Matching and Network Effects*

Marcel Fafchamps
University of Oxford†

Marco J. van der Leij
Tinbergen Institute and Erasmus University‡

Sanjeev Goyal
University of Essex§

May 2006

Abstract

This paper examines the existence and magnitude of network effects in the matching of workteams. We study the formation of co-author relations among economists over a thirty year period. Our principal finding is that a collaboration emerges faster among two authors if they are closer in the social network of economists. This proximity effect on collaboration is strong and robust but only affects initial collaboration. It has no positive influence on subsequent co-authorship. We also provide some evidence that matching depends on experience, junior authors being more likely to collaborate with senior authors.

JEL codes: J41, L14

Keywords: network formation, assortative matching, scientific collaboration, academia, job referral

*We are grateful for the useful comments received from Fernando Redondo-Vega, Tom Snijders, Markus Mobius, John Harrington, Jean Ensminger, and Jeff Johnson; from seminar participants in Harvard, Oxford, Utrecht, Alicante, and Essex; and from participants to the Conference on Networks, Behavior, and Poverty held in Oxford in December 2004. We thank the American Economic Association for making the data available.

†Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ. Email: marcel.fafchamps@economics.ox.ac.uk. Fax: +44(0)1865-281447. Tel: +44(0)1865-281446.

‡Tinbergen Institute, Erasmus University Rotterdam. E-mail: mvanderleij@few.eur.nl

§Department of Economics, University of Essex. E-mail: sgoyal@essex.ac.uk

1. Introduction

It is widely believed that the application of science to technology is responsible for the massive increases in prosperity the world has experienced since the onset of the industrial revolution. Modern growth theory identifies the accumulation of scientific knowledge as the primary source of growth (e.g. Aghion & Howitt 1992, Romer 1990, Lucas 1993). Little is know, however, on the process by which knowledge is created. At the heart of all knowledge-based models of growth is the idea of knowledge externality: old knowledge makes the creation of new knowledge easier. This feature is thought to account for persistent growth.

Because the capacity of any single individual to accumulate knowledge is limited, in practice knowledge externalities occur thanks to the non-rival nature of knowledge. Since knowledge is also complementary in the sense that new advances in knowledge are built upon earlier ones, it is natural to suspect that the knowledge creation process depends to a large extent on the mechanisms by which knowledge is shared between researchers. One key component of this sharing of knowledge is collaborative research. Only through collaborative research can scholars share intimate technical knowledge and bring together very specialized skills. Collaborative research is thus likely to play a central role in the knowledge creation process. It is therefore important to understand the factors that favor or hinder collaboration among researchers.

In this paper we study the forces that shape how collaborative research comes to be. We take co-authorship of an academic publication as our indicator of collaborative research. While this measure fails to capture other important forms of collaboration, such as participation in sponsored research projects, by focusing on tangible research output it offers the advantage of being unambiguous and easy to measure. We choose to focus on economists for two fundamental reasons. First, being economists ourselves, we feel we have a better understanding of the research context and are better armed to draw correct inference. Second and more importantly, economics

as a discipline has by and large resisted the temptation to include as coauthor all the members of a research team. This stands in sharp contrast with other disciplines such as medicine or biology where the number of coauthors is often very large and includes many people who did not participate directly to the research. This means that co-authorship in economics is a more accurate signal of actual pooling of knowledge and skills to produce an identifiable research output.

Using a database of all published articles in economic journals over the last 30 years, we construct a dataset containing all coauthored papers published during that period. By considering each author as a node and each co-authorship as a link between nodes, we define a research collaboration network. In this work, we investigate the determinants of co-authorship. Ultimately, these determinants affect the architecture of the network, a detailed analysis of which can be found in Goyal, van der Leij & Moraga-Gonzalez (2006)

We find robust evidence that the creation of new coauthorships is subject to network effects: a new collaboration emerges faster if the two authors are more closely connected, either directly or indirectly, through collaborations with others. Put differently, John and Jack are more likely to publish together if both have already published with Jill. We show that the network proximity effect extends quite far to include rather roundabout connections between authors. These results are obtained even though we control for pair-wise fixed effects and for a number of individual characteristics such as publication productivity, overlapping research interests, common affiliation, and total number of coauthorships. We also find that lower average productivity and large difference in productivity between authors favor co-authorship.

This paper fits in a growing literature on networks. It is increasingly recognized that some social phenomena are best understood as taking place within networks. Sociologists and anthropologists have long incorporated network in their conceptual toolbox and demonstrated

the usefulness of the concept in understanding certain market phenomena (e.g. Mitchell 1969, Granovetter 1995). In a seminar thought piece, Granovetter (1985) argues that most if not all economic transactions are embedded in social relationships that help shape them and are shaped by them. Without necessarily coming as far as Granovetter, Greif (2001) and North (2001) in their recent work of market institutions have recognized that perfect anonymity is seldom achieved in actual market transactions – and perhaps not even desirable as it would enable crooks to thrive. In a detailed study of market institutions in Sub-Saharan Africa, Fafchamps (2004) provides evidence that networks of business acquaintances play an important role in the operation of markets. Similar empirical evidence is provided by (e.g. McMillan & Woodruff 1999, Johnson, McMillan & Woodruff 2002, Fafchamps & Minten 2002, Fafchamps & Minten 2001, Fafchamps 2003, Fisman 2003) and many others. In a similar vein, Fafchamps & Lund (2003) and Dercon & de Weerdt (2002) examine the role of interpersonal networks in risk sharing among the rural poor while Munshi (2003) studies mutual assistance among Mexican migrant workers in the US. Gulati (1998) examines how existing social networks affect the chance of a new alliance among firms. Most of this empirical work examine the benefits individuals derive from their network of interpersonal contacts.

In the wake of these empirical advances, economic theorists have begun developing a body of economy theory focused on networks. Following early publications by Montgomery (1991) and Kranton (1996), recent examples of these efforts can be found, for instance, in the works of Kranton & Minehart (2001), Bala & Goyal (2000), Bala & Goyal (1998), Genicot & Ray (2003), and Bloch, Genicot & Ray (2004) on small networks. Vega-Redondo (2004) provides an excellent survey of the theoretical literature on large networks, which is largely inspired from the epidemiological literature. Much of this recent theoretical interest focuses on the formation of networks on which, with the exception of Goyal, van der Leij & Moraga-Gonzalez (2006),

Dercon & de Weerdt (2002) and Fafchamps & Gubert (2004), there is little empirical work by economists. This paper seeks to fill this gap.

This paper is organized as follows. The conceptual framework is provided in Section 2. The testing strategy is discussed in Section 3 together with various econometric issues. The data are presented in Section 4. Econometric analysis is summarized in Section 5.

2. Conceptual framework

We wish to understand how scientific collaborations are formed. Collaborating on a joint research project is fraught with dangers. Researchers may have insufficient information on the true ability of a potential coauthor, or be unable to fully predict their complementarities. Even if information problems can be solved, authors may come to disagree on the conduct of the research or they may resent an unfair distribution of the workload. There is also a risk of free riding or breach of promise, one author failing to provide sufficient input into the research venture. Because it is difficult if not impossible for an external party to assess researchers' input, joint research contracts are basically unenforceable by courts. Informal enforcement is the rule and probably rests on a combination of ethics, repeated interaction and reputational sanctions (e.g. Platteau 1994, Greif 1993).

In an environment characterized by asymmetric information and imperfect enforcement, it is natural to expect interpersonal relationships to matter because they convey information and facilitate enforcement (e.g. Granovetter 1995, Fafchamps 2004). In particular, it is reasonable to expect two researchers must know each other personally before they can collaborate: there are no anonymous collaborations. Our first assumption is thus that prior acquaintance is a necessary condition for collaboration. In particular we investigate whether information about previous coauthorship circulates through a network of professional acquaintances. If this is the

case, researchers who share coauthors are more likely to collaborate in the future. Testing this idea is the main purpose of this paper.

Prior acquaintance is not a sufficient condition for collaboration. In economics, it is always possible for a researcher to publish alone. Since collaboration is voluntary, parties to a scientific collaboration must expect more from working with together than what they could achieve in isolation. Our second assumption is thus that two researchers collaborate only if it is in their mutual interest at the time. Having outlined our two main assumptions, we now examine them in turn. We first model the factors affecting gains from collaboration before turning to the acquaintance process.

2.1. Scientific collaboration

In economics, scientific collaboration towards publication in a refereed journal takes the form of a work team created for a specific task. Success depends on the type of each researcher – their ability, experience, availability, and willingness to exert effort – and on the complementarity in their skills and interests. Since scientific collaboration is voluntary, it is natural to assume that two researchers collaborate if it is in their mutual interest.

To illustrate the factors influencing the decision to collaborate, we construct a simple model of research collaboration. We begin by postulating a research production function that relates the anticipated quality of joint research output R^{ij} to the effort e and ability a of researchers i and j:

$$R^{ij} = r(e_i, e_j, a_i, a_j)$$

We assume that function $r(.)$ is strictly increasing in all its arguments.

Each researcher is assumed to derive utility (and possibly other compensation in the form of salary or job promotion) from the quality of his or her research output. The utility derived

by i from a collaboration with j is thus:

$$\pi^{ij} = kR^{ij} = kr(e_i, e_j, a_i, a_j)$$

where $\frac{1}{2} \leq k \leq 1$ expresses the proportion of research output R^{ij} attributed to i by his or her peers. Researcher i compares the value of collaborating with j as against the possible returns from working with others and alone. Suppose that this outside option is given by $\bar{R}(E, a_i)$, where E is the total effort/time available to i, and a_i is his ability. Then researcher i chooses effort e_i in joint project with j to solve:

$$\max_{e_i^{ij}} U^i = kr(e_i, e_j, a_i, a_j) + \bar{R}(E - e_i, 0, a_i, 0)$$

which yields first order condition of the form:

$$k\frac{\partial r(e_i, e_j, a_i, a_j)}{\partial e_i} = \frac{\partial \bar{R}(E - e_i, a_i)}{\partial e_i} \tag{2.1}$$

Equation (2.1) implicitly defines an optimal choice of effort given respective abilities and the effort provided by the coauthor. Combining first order conditions for the two coauthors defines the Nash equilibrium level of collaborative effort e_i^* and e_j^*.

