OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 0305–9049
doi: 10.1111/obes.12360

Accounting for Heterogeneity in Network Formation Behaviour: An Application to Vietnamese SMEs*

Tadao Hoshino† Daichi Shimamoto‡ and Yasuyuki Todo§

†*Waseda University, Tokyo, Japan. (e-mail: thoshino@waseda.jp)*
‡*Kindai University, Higashi-Osaka, Japan (e-mail: d.shimamoto@eco.kindai.ac.jp)*
§*Research Institute of Economy, Trade and Industry, Waseda University, Tokyo, Japan (e-mail: yastodo@waseda.jp)*

Abstract

Network formation is often characterized by homophily, i.e. the tendency of agents to connect with others who have similar attributes. However, while most agents are homophilous, others could be heterophilous; they aim to create ties with dissimilar agents. This study provides empirical evidence supporting this hypothesis by applying a random coefficient approach to data on the information-sharing networks of small- and medium-sized Vietnamese enterprises. In particular, we find that firms tend to form heterophilous links with respect to business type and gender ratio. One possible reason for the heterophily is that firms can obtain useful and performance-improving information from such dissimilar partners.

I. Introduction

Over the past two decades, it has been increasingly recognized that social interaction plays an important role in a variety of economic activities by facilitating the diffusion of knowledge and technology (e.g. Durlauf and Fafchamps, 2005; Granovetter, 2005; Jackson, 2010). Recent studies have focused on how social networks, such as friendship networks and business partnerships, form and evolve. These studies have theoretically and empirically found that homophily, i.e. the tendency of agents to connect with others who are socially and economically similar or geographically close, is a major driving force of social network formation (e.g. McPherson, Smith-Lovin and Cook, 2001; Fafchamps and Gubert, 2007;

JEL Classification numbers: L14, D85, Z13.

*This research was conducted as part of a project entitled 'Empirical Analysis on Determinants and Impacts of Formation of Firm Networks' undertaken at the Research Institute of Economy, Trade, and Industry (RIETI). The authors thank the Japanese Society for the Promotion of Science (JSPS) for the KAKENHI Grant (No. 25101003 and 26245037), Waseda University for financial support and Tatsuo Hatta, Charles Y. Horioka, Eiichi Tomiura, Makoto Yano, and seminar participants at Asian Growth Research Institute, RIETI, XXXVII Sunbelt Conference of the International Network for Social Network Analysis for helpful comments. We are also grateful to Editor Climent Quintana-Domeque and the two anonymous referees for their constructive comments, which greatly improved the paper. The opinions expressed and the arguments provided in this paper are the sole responsibility of the authors and do not reflect those of Kindai University, RIETI, Waseda University, or any institution with which the authors are affiliated. Of course, all remaining errors are our own.

Currarini, Jackson and Pin, 2009; Baccara and Yariv, 2013; Currarini, Matheson and Vega-Redondo, 2016; Kets and Sandroni, 2019).

There is an on-going debate about whether homophily actually promotes the economic performance and welfare of agents in networks (Cowan and Jonard, 2004; Levine and Kurzban, 2006; Yavaş and Yücel, 2014; Luo *et al.*, 2015). A positive side of homophilous networks is that the agents may trust each other and be willing to share knowledge. Indeed, Coleman (1988) finds that when the agents are linked strongly and densely in a cluster, they tend to trust each other and form social capital to enhance economic development. A negative side is that such strongly tied homophilous network tends to be closed and prevent knowledge inflows from outside the network. Such negative effects of homophily have been empirically shown in McDonald and Westphal (2003) and McDonald, Khanna and Westphal (2008), who examine the networks of firms' chief executive officers (CEOs). Fafchamps and Gubert (2007) find that social networks to promote mutual help among farmers in rural Philippines are mostly homophilous and inefficiently reduce the risks of common shocks, such as bad weather, to farmers.

Considering the negative aspect of homophily explicitly, some theoretical studies indicate that the combination of homophilous and heterophilous ties can achieve the most active knowledge diffusion among agents and their highest performance (e.g. Cowan and Jonard, 2004; Kimura and Hayakawa, 2008; Yavaş and Yücel, 2014). However, to the best of our knowledge, there is no empirical study on network formation that has examined the possibility of the coexistence of homophily and heterophily. The main objective of our study is to fill this gap. Note that some existing papers (e.g. research collaborations (Moody, 2004) and syndicates of investment banks (Shipilov, Rowley and Aharonson, 2006)) have discovered heterophilous link formation in some social networks. However, they did not account for potential heterogeneity in the degree of homophily and heterophily across agents. We address this issue using a random coefficient (RC) approach, as described below.

In the statistical literature, a growing number of studies focus 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 behaviour itself (e.g. Christakis *et al.*, 2010; Mele, 2017; Leung, 2015; Sheng, 2016) and those that simply ignore such endogeneity and emphasize modelling a type of unobserved heterogeneity into each agent's behaviour (e.g. Krivitsky *et al.*, 2009; Graham, 2017; Jochmans, 2018).[1] For the former type, network formation is modelled as a game in which agents simultaneously form links, and the resulting estimator is typically very computationally demanding. Compared with the former, the latter type of model is more descriptive than structural but has great flexibility in its specification. In addition, as proposed by Goldsmith-Pinkham and Imbens (2013), Graham (2016) and Graham (2017, Section 3), if we can utilize network connections from the previous time period, then we can explain some form of interdependency in link formation, even within such a descriptive model. Thus, in terms of our research objective, it is reasonable to adopt this type of modelling. It should be emphasized that, although these models address unobserved heterogeneity of agents using fixed effects, they do not account for preference heterogeneity

[1] For a comprehensive summary of recent developments regarding econometric methods used for analyzing network formation, see, e.g. de Paula (2017) and Chandrasekhar (2016).

in terms of homophily or heterophily. In this paper, we develop RC dyadic link-formation logit models in which the effects of the given pairwise dissimilarity measures are allowed to distribute from negative (i.e. homophily) to positive (i.e. heterophily) values. First, we consider Normal-RC models in which the normality of the RCs is assumed; then, we relax the normality assumption with a Gaussian mixture approach.

The proposed estimator is applied to unique data set on small- and medium-sized enterprises (SMEs) in village industrial clusters in the apparel and textile industry in Vietnam. In particular, we focus on the networks among firms that exchange business information within the same village industrial cluster. These clusters are traditionally developed agglomerations of SMEs, including micro enterprises, in a particular industry, such as the apparel, wooden furniture, or ceramics industry. Because a lack of access to information may be an obstacle to firms' economic activities, informal information-sharing partners should play an important role in their economic activities. Our empirical results reveal that certain firm pairs show heterophilous behaviours in terms of attributes such as business type and gender ratio, whereas more than half of the pairs remain homophilous in terms of these attributes. Based on the earlier theoretical works (Cowan and Jonard, 2004; Kimura and Hayakawa, 2008; Yavaş and Yücel, 2014), we could interpret this phenomenon as follows: Homophilous and heterophilous links coexist so that agents can achieve better performance by exchanging information actively among strongly connected similar agents and learning new knowledge and experience from dissimilar ones.

The remainder of this paper is organized as follows: Section II gives a short literature review of homophily. Section III describes the data set used in the empirical analysis, and Section IV presents our network formation model and its estimation procedure. Our empirical results are presented in Section V with some discussions, and Section VI gives concluding remarks.

II. Literature review of homophily

Homophily is predominantly observed in many types of social networks of agents, such as individuals, households, and firms (McPherson *et al.*, 2001). Homophilous preference arises because, for example, agents face less uncertainty when interacting with those with similar backgrounds (Kets and Sandroni, 2019). This can be further amplified by the diffusion of preferences within the group (Kandel, 1978; Kossinets and Watts, 2009) and opportunities to meet similar agents (Currarini *et al.*, 2009, 2016).

However, it is still unclear whether homophilous networks result in more knowledge diffusion and higher performance of agents in the networks than heterophilous networks. It can be easily imagined that, because agents with similar attributes may be linked strongly in terms of trust, they are more willing to share information and knowledge than dissimilar agents. Then, homophily is likely to form 'dense' networks among similar agents. Coleman (1988) emphasizes the positive aspect of strong and dense social networks that create social capital. However, because homophilous preference often creates closed clusters of agents, knowledge of similar agents can be localized and overlapping (Golub and Jackson, 2012; Yavaş and Yücel, 2014). In fact, Granovetter (1973) reveals that job seekers obtain more useful information from people they meet only infrequently than from their close friends, emphasizing the importance of 'weak' ties. Burt (1992) also argues that individuals who

connect different groups, or fill 'structural holes', facilitate knowledge diffusion between groups.

In theoretical studies, while some evaluate the role of homophily positively (Baccara and Yariv, 2013, Jackson and Lopez-Pintado 2013), others, particularly those who incorporate knowledge diffusion into the model, point to different aspects of homophily. For example, Yavaş and Yücel (2014) show that homophily has inversed U-shaped effects on information diffusion as a result of the two opposing effects of homophily. Cowan and Jonard (2004) also find that the level of social knowledge can be maximized when homophilous ties are properly combined with heterophilous ties. Kimura and Hayakawa (2008) show that the presence of heterophily potentially leads to networks that exhibit a small-world property (i.e. most nodes can be reached from every other node through a small number of links). As Watts and Strogatz (1998) find, small-world networks promote rapid information dissemination. Thus, Kimura and Hayakawa (2008) theoretically support the role of heterophily in facilitating knowledge diffusion.

