange business information from the list of registered firms in the same cluster.

### *2.2 Construction of key variables*

The objective of our empirical analysis is to estimate the mechanism of information-sharing network formation among the SMEs in our sample. Therefore, the key dependent variable of interest is whether firms *i* and *j* within the same industrial cluster are exchanging information, i.e., whether they are connected by an information-exchange tie. We assume that information sharing is not a directed connection but is undirected between the firms such that when either one of a pair of firms nominates the other as a partner, there is a link between them. In addition, we focus only on links within the same village cluster because our sample firms are primarily SMEs in traditional village clusters for which local partners may be the most important information sources. In other words, we ignore links across different clusters. Network information was obtained for two time periods: one from the survey conducted in December 2014 to January 2015, and the other in August 2015.

In this analysis, we exclude firms with no information-sharing partners in both time periods and those with missing values in the independent variables described below. Consequently, our sample for empirical analysis consists of 203 firms from 11 village clusters, which create 2,835 "potential" network links within each cluster.[3] The number of actual links in 2015 is 225, implying that the network is not very dense. Among the 203 firms, the average number of information-sharing partners (i.e., the average degree) and the number of firms with no information-sharing partners during this period are 2.2 and 47, respectively. Figure 1 presents the networks from three typical village clusters in our sample. As shown in the figure, the density of links, which is the ratio of the number of actual links to the number of possible links in the network, varies across clusters, to an extent.

Table 1 shows the dynamics of links from 2014 to 2015. Among 370 firm pairs that had links in 2014, 302 removed their links in 2015, while 157 created new links. This evidence indicates that firms who intend to exchange business information in these village clusters often remove and create links. According to our interviews with some of these firms, one possible reason for this drastic change in firm networks is that once a firm changes its buyer, the firm

[3] Note that, since each firm seeks information sharing partners only within the same village cluster, the total number of potential links does not equal 203(203-1)/2.

has more opportunities to exchange information with clients of the new buyer.

Our independent variables that may affect the network connection are as follows (their definitions are in parentheses): **Years** (years since the firm's foundation), **Nworkers** (the number of workers), **Nsubcontractors** (the number of subcontractors), **Retail** (the percentage of retail sales out of total sales), **Wholesale** (the percentage of wholesale sales), **DirectEx** (the percentage of direct exports in total sales), **IndirectEx** (the percentage of indirect exports), **Age** (the age of the firm's CEO), **Female** (the indicator variable for whether the CEO is female), **Kinh** (the indicator variable for whether the CEO is Kinh, the major ethnicity in Vietnam), **Fboard** (the number of female board members), and **Fratio** (the proportion of female workers out of all workers).

Table 2 presents the summary statistics of the above variables in our sample. The average firm age is 9.1 years, whereas the average and median number of workers is 30 and 10, respectively, indicating that they are relatively young and small. Certain firms outsource a portion of their production processes to subcontractors; the number of subcontractors is 19.7 on average, while its median is two. Wholesale represents a substantial fraction of total sales (62.8 percentage on average). As some firms engage in exports, the average share of direct and indirect exports is 10 and 6 percent, respectively. The average age of managers is 43.7 years; 21 percent of the managers are females, and nearly all are Kinh, the largest and dominant ethnic group in Vietnam.

### *2.3 Factor analysis*

To mitigate the computational burden in the subsequent analysis and to account for the high correlation between the independent variables, we conduct a factor analysis and then use the resulting factors as the firm's attribute variables. The number of factors was determined in accordance with the conventional eigenvalue-one criterion. Thus, the number of eigenvalues larger than one in the correlation matrix of the variables listed above serves as the number of factors. Then, the criterion suggested that we should use five factors.

The results of the factor analysis with these five factors are presented in Table 3.[4] We can interpret the factors as follows. First, Factor 1 represents firm size as it assigns large weights to the number of workers and subcontractors. In addition, the large weight on **DirectEx** for this factor is consistent with our interpretation, since smaller firms typically lack their own export networks abroad. Next, Factor 2 is clearly interpreted as an index for retail-oriented firms. Factor 3 represents the index for indirect exports. Factor 4 refers to the age of the firm and of its CEO. Finally, Factor 5 is interpreted as the indicator for the firms with high

[4] Before conducting the factor analysis, each variable was standardized by subtracting its mean and dividing the difference by its standard deviation.

female participation. Accordingly, in the following, we refer to these five factors as **SCALE** (Factor 1), **RETAIL** (Factor 2), **INDIEX** (Factor 3), **AGE** (Factor 4) and **FEMALE** (Factor 5).

### *2.4 Descriptive examination of homophily*

Before conducting a detailed numerical investigation, we perform a rough check on the presence or absence of homophily and heterophily in network formation in terms of the five factors. In other words, we conduct a nonparametric kernel regression of the link connection between firms on the absolute difference of each factor values between them. The estimated conditional probability curves are provided in Figure 2. Notably, if homophily (resp. heterophily) exists in the network formation, the curves should exhibit an overall decreasing (resp. increasing) tendency.

The figures show that the probability curves do not have a clear tendency of increasing or decreasing, except for **AGE**. The probability of link formation apparently decreases as the difference in **AGE** values increases, implying the presence of a certain magnitude of homophily regarding this variable. For the **SCALE**, **RETAIL** and **FEMALE** variables, the probability of link formation does not always peak in the region close to zero; instead, it records the highest value for firms with certain differences. This phenomenon might be interpreted as evidence of heterophily in terms of these three variables. Nonetheless, if differences in the values of these variables are excessively large close to the upper boundary, the firms tend to avoid forming a link. For the **INDIEX** variable, we cannot observe any clear effects of the variable on the link formation.

## 3. The Model and Estimation Procedure

### *3.1 Model specification*

In this section, we describe our network formation model and its estimation procedure. First, suppose that there are $R$ village clusters in the population. The clusters are indexed by $r = 1, \dots, R$, and each firm belongs to one of these $R$ clusters (in our case, $R$ is equal to 11). The total number of firms located in the $r$-th village cluster is denoted by $n(r)$.

Let $g_{r,i,j} = 1$ if firms $i$ and $j$ in the $r$-th cluster ($i \neq j, i, j = 1, \dots, n(r)$) are connected and $g_{r,i,j} = 0$ otherwise in 2015, and let $Z_{r,i} = \left(Z_{r,i}^{(1)}, \dots, Z_{r,i}^{(d_z)}\right)'$ denote the $d_z \times 1$ attribute vector of firm $i$ (namely, the factors created in the previous section, with $d_z$ being five). In addition, we have information on the network connections from the previous time period, 2014, and we denote $\tilde{g}_{r,i,j}$ as the link status between $i$ and $j$ in 2014.