We now consider a number of special cases to illustrate some important factors affecting the decision to collaborate. The central concern of our analysis is the determinants of collaboration. An important determinant of collaboration is clearly the ability of the potential collaborator. We focus on the interaction between effort and ability in shaping the collaboration decision.

Collaboration among authors of similar ability: If the abilities of individuals improve the quality of research, there is a natural pressure towards individuals wanting to collaborate with others of

higher ability. The following example illustrates this intuition. Suppose research output takes the form $r = a_i + a_j$. This corresponds to the case where pooling abilities raises the quality of the research project. Here effort does not matter and can thus be ignored. In this setting an individual either chooses to work with j or to take an outside option given by (say) a_i. Author i prefers to collaborate whenever:

$$k(a_i + a_j) > a_i$$

and similarly for author j. If authors are of equal ability, i.e. if $a_i = a_j$, collaboration is optimal whenever $k \geq \frac{1}{2}$. This is intuitive: since pooling abilities raises research quality, it is optimal for researchers to collaborate as long as they receive sufficient credit for their joint work.

If authors are of unequal ability, collaborating is always attractive for the weaker author but need not be in the interest of the more able author. For instance, if $a_j = 0$, then i prefers not to collaborate for any $k < 1$. In general collaboration is optimal if and only if:

$$\frac{a_j}{a_i} > \frac{1-k}{k}$$

For instance, if $k = 2/3$, then j's ability must at least be equal to half of i's. In such a world, there is assortative matching: collaboration takes place between authors of similar ability level.

So far we have assumed that authors are fully complementary so that there is no overlap in their abilities. To allow for overlap, let each individual ability be made of two components: one that is shared by both authors, denoted a_{ij}, and one that is specific to each author. The condition for i's collaboration now is:

$$k(a_i + a_j + a_{ij}) > a_i + a_{ij}$$

which is satisfied whenever:

$$\frac{a_j}{a_i + a_{ij}} > \frac{1-k}{k}$$

It follows that if author j brings no special ability to the collaboration – i.e., if $a_j = 0$ – then the higher ability author i refuses to collaborate for an $k < 1$. This is also true even if author i has no special ability either – i.e., if $a_i = 0$. This shows that collaboration is more likely between authors whose abilities are complementary, that is, for whom the overlap in competence a_{ij} is a small component of their total ability.

To summarize, we have shown that if research output depends only on ability, collaboration is most likely between authors of a similar level of ability (assortative matching) but with non-overlapping competences (complementarity in competences). The above argument suggests that we should not expect to see much collaboration between researchers of very different abilities. We now examine whether this conclusion is also valid if we incorporate effort levels in the model.

Collaboration among authors with dissimilar ability: The example below explores the following idea: collaboration between high and low ability authors can arise if the low ability author provides more effort. In this manner the time-constrained high ability author can produce more research while the low ability researcher produces better quality output.

Suppose that research output takes the simple form:

$$R^{ij} = (a_i + a_j + a_{ij})r(e_i + e_j)$$

where a_{ij}, as before, represents overlapping ability and we assume decreasing returns to effort, i.e., $r'' < 0$. Suppose also that $(a_i + a_{ij})\bar{R}(x)$ is the return from allocating effort $x \in [0, E]$ to research alone. To simplify the exposition we assume that author j has no special ability – $a_j = 0$.

We begin by showing that, compared to i, author j allocates more effort to joint research than to own research. This is because effort on the joint research project is more productive for j thanks to i's high ability. Formally, when the authors collaborate we have first order conditions of the form:

$$\begin{aligned} k(a_i + a_{ij})r'(e_i + e_j) &= (a_i + a_{ij})\bar{R}'(E - e_j) \text{ for } i \\ k(a_i + a_{ij})r'(e_i + e_j) &= a_{ij}\bar{R}'(E - e_i) \text{ for } j \end{aligned}$$

from which we obtain:

$$\frac{\bar{R}'(E - e_i)}{\bar{R}'(E - e_j)} = \frac{a_i + a_{ij}}{a_{ij}} \tag{2.2}$$

Equation (2.2) shows that the marginal return to effort is higher for i than for j, which implies that $e_i < e_j$ since $r'' < 0$ by assumption. We also see that the ratio $\frac{e_j}{e_i}$ is increasing in a_i: the larger the ability gap between the two authors, the more unequally effort is divided between them. Using the first order conditions, it can also be shown that e_i^* is decreasing in e_j: author i provides less effort if j provides more.

We now ask whether collaboration takes place. The high ability author prefers to collaborate if:

$$\begin{aligned} k(a_i + a_{ij})r(e_i + e_j) + (a_i + a_{ij})\bar{R}(E - e_i) &> (a_i + a_{ij})\bar{R}(E) \\ \Leftrightarrow kr(e_i + e_j) + \bar{R}(E - e_i) &> \bar{R}(E) \end{aligned} \tag{2.3}$$

The low ability author prefers to collaborate so long as

$$k(a_i + a_{ij})r(e_i + e_j) + (a_{ij})\bar{R}(E - e_j) > (a_{ij})\bar{R}(E) \tag{2.4}$$

It is immediately clear that as long as $e_j > 0$ condition (2.3) is satisfied for e_i small enough: author i gets the benefit of an additional output without having to invest much effort. Clearly, the low ability author will prefer collaboration to the outside option for small values of $a_{i,j}$. Furthermore, from the first order condition (2.2) we see that e_j increases in a_i, and so from equations (2.3)-(2.4) it follows that the likelihood of collaboration increases in a_i. Given that $a_j = 0$ this means that the likelihood of collaboration increases in the ability difference between the two authors.

Let m_{ij} denote the likelihood that i and j collaborate given their type. If only ability matters, we expect assortative matching with researchers of similar quality working together: m_{ij} is decreasing in the absolute difference between the ability of researchers i and j. If effort matters as well, dissimilar matching can arise whereby a researcher with high ability – or experience – teams up with a less able or less experienced researcher who provides much of the grunt work. In that case, m_{ij} is increasing in the absolute difference between their abilities.

2.2. Matching and referral

We have seen that the likelihood of collaboration between two researchers depends on their type. In order to initiate a collaborative research project, however, two researchers first have to meet. One possibility is that researchers purposefully introduce themselves to those with whom they wish to collaborate. In this case the probability of collaborating simply depends on the mutual gains from collaborating: $P_t^{ij} = m_t^{ij}$.

For purposive matching to be feasible, type must be perfectly observable to all at little or no cost.[1] The model presented in the previous sub-section makes it clear that researchers have an incentive to overstate their ability in order to attract either high ability or hard working

[1] Assuming information processing can be done at reasonable cost: at any moment in time there are tens of thousands of economists actively publishing in refereed journals.

collaborators, depending on the kind of assortative matching.[2] If researchers can dissimulate their type, purposive matching is not feasible.

Another possibility is that authors are matched with each other according to some random process. They then observe each other's type, possibly at a cost, and decide whether to collaborate or not.[3] Let r denote an exogenously given matching probability and assume that, conditional on having met, two researchers i and j collaborate with a probability $m_t^{ij} \leq 1$. With these assumptions, the probability of collaboration between an arbitrary pair of authors i and j is:

$$P_t^{ij} = m_t^{ij} r \tag{2.5}$$

Such a system is not very efficient because screening has to take place for each potential pair of researchers. Since the total number of researchers is extremely large, the scope for duplication is enormous and the cost of search is very high, making it very unlikely that an efficient match will obtain. A more efficient outcome would arise if information about type circulates among researchers.

To model the information circulation process, we imagine a world in which researchers are introduced – or referred – to each other by common acquaintances. We assume that valuable information about type is conveyed, in both directions, by the referral process.[4] The literature has shown that referrals play an important role in matching workers and employers (Granovetter 1995). It is natural to expect referrals to play a similarly important role in matching researchers. In such a world, prior acquaintance matters: if i can be referred to j and vice versa, they have

[2] Even if ability and experience are perfectly observable, for instance through the publication record, researchers still have an incentive to underreport the number of collaborations in which they are involved, a point we did not discuss formally in the previous section but that follows immediately from the model.

[3] This process resembles the market formation process discussed in Fafchamps (2002).

[4] Here we do not model explicitly why accurate information is conveyed by the common acquaintance, but we can imagine that the referral game is embedded in a web of long-term relationships that serves to deter incorrect referral.

more chance of starting a scientific collaboration.

It is natural to assume that collaborating with someone reveals valuable information about their ability and motivation. It follows that a referral about a researcher i is particularly informative when it is provided by a previous coauthor of i. Referral by a coauthor can thus be construed as a vetting process, stating whether a coauthor is competent and can be trusted to do his or her share of the work.

To formalize these ideas, let S_t be the set of active researchers at time t. For the purpose of this paper, a researcher is considered active from the moment of his or her first publication. Some pairs of researchers have coauthored with each other, some are not. We describe the pattern of coauthorship as a graph in which each author is a node and each mutual acquaintance is a link between two nodes. Formally, let $l_t^{ij} = 1$ if at or before time t researchers i and j have coauthored with each other, and $l_t^{ij} = 0$ otherwise. The set of all $i \in S_t$ and l_t^{ij} forms the graph G_t. Because authors enter and exit and links are added as a result of joint publication, the graph changes over time.

To formalize the referral process, consider two authors i and j. Suppose that authors i and j share a common coauthor k. In the parlance of network theory, the network distance (or shortest path) d_t^{ij} between i and j in the coauthorship network is equal to 2. Assume that with probability $b < 1$ author k refers i and j to each other. Conditional on having been introduced, the researchers collaborate with probability $m_t^{ij} \leq 1$. The probability P_t^{ij} of observing a collaboration between i and j at time t is thus:

$$\begin{aligned} P_t^{ij} &= \Pr(i \text{ introduced to } j) \Pr(i \text{ collaborates with } j | i \text{ introduced to } j) \\ &= b m_t^{ij} \end{aligned}$$

Now suppose instead that the shortest path between i and j is of size 3: i has coauthored with k, j has coauthored with l, and l and k have coauthored with each other. Continue to assume that with probability b author k introduces coauthor i to coauthor l. Further assume that l in turn introduces i to his coauthor j, also with probability b. In this case we have:

$$P_t^{ij} = b^2 m_t^{ij}$$

Generalizing the above example, it follows that along any path of length d_t^{ij} the probability of i and j of being referred to each other is $b^{d_t^{ij}-1}$.