Empirical results on the role of homophily are also mixed. For example, Centola (2011) shows, using a social experiment, that a new health behaviour is more adopted in homophilous networks than in heterophilous networks. Caria and Fafchamps (2018) find that exogenous disclosure of the identities of agents in laboratory experiments leads to the creation of homophilous networks, but not always inefficient ones in terms of information diffusion. However, many others demonstrate negative effects of homophily on diffusion and performance. According to McDonald and Westphal (2003) and McDonald *et al.* (2008), when the CEOs of firms are linked with other CEOs of similar attributes, they often fail to seek for advice from outside the network. The CEOs' behaviours reduce firms' propensity to change corporate strategy in response to poor performance and thus generate downward spirals in firm performance. Fafchamps and Gubert (2007) find that farmers in the Philippines tend to form ties with neighbouring farmers rather than with distant farmers or non-farmers and that such homophilous networks can be fragile to economic risks because neighbouring farmers share the same weather condition.

These mixed findings can be observed in the literature on network density and diversity as well. That is, although some find positive effects of dense networks in terms of economic performance (Phelps, 2010), others find inversed U-shaped effects (Gilsing *et al.*, 2008) or negative effect (Zaheer and Bell, 2005; Fleming, King and Juda, 2007; Aral and Van Alstyne, 2011; Aral *et al.*, 2012; Gonzalez-Brambila, Veloso and Krackhardt, 2013; Todo, Matous and Inoue, 2016). For example, Watson (2007) indicates that, for SME firms, formal business networks with external partners, such as external accountants, are positively associated with survival and growth. This result supports the positive role of weak ties. Burt (2004) finds that workers perform better when they are linked with heterogeneous colleagues, confirming the role of structural holes (Burt, 1992). In addition, many studies revealed that weak ties with outsiders play an important role in many economic situations, including the diffusion of information related to job searches (Granovetter, 1973; Rogers, 2010), technical advice among employees (Constant, Sproull and Kiesler, 1996), and public information between firms (Uzzi, 1999; Uzzi and Lancaster, 2003). Recently, Cai and Szeidl (2017) and Fafchamps and Quinn (2018) conducted randomized experiments in which firms are exogenously provided opportunities to be linked

with other firms, and they found that new links with a variety of firms result in better performance.

In summary, the literature has found that, while homophily is noticeable in most types of networks, it has potentially negatively effects on knowledge diffusion and the performance of firms and individuals, most likely due to the closed nature of homophilous networks. These findings motivate us to examine the nature of network formation in detail. In particular, we hypothesize that the formation of a link between agents is mostly characterized by homophily to enjoy benefits from strong links, while a part of the link formation is characterized by heterophily to additionally enjoy benefits from outsiders. Using a unique data set and estimation method explained below, this study tests this hypothesis.

III. Data

Vietnamese SMEs in the apparel and textile industry

This study focuses on village industrial clusters of SMEs in the apparel and textile industry in the Red River Delta region of Vietnam. To identify these village clusters, we utilized data collected by the World Bank for the Vietnam Enterprise Survey (VES) in 2010. The VES is conducted on an annual basis by the General Statistical Office, and it covers all foreign-owned firms and randomly selected domestic private firms in Vietnam. We examined these data and extracted and identified the villages and communes (the smallest administrative unit in Vietnam) in the Red River Delta region with more than five registered firms in the textile and apparel industry, 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. A total of 354 SME firms operated in the apparel and textile industry in these 16 clusters.

We chose Vietnam as our study area because village industrial clusters have been traditionally developed in Vietnam; therefore, relatively dense ties between firms within the cluster can be observed. In addition, such firm ties are often removed and newly created, as we will see later, probably because firms need to address recent external shocks in the industry. For example, lowering trade barriers encourages exports but also results in more competition with Chinese imports. Therefore, the firms may need to seek new information-exchange partners. These situations in Vietnam provide us with an intriguing research environment for analysing that dynamic nature of network formation.

From December 2014 to January 2015, we conducted the first round of face-to-face interviews with the owners, managing directors, or highly ranked managers of 354 SMEs and obtained responses from 296 (resulting in a response rate of 84%). 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 included questions covering standard firm characteristics, such as sales, number of workers, number of subcontractors, main products, international trade activities, and ownership. In addition, we asked the interviewees to name partners, from a list of all registered firms in the same cluster, with whom they exchange business information; this variable is the dependent variable of interest, i.e. whether firms i and j that operate within the same industrial cluster are connected by an information-exchange tie.

The objective of our empirical analysis is to estimate the mechanism of information-exchange network formation among the SMEs in our sample.[2] We assume that the inter-firm network is undirected. In other words, when either one of a pair of firms nominates the other as a partner, we assume that the two firms exchange information. We assume so because our survey asked with which firm(s) the respondent firm exchanged business information, rather than from which firm(s) it received or to which firm(s) it provided information. In practice, some links are reported only by one of the paired firms, possibly because the role of the two firms in the information-exchange link is not symmetric. However, it is unclear whether more or less information flows from the firm that reported the link to the firm that did not report the link than flows in the opposite direction because of the question format. Therefore, we assume bidirectional flows even when only one firm reported the link between the two. In addition, we focus only on the 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.[3]

Summary statistics for the network data

The network data were obtained for two time periods: from the first round of the survey conducted from December 2014 to January 2015 and from the second round in August 2015. Our empirical analysis is performed on two samples: (i) a sample including firms with no information-sharing partners in both time periods (isolated firms) and (ii) a sample excluding isolated firms. In the following, when not stated explicitly, we will use the first sample. For both samples, if only one firm is available in a village cluster, the corresponding firm is not used in the analysis.

In this analysis, we exclude firms with missing values in the independent variables described below. Consequently, our sample for empirical analysis consists of 217 firms from 13 village clusters (including 14 isolated firms), which create 3,115 'potential' network links within each cluster. Figure 1 presents the networks from three typical village clusters in our sample and shows the density of links, i.e. the ratio of the number of active links to the number of possible links in the network. Figure 2 shows the scatter plot of each village cluster r's density (the ratio of the number of links in cluster r to the number of all possible links) in 2015 against $n(r)$, the total number of firms in village cluster r. From Figure 2, we can find an overall tendency: denser networks in smaller villages and sparser networks in larger villages.

Table 1 shows the dynamics of link formation from 2014 to 2015. In 2015, there were 226 active links, implying that the network was not very dense. Among all 217 firms, the average degree and the number of firms with no information-sharing partners during this period are

[2] Interfirm networks in SME clusters in Vietnam have been examined in the literature, including a seminal paper by McMillan and Woodruff (1999), who examine the effect of buyer-customer relationships among SMEs in Vietnam on provision of informal credit. Our focus is on information-sharing networks, rather than trading networks, and thus is different from theirs.

[3] Surveyed firms reported 332 firms that did not appear in the list of registered firms in the same cluster as their information-exchange partners. Among the 332 firms, 9.6%, 19.9%, 57.7%, and 3.3% are unregistered firms in the same cluster, firms in different clusters in the same province, firms in other provinces, and firms in foreign countries, respectively. Since it is almost impossible to interview all potential partner firms of the 332 firms in our sample region, our analysis focuses only on the network formation among the registered firms in the same cluster.

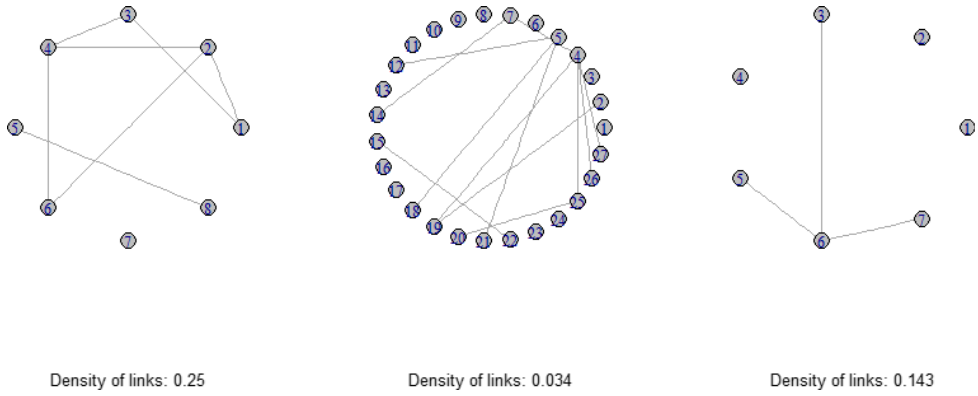

Figure 1. Typical information-sharing networks.

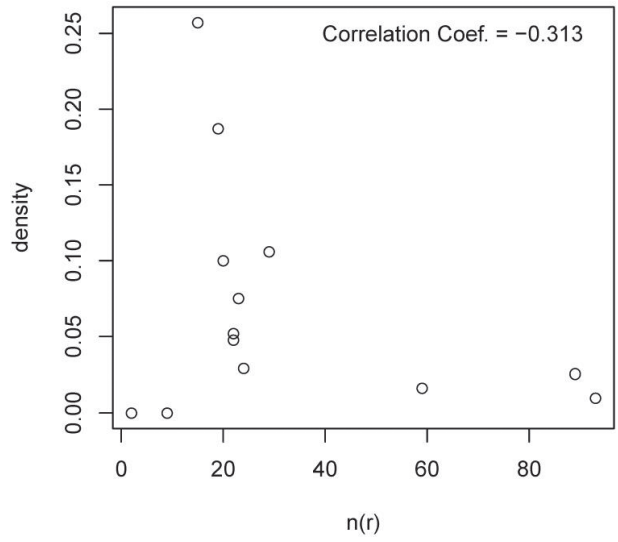

Figure 2. Scatter plot of (Village size, Density).