In the following, for notational simplicity, we omit the subscript $r$ when there is no confusion. Suppose that the gain involved in forming a link between firms $i$ and $j$ for firm $i$ is given by the following,

$$(1)\quad U_i(g_{i,j}=1)-U_i(g_{i,j}=0)=u_i(Z_i,Z_j,\tilde{G}_{n(r)},\varepsilon_{i,j}),$$

where $\tilde{G}_{n(r)}$ is an $n(r)\times n(r)$ adjacency matrix with its $(i,j)$-th element being $\tilde{g}_{i,j}$, and $\varepsilon_{i,j}$ is $i$'s unobserved preference for forming a link with $j$. We introduce the previous—as opposed to the current—network connections in the utility function. Thus, we can avoid addressing the complexity of a many-player network formation game with (potentially) multiple equilibria, while incorporating a form of "dynamic" interdependencies in the link formation, following Graham (2015, Section 4), Graham (2016) and Goldsmith-Pinkham and Imbens (2013).

Firms $i$ and $j$ form a link if the sum of their marginal gains from the link is positive, including even when the marginal gain for one of the two firms is negative. For example, consider a case in which $u_i>0$, $u_j<0$, and $u_i+u_j>0$. Here, firm $i$ has an incentive to transfer a portion of its gain to firm $j$ such that firm $j$ can obtain a non-negative profit; thus, both firms can benefit from forming the link. Assuming that the benefit of forming an information-sharing partnership is transferable with one another in this manner, our network formation model is given by

$$(2)\ g_{i,j}=\mathbf{1}\{u_i(Z_i,Z_j,\tilde{G}_{n(r)},\varepsilon_{i,j})+u_j(Z_j,Z_i,\tilde{G}_{n(r)},\varepsilon_{j,i})>0\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function, which is one if the argument is true and zero otherwise,

$$\begin{aligned}&u_i(Z_i,Z_j,\tilde{G}_{n(r)},\varepsilon_{i,j})+u_j(Z_j,Z_i,\tilde{G}_{n(r)},\varepsilon_{j,i})\\&=\alpha+|Z_i'-Z_j'|\beta_{i,j}+(Z_i'+Z_j')\gamma+\delta_1\tilde{g}_{i,j}+\delta_2\textstyle\sum_{k=1}^{n}\tilde{g}_{i,k}\tilde{g}_{j,k}+\delta_3\ln n(r)+\delta_4 d_{i,j}-\epsilon_{i,j},\end{aligned}$$

and where $\alpha$, $\beta_{i,j}$, $\gamma$, $\delta_1$, $\delta_2$, $\delta_3$ and $\delta_4$ are unknown parameters; $d_{i,j}$ is a geographical distance between $i$ and $j$, and $\epsilon_{i,j}=\epsilon_{i,j}(\varepsilon_{i,j},\varepsilon_{j,i})$ is an unobserved random variable. As emphasized in the previous section, $\beta_{i,j}$, the coefficients on the absolute difference in the attributes between $i$ and $j$, can take pair-specific values to capture the heterogeneity in the effect of homophily / heterophily of firms' attributes on their network formation. The terms $\delta_1\tilde{g}_{i,j}$ and $\delta_2\sum_{k=1}^{n}\tilde{g}_{i,k}\tilde{g}_{j,k}$ capture the effects of whether $i$ and $j$ were information-sharing partners in 2014 and that of the number of information-sharing partners they shared in 2014, respectively. The inclusion of the latter term is motivated by a common suggestion

in the literature regarding social network formation: if two individuals have friends in common, they are likely to become friends (see, e.g., Jackson and Rogers, 2007).

In the following, we assume that $\epsilon_{i,j}$'s are independent across all firm pairs and that they follow a logistic distribution. Then, the probability of $g_{i,j}$ conditional on $\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}\right)$ is given by

$$(3) \qquad P\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta\right) = p\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta\right)^{g_{i,j}} \left[1 - p\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta\right)\right]^{1-g_{i,j}},$$

where $\theta = (\alpha, \gamma, \delta_1, \delta_2, \delta_3, \delta_4)$, and

$$p\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta\right) = \frac{\exp\left(\alpha + \left|Z_i' - Z_j'\right|\beta_{i,j} + \left(Z_i' + Z_j'\right)\gamma + \delta_1 \tilde{g}_{i,j} + \delta_2 \sum_{k=1}^{n} \tilde{g}_{i,k}\tilde{g}_{j,k} + \delta_3 \ln n(r) + \delta_4 d_{i,j}\right)}{1 + \exp\left(\alpha + \left|Z_i' - Z_j'\right|\beta_{i,j} + \left(Z_i' + Z_j'\right)\gamma + \delta_1 \tilde{g}_{i,j} + \delta_2 \sum_{k=1}^{n} \tilde{g}_{i,k}\tilde{g}_{j,k} + \delta_3 \ln n(r) + \delta_4 d_{i,j}\right)}.$$

*3.2 Normal random-coefficient model*

We first consider the case where the random coefficients $\beta_{i,j} = \left(\beta_{i,j}^{(1)}, \ldots, \beta_{i,j}^{(d_z)}\right)'$ are normally distributed. Further, for simplicity of exposition, let us assume that the elements of $\beta_{i,j}$ are mutually independent.[5] Then, we obtain the conditional probability of $g_{i,j}$ on $\left(Z_i, Z_j, \tilde{G}_{n(r)}\right)$ by

$$(4) \qquad K^N\left(Z_i, Z_j, \tilde{G}_{n(r)}; \theta, b, s\right) = \int P\left(Z_i, Z_j, \tilde{G}_{n(r)}, \beta; \theta\right) \prod_{l=1}^{d_z} \phi\left(\beta^{(l)} | b^{(l)}, s^{(l)}\right) \mathrm{d}\beta^{(l)},$$

where $b = \left(b^{(1)}, \ldots, b^{(d_z)}\right)'$, $s = \left(s^{(1)}, \ldots, s^{(d_z)}\right)'$, $\phi(\cdot\,|a_1, a_2)$ is the normal density function with mean $a_1$ and standard deviation $a_2$; thus, $\phi\left(\beta^{(l)} | b^{(l)}, s^{(l)}\right)$ serves as the density function for $\beta_{i,j}^{(l)}$, $l = 1, \ldots, d_z$. Hence, we can estimate the unknown parameters $(\theta, b, s)$ as the maximizer of the log-likelihood function

$$(5) \qquad Q^N(\theta, b, s) = \sum_{r=1}^{R} \sum_{i=1}^{n(r)} \sum_{j>i} \ln K^N\left(Z_{r,i}, Z_{r,j}, \tilde{G}_{n(r)}; \theta, b, s\right).$$

Notably, to evaluate the log-likelihood function (5), the multi-dimensional integration in the probability function (4) must be solved, which has no closed-form solution. Thus, in application, we use a simulated maximum likelihood (SML) method to estimate $(\theta, b, s)$ with a Monte Carlo approximation to (4) (see, e.g., Train, 2003). Clearly, the presence of pair-

[5] In the empirical analysis below, we first estimated a model that allows for non-zero covariances among random coefficients. However, the estimated covariances were small in magnitude and statistically not different from zero at any reasonable significance level. Accordingly, we confine our analysis to the simple case of independent random coefficients.

specific heterogeneity in network formation can be statistically tested by checking the significance of the estimated standard deviations $s$ of the random coefficients.