So far we have focused on a single network path between i and j. In practice, there might be multiple paths linking them. Consider Figure 1, for instance. There are four paths linking i to j, but they share a common segment kl. We want to find out the value of P_t^{ij} in this case. Let us assume that with probability b, each node refers i to the next node along each path originating from it. This means that nodes a, b, c, k and d each refer i once to the next node with probability b while node l refers him to both c and d, in each case with probability b. The same thing happens in the other direction regarding j.With these assumptions, the probability of i and j being referred to each other is $2b \times b \times (b^2 + b^2) = 4b^2$, that is, to $b^{d-1} \times$ the number of paths. It can be verified that this example generalizes to all configurations. The total probability of observing a collaboration between i and j at t can thus be written:

$$P_t^{ij} = m_t^{ij} \sum_{d=2}^{\infty} C_t^{ij}(d) b^{d-1} \tag{2.6}$$

$$= m_t^{ij} C_t^{ij}(d_t^{ij}) b^{d_t^{ij}-1} + m_t^{ij} \sum_{d=d_t^{ij}+1}^{\infty} C_t^{ij}(d) b^{d-1} \tag{2.7}$$

where $C_t^{ij}(d)$ denotes the number of paths of length d between i and j.

In practice, calculating all possible paths at all distances is an extremely cumbersome process for a network as large as the one we are studying. A closer inspection of (2.6) reveals that the term b^{d-1} falls rapidly with distance provided that b is *small*. If $C_t^{ij}(d)$ does not increase too rapidly with distance, the value of P_t^{ij} is determined primarily by the first term $b^{d_t^{ij}-1}$ where as before d_t^{ij} is the shortest path between i and j. In this case, P_t^{ij} can be approximated by:

$$P_t^{ij} \approx m_t^{ij} c_t^{ij} b^{d_t^{ij}-1} \tag{2.8}$$

where we have defined $c_t^{ij} = C_t^{ij}(d_t^{ij})$, that is, c_t^{ij} is the number of shortest paths between i and j.

Equation (2.8) forms the basis of our testing strategy. If coauthorship networks serve to introduce potential coauthors to each other and a referral is a prerequisite for collaboration, we should observe a relationship of the form depicted by (2.8). For estimation purposes, equation (2.8) is estimated using logit. It is then useful to derive the logit functional form that best corresponds to (2.8). The logit regression takes the form:

$$P_t^{ij} = \frac{e^{\beta X_t^{ij}}}{1 + e^{\beta X_t^{ij}}} \tag{2.9}$$

We want to know how to write βX_t^{ij}. We begin by noting that, for P_t^{ij} small – as is the case in our data – equation (2.9) is approximatively equal to:

$$P_t^{ij} \approx e^{\beta X_t^{ij}} = m_t^{ij} c_t^{ij} b^{d_t^{ij}-1}$$

Taking logs, we obtain:

$$\beta X_t^{ij} = -\log b + \log b(d_t^{ij}) + \log c_t^{ij} + \log m_t^{ij} \tag{2.10}$$

We thus need to estimate a logit model in which the regressors are the length of the shortest path, which enters linearly, the number of shortest paths, and $\log m_t^{ij}$. The dependent variable takes value 1 if i and j collaborate and 0 otherwise. Equation (2.10) predicts that the coefficient of d_t^{ij} is the log of unknown probability b and the coefficient of $\log c_t^{ij}$ should be 1.

In contrast, if referral does not matter – either because information circulates freely or because researchers screen each other directly – then equation (2.5) applies[5] and the model boils down to

$$\beta X_t^{ij} = -\log r + \log m_t^{ij}$$

Testing network referral thus boils down to testing whether the coefficients of d_t^{ij} and $\log c_t^{ij}$ are significant.

So far we have assumed that referral information only circulates between coauthors. This is probably too restrictive. We now discuss what happens if we relax this assumption and allow referrals to circulate more broadly. Suppose there exists a network of personal acquaintance among economists. In this network, a link exists between i and j if i and j know each other well enough to transmit accurate and trustworthy information about other researchers' type. This network is denser – i.e., has more links – than the coauthorship network but, and this is the important point, it includes it since people who have coauthored a paper together by definition know each other.[6] Other assumptions remain unchanged.

[5] In the case of purposive matching, we simply have $r = 1$.

[6] This need not be the case in other sciences where the number of authors on a single paper can be very large. But for economics it is a reasonable assumption.

We have seen that the probability that two researchers are referred to each other is a decreasing function of the network distance between them. Let d_a^{ij} and d_c^{ij} denote the shortest path between i and j in the acquaintance and coauthorship networks, respectively. Define c_a^{ij} and c_c^{ij} similarly. Dropping time and individual subscripts to improve readability, we now have $P \approx mc_a b^{d_a - 1}$ and hence:

$$\beta X = -\log b + \log b(d_a) + \log c_a + \log m$$

We observe d_c but we do not observe d_a. However, d_c provides some useful information regarding d_a. Since the coauthorship network is included in the acquaintance network, we have:

$$d_a \leq d_c$$

Consequently, the lower d_c is, the lower d_a must be. To illustrate this, Figure 2 presents the unconditional distribution of $f(d_a)$. Observation d_c provides an upper bound statistic on d_a. The conditional distribution of d_a is the truncated distribution below d_c. It follows that $E[d_a|d_a \leq d_c]$ increases with d_c, as shown in the Figure. Put differently, d_c provides information about unknown d_a since the average value of unobserved d_a increases monotonically with observed d_c. We can therefore regard d_c as a valid proxy variable for d_a (Wooldridge 2002). The requirement is that d_c not be so much above the distribution of d_a that $\partial E[d_a|d_a \leq d_c]/\partial d_c \to 0$. This is illustrate in Figure 1 where we see that as d_c increases, $E[d_a|d_a \leq d_c] \to E[d_a]$. To summarize, if we regress P^{ij} on d_c^{ij} and find a significant relationship, this means that network referral matters. If we do not find a significant relationship, it could be either because there is none or because our proxy variable is too crude.

Next we note that the information content of d_c increases as d_c falls. This is because as d_c

falls, the conditional distribution of d_a gets 'squeezed' around its lower bound (the lowest value of $d_a = 1$ when the two researchers are already acquainted). A contrario, when d_c is large, e.g., well above the distribution of d_a, it conveys little if any information about the likely value of d_a. The difference between d_a and d_c thus falls with d_c. Put differently, d_c becomes a better measure of d_a at low values of d_c. Consequently, we expect measurement error bias to be smaller at small values of d_c.[7] This can be investigated by regressing P^{ij} on a series of dummy variables, one for each value of d_c. We expect dummy coefficients to be strongest and most significant at low values of d_c while coefficients should be negligible and non-significant for values of d_c above a certain threshold.

Turning to the number of paths, we also note that c_c constitutes an imperfect measure of c_a. To see this, note that if $d_c = d_a$ then $c_a \geq c_c$: if the coauthorship distance is the same as acquaintance distance, then the number of paths between i and j in the coauthorship network provides a lower bound for the number of paths in the acquaintance network. We have seen that the likelihood that $d_c = d_a$ increases at low values of d_c. Combining the two observations, it follows that c_c constitutes a proxy variable for c_a and that the accuracy of this proxy variable is higher at low values of d_c. If, however, referrals only circulate via the coauthorship network, then equation (2.10) is the correct model and there is not attenuation bias as d_c increases.

This suggests a way of testing whether referrals only circulate in the coauthorship network. Add an interaction term of the form distance $\times \log c_c$ to equation (2.10). If the coauthorship network is embedded inside a denser acquaintance network, attenuation bias implies that the coefficient of the interaction term is negative: c_c becomes a worse proxy for c_a as d_c increases. If referral circulates only in the coauthorship network, then the interaction terms is non-significant.

[7] In the univariate linear case $y = \alpha + \beta x + v$, it can be shown that $p\lim \widehat{\beta} = \beta \frac{1}{1+\sigma_v^2/\sigma_{x^*}^2}$ where σ_v^2 is the variance of the measurement error, $\sigma_{x^*}^2$ is the variance of the true regressor (without measurement error), and β is the true coefficient of x. This shows that the bias in $\widehat{\beta}$ falls when σ_v^2 falls: the less measurement error, the less bias there is.

3. Testing strategy

We have seen in the previous section that the likelihood that two researchers collaborate depends on characteristics such as ability and skill complementarities. In addition, we argued that two researchers must know each other before they can collaborate. To the extent that referral circulates through interpersonal networks, network proximity is expected to affect the likelihood that two researchers begin collaborating with each other. Once they have begun collaborating, however, referral no longer matters and subsequent collaborations should thus depend exclusively on m_t^{ij}. This constitutes the basis for our testing strategy. In this section we describe how estimation and identification problems are dealt with.

Let S_t denote the set of active researchers at time t. For the purpose of this paper, a researcher is considered active from the moment of his or her first publication. At time t, each researcher $i \in S_t$ can potentially coauthor an article with any other researcher $j \in S_t$. Let y_t^{ij} be a dichotomous variable taking value 1 if authors i and j publish a article together in year t, and 0 otherwise. The collection of y_t^{ij} can be represented as a graph or network N_t where each author is a node and each co-authorship is a link.