2.08 (226×2/217) and 61, respectively. Among the 370 firm pairs that had links in 2014, 302 removed their links in 2015, while 158 created new links, which indicates that firms intending to exchange business information in these village clusters relatively frequently remove old links and create new ones. According to our interviews with some of the firms' owners, one possible reason for this phenomenon is that, once a firm changes its buyer, the firm has more opportunities to exchange information with other clients of the new buyer.

To visually understand the relationship between the network in 2014 and the link connection between firms in 2015, we perform a non-parametric kernel regression of link status in 2015 on the number of common partners and on the inverse path distance in 2014,[4] whose results are shown in Figure 3. In each figure, the dashed lines show the

[4]The path distance between agents is defined by the length of the shortest path between them. Following the convention, we define the distance to be infinite if they belong to disjoint networks; thus, the inverse distance is zero in this case.

TABLE 1

Link dynamics

Number of links		*Linked in 2015*		
		Yes	*No*	*Total*
Linked in 2014	Yes	68	302	370
	No	158	2,587	2,745
Total		226	2,889	3,115

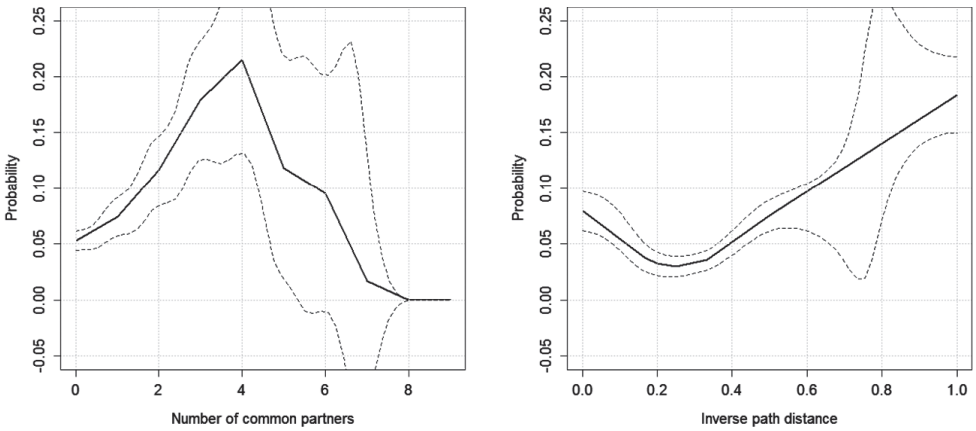

Figure 3. Probability of link formation with respect to network statistics.

95% confidence interval. From the figure on the left-hand side, we can observe that the probability of link formation in 2015 is low if the pair of firms share only a small number of common partners, as expected. The probability of link formation is maximized when the number of common links is approximately four in 2014, and, interestingly, from this point, the probability decreases as the number of common links increases. The figure on the right-hand side shows an overall persistency of the network formation: the shorter the path distance in 2014, the higher the probability of link formation in 2015. An interesting observation is that the probability of forming a link for firms that are totally disconnected in 2014 (such that inv. path distance = 0) is higher than that for firms that are only indirectly connected in 2014 with a certain length (e.g. inv. path distance = 0.2), which would imply that firms prefer to make connections with those unfamiliar to them than with those only slightly familiar to them.

Factor analysis

The independent variables that may affect the network connection are as follows (with their definitions 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)

TABLE 2

Summary statistics

Variable	*Mean*	*Median*	*Standard deviation*	*Minimum*	*Maximum*
Years	8.9263	8	5.8582	1	26
Nworkers	32.1982	10	84.5186	1	1,000
Nsubcontractors	19.0737	2	54.1116	0	450
Retail	20.8295	0	34.4940	0	100
Wholesale	62.5714	90	41.8967	0	100
DirectEx	9.8848	0	27.4400	0	100
IndirectEx	6.6682	0	22.6043	0	100
Age	43.5161	43	10.1046	25	69
Female	0.2074	0	0.4064	0	1
Kinh	0.9677	1	0.1771	0	1
Fboard	0.7373	1	0.6087	0	4
Fratio	0.6460	0.7	0.3081	0	1

Note: Sample size = 217; Number of isolated firms = 14.

- **Wholesale** (the percentage of wholesale sales)
- **DirectEx** (the percentage of direct exports in total sales)
- **IndirectEx** (the percentage of indirect exports in total sales)
- **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 over all workers).

In addition, possible determinants of network connection include the level of productivity and the types of production and management technology used. Although our survey questions ask about sales, many firms did not respond to them maybe because they were afraid of information leakage to the tax offices and competitors.[5] However, variation in productivity levels and technologies may be captured, at least to some extent, by the independent variables above, such as the number of workers, the number of subcontractors, firm age, and the share of direct and indirect exports.

Table 2 presents the summary statistics of the above variables in our sample. The average firm age is 8.9 years, whereas the average and median numbers of workers are 32 and 10, respectively, indicating that our sample includes relatively young and small firms. Certain firms outsource a portion of their production processes to subcontractors; the number of subcontractors is 19.1 on average, while the median is two. Wholesale sales represent a substantial fraction of total sales (62.6% on average). As some firms engage in exports, the average shares of direct and indirect exports are 9.9% and 6.7%, respectively. The average age of managers is 43.5 years; 20.8% of the managers is females, and nearly all are Kinh.

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

[5] We do not use sales as an independent variable because the number of units with missing sales data is 59 among 284.

TABLE 3

Factor analysis (sample size = 217)

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*
Variable	***SCALE***	***RETAIL***	***INDIEX***	***AGE***	***FEMALE***
Years	0.0715	0.0490	0.0249	0.4770	0.0431
ln(**Nworkers** + 1)	0.5254	−0.0721	0.1241	0.2239	−0.1694
ln(**Nsubcontractors** + 1)	0.3882	0.0668	0.0779	−0.0046	0.2198
Retail	−0.3025	0.9368	−0.1172	0.0390	0.1099
Wholesale	−0.4240	−0.8138	−0.3682	−0.1101	−0.0850
DirectEx	0.9837	0.0816	−0.1103	0.0958	−0.0139
IndirectEx	0.0539	−0.0198	0.9957	0.0285	0.0053
Age	0.0865	−0.0352	−0.0015	0.7842	−0.0674
Female	−0.0040	−0.0476	−0.0157	−0.2927	0.4331
Kinh	−0.0972	−0.1710	0.0549	0.0203	0.1101
Fboard	−0.0527	0.0215	−0.0479	−0.0037	0.4390
Fratio	0.2128	0.0089	0.1531	0.2189	0.4786
Loadings	1.739	1.592	1.204	1.050	0.724
Proportion of variance	0.145	0.133	0.100	0.088	0.060
Cumulative variance	0.145	0.278	0.378	0.465	0.526

then use the resulting factor scores (generated by the regression method) as the firm's attribute variables. The number of factors was determined in accordance with the conventional eigenvalue-one criterion. That is, the number of eigenvalues larger than one in the correlation matrix of the variables listed above serves as the number of factors. The criterion suggests that we use five factors. The results of the factor analysis for these five factors are presented in Table 3.[6] 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 considerable weight of **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 that of its CEO. Finally, Factor 5 is interpreted as an indicator for firms with high 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).

Descriptive examinations of homophily

Before conducting a detailed econometric investigation, we perform two simple examinations on the presence or absence of homophily and heterophily.

First, we compute the 'inbreeding' homophily index developed by Currarini *et al.* (2009) for each of the variables presented in section Factor analysis. Suppose that there are K types of agents in a cluster, and let N_i denote the number of type-i agents in the cluster. Then,

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

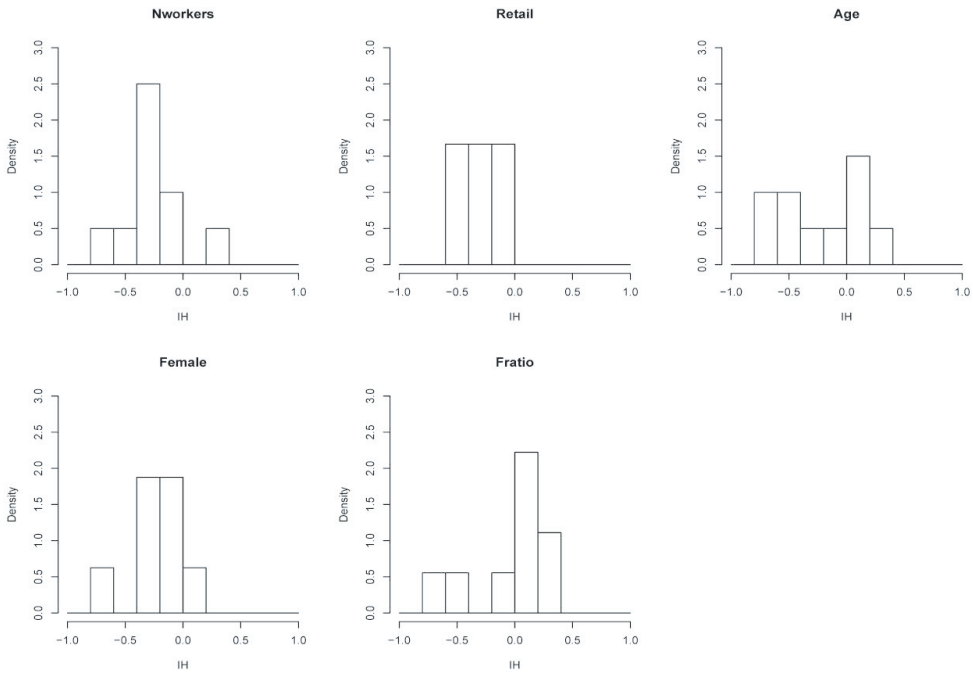

Figure 4. Distribution of Inbreeding Homophily Index.