*3.3 Gaussian mixture sieve random-coefficient model*

In general, the assumption of normally distributed random coefficients is restrictive. Thus, we consider relaxing the normality assumption for the random coefficients, while the assumption of logistic errors remains unchanged. Such a model is very useful and garners focus in the literature (Fox et al., 2011; Fox et al. 2016) due to its computational tractability, compared with fully nonparametric random-coefficient models such as those in Ichimura and Thompson (1998) and Gautier and Kitamura (2013); however, it retains a high degree of flexibility in the distribution of random coefficients.

For simplicity, we continue to assume independence among the random coefficients. Then, the joint density function of $\beta_{i,j}$ can be written in general as $f(\cdot) = \prod_{l=1}^{d_z} f_l(\cdot)$ with $f_l(\cdot)$ being the marginal density function of $\beta_{i,j}^{(l)}$, $l = 1, \dots, d_z$. A convenient method to estimate an unknown density function is to use the Gaussian mixture sieve (GMS) approximation. As shown in Genovese and Wasserman (2000), Ghosal and Van Der Vaart (2001), and Norets (2010), a wide class of density functions can be approximated by Gaussian mixtures. A typical Gaussian mixture density function for a random variable $x$ can be expressed as

$$f_n(x|\xi) = \sum_{m=1}^{M_n} \pi_m \phi(x|\mu_m, \sigma), \sum_{m=1}^{M_n} \pi_m = 1, \ \pi_m \geq 0 \ \text{ for } \ m = 1, \dots, M_n$$

where $\xi = (\pi_1, \dots, \pi_{M_n-1}, \mu_1, \dots, \mu_{M_n}, \sigma)'$, and $M_n$ is a positive integer allowed to increase as the sample size $n$ (in our case, $n = \sum_{r=1}^{R} n(r)(n(r)-1)/2$) increases. Then, for each $l = 1, \dots, d_z$, the density function $f_l(\cdot)$ can be approximated by $f_n(\cdot\,|\xi^{(l)})$ for a parameter vector $\xi^{(l)}$ for a large $M_n = M_n(l)$. Hence, similar to (4), we can approximate the conditional probability of $g_{i,j}$ on $(Z_i, Z_j, \tilde{G}_{n(r)})$ by

$$(6) \quad K^{GMS}(Z_i, Z_j, \tilde{G}_{n(r)}; \theta, \xi^{(1)}, \dots, \xi^{(d_z)}) = \int P(Z_i, Z_j, \tilde{G}_{n(r)}, \beta; \theta) \prod_{l=1}^{d_z} f_n(\beta^{(l)}|\xi^{(l)}) \, \mathrm{d}\beta^{(l)},$$

and the resulting log-likelihood function is

$$(7) \quad Q^{GMS}(\theta, \xi^{(1)}, \dots, \xi^{(d_z)}) = \sum_{r=1}^{R} \sum_{i=1}^{n(r)} \sum_{j>i} \ln K^{GMS}(Z_{r,i}, Z_{r,j}, \tilde{G}_{n(r)}; \theta, \xi^{(1)}, \dots, \xi^{(d_z)}).$$

Again, since solving the maximization problem for the objective function (7) is computationally intractable due to the multi-dimensional integration in (6), we use the SML

method to estimate $\left(\theta, \xi^{(1)}, \ldots, \xi^{(d_z)}\right)$. In doing so, to effectively restrict the parameter space of $\pi_m$'s, we set $\pi_m = \exp(a_m) / \sum_{m'=1}^{M_n} \exp(a_{m'})$ with $a_{M_n} = 0$, and we do not directly estimate $\pi_m$ but estimate $a_m$. Once the estimate of the density function of a random coefficient is available, we can directly simulate the moments of the random coefficient, which can be further used to statistically check the presence of a heterogeneous preference in the network formation.

## 4. Estimation Results

### *4.1 Regression analysis*

In addition to the two aforementioned random-coefficient (RC) models, the Normal-RC logit model and the GMS-RC logit model, we also estimate a simple logit model as a benchmark in which the coefficients of the absolute difference variables are assumed to be constant. The estimation results from these models are summarized in Table 4.

First, we focus on the effects of the absolute difference variables. Recall that if the coefficients of these variables are significantly negative (resp. positive), it implies the presence of homophily (resp. heterophily) for the corresponding variables. Then, the results from the simple logit model, reported in column (1), indicates that the **AGE** variable presents statistically significant homophily, i.e., firms with dissimilar **AGE** values are less likely to form an information-exchange link. Conversely, for the other variables, we observe no significant impacts on link formation, either for homophily or heterophily.

The results from the Normal-RC model and the GMS-RC model are reported in columns (2) and (3), respectively. For these models, the effect of the **SCALE** variable, which is found to be insignificant in the simple logit case, is negative but insignificant in its mean, while its standard deviation is estimated to be significantly different from zero. In other words, some pairs of firms have homophily with regard to this variable, while other pairs of firms may have a heterophilous link-formation pattern. To visualize this situation more clearly, we present the estimated density functions of the random coefficients in Figure 3, in which the estimates from the Normal-RC model are depicted in gray lines, and those from the GMS-RC model are depicted in black. The estimated density for the coefficient of the **SCALE** variable is provided in panel (A). This figure clearly shows that the larger portion of the support of the estimated density is included in the negative region, implying that most firm pairs tend to form links if their **SCALE** values are similar (and not dissimilar) to one another. We find similar heterogeneous effects of the **RETAIL** variable and the **FEMALE** variable as well; however, such heterogeneity does not significantly exist for **INDIEX** and **AGE**. For the **INDIEX** variable, the effect is insignificant, including in the mean effect. Conversely, the mean effect

of the **AGE** variable is also statistically significantly negative in the random-coefficient models, suggesting the presence of a pure homophilous pattern on this variable.