We wish to investigate whether the likelihood of co-authorship falls with network distance, that is, whether authors who are closer in the co-authorship network and who share more common coauthors are more likely to begin publishing together. Formally, we want to test whether, conditional $y_{t-s}^{ij} = 0$ for all s, the likelihood that $y_t^{ij} = 1$ increases in d_t^{ij} and c_t^{ij}, i.e., whether for first collaborations:

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 0 \text{ for all } s \geq 1) = f(d_t^{ij}, c_t^{ij}, m_t^{ij}) \tag{3.1}$$

with $\partial f/\partial d > 0$ and $\partial f/\partial c > 0$. For subsequent collaborations, we write:

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 1 \text{ for some } s \geq 1) = g(d_t^{ij}, c_t^{ij}, m_t^{ij}) \tag{3.2}$$

If network effects capture referral, we expect that $\partial g/\partial d = 0$ and $\partial g/\partial c = 0$ since once two researchers have collaborated referral is no longer necessary. Estimating equations (3.1) and (3.2) is the objective of this paper.

For estimation of (3.1) to yield meaningful inference about network effects, we must control for factors that could create a false correlation between y_t^{ij} and d_t^{ij} or c_t^{ij}. Our biggest concern is unobserved heterogeneity. Researchers choose to work together because they share common interests or complementary abilities. Since skill complementarity is specific to each pair of researchers, meaningful inference requires that we control for a pairwise-specific fixed effect μ^{ij}. The models to be estimated are of the form:

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 0 \text{ for all } s \geq 1) = f(d_t^{ij}, c_t^{ij}, m_t^{ij}, \mu^{ij}) \tag{3.3}$$

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 1 \text{ for some } s \geq 1) = g(d_t^{ij}, c_t^{ij}, m_t^{ij}, \mu^{ij}) \tag{3.4}$$

where μ^{ij} is a fixed effect corresponding to each researcher pair. Fixed effect controls for many possible time-invariant determinants of scientific collaboration, such as innate ability, education, gender, ethnicity, date and place of birth, etc. Only time-vaying regressors d_t^{ij}, c_t^{ij} and m_t^{ij} are identified.[8]

[8] For equation (3.3), variation in duration is essential to identification. To see why, imagine a contrario that all collaborations happen in two periods. In a fixed effect logit context, each observation has likelihood function:

$$\Pr(y_0^{ij} = 0, y_1^{ij} = 1) = \frac{e^{\alpha + \beta p_2^{ij} + \gamma c_2^{ij}}}{e^{\alpha + \beta p_1^{ij} + \gamma c_1^{ij}} + e^{\alpha + \beta p_2^{ij} + \gamma c_2^{ij}}}$$

Since, by construction, all collaborations take place in period 2, all likelihood functions have the same form. From this, it is immediately obvious that the only identiable parameter is α, the constant term.

By contrast, imagine that some collaborations take place in two periods, with the likelihood above, while others

We estimate equations (3.3) and (3.4) using a fixed effect logit estimator. Doing so raises a well known identification problem. Both equations are equivalent to duration models with fixed effects, except that they are estimated in discrete time. It is well known that in single spell duration models duration dependence and fixed effects cannot be separately estimated.[9] In practice this means that we cannot include time effects in equations (3.3) and (3.4) since duration dependence is subsumed in the fixed effect.

Estimation of the first collaboration model (3.3) raises an additional difficulty that has been noted by Allison & Christakis (2005). To understand the problem, assume that both authors begin publishing at time t_0 and coauthor their first paper together at time t_1. This means that $y_t^{ij} = 0$ for all $t \in [t_0, t_1)$ and $y_t^{ij} = 1$ for $t = t_1$. Thus for each pair ij the time sequence of dependent variables takes the form $y^{ij} = \{0, ..., 0, 1\}$. The only thing that varies across pairs is the number of 0 observations. This mecanically generates a spurious correlation between the dependent variable and any regressor that exhibits a time trend. The nature of the problem is illustrated in Appendix using a Monte Carlo simulation.

The solution we adopt is to eliminate any time trend in the regressors by detrending them. This is achieved by first regressing each regressor on a pairwise-specific fixed effect and a linear time trend. Residuals from this regression are then used in (3.3) in lieu of the original regressors.[10] In Appendix we show that this method yields consistent estimates.

---

take place after three periods with likelihood of the form:

$$\Pr(y_0^{ij} = 0, y_1^{ij} = 0, y_2^{ij} = 1) = \frac{e^{\alpha+\beta p_2^{ij}+\gamma c_2^{ij}}}{e^{\alpha+\beta p_0^{ij}+\gamma c_0^{ij}} + e^{\alpha+\beta p_1^{ij}+\gamma c_1^{ij}} + e^{\alpha+\beta p_2^{ij}+\gamma c_2^{ij}}}$$

Identification of β and γ is obtained by combining both types of observations.

[9] Identification is possible in multiple spell duration models when the fixed effect is the same across time (Chamberlain 1985). In our case, however, we expect the fixed effect to be different for first and subsequent collaborations. This is because at the time of first collaboration both researchers only have limited information about each other but are better informed when deciding whether to continue collaborating. For this reason we estimate first collaboration and subsequent collaborations separately.

[10] We also apply this procedure to model (3.4) even though in this case correction is not required since the dependent variable does not exhibit any systematic time trend. As we will see in this case detrending does not affect results much.

4. The data

The data used for this paper come from the Econlit data base, compiled by the editors of the *Journal of Economic Literature*. The data base contains information on all articles published in economic journals between 1969 and 1999. Only limited information is available on each paper. We use this database to construct the variables of interest as follows.

4.1. Definition of variables

We begin by discussing how variables measuring network proximity and ability are constructed. The co-authorship variable y_t^{ij} is defined as follows. Suppose authors i and j coauthor a paper in year t_1^{ij}. We create a variable y_t^{ij} that takes value 1 at t_1^{ij} and 0 otherwise. To determine whether i and j are active at time $t \neq t_1^{ij}$, we look in the database for the earliest year of publication for each author separately, say t_0^i and t_0^j. We then define $t_0^{ij} = \max\{t_0^i, t_0^j\}$. We thus have $y_t^{ij} = 0$ for all $t_0^{ij} \leq t < t_1^{ij}$ and $y_t^{ij} = 1$ for $t = t_1^{ij}$. We proceed similarly for subsequent joint publications. For instance, suppose i and j publish another paper at time t_2^{ij}. We then let $y_t^{ij} = 0$ for all $t_1^{ij} < t < t_2^{ij}$ and $y_t^{ij} = 1$ at $t = t_2^{ij}$.

Our main regressor of interest is network distance d_t^{ij} between i and j – that is, the shortest path between i and j in the coauthorship network. To construct d_t^{ij}, we proceed as follows. We begin by constructing the coauthorship network N_t using authors as nodes and coauthorships as network links and including all publications from year $t-10$ until $t-1$. The reason for combining 10 years of publications is that the effect of network proximity on co-authorship does not die off instantaneously.[11] Since 10 years of data are necessary to construct N_t, observations from 1969 to 1979 cannot be used in the regression analysis.

[11] We experimented with different time lags and found a 10 year window to yield stable results. The lag is long enough to allow memory but at the same time it is sufficiently short to ensure enough observations to allow estimation.

Having obtained the coauthorship network, we compute the shortest network distance d_t^{ij} from i to j in N_t. For instance, if i and j have both published with k, then $d_t^{ij} = 2$. Variable c_t^{ij} is the number of shortest paths between i and j in N_t; it is 0 if i and j are unconnected. When computing the distance from i to j, any direct link between i and j is ignored.

If i were j are not connected, i.e., if there was no chain of coauthors leading from i to j in the 10 years prior to t, then d_t^{ij} is not defined (it is de facto infinite). For this reason, we find it easier to work with the inverse of distance, which we call network proximity p_t^{ij} defined as:

$$p_t^{ij} = \frac{1}{d_t^{ij}}$$

By construction, p_t^{ij} varies between 0 and 1/2. It is 0.5 if i and j share a common coauthor and it is 0 if i and j are unconnected.[12] Variable p_t^{ij} is the distance measure used in the estimation of equation (3.3) and (3.4).

Turning to m_t^{ij}, we begin by noting that fixed effects capture most individual or pairwise factors that might affect the likelihood of forming a scientific collaboration, such as having gone to the same graduate school, having similar abilities, or sharing common interests. We nevertheless recognize that certain elements of m_t^{ij} may change over time, such as research interests and productivity. For instance, we expect research productivity to increase as the beginning of a researcher's career and to fall as retirement approaches. To capture this idea, we look at the publication record of each author i and j.

The number of published papers is an important but imperfect measure of a researcher's productivity. The quality of research also matters. To construct a simple quality-corrected index of research productivity q_t^{ij}, we make use of the point system developed by the Tinbergen Institute in the Netherlands for its tenuring process. According to this system, each journal is

[12] Since own link is ignored in the computation of d_t^{ij}, p_t^{ij} never takes the value 1.

given a number of points. Publishing in the top rank journals, for instance, yields four points, compared to 1 point for a low rank journal. Publishing in intermediate journals yield 2 or 3 points. Tenure decisions are taken based on the number of points a researcher has accumulated. We mimic this process for all authors in our database. For each author variable q_t^i is simply the number of points author i has earned at year t.

Unlike network distance d_t^{ij}, which is a characteristic of a link or pair, research output q_t^i is author-specific. Here we encounter a practical difficult that arises in all symmetric (undirectional) network regressions: since both authors occupy a symmetrical position in the coauthor pair, regressors must not depend on the order of indexation. This means that the same regressors must obtain if we reverse the order of i and j. There are several equivalent ways of dealing with this difficulty.[13] Here we simply choose the mean and the absolute difference:

$$\begin{aligned} \overline{q}_t^{ij} &\equiv \frac{q_t^i + q_t^j}{2} \\ \Delta q_t^{ij} &\equiv \left| q_t^i - q_t^j \right| \end{aligned}$$

Variables $\overline{q}_t^{ij}$ and Δq_t^{ij} capture research productivity effects as follows. If producers of a large quantity of high quality research are attractive research partners for each other, the coefficient of $\overline{q}_t^{ij}$ will be positive and significant. In contrast, if highly productive researchers find that they are more productive on their own, $\overline{q}_t^{ij}$ will have a negative coefficient. If coauthorship is dominated by assortative matching, the coefficient of Δq_t^{ij} should be negative and significant: the more dissimilar the authors become, the less likely they are to collaborate. In contrast, if junior-senior collaborations dominate, we expect the coefficient of Δq_t^{ij} to be positive: the more dissimilar the authors, the more likely they are to collaborate.