$\omega_i = \frac{N_i}{N}$ represents the share of type-i agents in the cluster, where $N = \Sigma_{k=1}^{K} N_k$. Furthermore, let s_i denote the average number of links that type-i agents form with agents of the same type within the same cluster, and let d_i denote the average number of links that type-i agents form with agents of different types within the same cluster. Then, the 'basic' homophily index (Currarini *et al.*, 2009) for type-i agents can be defined as the share of the links that type-i agents form with those of the same type: $H_i = \frac{s_i}{s_i + d_i}$. If the basic homophily index for type-i agents is greater than their actual share, i.e. $\omega_i < H_i$, they are more likely to form links with the same type than with other types. Currarini *et al.* (2009) further define the inbreeding homophily index for type-i agents by $IH_i = \frac{H_i - \omega_i}{1 - \omega_i}$. The advantage of IH_i over H_i is that the former can measure the bias in link formation within group i, or the deviation from random formation, $(H_i - \omega_i)$, relative to the maximum potential bias $(1 - \omega_i)$. According to the definition, a positive (resp. negative) value of the inbreeding homophily index IH_i indicates homophily (resp. heterophily).

We classify the firms in our sample into two or three types based on each of the variables listed in section Factor analysis. When the variable used is a dummy variable, such as **Female** and **Kinh**, we simply divide the sample into two groups according to its value. When the variable is continuous, we create three groups based on the 1/3 and 2/3 quantiles. Then, we calculate IH_i for all village clusters for each variable and present the histograms for some selected variables in Figure 4.[7] These histograms show that the level of homophily/heterophily within the cluster varies substantially across clusters. For example,

[7] When a dummy variable is used to calculate IH_i, it cannot be defined in some clusters because all firms belong to either of the two types. In this case, we simply treat IH_i for these clusters as missing.

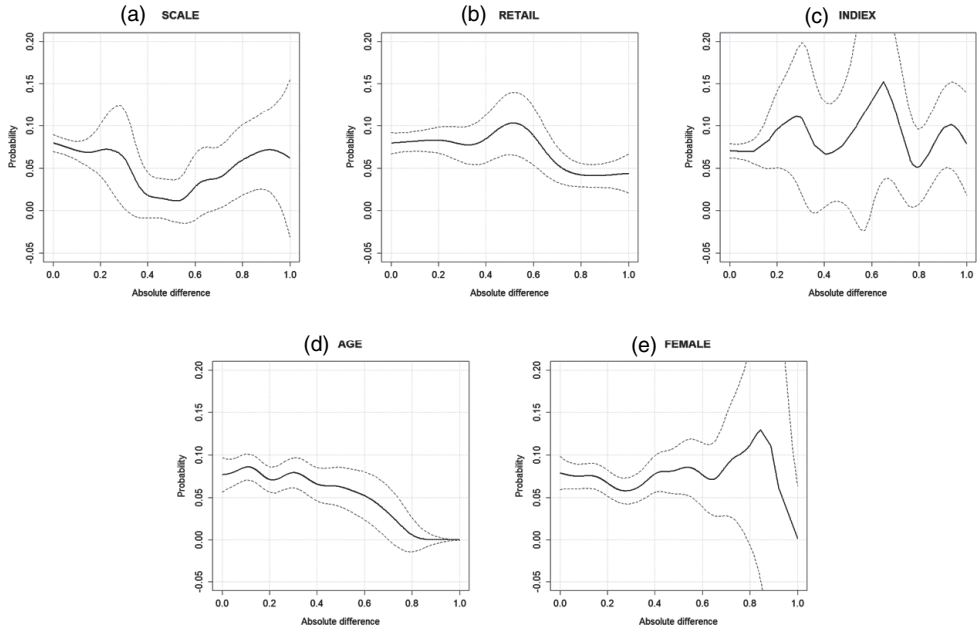

Figure 5. Probability of link formation with respect to the factor scores.
Note: The domain of the probability function is rescaled to [0, 1].

the inbreeding homophily index of **Age** is positive (homophily) for approximately a half of the clusters, while it is negative (heterophily) for the rest. In terms of **Nworkers**, **Retail**, and **Female**, the homophily index is negative for most clusters. These results imply that, depending on the choice of characteristic variable, there are different distributional patterns of homophily and heterophily.

Second, we conduct a non-parametric kernel regression of the link connection between firms on the absolute difference of each of the calculated factor scores between them. The estimated conditional probability curves are provided in Figure 5. Note that if homophily (resp. heterophily) exists in the network formation, then the curves should exhibit an overall decreasing (resp. increasing) tendency.

The figure shows that, except for **AGE**, the probability curves do not have a clear tendency to either increase or decrease. 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, which is contrary to the finding from the inbreeding homophily index. This discrepancy is understandable because the homophily index used above is an intra-network measure while the kernel regressions here are performed globally on all networks. For the **SCALE** variable, interestingly, the probability increases when the potential partner is either very similar or dissimilar. For the **RETAIL** and **FEMALE** variables, the probability of link formation does not peak in the region close to zero but rather records the highest value for firms with certain differences. Nonetheless, if differences in the values of these variables are excessively large near the upper boundary, then the firms tend to avoid forming a link. For the **INDIEX** variable, we cannot observe any clear effects of the variable on link formation.

IV. The model and estimation procedure

Model specification

In this section, we describe our network formation model and its estimation procedure. For 2015 data, let $g_{r,i,j}=1$ if firms i and j in the r-th village cluster ($i\neq j; i,j=1,\ldots,n(r)$) are connected, let $g_{r,i,j}=0$ otherwise, and let $Z_{r,i}=(Z_{r,i}^{(1)},\ldots,Z_{r,i}^{(d_z)})'$ denote the $d_z\times 1$ attribute vector of firm i (namely, the factor scores created in the previous section with $d_z=5$). 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

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

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 the unobserved preference of i for forming a link with j. We introduce the previous network connections in the utility function to incorporate 'dynamic' interdependencies in the link formation, as in Goldsmith-Pinkham and Imbens (2013), Graham (2016) and Graham (2017, Section 3).

Firms i and j form a link if the sum of their marginal gains from the link is positive. 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 so 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 in this manner, the current link status is determined by

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

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Let $T_{i,j}(\tilde{G}_{n(r)})$ be a 3×1 vector defined by [the average of the degree of i and the degree of j in 2014, the number of common partners between i and j in 2014, the inverse path distance between i and j in 2014]. We assume that

$$u_i\left(Z_i,Z_j,\tilde{G}_{n(r)},\varepsilon_{i,j}\right)+u_j\left(Z_j,Z_i,\tilde{G}_{n(r)},\varepsilon_{j,i}\right)=V_{i,j}\left(Z_i,Z_j,\tilde{G}_{n(r)},\beta_{i,j};\theta\right)+\epsilon_{i,j},$$

where $V_{i,j}(Z_i,Z_j,\tilde{G}_{n(r)},\beta_{i,j};\theta)=\alpha+|Z_i'-Z_j'|\beta_{i,j}+(Z_i'+Z_j')\gamma+T_{i,j}(\tilde{G}_{n(r)})'\delta_1+\delta_2 d_{i,j}+\delta_3\ln n(r)$, and α, $\beta_{i,j}$, γ, δ_1, δ_2, and δ_3 are unknown parameters with $\theta=(\alpha,\gamma,\delta_1,\delta_2,\delta_3)$; $d_{i,j}$ represents the geographical distance between i and j; and $\epsilon_{i,j}=\epsilon_{i,j}(\varepsilon_{i,j},\varepsilon_{j,i})$ is an unobserved random variable.[8]

[8] In the recent literature of network formation models, the importance of controlling for (degree and other forms of) unobserved heterogeneity among agents has been emphasized, e.g. Graham (2017) and Jochmans (2018). These papers consider models with agent-specific fixed effects and propose estimators based on differencing out the fixed effects by utilizing the functional form of the logistic distribution. Determining whether such a differencing-out approach can be applied to our framework with RCs is not a straightforward task. If the number of potential partners is sufficiently large for each firm, then it is possible to directly estimate the fixed effects as parameters; however, this is not the case for our data. Then, instead of introducing firm-specific fixed effects, we introduce the term $Z_i'\gamma$ to measure the popularity or productivity of each firm.

The coefficients of the absolute difference in the attributes between i and j, $\beta_{i,j}$, can take pair-specific values to capture the heterogeneity in the effect of homophily and heterophily of firms' attributes on their network formation. In a special case when the pair-specific RCs $\beta_{i,j}$ can be decomposed into the sum of individual-specific coefficients, i.e. $\beta_{i,j} = \beta_i + \beta_j$, and the individual parameters are independent and identically distributed, the mean and the variance of the individual parameters are identical to half those of $\beta_{i,j}$. Hence, in this case, we can still infer the behavioural patterns of individual firms by estimating the mean and variance of $\beta_{i,j}$.

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

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

where $p(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta) = \dfrac{\exp(V_{i,j}(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta))}{1 + \exp(V_{i,j}(Z_i, Z_j, \tilde{G}_{n(r)}, \beta_{i,j}; \theta))}$.