Next, we turn to the results for the firm pair's sum of their attribute variables. For all three models, we find that the **SCALE** and **FEMALE** variables have negatively significant effects, that the **INDIEX** variable has a positively significant effect, and that the remaining two variables are not significantly related to link-formation behavior. For the **SCALE** variable, it should be understandable that large-scale firms have fewer incentives to construct new information-sharing connections as they already have their own rich business know-how and enjoy the latest technologies; thus, they are more reluctant to impart their information to others than smaller firms are. The result for the **INDIEX** variable would also be reasonable because the presence of local information-sharing partners is important, particularly for indirect-exporting domestic-oriented firms rather than for direct-exporting firms. Firms with female managers and workers are less likely to form links possibly because females may have to work more at home and thus have fewer opportunities to meet others. Notably, while we have observed that the similarity of the **AGE** variable is an important factor in the link formation, the value of the variable itself is not.

Finally, let us explain the effects of the remaining independent variables. The effect of the link status in 2014 (Link 2014) is positive and statistically significant in all model specifications, indicating that the information-sharing partnership tends to be persistent. The effects of the mutual link variable (Mutual Link 2014), the number of partners the two firms share in common in 2014, is positive (although its statistical significance is not strong), supporting previous findings, as in Jackson and Rogers (2007). Notably, the increase in the total number of firms in the same village decreases the probability of the network formation being statistically significant in all specifications. It should be natural to suppose that forming a partnership link and its preservation is costly and thus that the capacity of the number of partners a firm can hold is limited. If this supposition is true, we would observe denser networks in smaller villages and sparser networks in larger villages. Finally, the geographical distance between two firms does not affect their link status significantly, probably because we focus on link formation among those firms in the same small village clusters.

### *4.2 Simulation analysis*

For illustrative purposes, we now quantitatively examine how the probability of making a link changes as the attributes of partner firms change. In this analysis, we focus on the raw firm characteristics (12 variables listed in Table 2), rather than on the five-factor variables generated. Each factor variable is created by a linear combination of the firm's 12 characteristic variables, where the corresponding weights are shown in Table 3. By combining

these weight values and the estimation results provided by the GMS-RC model, we can calculate the probability of building a link between a pair of firms for particular values of characteristic variables.

Specifically, we consider a representative firm $i^*$, whose characteristic variables are medians of these, i.e., $\mathbf{Years}_{i^*} = 8$, $\mathbf{Nworkers}_{i^*} = 10$, and so forth (see Table 2). Consider a firm $j^*$, which is a potential information-sharing partner for $i^*$. Suppose that the firm $j^*$ has exactly the same value of characteristic variables as $i^*$. In this case, the probability of link formation between the two firms is equal to 0.0513.[6] Then, we track the changes in the probability of link formation between $i^*$ and $j^*$ by shifting only the value of a particular characteristic variable for $j^*$, while the other characteristic variables remain the same.

In the following, we focus on five characteristic variables including **Years** (firm age), **Age** (manager age), **Nworkers** (the number of workers), **Nsubcontractors** (the number of subcontractors), and **Fratio** (the proportion of female workers in all workers). The results are summarized in Figure 4. The solid line in the figure indicates the mean probability of link formation between $i^*$ and $j^*$, and the upper and lower dashed lines are the 95th and 5th percentile probability curves, respectively. The x-axes denote the value of the characteristic variables for the potential partner $j^*$, and the vertical line indicates the median value, namely the value of the characteristic variables for $i^*$. Note that the asymmetry of the probability curves is due to the term of $\left(Z_i' + Z_j'\right)\gamma$.

First, panel (A) clearly shows that as the absolute difference in **Years** between $i^*$ and $j^*$ increases, the probability of link formation decreases nearly linearly, without showing an increasing (i.e., heterophilous) tendency even in the 95th percentile curve. According to this result, an increase in the difference in years of operation (i.e., **Years**) by one standard deviation (5.95) leads to a decrease in the probability of link formation by approximately only 0.67%, on average. Thus, we can conclude that homophily in terms of **Years** is not numerically huge. The width between the 95th and 5th percentile curves is relatively narrow, indicating that firms are homogenously homophilous in **Years**.

Similarly, the panels (B), (C), (D) and (E) show simulation results for the variables **Age**, **Nworkers**, **Nsubcontractors** and **Fratio**, respectively. For all these variables, we can observe that the mean probability of link formation decreases as their absolute difference enlarges, implying the presence of homophily, on average. However, in contrast to **Years**, our results suggest that the probability curves at the upper percentile can exhibit an increasing trend for these four variables. For example, when the number of workers (i.e., **Nworkers**)

---

[6] In this and the following simulation analysis, we set $\tilde{g}_{i^*,j^*} = 0$, $\sum_{k=1}^{n} \tilde{g}_{i^*,k}\tilde{g}_{j^*,k} = 0$, $\ln n(r) = 4.4886$ (log of the median of $n(r)$'s), and $d_{i^*,j^*} = 1.2027$ (the median of $d_{i,j}$'s).

of partner firm $j^*$ increases by its standard deviation (80.95) from the baseline, the link-formation probability at the 95th percentile increases to approximately 8.2%, which is 3.1% larger than the baseline case with no difference in the value of **Nworkers**. In contrast, the probability at the 5th percentile decreases to approximately 1.1%, which is 4% smaller than the baseline case. This substantial variation in the change in the probability of link formation indicates that differences in the degree of homophily / heterophily in terms of firm size among firms are quantitatively large. This finding also applies to **Age, Nsubcontractors** and **Fratio.**

*4.3 Discussions*

Table 5 summarizes the patterns of homophily and heterophily for the information-sharing network in our data. The presence of heterophilous patterns for the factors **SCALE**, **RETAIL** and **FEMALE** highlight the importance of incorporating the heterogeneous patterns of homophily and heterophily into analyses of network formation. If a network formation model is estimated without assuming such heterogeneity, as done in the conventional empirical studies, one would overlook the influence of heterogeneity and thus ultimately find only the presence of homophily averaged over the agents.

The reason for the presence of heterophilous patterns could be explained in accordance with the social network studies that show the role of heterophilous links in knowledge diffusion. For example, a seminal work by Granovetter (1973) shows that people who job seekers meet less frequently are more important information sources than job seekers' close friends, thus emphasizing the strength of weak ties. Burt (2004) finds that workers perform better when they are linked with heterogeneous colleagues, confirming the role of structural holes proposed by Burt (1992) that connect different groups in facilitating knowledge diffusion. In our case, developing a heterophilous link can increase a chance to receive new knowledge because firms of different size, with different products or characterized by different gender ratios are likely to have different production schemes and management technologies or business resources.