[13] Sociologists, for instance, have proposed using the difference and absolute difference between the characteristics of i and j.

To control for changes in research interests, we use the JEL codes contained in the database to define an index of overlapping interests ω_t^{ij}. We categorize the articles into 19 subfields corresponding to the first digit of the JEL codes[14]. If for an article multiple JEL codes are given, then this article is 'divided' and assigned proportionally to the corresponding fields[15]. The index is then constructed as follows. Suppose that $x_{t,f}^i$ is the fraction of articles written by i in field f in the period from $t-10$ to $t-1$ (such that $\sum_f x_{t,f}^i = 1$). We then consider the following measure of field overlap between i and j in year t:

$$\omega_t^{ij} = \frac{\sum_f x_{t,f}^i x_{t,f}^j}{\sqrt{\left(\sum_f (x_{t,f}^i)^2\right)\left(\sum_f (x_{t,f}^j)^2\right)}}$$

This measure ranges from 0 if i and j did not write any paper in the same field, to 1 if i and j wrote in exactly the same fields and in exactly the same proportion. Together, ω_t^{ij}, $\overline{q}_t^{ij}$ and Δq_t^{ij} measure changes in the gains from potential collaboration m_t^{ij}.

4.2. Controls

Since one of our main objectives is to identify the effect of the coauthorship network on the likelihood of doing collaborative research, it is important that we control for factors that may affect the likelihood of coauthorthip and be correlated with network distance. Most of these factors are captured by the pairwise fixed effect, but some time-varying effects remain a cause for concern, notably variation in individual network size and changes in affiliation.

Some researchers have a higher propensity to collaborate than others. To the extent that this trait is time-invariant, it is captured in the fixed effect. But a researcher's propensity

[14] The JEL classification changed in 1990. For articles before 1990 we matched old JEL codes to new JEL codes on the basis of the code descriptions. A correspondence table between old and new JEL codes can be obtained from the authors on request.

[15] To give an example, if for one article the JEL codes A10, A21 and B31 are given, then 2/3 of the article is assigned to field A, while 1/3 of the article is assigned to field B.

to collaborate may also vary over time: as authors build up co-authoring links with a large number of other authors, new collaboration opportunities probably arise at a higher rate. A researcher's network of past collaborators may thus measure a time-varying propensity to collaborate. Because authors with many collaborators have a higher degree in the coauthorship network, their distance to other authors is on average smaller. This may generate a spurious correlation between changes in network distance and coauthorship.

To capture this effect, we calculate the total number of coauthors n_t^i of author i, computed over the ten years preceding time t, and similarly for author j. Because of symmetry, we transform n_t^i and n_t^j in the same fashion as we did for q_t^i and q_t^j, that is, we compute their mean $\overline{n}_t^{ij}$ and absolute difference Δn_t^{ij}.

The propensity to collaborate may also vary with departmental or employer affiliation: close physical proximity may bring researchers together – or it may pull them apart as they seek to distinguish themselves from their colleagues. If researchers collaborate primarily with colleagues, network proximity may simply capture common affiliation. It is therefore important to control for affiliation. Oyer (2005) has shown that economists who start their academic career in a better deparment have a higher research productivity. This may in part be due to contacts junior researchers form with colleagues.

The JEL database contains information about author affiliation, but only after 1989 and occasionally 1988. Moreover the data is spotty and incomplete.[16] It is nevertheless informative to test whether our results are robust to the inclusion of affiliation data.

We construct common affiliation variables as follows. Let F_t^i be the set of all affiliations of author i that are mentioned in i's articles published in year t. Note that F_t^i will be empty if i

[16] Affiliation data is recorded as strings. Much time was invested cleaning the data, for instance to correct spelling mistakes, and differences in language, and irrelevant name variation – e.g., U Harvard or University of Harvard. The bulk of our data cleaning effort was devoted to ensure that individuals coming from the same university are identified as having the same affiliation.

did not publish in year t or if no affiliations were mentioned. To fill these empty gaps, we define

$$\widetilde{F}_t^i = \begin{cases} F_t^i & \text{if } F_t^i \neq \emptyset \\ \widetilde{F}_{t-1}^i & \text{if } F_t^i = \emptyset. \end{cases}$$

This definition of an author's affiliation assumes that an author's affiliation remains unchanged until information to the contrary is given. The common affiliation variable, f_t^{ij}, is then defined as follows. If both $\widetilde{F}_{t-1}^i$ and $\widetilde{F}_{t-1}^j$ are non-empty, then

$$f_t^{ij} = \begin{cases} 1 & \text{if } \widetilde{F}_{t-1}^i \cap \widetilde{F}_{t-1}^j \neq \emptyset \\ 0 & \text{if } \widetilde{F}_{t-1}^i \cap \widetilde{F}_{t-1}^j = \emptyset. \end{cases}$$

If either $\widetilde{F}_{t-1}^i$ or $\widetilde{F}_{t-1}^j$ is missing, then f_t^{ij} is missing as well. Observations for the common affiliation variable start only in 1988. Including affiliation data seriously reduces panel length, making estimation more problematic and weakening inference.

4.3. Descriptive statistics

Descriptive statistics for all variables of interest are displayed in Table 1. The first panel the Table presents summary statistics for the data up to the first collaboration, while the second panel presents statistics for the data after the first collaboration. In the first columns we show summary statistics for all authors. This includes many authors who have published relatively little and appear infrequently in the EconLit database. To capture actively publishing economists, we drop all authors with fewer than 20 publications in the whole database and we recompute all statistics. These are presented at the right of the Table.

We focus first on data up to the first collaboration. The duration to the first collaboration is 5 years on average – 9 years when we limit the sample to actively publishing economists. The

Table 4.1: Summary statistics of the data

Variable	Description	All authors			Authors with > 20 papers		
		Mean	Std.Dev.	Max	Mean	Std.Dev.	Max
Number of pairs		37418			3821		
Number of observations		344085			59515		
	Data before first collaboration						
Number of observations		163306			24719		
$t_1^{ij} - t_0^{ij}$	Duration to first collab.	4.83	4.80	20	9.11	5.49	20
p_t^{ij}	Proximity	.086	.133	.5	.131	.148	.5
$\Pr(p_t^{ij} > 0)$	Connected	.426	.494	1	.616	.486	1
$d_t^{ij} \vert p_t^{ij} > 0$	Distance if connected	3.01	4.24	35	4.04	4.19	25
c_t^{ij}	Number of shortest paths	.897	1.791	69	1.28	1.91	34
$\overline{n}_t^{ij}$	Avg. degree	3.18	2.62	30	4.87	3.36	30
Δn_t^{ij}	Dif. in degree	3.32	3.71	47	4.13	3.91	40
$\overline{q}_t^{ij}$	Avg. productivity	6.32	10.63	187.25	13.75	15.85	187.25
Δq_t^{ij}	Dif. in productivity	8.99	15.64	282.33	16.74	21.71	265.67
ω_t^{ij}	Field overlap	.489	.352	1	.534	.320	1
f_t^{ij}	Common affiliation	.215	.411	1	.176	.381	1
	Data after first collaboration						
Number of observations		180779			34796		
y_t^{ij}	Subsequent collaboration	.146	.353	1	.146	.354	1
p_t^{ij}	Proximity	.278	.221	.5	.334	.187	.5
$\Pr(p_t^{ij} > 0)$	Connected	.721	.449	1	.906	.292	1
$d_t^{ij} \vert p_t^{ij} > 0$	Distance if connected	2.43	2.91	30	2.92	2.73	23
c_t^{ij}	Number of shortest paths	1.038	.562	35	1.105	.810	29
$\overline{n}_t^{ij}$	Avg. degree	5.80	3.57	39	8.86	4.19	39
Δn_t^{ij}	Dif. in degree	4.13	4.30	46	5.05	4.62	43
$\overline{q}_t^{ij}$	Avg. productivity	7.64	12.58	209.33	15.38	17.80	209.33
Δq_t^{ij}	Dif. in productivity	8.99	16.17	282.33	16.48	21.68	282.33
ω_t^{ij}	Field overlap	.714	.259	1	.685	.260	1
f_t^{ij}	Common affiliation	.257	.437	1	.212	.409	1

difference reflects the fact that the careers of active economists are longer, and therefore they have more collaborations initiated later in their career.

Network proximity prior to the first collaboration is .086 on average, which is rather small. This corresponds to an average of 11 degrees of separation between authors in N_t. The probability of being (indirectly) connected before the first collaboration is 42%. The number of shortest paths is on average less than 1. Thus, in most cases there is no path or only a single shortest path between the two economists. Active economists are on average closer in terms of network proximity: p_t^{ij} is larger on average and $\Pr(p_t^{ij} > 0)$ is larger as well. This is normal since, by definition, active authors have a higher degree in the co-authorship network and thus have a larger network.

When connected, authors are rather close, with an average distance of 3 (i.e., 2 degrees of separation). This is a remarkably short distance. For instance, Goyal, van der Leij & Moraga-Gonzalez (2006) found that the average degree of separation of all connected pairs is around 8 to 12 in the co-author network. Hence, the summary statistics tell us that the network distance is much smaller for pairs that eventually start a collaboration. This suggests that collaboration is associated with 'closeness' in the network.

Productivity and connectedness variables are shown next. We see that for the average author in our database the average number of past coauthors is fairly large. As could be expected, the average is much higher – nearly twice – for actively publishing economists. The average difference in connectedness is also quite large. This suggests that, as predicted by the model, certain authors adopt a strategy whereby they seek many collaborations with authors who co-author articles with few people.

The average value of the research productivity index is shown in next two rows of Table 1. We find a large average difference between authors, suggesting that at the time of first collaboration,

authors differ widely in terms of productivity. This is true even if we limit the sample to pairs of coauthors who, over the entire span of the EconLit database, have published a lot. This constitutes prima facie evidence of dissimilar matching, that is, that many first collaborations take place between senior and junior authors to take advantage of complementarities between them.