Normal-RC model

First, we consider the case where the RCs $\beta_{i,j} = (\beta_{i,j}^{(1)}, \ldots, \beta_{i,j}^{(d_z)})'$ are normally distributed. Furthermore, for simplicity of exposition, let us assume that the elements of $\beta_{i,j}$ are mutually independent.[9] Then, we obtain the conditional probability of $g_{i,j}$ on $(Z_i, Z_j, \tilde{G}_{n(r)})$ by

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

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

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

It should be noted that, to evaluate the log-likelihood function (5), the multi-dimensional integration in the probability function given in Equation (4), which has no closed-form solution, must be solved. Thus, in application, we use a simulated maximum likelihood (SML) method to estimate (θ, b, s) with a Monte Carlo approximation to function (4) (see, e.g. Train, 2003). Clearly, the presence of pair-specific preference heterogeneity in network formation can be statistically tested by checking the significance of the estimated standard deviations s of the RCs.

[9] In the empirical analysis below, we first estimated a model that allows for non-zero covariances among the RCs. However, the estimated covariances were small in magnitude, and they were not statistically different from zero at any reasonable significance level. Accordingly, we confine our analysis to a simple case of independent RCs. The estimation results for the full covariance model can be obtained from the authors upon request.

Gaussian mixture RC model

In general, the distribution of the preference parameters is not always unimodal and symmetric. In addition, if the specification of the preference distribution is not correct, then the resulting ML estimators will be inconsistent. Thus, the assumption of normally distributed RCs may be restrictive. To overcome this issue, we consider relaxing the normality assumption for the RCs, while the assumption of logistic errors remains unchanged. Such a model is very useful and has garnered focus in the literature (Fox, Ryan and Bajari, 2011; Fox, Kim and Yang, 2016) due to its computational tractability compared with fully non-parametric RC models such as those in Ichimura and Thompson (1998) and Gautier and Kitamura (2013) while retaining a high degree of flexibility in the distribution of RCs.

For simplicity, we continue to assume independence among the RCs. Thus, the joint density function of $\beta_{i,j}$ can be written in general as $f(\cdot) = \prod_{l=1}^{d_z} f_l(\cdot)$, where $f_l(\cdot)$ represents the marginal density function of $\beta_{i,j}^{(l)}$, $l = 1,\ldots,d_z$. A convenient method of estimating an unknown density function is Gaussian mixture (GM) approximation. A typical Gaussian mixture density function for a random variable x can be expressed as

$$f_n(x|\xi) = \sum_{m=1}^{M} \pi_m \phi\left(x|\mu_m, \sigma\right), \sum_{m=1}^{\mathrm{M}} \pi_m = 1, \pi_m \geqslant 0 \text{ for } m = 1,\ldots,M,$$

where $\xi = (\pi_1,\ldots,\pi_{M-1},\mu_1,\ldots,\mu_M,\sigma)'$, and M is a positive integer that is allowed to increase as the sample size increases. Then, for each $l = 1,\ldots,d_z$, the density function $f_l(\cdot)$ can be well-approximated by $f_n(\cdot|\xi^{(l)})$ for a parameter vector $\xi^{(l)}$. Hence, we can approximate the conditional probability of $g_{i,j}$ on $(Z_i, Z_j, \tilde{G}_{n(r)})$ by

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

and the resulting log-likelihood function is

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

Again, since solving the maximization problem for the objective function in Equation (7) is computationally intractable, we use the SML method to estimate $(\theta, \xi^{(1)},\ldots,\xi^{(d_z)})$. However, it is well known that the estimation of mixture models is quite computationally burdensome and unstable. To overcome this difficulty, we use an EM (expectation-maximization) algorithm to maximize the simulated log-likelihood function (see, e.g. Train, 2008). Once the estimate of the density function of a RC is available, we can directly simulate the moments of the RC, and the simulation can be further used to statistically check the presence of a heterogeneous preference in the network formation.

V. Empirical results

Regression analysis

In addition to the two aforementioned RC logit models, 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 the simple logit models and those from the RC logit models are summarized in Tables 4 and 5, respectively. As mentioned in Fafchamps and Gubert (2007) and Comola and Fafchamps (2013), the correlation of the unobserved error terms is a concern for dyadic logit analysis. Thus, we account for the potential dyadic correlation following their approach,[10] and we report below not only the conventional standard errors but also the correlation robust standard errors of the estimates. In the tables, the parameter estimates that are significant at the 5% level based on either the conventional standard errors or the robust standard errors are underlined, and those significant based on both are bolded.

We first describe the results from the simple logit models by focusing on the effects of the absolute difference variables. Recall that, if the coefficients of these variables are significantly negative (resp. positive), then homophily (resp. heterophily) is present for the corresponding variables. The results indicate that the **AGE** variable presents statistically significant homophily, i.e. firms with dissimilar **AGE** values are less likely to form an information-exchange link. However, it should be noted that the **AGE** variable is 'not' significant at the 5% level under the robust standard error, which is a reminiscent of the discrepancy between the homophily index and the kernel regression result observed in section Descriptive examinations of homophily. This result implies that the observed homophily might be simply due to the fact that firms with similar ages tend to form a village cluster. For the other absolute difference variables, we observe no significant impacts on link formation.

Next, we review the results for the firm pair's sum of their attribute variables. 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 behaviour. Regarding the **SCALE** variable, it is understandable that large-scale firms have few incentives to construct new information-sharing connections, as they already have rich business competencies and enjoy the latest technologies. Furthermore, the result for the **INDIEX** variable is reasonable because the presence of local information-sharing partners is important, particularly for indirect-exporting domestic-oriented firms rather than direct-exporting firms. Interestingly, while we have observed that the similarity of the **AGE** variable can be an important factor for link formation, the value of the variable itself is not.

For the effects of the remaining independent variables, the inverse path distance has a positive influence, indicating that the information-sharing partnership tends to be persistent and that the closer in the network in 2014, the more likely they are connected in 2015. Although its statistical significance is weak, the effect of the common link variable is also positive, supporting previous findings (e.g. Jackson and Rogers, 2007). Notably, an increase in the total number of firms in the same village decreases the probability of the network formation statistically significantly. It is imaginable that it is costly to form a

[10] For details, see Equation (10) and footnote 12 in Comola and Fafchamps (2013).

TABLE 4

Estimation results (simple logit models)

	(1) Logit			*(2) Logit*		
Variable	Estimate	S.E.	Robust S.E.	Estimate	S.E.	Robust S.E.
Absolute difference of:						
SCALE	0.056	0.143	0.232	0.073	0.141	0.227
RETAIL	−0.124	0.096	0.090	−0.135	0.094	0.082
INDIEX	−0.131	0.112	0.105	−0.172	0.112	0.104
AGE	−0.298	0.126	0.188	−0.316	0.129	0.190
FEMALE	0.217	0.135	0.206	0.221	0.131	0.190
Sum of:						
SCALE	−0.220	0.084	0.116	−0.221	0.086	0.117
RETAIL	−0.048	0.063	0.091	−0.029	0.063	0.092
INDIEX	0.154	0.083	0.066	**0.197**	0.083	0.063
AGE	0.075	0.064	0.102	0.036	0.064	0.099
FEMALE	−0.194	0.070	0.122	−0.218	0.069	0.118
Intercept	0.428	0.532	0.837	0.614	0.543	0.826
Degree 2014	0.022	0.039	0.049	0.005	0.039	0.051
Mutual Link 2014	0.070	0.070	0.090	0.084	0.070	0.090
Inv. Path Length 2014	0.833	0.286	0.472	0.650	0.288	0.476
Distance in km	0.010	0.126	0.217	0.026	0.125	0.210
ln n(r)	**−0.812**	0.123	0.204	**−0.799**	0.125	0.202
Log-likelihood	−734.735			−718.111		
Inclusion of isolated firms	Yes			No		
Sample size	3,115			2,835		

Note: The estimates significant at the 5% level based on either the conventional S.E. or the robust S.E. are underlined, and those significant based on both are bolded.

partnership link and ensure its preservation, which would indicate that the capacity of the number of partners a firm can hold is limited and is consistent with the finding in Figure 2. Finally, the geographical distance between two firms does not affect their link status significantly, probably because we focus on link formation among firms in the same small village clusters.

We next report the results from the RC logit models. We first estimated models where all five absolute difference variables have RCs. Then, for both Normal and GM-RC logit models, we found that the standard deviations of the RCs of **SCALE**, **INDIEX**, and **AGE** were insignificant at any reasonable significance levels in terms of both types of standard errors. This finding indicates that there is no variation in the degree of homophily and heterophily across firms in terms of these firm attributes. Thus, to improve the efficiency, we re-estimate the models under the assumption that these three variables have constant coefficients, as reported in Table 5.[11]

As shown in Table 5, while the absolute difference of **RETAIL** and that of **FEMALE** are found to be insignificant in the simple logit models, their standard deviations are es-

[11] The estimation results of the full models are available upon request.