However, we also find that nearly all firm pairs are homophilous in terms of **INDIEX** and **AGE** and that the majority are homophilous in terms of the other three factors. A possible interpretation of these findings is that although heterophily may be beneficial due to the diffusion of new knowledge, firms are primarily homophilous probably because of the lower costs of constructing a link with similar and socially and economically close firms. This finding is similar to that of Fafchamps and Gubert (2007) who observe that farmers in rural Philippines seeking mutual help form links primarily with farmers, rather than with people in other occupations such as factory workers, retailors, and taxi drivers. Fafchamps and Gubert

(2007) argue that mutual help through such homophilous networks is inefficient because negative shocks in agriculture homogeneously affect farmers within a region and farmers could share risks more efficiently through ties with non-farmers.

It is difficult to conclude whether the firm networks in our sample are efficiently formed to maximize their profits through knowledge diffusion within the network because the costs of creating homophilous and heterophilous ties are not available in our data. However, our results indicate that firms (or firms' decision makers) are well diversified; therefore, some act as brokers of knowledge among different groups, improving the aggregate performance of the economy.

## 5. Conclusion

In this study, we proposed an estimation procedure for a network formation model that allows us to identify the heterogeneous behavioral patterns of homophily and heterophily. In particular, we developed a dyadic logit model with random coefficients in which the random coefficients are assumed to be distributed in either a normal distribution or more general distribution that we approximate as the Gaussian mixture. Then, we applied the proposed method to the data on the network formation of business information-sharing partners for SMEs in the textile industry in Vietnam. The obtained estimation results were, in turn, used to conduct a set of simulation analyses to demonstrate how the probability of link formation varies with changes in the values of partner firm characteristic variables.

We found that the firms homogeneously show homophily patterns in terms of the age of firms and their CEOs. By contrast, a portion of firm pairs show heterophilous patterns for factors such as size, type of clients, and gender composition of the firm although homophily, on average, remains dominant in these aspects. The heterophilous link formation can be explained by the argument in the previous literature on social networks contending that agents can benefit more from heterophilous links through the diffusion of new knowledge than from homophilous links (Granovetter, 1973; Burt, 1992). To our knowledge, such behavioral heterogeneity of homophily / heterophily in the link formation across agents has not been numerically clarified in the literature.

## References


Beugelsdijk S, Smulders S. 2004 Bridging and Bonding Social Capital: Which Type Is Good for Growth? In: Arts WA, Hagenaars JA, Halman L (Eds), The Cultural Diversity of European Unity. Findings, Explanations and Reflections from the European Values Study. Brill Academic Publishing. p. 147-185.

Burt RS. 1992 Structural Holes: The Social Structure of Competition. Harvard University Press: Cambridge.

Burt RS. 2004 Structural Holes and Good Ideas. American journal of sociology. 110; 349-399.

Currarini S, Jackson MO, Pin P. 2009 An Economic Model of Friendship: Homophily, Minorities, and Segregation. Econometrica. 77; 1003-1045.

Fafchamps M, Gubert F. 2007 The Formation of Risk Sharing Networks. Journal of Development Economics. 83; 326-350.

Gautier E, Kitamura Y. 2013 Nonparametric Estimation in Random Coefficients Binary Choice Models. Econometrica. 81; 581-607.

Genovese CR, Wasserman L. 2000 Rates of Convergence for the Gaussian Mixture Sieve. Annals of Statistics. 1105-1127.

Ghosal S, Van Der Vaart AW. 2001 Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities. Annals of Statistics. 1233-1263.

Goldsmith-Pinkham P, Imbens GW. 2013 Social Networks and the Identification of Peer Effects. Journal of Business & Economic Statistics. 31; 253-264.

Graham BS. 2016 Homophily and Transitivity in Dynamic Network Formation. NBER Working Paper No.22186.

Granovetter MS. 1973 The Strength of Weak Ties. American journal of sociology. 78; 1360-1380.

Hite JM, Hesterly WS. 2001 The Evolution of Firm Networks: From Emergence to Early Growth of the Firm. Strategic Management Journal. 22; 275-286.

Ichimura H, Thompson TS. 1998 Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution. Journal of Econometrics. 86; 269-295.

Jackson MO. 2010 Social and Economic Networks. Princeton University Press.

Jackson MO, Rogers BW. 2007 Meeting Strangers and Friends of Friends: How Random Are Social Networks? American Economic Review. 97; 890-915.

McFadyen MA, Cannella AA. 2004 Social Capital and Knowledge Creation: Diminishing Returns of the Number and Strength of Exchange Relationships. Academy of Management journal. 47; 735-746.

McPherson M, Smith-Lovin L, Cook JM. 2001 Birds of a Feather: Homophily in Social Networks. Annual review of sociology. 415-444.

Norets A. 2010 Approximation of Conditional Densities by Smooth Mixtures of Regressions. The Annals of statistics. 38; 1733-1766.

Perry-Smith JE. 2006 Social yet Creative: The Role of Social Relationships in Facilitating Individual

Creativity. Academy of Management journal. 49; 85-101.
Todo Y, Matous P, Inoue H. 2016 The Strength of Long Ties and the Weakness of Strong Ties: Knowledge Diffusion through Supply Chain Networks. Research Policy. 45; 1890-1906.
Train K. 2003. Discrete Choice with Simulation. Cambridge University Press, Cambridge.

Table 1: Dynamics of links

| Number of links | | Linked in 2015 | | Total |
|---|---|---|---|---|
| | | Yes | No | |
| Linked in 2014 | Yes | 68 | 302 | 370 |
| | No | 157 | 2,308 | 2,465 |
| Total | | 225 | 2,610 | 2,835 |

Table 2: Summary Statistics (sample size = 203)

| Variable | Mean | Median | Std. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| Years | 9.0690 | 8 | 5.9507 | 1 | 26 |
| Nworkers | 29.9754 | 10 | 80.9504 | 1 | 1,000 |
| Nsubcontractors | 19.6650 | 2 | 55.5120 | 0 | 450 |
| Retail | 20.9163 | 0 | 34.5206 | 0 | 100 |
| Wholesale | 62.8374 | 90 | 41.6823 | 0 | 100 |
| DirectEx | 10.0542 | 0 | 27.5487 | 0 | 100 |
| IndirectEx | 6.1429 | 0 | 21.3823 | 0 | 100 |
| Age | 43.7340 | 43 | 9.8563 | 25 | 69 |
| Female | 0.2118 | 0 | 0.4096 | 0 | 1 |
| Kinh | 0.9704 | 1 | 0.1698 | 0 | 1 |
| Fboard | 0.7340 | 1 | 0.6038 | 0 | 4 |
| Fratio | 0.6520 | 0.7037 | 0.3010 | 0 | 1 |