Statistics on field overlap and common affiliation are presented next. Field overlap is around 50%, suggesting that most economists collaborate with someone in their field. Around 20% of economists had a common affiliation at some point during the 10 years prior to their first collaboration. Actively publishing researchers have a slightly higher field overlap and less often a common affiliation. This last finding might be due to greater travel opportunities for senior authors, making it easier for them to establish contacts outside their own department.

The second panel of Table 1 presents similar information for subsequent collaborations. We observe that authors repeat their collaboration in 15% of the years that follow their first joint publication. When we compare other variables in the lower panel to the variables in the upper panel we observe that authors get closer after the first collaboration. Field overlap and the likelihood of a common affiliation also increase. Collaboration thus seems to bring co-authors closer in terms of network and affiliation.

5. Econometric results

5.1. First collaboration

We now turn to the estimation of our models. We begin with equation (3.3) which analyses the determinants of the first collaboration between a pair of researchers. The complete regression

model for (3.3) is of the form:

$$\Pr(y_t^{ij}=1) = f(\beta p_t^{ij} + \gamma_1 \log c_t^{ij} + \gamma_2 p_t^{ij} \log c_t^{ij} + \theta_1 \overline{q}_t^{ij} + \theta_2 \Delta q_t^{ij} + \theta_3 \omega_t^{ij} + \lambda z_t^{ij} + \mu^{ij}) \quad (5.1)$$

where z_t^{ij} stands for various controls such as $\overline{n}_t^{ij}, \Delta n_t^{ij}$ and f_t^{ij}. As explained earlier, the interaction term $p_t^{ij} \log c_t^{ij}$ is included to test whether referrals only circulate in the coauthorship network. If the coauthorship network is embedded inside a denser acquaintance network, attenuation bias implies that $\gamma_2 > 0$; if referral circulates only in the coauthorship network, then $\gamma_2 = 0$.[17] Equation (5.1) is estimated using conditional logit to eliminate the fixed effect μ^{ij}. All regressors are detrended to eliminate spurious correlation with the dependent variable.[18]

We estimate model (5.1) on the entire dataset as well as on the restricted dataset of authors with at least 20 publications. The dataset indeed contains a very large number of authors who have only published a small number of journal articles. We suspect that these individuals are not committed researchers. For occasional authors, it is conceivable that network interaction does not work in the same fashion. To test the robustness of our findings to the inclusion of occasional authors, we estimate the model using both datasets.

We first present results based on first collaborations only. Coefficient estimates, reported in Table 2, show a very strong positive effect of network proximity p_t^{ij}: the magnitude of the coefficient is very large and the z-statistics is highly significant in the full sample as well as in the restricted sample with only high productivity researchers. This suggests that network proximity plays an important role in research collaborations.

The coefficient γ_1 of the (log of the) number of shortest paths $\log c_t^{ij}$ is not significant in the full regression but the coefficient γ_2 of the interaction term is positive and significant in

[17] Since we used proximity p_t^{ij} instead of distance d_t^{ij}, the sign of the interaction term is reversed.

[18] It is essential to detrend regressors using only observations entering in the estimation of (5.1). So detrending is redone each time the inclusion of a new regressor, such as a_t^{ij}, results in a loss of valid observations.

both regressions. Put differently, the number of shortests paths does not matter except at short network distances. This is consistent with the idea that, as network proximity increases, distance in the co-authorship network becomes a better measure of distance in an (unobserved) acquaintance network. This result suggest that co-authorship referrals circulate in an unobserved acquaintance network that is denser than the observed co-authership network.

Results also inform us regarding the type of complementarities that matter in economic research. The coefficient of the field overlap index ω_t^{ij} is positive and significant in the entire sample, indicating that, as expected, authors are more likely to initiate a collaboration if their research interests converge. The coefficient of the difference in research output Δq_t^{ij} is positive in the full data set (first column of Table 2), suggesting that authors are more likely to initiate a research collaboration if they differ in ability or experience. This is indicative of dissimilar matching, as would arise for instance when a well established researcher teams up with a junior researcher. We also see that authors are more likely to initiate a collaboration when their average research output falls. These results hold for the entire sample but not for the sample of high-productivity authors where productivity variables are non-significant. This is probably because dropping less productive authors loses the kind of collaborations described above.

To better understand productivity effects, iso-likelihood curves are depicted in Figure 3. Each curve gives a combination of author output that yields the same likelihood of publishing jointly at time t, conditional on not having published together until then. Likelihood falls as one gets closer to the origin. The Figure shows that the iso-likelihood curves are sharply kinked. This means that the likelihood of collaboration is first and foremost a function of the productivity of the most productive of the two authors: the more productive this author is, the longer it takes for the pair to publish together. This makes sense: a productive author is someone who publishes a lot in good journals. The more productive author i is, the less time i has to devote

to publishing with j. The fact that the iso-likelihood curve is sharply kinked suggests that it is the time of the most productive author that is the binding constraint. How productive the less productive author is does not matter much. This is consistent with a model of scientific collaboration in which the more productive author provides guidance while the less prominent author provides his or her time.

Turning to controls, we see that, as anticipated, in both samples the likelihood of initiating collaboration increases with the number of coauthors: the more coauthors researchers have, the faster they are likely to collaborate with each other. The difference in the number of collaborators has a negative sign and is significant in both samples as well.

Table 2 indicates the existence of a strong network proximity effect. It is of interest to ascertain whether this result is driven by a local effect over very short network distances, or whether it is a more diffuse effect spreading over long network distances. Network effects at very short distances could be interpreted as evidence of common socialization, as would be the case if researchers introduce their respective coauthors to each other. More diffuse network effects, say, at network distances over 4, are unlikely to be the result of such explicit socialization. But, as explained earlier, it could indicate the existence of a denser acquaintance network. To investigate this idea, we reestimate model (5.1) by replacing p_t^{ij} with a series of dummy variables representing network distance. Coefficient estimates are presented in Figure 4 together with their 95% confidence interval (the dashed lines). We again show results for the full and restricted samples. When interpreting the Figure, it is useful to remember that, prior to the first collaboration, network distance is a least 2.

Results show that network effects are diffuse and are certainly not limited to short network distances: in the full sample, network proximity has a significantly positive effect on the likelihood of collaboration for distances up to 11 degrees of separation. It is extremely unlikely

that this results from explicit socialization among coauthors, i.e., from j introducing i to k, who in turn introduces i to l, who introduces him to m, and so on 11 times. But it is consistent with the existence of 'dark matter', that is, a denser but unobserved acquaintance network. In the case of the sample restricted only to high productivity authors, coefficients are estimated less precisely, which may explain why network effects are not significant above four degrees of separation.

As explained earlier, we worry that our network proximity results may be due to joint affiliation: as researchers join the same department or institute, they may be more inclined to work together. This local in-breeding effect may generate a spurious relationship between network proximity and co-authorship. To investigate this possibility, we reestimate the model with the joint affiliation variable f_t^{ij}. As pointed out earlier, affiliation information is only available after 1988. This means that adding f_t^{ij} results in a massive loss of observations – more than half.

Estimation results are presented in Table 3. Our main results are unchanged: network proximity remains significant in the full and restricted samples; the interaction term $p_t^{ij} \log c_t^{ij}$ remains significant in the restricted sample; and average research output remains negative and significant in the full sample. Contrary to expectations, common affiliation has a negative effect on the likelihood of initiating a first collaboration – the effect is only significant in the full sample.[19] When interpreting this finding, it is important to remember that estimation controls for pairwise fixed effects. This means that identification is obtained only from pairs of authors who, at some point in time, had a common affiliation but subsequently moved away from each other. A negative sign means that such researchers are more likely to start collaborating after moving apart. Why this is the case is unclear – perhaps collaborating is a way to keep interacting

[19] The importance of de-trending all regressors is best illustrated by noting that, when common affiliation is not detrended, it has a positive and significant coefficient.

with each other, perhaps being in the same department engenders tensions and turf battles that make collaboration more difficult. For our purpose, the main point is that our proximity results are not due to omitting affiliation data. Moreover, since a_t^{ij} is either non-significant or negative, in-breeding bias is unlikely and consequently the results presented in Table 2 are probably safe.

5.2. Subsequent collaborations

Turning to subsequent collaborations, we reestimate equation (5.1) using data on subsequent collaborations between two authors who have published one paper together. The form of the regression is the same. Network proximity is defined as the distance between two authors in the co-authorship network that ignores their own joint work. If network proximity is a significant determinant of first collaboration because of a referral effect, we would expect network proximity not to be significant for subsequent collaborations: since the two authors have published a paper together, they no longer need to be introduced to each other.

Results are summarized in Table 4. As anticipated, network proximity no longer has a positive effect on the likelihood to collaborate. In fact, it is now negative and statistically significant in both the full sample and the sample restricted to highly productive researchers.

To investigate why this is the case, we reestimate the model with distance dummies, as we did for Figure 4. Results are not shown here to save space. We find that the only negative and significant dummy is for distance 2; other distance dummies are not significant. What this means is that two researchers who have published jointly in the past are less likely to publish again if both of them are separately publishing with the same coauthor. This is probably not surprising: both authors are busy publishing with the same co-author and thus have less time to publish together. This interpretation is reinforced by noting that the coefficient of the interaction term $p_t^{ij} \log c_t^{ij}$ is significant and negative in the full and restricted samples. This means that the more

common co-authors the two researchers have, the less likely they are to write together. These findings are consistent with an observation made by Goyal & van der Leij (2005), namely that researchers with lots of different co-authors often appear in the network as stars, that is, their co-authors seldom publish with each other separately. What our results suggest is that this may be because they are kept busy by the star author and do not have the time to work with each other.

We again find that collaboration is more likely between authors who differ in the number and quality of their publications, a result consistent with dissimilar matching. This result is only observable in the full sample from which low productivity authors have not been dropped. The likelihood of collaboration also falls with average research output, confirming that repeated research collaboration is more likely among low productivity researchers. When we draw iso-likelihood curves as in Figure 3, they nearly form a right angle.[20] This suggests that the time constraint of the most productive of the two authors is even more determinant in subsequent collaborations than for the first co-authored publication.