TABLE 5

Estimation results (RC logit models with $M = 2$)

Variable		*(1) Normal-RC Logit*			*(2) Normal-RC Logit*			*(3) GM-RC Logit*			*(4) GM-RC Logit*		
		Estimate	*S.E.*	*Robust. S.E.*	*Estimate*	*S.E.*	*Robust. S.E.*	*Estimate*	*S.E.*	*Robust. S.E.*	*Estimate*	*S.E.*	*Robust. S.E.*
Absolute difference of:													
SCALE	Mean	0.143	0.169	0.249	0.169	0.165	0.238	0.162	0.180	0.264	0.189	0.175	0.248
RETAIL	Mean	−0.688	0.381	0.451	−0.810	0.383	0.423	−0.725	0.450	0.482	−0.828	0.584	0.577
	S.D.	**0.889**	0.359	0.417	**1.011**	0.349	0.369	0.567	0.775	0.823	0.699	0.706	0.676
INDIEX	Mean	−0.116	0.124	0.105	−0.172	0.133	0.110	−0.096	0.129	0.119	−0.132	0.133	0.116
AGE	Mean	−0.328	0.165	0.255	−0.363	0.177	0.264	−0.336	0.167	0.275	−0.369	0.188	0.282
FEMALE	Mean	−0.518	0.531	0.374	−0.477	0.572	0.493	−0.952	0.799	0.695	−0.677	0.856	0.777
	S.D.	**1.452**	0.570	0.474	**1.448**	0.627	0.602	**1.755**	0.705	0.616	**1.491**	0.759	0.710
Sum of:													
SCALE		**−0.273**	0.103	0.139	**−0.283**	0.108	0.142	**−0.305**	0.112	0.149	**−0.315**	0.116	0.148
RETAIL		−0.033	0.070	0.095	−0.004	0.070	0.091	−0.045	0.074	0.104	−0.015	0.072	0.099
INDIEX		0.156	0.092	0.072	**0.209**	0.103	0.078	0.128	0.095	0.081	0.177	0.101	0.074
AGE		0.098	0.086	0.137	0.058	0.088	0.135	0.122	0.091	0.143	0.072	0.091	0.138
FEMALE		−0.203	0.081	0.127	−0.243	0.084	0.130	−0.192	0.089	0.134	−0.242	0.090	0.134
Intercept		1.231	0.657	0.862	1.597	0.716	0.944	1.207	0.706	0.891	1.316	0.750	0.959
Degree 2014		0.021	0.043	0.048	−0.004	0.044	0.049	0.018	0.046	0.054	−0.011	0.047	0.052
Mutual Link 2014		0.097	0.085	0.103	0.117	0.086	0.104	0.114	0.091	0.111	0.136	0.091	0.109
Inv. Path Length 2014		**1.099**	0.322	0.380	**0.926**	0.325	0.392	**1.293**	0.348	0.396	**1.168**	0.349	0.403
Distance in km		−0.002	0.127	0.182	0.004	0.132	0.186	0.044	0.023	0.024	0.044	0.024	0.026
ln n(r)		**−0.989**	0.154	0.207	**−1.011**	0.164	0.223	**−1.010**	0.168	0.218	**−0.983**	0.176	0.235
Log-likelihood		−730.752			−714.397			−725.310			−712.282		
Inclusion of isolated firms		Yes			No			Yes			No		
Sample size		3,115			2,835			3,115			2,835		

Note: The number of Monte Carlo repetitions to approximate the multidimensional integration in Equations (4) and (6) was set to 500. The (robust) standard errors of the mean and standard deviation of the random coefficients in GM-RC models are obtained via parametric bootstrap with 5,000 replications based on the asymptotic normal distribution of the estimates of $\xi^{(l)}$s. We have also estimated models with $M = 3$. However, the efficiency gains were minor, and the models with $M = 2$ outperformed them in terms of AIC. The estimates significant at the 5% level based on either the conventional S.E. or the robust S.E. are underlined, and those significant based on both are bolded.

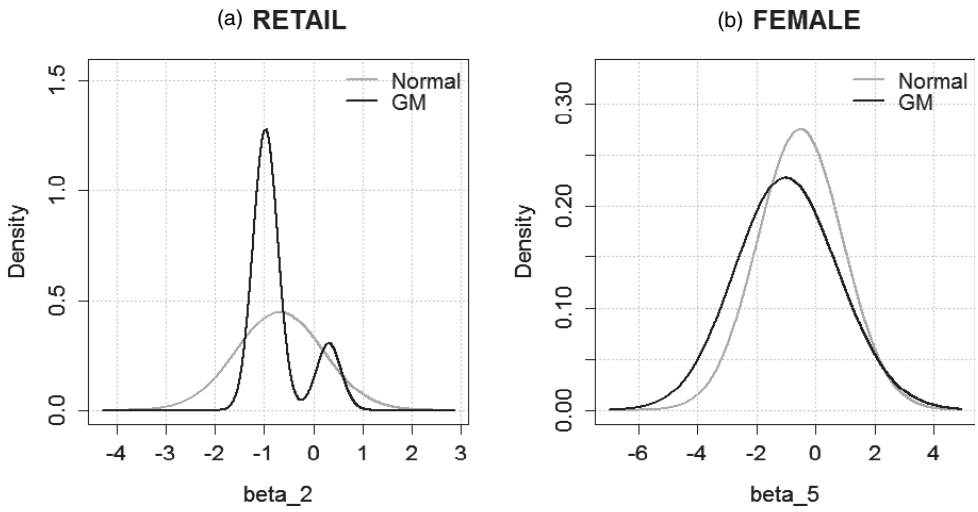

Figure 6. Estimated density of the random coefficients.

timated to be significantly different from zero in the RC logit models. Figure 6 shows the estimated density functions of the RCs of these variables. Here, the estimates from the Normal-RC model are depicted with grey lines, and those from the GM-RC model are shown in black. Panel (A) in the figure shows that, although more than a half of the support of the coefficients is included in the negative region, a certain portion of firms prefers to link with dissimilar partners; that is, most retailing (non-retailing) firms want to share information with other retailing (non-retailing) firms, while some are willing to be linked with non-retailing (retailing) firms. Additionally, some female (male) dominant firms form ties with male (female) dominant firms, although most of them are homophilous, as found in the literature (McPherson *et al.*, 2001). In Vietnam, female-owned SMEs tend to be less productive than male-owned ones, possibly because the former have less access to finance than the latter (Tuan, 2012). By contrast, Bjerge, Torm and Trifkovic (2016) find that enterprise training in Vietnamese SMEs improves skills of female workers more than those of male workers. These findings suggest differences in ability and knowledge between Vietnamese SMEs with large female participation and those with large male participation. Therefore, heterophilous ties between firms would emerge possibly because the firms seek for new knowledge with greater opportunities to meet firms with dissimilar ability and knowledge, as suggested by Currarini *et al.* (2009) and Currarini *et al.* (2016).

For other absolute difference variables, similarly to the results of the simple logit models, the **AGE** variable has a significant negative impact under the conventional (non-robust) standard error, while its effect is insignificant using the cluster-robust standard errors. **SCALE** and **INDIEX** are not statistically significant, implying that the link connections are characterized by neither homophily nor heterophily in terms of **SCALE** and **INDIEX**. The results of the other independent variables are also quite similar to those obtained in the simple logit models, and thus we omit the details.

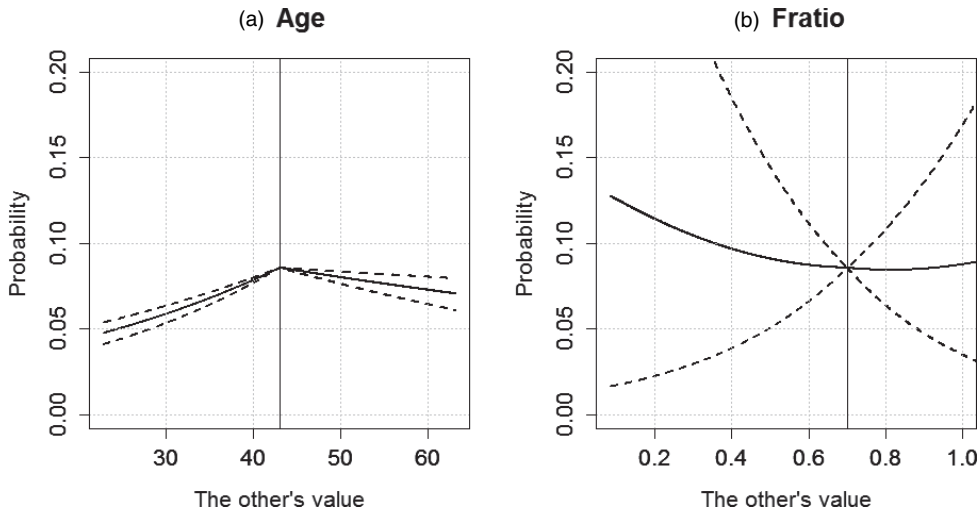

Figure 7. Heterogeneity in the probability of link formation.

Simulation analysis

For illustrative purposes, we now quantitatively examine how the probability of linking 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 the factor scores. Since our factor scores are created by linear combination of the firm's 12 characteristic variables, we can easily calculate the probability of building a link between a pair of firms with particular values of the raw characteristic variables.

Specifically, we consider a representative firm i^*, whose characteristic variables are medians of the raw variables, i.e. $\mathbf{Years}_{i^*} = 8$, $\mathbf{Nworkers}_{i^*} = 10$, etc (see Table 2). Consider firm j^*, which is a potential partner for i^*. Suppose that firm j^* has exactly the same values for the characteristic variables as i^*. In this case, the probability of link formation between these two firms is approximately 8.5% based on the GM-RC model (3) in Table 5.[12] 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 variables remain unchanged.

In the following, we focus on two characteristic variables: **Age** and **Fratio**; the former represents a variable that exhibits weak homophily, while the latter represents a variable that exhibits the coexistence of homophily and heterophily in the above analysis. The results are summarized in Figure 7. 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 lines indicate the median values, namely, the values of the characteristic variables for i^*. Note that the asymmetry of the probability curves is due to the term $(Z_i' + Z_j')\gamma$.