Table 3: Factor Analysis (sample size = 203)

| Variable | Factor 1<br>SCALE | Factor 2<br>RETAIL | Factor 3<br>INDIEX | Factor 4<br>AGE | Factor 5<br>FEMALE |
|---|---|---|---|---|---|
| Years | 0.0259 | 0.0770 | -0.0066 | 0.5063 | 0.0655 |
| ln(Nworkers + 1) | 0.5484 | -0.1083 | 0.1016 | 0.1087 | -0.1998 |
| ln(Nsubcontractors + 1) | 0.3961 | 0.0738 | 0.0797 | 0.0331 | 0.2181 |
| Retail | -0.2398 | 0.9573 | -0.1394 | 0.0508 | 0.0207 |
| Wholesale | -0.4826 | -0.8038 | -0.3274 | -0.0991 | -0.0326 |
| DirectEx | 0.9904 | 0.0114 | -0.1016 | 0.0651 | -0.0160 |
| IndirectEx | 0.0524 | 0.0069 | 0.9948 | 0.0277 | 0.0491 |
| Age | 0.1269 | -0.0430 | 0.0475 | 0.8210 | -0.2106 |
| Female | -0.0188 | -0.0380 | -0.0198 | -0.1911 | 0.4689 |
| Kinh | -0.1299 | -0.1037 | 0.0495 | 0.0033 | 0.1722 |
| Fboard | -0.0277 | 0.0589 | -0.0174 | -0.0112 | 0.4322 |
| Fratio | 0.2070 | 0.0529 | 0.1095 | 0.2493 | 0.4840 |
| Loadings | 1.809 | 1.606 | 1.161 | 1.059 | 0.811 |
| Proportion of variance | 0.151 | 0.134 | 0.097 | 0.088 | 0.068 |
| Cumulative variance | 0.151 | 0.285 | 0.381 | 0.470 | 0.537 |

Table 4: Estimation Results (sample size = 2,835)

| Variable | | (1) Logit Estimate | (1) Logit t-value | (2) Normal-RC Logit [a] Estimate | (2) Normal-RC Logit [a] t-value | (3) GMS-RC Logit [a] Estimate | (3) GMS-RC Logit [a] 95% CI [b] |
|---|---|---|---|---|---|---|---|
| Absolute difference of: | | | | | | | |
| **SCALE** | (Mean) | 0.0358 | 0.5138 | -0.2230 | -0.9519 | -0.2986 | [-1.0369, 0.7853] |
| | (Std. Dev.) | | | 0.5135 | 2.4461 | 0.5626 | [0.2904, 1.4425] |
| **RETAIL** | (Mean) | -0.0555 | 0.9134 | -0.5181 | -1.7929 | -0.6439 | [-1.1752, 0.5138] |
| | (Std. Dev.) | | | 0.5966 | 2.5756 | 0.7345 | [0.1644, 1.5979] |
| **INDIEX** | (Mean) | -0.1545 | 1.5901 | -0.1485 | -1.0747 | -0.1814 | [-0.2632, -0.0192] |
| | (Std. Dev.) | | | 0.0579 | 0.1757 | 0.0804 | [0.0353, 0.1634] |
| **AGE** | (Mean) | -0.1940 | 2.3431 | -0.2576 | -2.3356 | -0.2603 | [-0.2971, -0.0217] |
| | (Std. Dev.) | | | 0.0175 | 0.1064 | 0.0515 | [0.0294, 0.1341] |
| **FEMALE** | (Mean) | 0.1227 | 1.4573 | -0.3167 | -0.8888 | -0.2530 | [-0.6875, 0.6034] |
| | (Std. Dev.) | | | 0.8235 | 2.2247 | 0.7498 | [0.2518, 1.2797] |
| | | | | | | | t-value |
| Intercept | | 0.4506 | 0.8429 | 1.6027 | 2.0957 | 1.8726 | 2.4889 |
| Sum of: | | | | | | | |
| **SCALE** | | -0.1067 | -2.6711 | -0.1704 | -2.9237 | -0.1715 | -2.2245 |
| **RETAIL** | | -0.0356 | -0.8435 | -0.0195 | -0.3868 | -0.0317 | -0.6352 |
| **INDIEX** | | 0.1935 | 2.6995 | 0.2380 | 2.6527 | 0.2717 | 2.5037 |
| **AGE** | | 0.0077 | 0.1770 | 0.0288 | 0.5192 | 0.0318 | 0.5183 |
| **FEMALE** | | -0.1402 | -2.9161 | -0.1798 | -2.9213 | -0.1852 | -2.7613 |
| Link 2014 | | 0.8090 | 4.5552 | 1.1372 | 4.4946 | 1.2457 | 4.6561 |
| Mutual Link 2014 | | 0.0879 | 1.6009 | 0.1265 | 1.7334 | 0.1275 | 1.6882 |
| $\ln n(r)$ | | -0.7202 | -5.8764 | -0.9463 | -5.4558 | -1.0156 | -6.1538 |
| Distance in km | | 0.0029 | 1.3852 | 0.0042 | 1.5108 | 0.0045 | 1.4451 |
| Log-likelihood | | -714.8963 | | -709.6927 | | -704.7151 | |

a: The number of Monte Carlo repetitions to approximate the multi-dimensional integration in (4) and (6) was set at 500.

b: The 95% confidence intervals (CI) for the mean and the standard deviation of the random coefficients are those obtained by parametric bootstrap with 1,000 replications based on the asymptotic normal distribution of the estimates of $\xi^{(l)}$s.

Table 5: Patterns of Homophily and Heterophily

| Variable | Homophily or Heterophily | Main components [a] |
|---|---|---|
| **SCALE** | Both | **Nworkers**, **Nsubcontractors**, **DirectEx** |
| **RETAIL** | Both | **Retail** |
| **AGE** | Homophily | **Years**, **Age** |
| **FEMALE** | Both | **Female**, **Fboard**, **Fratio** |

a: Definitions of the component variables are provided in subsection 2.2.