Turning to field overlap, we find that authors are more likely to publish together again if their fields of interest drift apart. The effect is large in magnitude and highly significant in the full and restricted samples. One possible explanation is that as their research interests evolve, the two researchers acquire skills that may be complementary, inducing them to collaborate again.

Control variables $\overline{n}_t^{ij}$ and Δn_t^{ij} measuring number of coauthors have the opposite sign compared to Table 2. This is true for the full and restricted samples. Authors with more coauthors appear less likely to continue collaborating with the same person, while the difference in number of coauthors Δn_t^{ij} now has a positive sign. The negative sign for $\overline{n}_t^{ij}$ suggests that certain

[20] This is immediately apparent form the fact that the coefficient of average output is nearly twice, in absolute value, the coefficient on output difference.

researchers have a higher propensity to seek new coauthors. As they do so, it leaves them less time to continue collaborating with earlier coauthors. Such researchers tend to be less 'faithful' to earlier relationships as they seek an ever expanding number of coauthors. The positive sign on Δn_t^{ij} may be because, if one of the two coauthors initiates fewer collaborations with new people, he or she may be able to rekindle a pre-existing collaboration, even if the other author has continued to expand his or her n_t^i.

In Table 5 we present the results of similar estimation where we have added a regressor for common affiliation f_t^{ij}. We see that, although the size of the sample drops fairly dramatically, results remain basically unchanged: all coefficients retain the same sign and most gain in significance. This shows that the results presented in Table 4 are not an artifact of omitting affiliation information. We again find that common affiliation has a negative sign: co-authors are more likely to continue writing together if they move apart. Why this is the case is not entirely clear – it may be because researchers who move across departments or institutions tend to be more able and hence more desirable co-authors, it may be because collaboration is a way for people with similar interests to continue interacting. What matters here is that our conclusions regarding network proximity are not affected by changes in affiliation.

6. Conclusions

We have examined the process by which scientific collaborations are formed. In particular we have investigated two hypotheses: do referrals by co-authors play a role in the initiation of research collaborations, and do research collaborations bring together authors that are similar or dissimilar. Using a simple model of co-authorship, we have shown that collaboration between a low ability and a high ability author can arise if the low ability author provides labor while the other provides guidance. We also devised a test of whether referrals travel only through the

co-authorship network, or whether they travel through a denser but unobserved acquaintance network in which the co-authorship network is embedded.

Because of the importance of unobserved heterogeneity, our testing strategy controls for pair-wise fixed effects and relies solely on time-varying changes to identify factors that affect the likelihood of co-authorship. We develop an original way of dealing with potential bias in estimating a discrete time duration model with fixed effects. Monte Carlo simulations demonstrate that our method eliminates the bias inherent in this category of models, thereby opening the door to the estimation of duration models with time-varying regressors and individual fixed effects.

We applied this methodology to the Econlit database of joint publications in economic journals over the period 1969 to 1999. We use the databased to construct an index of field overlap as well as an indicator of common affiliation which is used as control variable. Our results indicate that network proximity is a strong determinant of first collaboration, suggesting that co-author referrals play a role in the formation of scientific collaborations. We also conclude that referrals travel through a network that is denser than the co-authorship, but in which the latter is embedded. We call this the acquaintance network. We show that network proximity does not increase the likelihood of subsequent collaboration, a result that is consistent with the referral hypothesis.

Regarding the second hypothesis, we find that co-authorship is more likely between dissimilar authors, where ability is proxied by the quality and quantity of past publications. This is true for first as well as subsequent collaborations. This finding is consistent with the idea that one author provides labor while the other provides guidance – as is common, for instance, when a doctoral student writes an article with his or her thesis supervisor.

Our results also throw additional light on the architecture of the scientific collaboration

network studied by Goyal, van der Leij & Moraga-Gonzalez (2006) and Goyal & van der Leij (2005). In particular, they indicate that clustering may be due to in-breeding bias – co-author referral circulates information locally in the acquaintance network, making it easier for proximate authors to collaborate. In-breeding bias has important implications regarding the efficiency of research collaborations in the sense that isolated researchers are less likely to collaborate with others. Researchers at the center of a dense web of collaborative relationships are much more likely to form new collaborations.

The analysis presented here leaves important questions unanswered. It is unclear, for instance, whether collaborations are beneficial to researchers. It is indeed conceivable that collaborations are inefficient because they entail coordination costs. Researchers, for instance, may prefer to work alone but be compelled to accept to collaborate either for altruistic reasons (e.g., assisting less experienced researchers) or due to social pressure (e.g., sharing the fruits of academic success). Testing whether collaboration makes scientists more or less productive is the object of future research.

7. Appendix

In this appendix we illustrate the difficulty inherent in estimating a fixed effect logit model for first collaborations, and show how detrending regressors solves the problem. To this effect, we construct a Monte Carlo simulation that reproduces the kind of data we have. We begin by generating pair-wise fixed effects $u_i \sim N(0,5)$.[21] We then create two potential regressors x_{it} and z_{it} indexed over individual (e.g., pair of authors) i and time t. Each regressor is constructed as a trend with noise:

$$\begin{aligned} x_{it} &= t + \varepsilon_{it}^x \\ z_{it} &= t + \varepsilon_{it}^z \end{aligned}$$

with $\varepsilon_{it}^x \sim N(0,100)$ and $\varepsilon_{it}^x \sim N(0,100)$. A latent variable y_{it}^* is then generated as:

$$y_{it}^* = -2 + x_{it} + u_i + \varepsilon_{it} \tag{7.1}$$

with $\varepsilon_{it} \sim N(0,400)$. The dichotomous dependent variable is defined as $y_{it}^a = 1$ if $y_{it}^* > 0$, 0 otherwise. Since z_{it} does not enter equation (7.1), any correlation observed between z_{it} and y_{it}^a must be regarded as spurious. We then define $y_{it} = y_{it}^a$ except if $y_{it-s}^a = 1$ for any $s > 0$, in which case y_{it} is defined as missing. Variable y_{it} thus has the same form as the dependent variable in the first collaboration case: a series of 0 ending with a single 1.

We generate 1000 samples of y_{it}^a, y_{it}, x_{it} and z_{it}, each with $t = \{1, ...20\}$ and $i = \{1, ...100\}$. We begin by regressing y_{it}^a and y_{it} on x_{it} and z_{it} using fixed effect logit. In the case of y_{it}^a, the dependent variable switches back and forth from 0 to 1 with no clear trend. The fixed effect

[21] Variances a chosen so as to generate a distribution of the dependent variable that resembles that of the paper.

Table 7.1: Monte Carlo results without detrending.

A. y^a_{it} is the dependent variable	E[coef]	σ[coef]	E[t-value]	% significant
coefficient of x_{it}	0.088	0.008	10.95	100%
coefficient of z_{it}	0.000	0.007	-0.04	5%
Number of observations	2000			
B. y_{it} is the dependent variable				
coefficient of x_{it}	0.131	0.032	4.53	100%
coefficient of z_{it}	0.032	0.024	1.37	28%
Average number of usable observations	237			

logit regressor therefore yields consistent coefficient estimates and correct inference. In the case of y_{it}, however, for each i, the sequence of dependent variables ends with a 1. This creates a spurious correlation with any regressor that includes a trend component. As a result, variable x_{it} may erroneously test significant, leading to incorrect inference.

Results are shown in Table A1. The % significant column gives the percentage of Monte Carlo replications in which the coefficient is significantly different from 0 at the 5% level. As anticipated, the fixed effect logit applied to the full data y^a_{it} yields a consistent 0 coefficient for z_{it}. Moreover we see that the z_{it} coefficient is found significant only in 5% of the regressions, a proportion commensurate with the 5% significance level used for the test. In contrast, results for y_{it} yield noticeably different coefficients for z_{it} and x_{it}. Since coefficients estimates for y^a_{it} are consistent, this indicates that the coefficients of both x_{it} and z_{it} are inconsistently estimated by applying fixed effect logit to first collaboration-style data. Moreover, we see that in 28% of the simulations we reject the (correct) null hypothesis that the coefficient of z_{it} is 0. In contrast, when we perform this simulation without trend in x_{it} and z_{it}, results show no bias. The trend element included in the regressors is what generates inconsistent estimates and incorrect inference.

This simple observation suggests that removing the trend in x_{it} and z_{it} may get rid of the problem. Of the reader may worry that detrending the regressors would lose valuable information that is essential to estimation. While this may be true in general, it is not the case here because we are implicitly estimating a fixed effect duration model in which duration

Table 7.2: Monte Carlo results with detrending.

A. y^a_{it} is the dependent variable	E[coef]	σ[coef]	E[t-value]	% significant
coefficient of x^d_{it}	0.085	0.009	9.91	100%
coefficient of z^d_{it}	0.000	0.008	-0.03	5%
Number of observations	2000			
B. y_{it} is the dependent variable				
coefficient of x^d_{it}	0.089	0.025	3.64	98%
coefficient of z^d_{it}	0.000	0.021	0.00	4%
Average number of usable observations	237			

dependence cannot be estimated independently from the fixed effect. Put differently, we cannot estimate the time dependence of the hazard. Consequently, it is intuitively clear that the trend information contained in the regressors provides no information that is useful in identifying coefficients. For this reason, partially out the effect of time is a valid solution to our inconsistent estimation problem. We therefore estimate the following regressions:

$$\begin{aligned} x_{it} &= \gamma_x t + v^x_i + e^x_{it} \\ z_{it} &= \gamma_z t + v^z_i + e^z_{it} \end{aligned}$$

and obtain $x^d_{it} = x_{it} - \widehat{\gamma}_x t$ and $x^d_{it} = x_{it} - \widehat{\gamma}_z t$.

We then regress y_{it} on x^d_{it} and z^d_{it}. If detrending solves the spurious correlation problem, coefficient estimates and inference should be similar to the results obtained in the first panel of Table A1. For the sake of comparison, we also regress y^a_{it} on x^d_{it} and z^d_{it}. Results are presented in Table A2. They show that detrending eliminates the bias in both coefficients in the y_{it} – i.e., first collaboration – regression while keeping things basically unchanged in the y^a_{it} – i.e., repeated collaboration – regression. There is of course a large loss of precision between the y^a_{it} regression and the detrended y_{it} regression, but this is due to the massive loss of observations that results from throwing away all observations of y^a_{it} after the first 1 realization. There is, however, a slight loss of efficiency when applying detrending to the repeated collaboration data.