[12] In this and the following simulation analysis, we set the elements of $T_{ij}(\tilde{G}_{n(r)})$ to zero, $\ln n(r) = 4.4886$ (log of the median of $n(r)$'s), and $d_{i^*,j^*} = 1.0806$ (the median of $d_{i,j}$'s).

Panel (A) of Figure 7 shows that, as the absolute difference in **Age** increases, the mean probability of link formation decreases. The figure shows that a difference in the manager's age by 10 years leads to a decline in the probability of link formation between the two by approximately 1–2%. Panel (B) of Figure 7 illustrates the simulation result for the variable **Fratio**. This figure shows that, when the share of female workers of a potential partner becomes larger than the median share, the mean probability that the two firms form a link does not change much. However, the 90% interval of the link formation probability is quite wide. For example, the firm's probability of forming a link with a firm with only female workers (**Fratio** = 1) ranges from approximately 3.5% to 17.5%. This finding clearly points to a large degree of heterogeneity of homophily and heterophily in terms of the share of female workers.

Potential sources of heterogeneity

In this subsection, we investigate the potential sources of the preference heterogeneity in the network formation behaviour. Note that since our network formation model is based on dyadic covariates, we cannot identify individual firm-specific reasons for the heterogeneity. Instead, we can identify pair-specific factors that describe the heterogeneity. For such covariates, we focus on the following two 'distance' variables: the log the geographical distance between two firms, ln(Distance in km + 1), and the log of the inverse network distance in 2014 between the two, ln(Inv. Path Length 2014 + 1).[13] Using these two variables, we modify our RC model in the following manner:

$$\text{Mean of } \beta_{i,j} = b_0 + b_1 \ln(\text{Distance in km} + 1) + b_2 \ln(\text{Inv. Path Length } 2014 + 1).$$

Considering the complexity of the model, we apply the above formulation only for the Normal-RC model (1) in Table 5.

A short summary of the estimation result is presented in Table 6. Firstly, when the two distance variables are controlled, the standard deviations for the RCs are no longer significant for both **RETAIL** and **FEMALE**. This would imply that a certain portion of the preference heterogeneity regarding these firm characteristics can be explained by geographic and network distance between firms.

The result shows that ln(Inv. Path Length 2014 + 1) has significant negative impacts on the means of the RCs for both **RETAIL** and **FEMALE**. That is, two neighbouring firms in the network are more likely to be linked when their attributes are similar, or their link formation is homophilous. By contrast, because the coefficients for the pair of two sufficiently distant firms in the network is positive, they are more likely to be linked when their attributes are dissimilar, or their link formation is heterophilous. Therefore, it is implied that firms tend to be linked with similar firms in the neighbourhood of their network to form a cluster of densely connected similar agents. However, firms may also be linked with dissimilar firms that are far away in the network, possibly to learn new knowledge and information from the dissimilar and distant firms, as suggested by the literature in Section II.

[13] One is added before taking logs because the smallest values of these distance variables are zero.

TABLE 6

Sources of heterogeneity (Normal-RC logit model)

Variable		*Estimate*	*S.E.*	*Robust. S.E.*
Absolute difference of:				
RETAIL	Coef. of ln(Distance in km + 1)	0.027	0.105	0.069
	Coef. of ln(Inv. Path Length 2014 + 1)	**−0.735**	0.372	0.353
	S.D.	0.390	0.444	0.494
FEMALE	Coef. of ln(Distance in km + 1)	−0.280	0.227	0.247
	Coef. of ln(Inv. Path Length 2014 + 1)	**−1.461**	0.621	0.706
	S.D.	0.973	0.690	0.985
Log-likelihood			−724.715	
Inclusion of isolated firms			Yes	
Sample size			3,115	

Note: The estimation results for the other parameters are omitted to save space. The number of Monte Carlo repetitions to approximate the multidimensional integration was set to 500. The estimates significant at the 5% level based on either the conventional S.E. or the robust S.E. are underlined, and those significant based on both are bolded.

Discussions

Our empirical results reveal that most links are homophilous while others are heterophilous in terms of some particular attributes, highlighting the importance of incorporating heterogeneous patterns of homophily and heterophily into analyses of network formation. If a network formation model is estimated without assuming such heterogeneity, as in conventional empirical studies, then one would have overlooked the influence of heterogeneity and thus eventually found only the presence of homophily (or nothing) averaged over the firms. Further, our analysis finds that heterophily arises, or firms are more likely to be linked with dissimilar ones, when the network distance between them is long.

The presence of heterophilous patterns could be explained in accordance with social network studies that note the role of heterophilous links in knowledge diffusion, as presented in Section II. In our case, developing a heterophilous link can increase opportunities to receive new knowledge because firms of different business types that have different products or those characterized by different gender ratios are likely to have different production schemes, management technologies and business resources (McDonald and Westphal, 2003; McDonald *et al.*, 2008).

It is an interesting question whether the firms' networks in our sample are efficiently formed to maximize their profits through knowledge diffusion within the network. Answering this question is challenging. As suggested by Cowan and Jonard (2004) and Yavaş and Yücel (2014), a certain combination of homophilous and heterophilous ties helps improve economic performance because individuals in such a network can simultaneously enjoy the benefits of homophily (e.g. strong trust relationships) and heterophily (e.g. inflows of new knowledge). Therefore, the result for our sample, i.e. that most of the links are homophilous but partly heterophilous, could be a consequence of 'economically rational' network formation behaviour.

VI. Concluding remarks

In this study, we proposed an estimation procedure for a network formation model that allows us to identify the heterogeneous behavioural patterns of homophily and heterophily. In particular, we developed a dyadic logit model with RCs in which the RCs are assumed to be distributed in either a normal distribution or a general distribution that we approximate as the Gaussian mixture. Then, we applied the proposed method to data on the network formation of business information-sharing partners of SMEs in the textile industry in Vietnam. The obtained estimation results were 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 a portion of firm pairs shows heterophilous patterns according to the business type and gender composition of the firm, although, on average, homophily remains dominant in these aspects. To the best of our knowledge, this study is the first in the literature to succeed in numerically documenting the behavioural heterogeneity of homophily and heterophily in network formation using a specific statistical model. The heterophilous link formation could be explained by the argument in 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).

Finally, we note several caveats regarding our analysis. First, we focused on link formation within each cluster and ignored links with firms outside the cluster, primarily because of data limitations. To investigate the formation of links with firms outside the cluster, we need to identify all potential partners of our sample firms, regardless of their locations, and collect information regarding their attributes, which is simply impossible. However, it should be noted that the omission of external links may bias our estimation. For example, if firms strategically form heterophilous links with firms outside the cluster to obtain new information while forming homophilous links with those within the cluster to avoid high link formation costs, our analysis may underestimate the existence of heterophilous links.

Second, although the dyadic RC logit models allow us to investigate the presence of heterogeneity in the effect of homophily and heterophily on link formation between firm pairs, we cannot directly identify each individual firm's preference pattern. Hence, it is difficult to examine what types of firms benefit most from these links and to what extent heterophilous and/or homophilous ties are helpful for each firm's performance.

Third, it should be emphasized that our estimation framework does not always identify the causal relationship of network formation. Our finding is simply that network formation is mostly described by the similarity of the agents and partly described by the dissimilarity of the agents. We interpret this phenomenon as a positive sign for the coexistence of homophilous and heterophilous preference patterns.

Finally, related to this issue, our data set cannot distinguish between homophilous 'preferences' and homophilous 'actions'. In other words, the observed homophilous actions, rather than being caused by homophilous preferences, may be simply due to the fact that firms with particular attributes formed a cluster for some economic, social or geographic reasons. To clearly identify the firms' preferences on their link formation, our model needs to be modified by considering the firms' location choices. To develop a comprehensive

understanding of the network formation behaviour of firms, these issues must be incorporated; however, they are left for future research.

Final Manuscript Received: January 2020.

References

Aral, S., Brynjolfsson, E. and Van Alstyne, M. (2012). 'Information, technology, and information worker productivity', *Information Systems Research*, Vol. 23, pp. 849–867.

Aral, S. and Van Alstyne, M. (2011). 'The diversity-bandwidth trade-off', *American Journal of Sociology*, Vol. 117, pp. 90–171.

Baccara, M. and Yariv, L. (2013). 'Homophily in peer groups', *American Economic Journal: Microeconomics*, Vol. 5, pp. 69–96.

Bjerge, B., Torm, N. and Trifkovic, N. (2016). *Gender Matters: Private Sector Training in Vietnamese SMEs*. WIDER Working Paper, No. 2016/149, The United Nations University World Institute for Development Economics.

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*, Harvard University Press: Cambridge, MA.

Burt, R. S. (2004). 'Structural Holes and Good Ideas', *American Journal of Sociology*, Vol. 110, pp. 349–399.

Cai, J. and Szeidl, A. (2017). 'Interfirm Relationships and Business Performance', *The Quarterly Journal of Economics*, Vol. 133, pp. 1229–1282.

Caria, S., and Fafchamps, M. (2018). *Can People Form Links to Efficiently Access Information*?, Unpublished. Available at https://web.stanford.edu/∼fafchamp/LinkFormation.pdf.

Centola, D. (2011). 'An Experimental Study of Homophily in the Adoption of Health Behavior', *Science*, Vol. 334, pp. 1269–1272.

Chandrasekhar, A. (2016). *Econometrics of Network Formation*, Oxford University Press, Oxford.

Christakis, N. A., Fowler, J. H., Imbens, G. W. and Kalyanaraman, K. (2010). *An Empirical Model for Strategic Network Formation*, NBER Working Paper No. 16039.