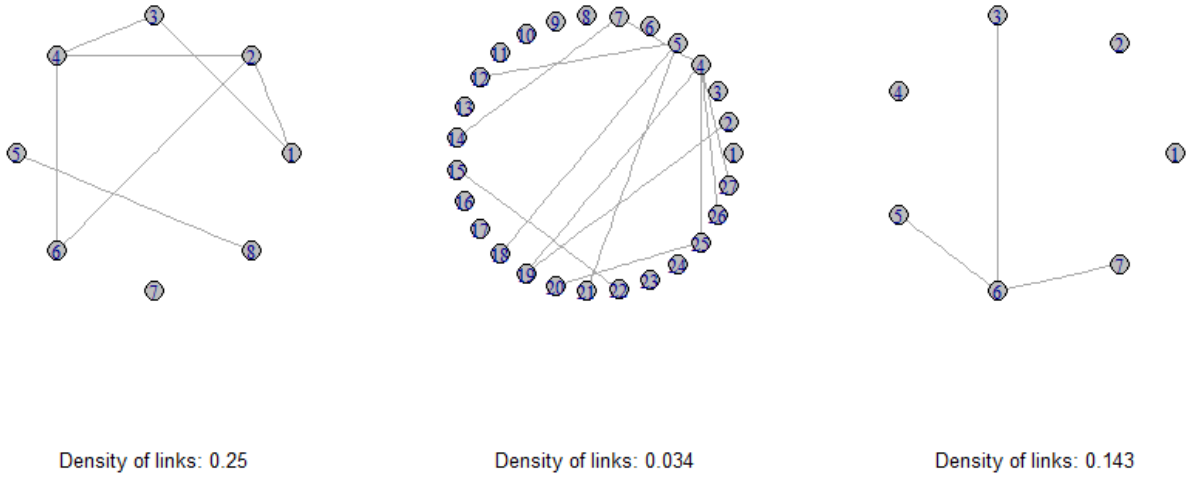


Figure 1: Typical information-sharing networks

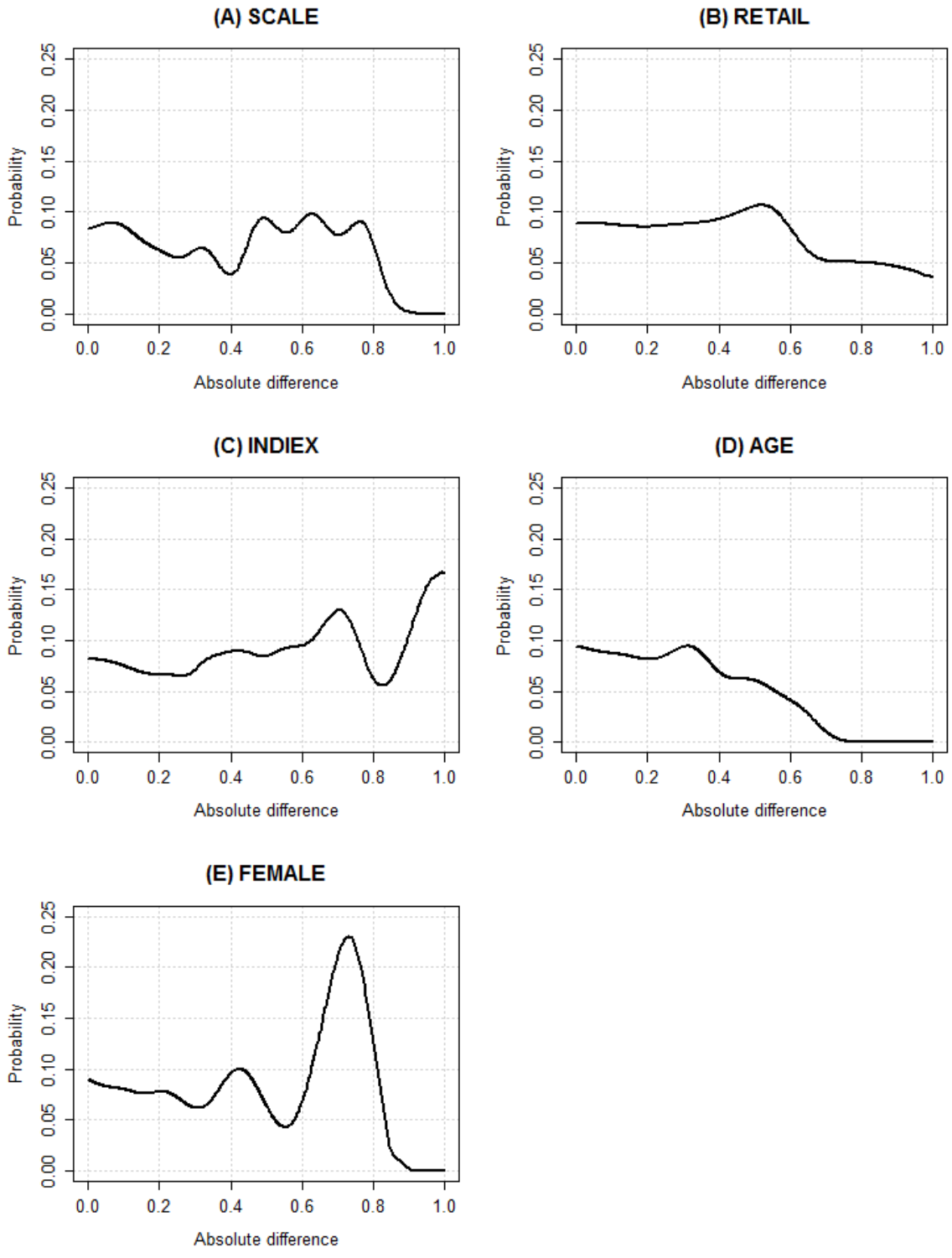


Figure 2: Conditional probability of link formation

NOTE: The domain of the probability function is re-scaled to [0, 1].