Coleman, J. S. (1988). 'Social Capital in the Creation of Human Capital', *American Journal of Sociology*, Vol. 94, pp. S95–S120.

Constant, D., Sproull, L. and Kiesler, S. (1996). 'The kindness of strangers: the usefulness of electronic weak ties for technical advice', *Organization Science*, Vol. 7, pp. 119–135.

Comola, M. and Fafchamps, M. (2013). 'Testing Unilateral and Bilateral Link Formation', *The Economic Journal*, Vol. 124, pp. 954–976.

Cowan, R. and Jonard, N. (2004). 'Network Structure and the Diffusion of Knowledge', *Journal of Economic Dynamics and Control*, Vol. 28, pp. 1557–1575.

Currarini, S., Jackson, M. O. and Pin, P. (2009). 'An economic model of friendship: homophily, minorities, and segregation', *Econometrica*, Vol. 77, pp. 1003–1045.

Currarini, S., Matheson, J. and Vega-Redondo, F. (2016). 'A Simple Model of Homophily in Social Networks', *European Economic Review*, Vol. 90, pp. 18–39.

de Paula, A. (2017). 'Econometrics of Network Models', in Honoré B, Pakes A, Piazzesi M. and Samuelson L (eds), *Advances in Economics and Econometrics: Eleventh World Congress, Econometric Society Monographs*. Cambridge: Cambridge University Press, pp. 268–323.

Durlauf, S. N. and Fafchamps, M. (2005). Social capital, in Philippe A. and Steven N. D. (eds), *Handbook of Economic Growth*, Amsterdam: Elsevier B.V. 2005. pp. 1639–1699.

Fafchamps, M., and Gubert, F. (2007). 'The Formation of Risk Sharing Networks', *Journal of Development Economics*, Vol. 83, pp. 326–350.

Fafchamps, M. and Quinn, S. (2018). 'Networks and manufacturing firms in africa: results from a randomized field experiment', *The World Bank Economic Review*, Vol. 32, pp. 656–675.

Fleming, L., King III, C. and Juda, A. I. (2007). 'Small worlds and regional innovation', *Organization Science*, Vol. 18, pp. 938–954.

Fox, J. T., Kim, K. and Yang, C. (2016). 'A Simple nonparametric approach to estimating the distribution of random coefficients in structural models', *Journal of Econometrics*, Vol. 195, pp. 236–254.
Fox, J. T., Ryan, S. P. and Bajari, P. (2011). 'A simple estimator for the distribution of random coefficients', *Quantitative Economics*, Vol. 2, pp. 381–418.
Gautier, E. and Kitamura, Y. (2013). 'Nonparametric estimation in random coefficients binary choice models', *Econometrica*, Vol. 81, pp. 581–607.
Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., and den Oord, A. (2008). 'Network embeddedness and the exploration of novel technologies: technological distance, Betweenness Centrality and Density', *Research Policy*, Vol. 37, pp. 1717–1731.
Goldsmith-Pinkham, P. and Imbens, G. W. (2013). 'Social networks and the identification of peer effects', *Journal of Business & Economic Statistics*, Vol. 31, pp. 253–264.
Golub, B. and Jackson, M. O. (2012). 'How homophily affects the speed of learning and best-response dynamics', *The Quarterly Journal of Economics*, Vol. 127, pp. 1287–1338.
Gonzalez-Brambila, C. N., Veloso, F. M. and Krackhardt, D. (2013). 'The impact of network embeddedness on research output', *Research Policy*, Vol. 42, pp. 1555–1567.
Graham, B. S. (2016). *Homophily and Transitivity in Dynamic Network Formation*', NBER Working Paper No. 22186.
Graham, B. S. (2017). 'An Econometric model of network formation with degree heterogeneity', *Econometrica*, Vol. 85, pp. 1033–1063.
Granovetter, M. S. (1973). 'The strength of weak ties', *American Journal of Sociology*, Vol. 78, pp. 1360–1380.
Granovetter, M. (2005). 'The Impact of Social Structure on Economic Outcomes', *Journal of Economic Perspectives*, Vol. 19, pp. 33–50.
Ichimura, H. and Thompson, T. S. (1998). 'Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution', *Journal of Econometrics*, Vol. 86, pp. 269–295.
Jackson, M. O. (2010). *Social and Economic Networks*, Princeton University Press, Princeton, NJ.
Jackson, M. O. and López-Pintado, D. (2013). 'Diffusion and contagion in networks with heterogeneous agents and homophily'. *Network Science*, Vol. 1, pp. 49–67.
Jackson, M. O. and Rogers, B. W. (2007). 'Meeting strangers and friends of friends: how random are social networks?', *American Economic Review*, Vol. 97, pp. 890–915.
Jochmans, K. (2018). 'Semiparametric analysis of network formation', *Journal of Business & Economic Statistics*, Vol. 36, pp. 705–713.
Kandel, D. B. (1978). 'Homophily, selection, and socialization in adolescent friendships', *American Journal of Sociology*, Vol. 84, pp. 427–436.
Kimura, D. and Hayakawa, Y. (2008). 'Coevolutionary networks with homophily and heterophily', *Physical Review E*, Vol. 78, pp. 016103.
Kets, W. and Sandroni, A. (2019). 'A belief-based theory of homophily', *Games and Economic Behavior*, Vol. 115, pp. 410–435.
Kossinets, G. and Watts, D. J. (2009). 'Origins of homophily in an evolving social network', *American Journal of Sociology*, Vol. 115, pp. 405–450.
Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). 'Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models', *Social Networks*, Vol. 31, pp. 204–213.
Leung, M. P. (2015). 'Two-step estimation of network-formation models with incomplete information', *Journal of Econometrics*, Vol. 188, pp. 182–195.
Levine, S. S. and Kurzban, R. (2006). 'Explaining clustering in social networks: towards an evolutionary theory of cascading benefits', *Managerial and Decision Economics*, Vol. 27, pp. 173–187.
Luo, S., Du, Y., Liu, P., Xuan, Z., and Wang, Y. (2015). 'A study on coevolutionary dynamics of knowledge diffusion and social network structure', *Expert Systems with Applications*, Vol. 42, pp. 3619–3633.
McDonald, M. L., Khanna, P., and Westphal, J. D. (2008). Getting them to think outside the circle: corporate governance, ceos external advice networks, and firm performance', *Academy of Management Journal*, Vol. 51, pp. 453–475.
McDonald, M. L., and Westphal, J. D. (2003). Getting by with the advice of their friends: ceos advice networks and firms' strategic responses to poor performance', *Administrative Science Quarterly*, Vol. 48, pp. 1–32.

McMillan, J., and Woodruff, C. (1999). 'Interfirm relationships and informal credit in vietnam', *The Quarterly Journal of Economics*, Vol. 114, pp. 1285–1320.
McPherson, M., Smith-Lovin, L. and Cook J.M. (2001). 'Birds of a feather: homophily in social networks', *Annual Review of Sociology*, Vol. 27, pp. 415–444.
Mele, A. (2017). 'A structural model of dense network formation', *Econometrica*, Vol. 85, pp. 825–850.
Moody, J. (2004). 'The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999', *American Sociological Review*, Vol. 69, pp. 213–238.
Phelps, C. C. (2010). 'A longitudinal study of the influence of alliance network structure and composition on firm exploratory innovation', *Academy of Management Journal*, Vol. 53, pp. 890–913.
Rogers, E. M. (2010). *Diffusion of Innovations*, Simon and Schuster, New York, NY.
Sheng, S. A. (2016). *Structural Econometric Analysis of Network Formation Games*. Working paper.
Shipilov, A. V., Rowley, T. J. and Aharonson, B. S. (2006). When do networks matter? a study of tie formation and decay, in: Baum, J. A. C., Dobrev, S. D., Van Witteloostuijn, A. (eds), *Ecology and Strategy*. Bingley, UK: Emerald Group Publishing Limited, pp. 481–519.
Todo, Y., Matous, P. and Inoue, H. (2016). 'The strength of long ties and the weakness of strong ties: knowledge diffusion through supply chain networks', *Research Policy*, Vol. 45, pp. 1890–1906.
Train, K. (2003). *Discrete Choice with Simulation*, Cambridge University Press, Cambridge.
Train, K. E. M. (2008). 'Algorithms for nonparametric estimation of mixing distributions', *Journal of Choice Modelling*, Vol. 1, pp. 40–69.
Tuan, N. P. (2012). 'Gender, innovation and the growth of small medium enterprises: an empirical analysis of Vietnam's manufacturing firms', *VNU Journal of Science: Economics and Business*, Vol. 2, pp. 87–102.
Uzzi, B. (1999). 'Embeddedness in the making of financial capital: how social relations and nnetworks benefit firms seeking financing', *American Sociological Review*, Vol. 64, pp. 481–505.
Uzzi, B., and Lancaster, R. (2003). 'Relational embeddedness and learning: the case of bank loan managers and their clients', *Management Science*, Vol. 49, pp. 383–399.
Watson, J. (2007). 'Modeling the relationship between networking and firm performance', *Journal of Business Venturing*, Vol. 22, pp. 852–874.
Watts, D. J., and Strogatz, S. H. (1998). 'Collective dynamics of small-world networks', *Nature*, Vol. 393, pp. 440–442.
Yavaş, M, Yücel, G. (2014). 'Impact of homophily on diffusion dynamics over social networks', *Social Science Computer Review*, Vol. 32, pp. 354–372.
Zaheer, A., and Bell, G. G. (2005). 'Benefiting from network position: firm capabilities, structural holes, and performance', *Strategic Management Journal*, Vol. 26, pp. 809–825.