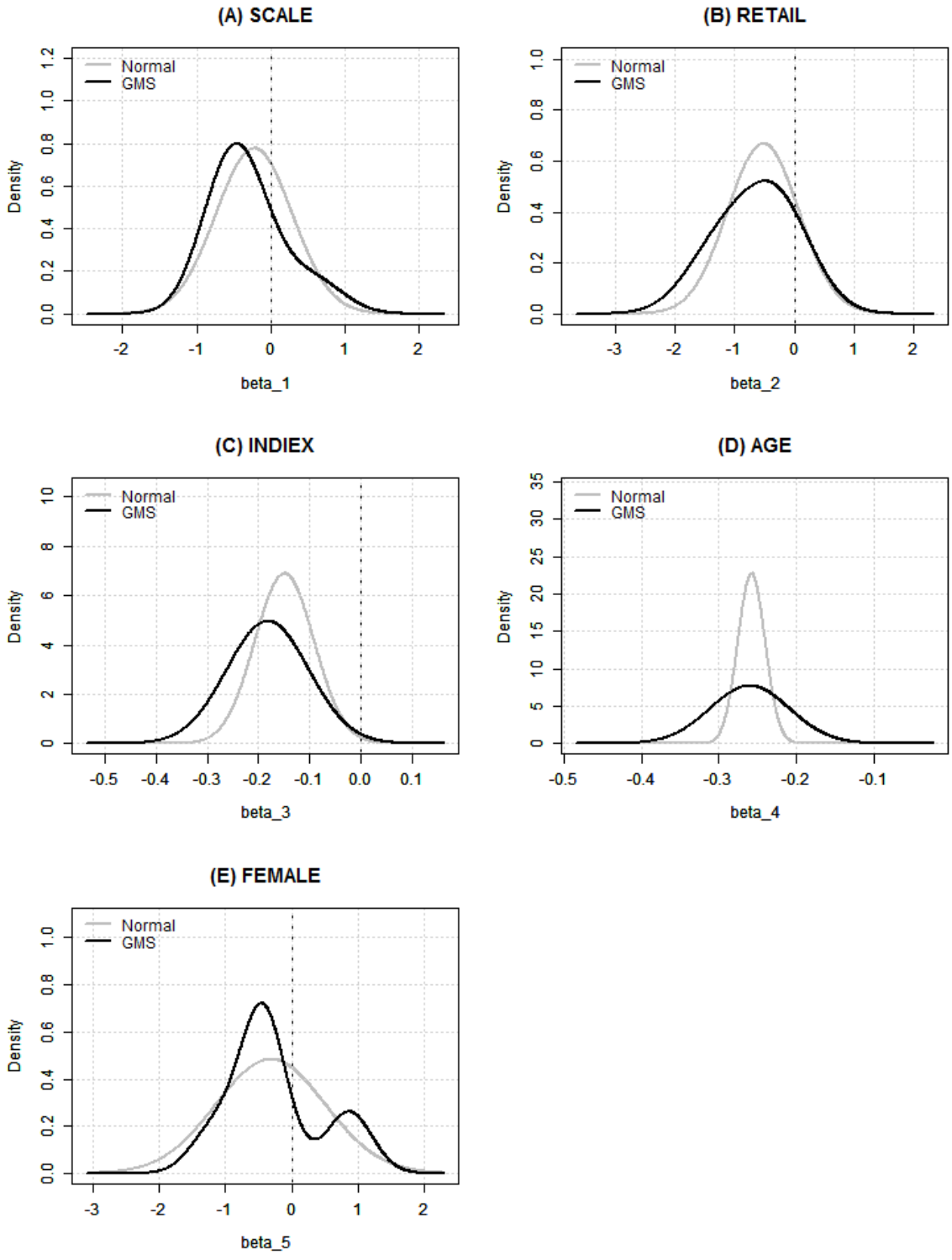


Figure 3: Estimated density of the random coefficients

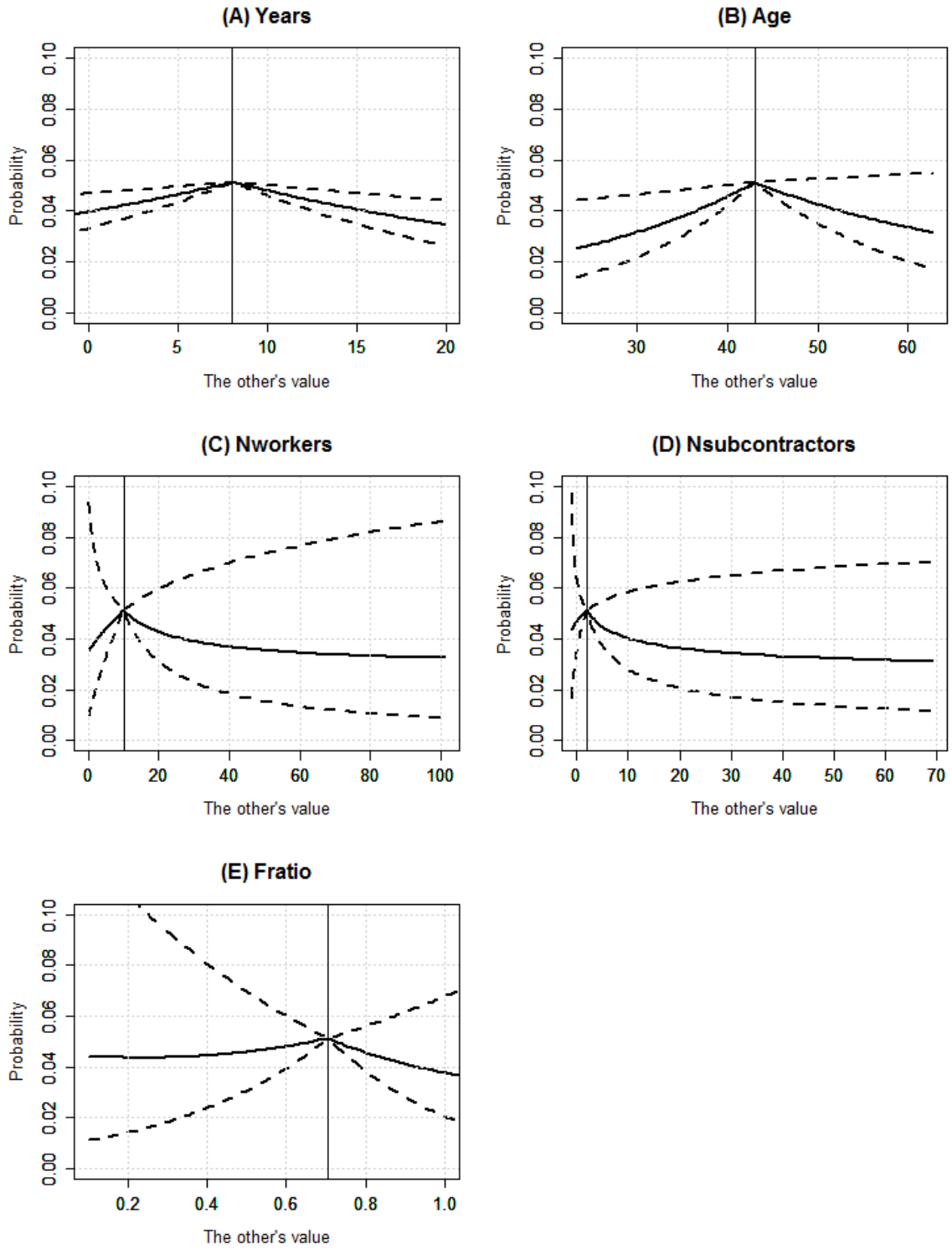


Figure 4: Heterogeneity in the probability of link formation