What these results show is that detrending regressors ensures consistent estimates and correct inference in the first collaboration regression while it still ensure consistent results in the repeated collaboration regression.

References

Aghion, Philippe & Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60(2):323–351.

Allison, Paul D. & Nicholas A. Christakis. 2005. "Fixed Effects Methods for the Analysis of Non-Repeated Events." (mimeograph).

Bala, Venkatesh & Sanjeev Goyal. 1998. "Learning from Neighbors." *Review of Economic Studies* 65(3):595–621.

Bala, Venkatesh & Sanjeev Goyal. 2000. "A Non-Cooperative Model of Network Formation." *Econometrica* 68(5):1181–1229.

Bloch, F., Garance Genicot & Debraj Ray. 2004. "Social Networks and Informal Insurance." (mimeograph).

Chamberlain, Gary A. 1985. Heterogeneity, Omitted Variable Bias, and Duration Dependence. In *Longitudinal Analysis of Labor Market Data*. Cambridge: James J. Heckman and Burton Singer (eds.), Cambridge University Press pp. 3–38.

Dercon, Stefan & Joachim de Weerdt. 2002. Risk-Sharing Networks and Insurance Against Illness. Technical report CSAE Working Paper Series No. 2002-16, Department of Economics, Oxford University Oxford: .

Fafchamps, Marcel. 2002. Spontaneous Market Emergence. In *Topics in Theoretical Economics*. Vol. 2(1), Article 2 Berkeley Electronic Press at www.bepress.com.

Fafchamps, Marcel. 2003. Ethnicity and Networks in African Trade. In *Contributions to Economic Analysis and Policy*. Vol. 2(1) Berkeley Electronic Press at www.bepress.com p. article 14.

Fafchamps, Marcel. 2004. *Market Institutions in Sub-Saharan Africa*. Cambridge, Mass.: MIT Press.

Fafchamps, Marcel & Bart Minten. 2001. "Social Capital and Agricultural Trade." *American Journal of Agricultural Economics* 83(3):680–685.

Fafchamps, Marcel & Bart Minten. 2002. "Returns to Social Network Capital Among Traders." *Oxford Economic Papers* 54:173–206.

Fafchamps, Marcel & Flore Gubert. 2004. *The Formation of Risk Sharing Networks: Evidence from the Philippines*. Oxford: Department of Economics, Oxford University. (in preparation).

Fafchamps, Marcel & Susan Lund. 2003. "Risk Sharing Networks in Rural Philippines." *Journal of Development Economics* 71:261–87.

Fisman, Raymond. 2003. "Ethnic Ties and the Provision of Credit: Relationship-Level Evidence from African Firms." *Advances in Economic Analysis and Policy* 3(1) Article 4.

Genicot, Garance & Debraj Ray. 2003. "Group Formation in Risk-Sharing Arrangements." *Review of Economic Studies* 70(1):87–113.

Goyal, Sanjeev & Marco van der Leij. 2005. "Strong Ties in a Small World." (mimeograph).

Goyal, Sanjeev, Marco van der Leij & Jose Luis Moraga-Gonzalez. 2006. "Economics: An Emerging Small World." *Journal of Political Economy* . (forthcoming).

Granovetter, M. 1985. "Economic Action and Social Structure: The Problem of Embeddedness." *Amer. J. Sociology* 91(3):481–510.

Granovetter, Mark S. 1995. *Getting a Job: A Study of Contacts and Carreers*. Chicago: University of Chicago Press. 2nd edition.

Greif, Avner. 1993. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *Amer. Econ. Rev.* 83(3):525–548.

Greif, Avner. 2001. Impersonal Exchange and the Origin of Markets: From the Community Responsibility System to Individual Legal Responsibility in Pre-modern Europe. In *Communities and Markets in Economic Development*. Oxford: Masahiko Aoki and Yujiro Hayami (eds.), Oxford University Press pp. 1–41.

Gulati, Ranjay. 1998. "Alliances and Networks." *Strategic Management Journal* 19:293–317.

Johnson, Simon, John McMillan & Christopher Woodruff. 2002. "Courts and Relational Contracts." *Journal of Law, Economics, and Organization* 18(1):221–77.

Kranton, Rachel & Deborah Minehart. 2001. "A Theory of Buyer-Seller Networks." *American Economic Review* 91(3):485–508.

Kranton, Rachel E. 1996. "Reciprocal Exchange: A Self-Sustaining System." *Amer. Econ. Rev.* 86(4):830–851.

Lucas, Robert E. 1993. "Making a Miracle." *Econometrica* 61(2):251–272.

McMillan, John & Christopher Woodruff. 1999. "Interfirm Relationships and Informal Credit in Vietnam." *Quarterly Journal of Economics* 114(4):1285–1320.

Mitchell, J. Clyde. 1969. *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*. Manchester: Manchester U. P.

Montgomery, James D. 1991. "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis." *Amer. Econ. Rev.* 81(5):1408–1418.

Munshi, Kaivan. 2003. "Networks in the Modern Economy: Mexican Migrants in the US Labor Market." *Quarterly Journal of Economics* 118(2):549–99.

North, Douglas. 2001. Comments. In *Communities and Markets in Economic Development*. Oxford: Masahiko Aoki and Yujiro Hayami (eds.), Oxford University Press pp. 403–8.

Oyer, Paul. 2005. "The Macro-Foundations of Microeconomics: Initial Labor Market Conditions and Long-Term Outcomes for Economists and MBAs." (mimeograph).

Platteau, Jean-Philippe. 1994. "Behind the Market Stage Where Real Societies Exist: Part II - The Role of Moral Norms." *J. Development Studies* 30(4):753–815.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5) pt.2:S71–102.

Vega-Redondo, Fernando. 2004. "Diffusion, Search, and Play in Complex Social Networks." (mimeograph).

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.

Table 2. New collaborations

Estimator is fixed effect logit on detrended regressors

	All co-authored papers		**Papers coauthored by authors with more than 20 publications**	
	Coef.	z-stat.	Coef.	z-stat.
Network proximity	1.464	**13.16**	1.175	**4.53**
Number of shortest paths (log)	-0.062	-1.61	-0.290	**-3.31**
Network proximity x N. of paths	0.998	**3.36**	2.235	**3.82**
Difference in research output	0.003	**3.27**	0.002	1.59
Average research output	-0.009	**-5.91**	-0.001	-0.57
Field overlap index	0.248	**4.68**	0.050	0.35
Difference in number of coauthors	-0.017	**-2.80**	-0.024	**-2.08**
Average number of coauthors	0.107	**10.45**	0.136	**7.02**
Number of observations	159170		24446	
Number of articles	26601		3103	
Minimum number of years	2		2	
Average number of years	6		7.9	
Maximum number of years	20		20	

Table 3. New collaborations, with affiliation control

Estimator is fixed effect logit on detrended regressors

	All co-authored papers		**Papers coauthored by authors with more than 20 publications**	
	Coef.	z-stat.	Coef.	z-stat.
Network proximity	0.882	**6.36**	0.583	**1.70**
Number of shortest paths (log)	-0.009	-0.20	-0.182	**-1.67**
Network proximity x N. of paths	0.280	0.80	1.264	**1.80**
Difference in research output	0.002	1.35	0.001	0.70
Average research output	-0.006	**-3.11**	-0.001	-0.36
Field overlap index	0.119	**1.65**	0.022	0.10
Difference in number of coauthors	-0.001	-0.16	-0.005	-0.36
Average number of coauthors	0.029	**2.19**	0.033	1.32
Common affiliation	-0.259	**-6.88**	-0.118	-1.20
Number of observations	86964		11949	
Number of articles	17272		1983	
Minimum number of years	2		2	
Average number of years	5		6	
Maximum number of years	12		12	

Table 4. Subsequent collaborations

Estimator is fixed effect logit on detrended regressors

	All co-authored papers		**Papers coauthored by authors with more than 20 publications**	
	Coef.	z-stat.	Coef.	z-stat.
Network proximity	-0.622	**-8.44**	-0.757	**-5.05**
Number of shortest paths (log)	0.178	**4.61**	0.034	0.44
Network proximity x N. of paths	-1.129	**-8.03**	-0.509	**-2.12**
Difference in research output	0.006	**6.51**	0.001	0.81
Average research output	-0.012	**-10.27**	-0.001	-0.33
Field overlap index	-1.518	**-20.36**	-0.761	**-4.44**
Difference in number of coauthors	0.034	**6.64**	0.022	**2.95**
Average number of coauthors	-0.108	**-13.45**	-0.052	**-4.51**
Number of observations	105427		20895	
Number of articles	14449		1860	
Minimum number of years	2		2	
Average number of years	7.3		11.2	
Maximum number of years	20		20	

Table 5. Subsequent collaborations, with affiliation control

Estimator is fixed effect logit on detrended regressors

	All co-authored papers		**Papers coauthored by authors with more than 20 publications**	
	Coef.	z-stat.	Coef.	z-stat.
Network proximity	-1.524	**-9.38**	-0.765	**-2.60**
Dummy whether connected	0.231	**3.30**	-0.125	-0.87
Number of shortest paths (log)	0.105	**1.97**	0.193	**1.89**
Network proximity x N. of paths	-1.116	**-5.99**	-0.997	**-3.13**
Difference in research output	0.007	**6.37**	0.002	1.38
Average research output	-0.014	**-10.24**	-0.003	**-1.65**
Field overlap index	-2.103	**-19.03**	-1.210	**-4.80**
Difference in number of coauthors	0.045	**6.91**	0.023	**2.36**
Average number of coauthors	-0.140	**-12.86**	-0.042	**-2.72**
Common affiliation	-0.220	**-5.73**	0.120	1.54
Number of observations	61336		13372	
Number of articles	9919		1523	
Minimum number of years	2		2	
Average number of years	6.2		8.8	
Maximum number of years	12		12	

Figure1. Referral paths

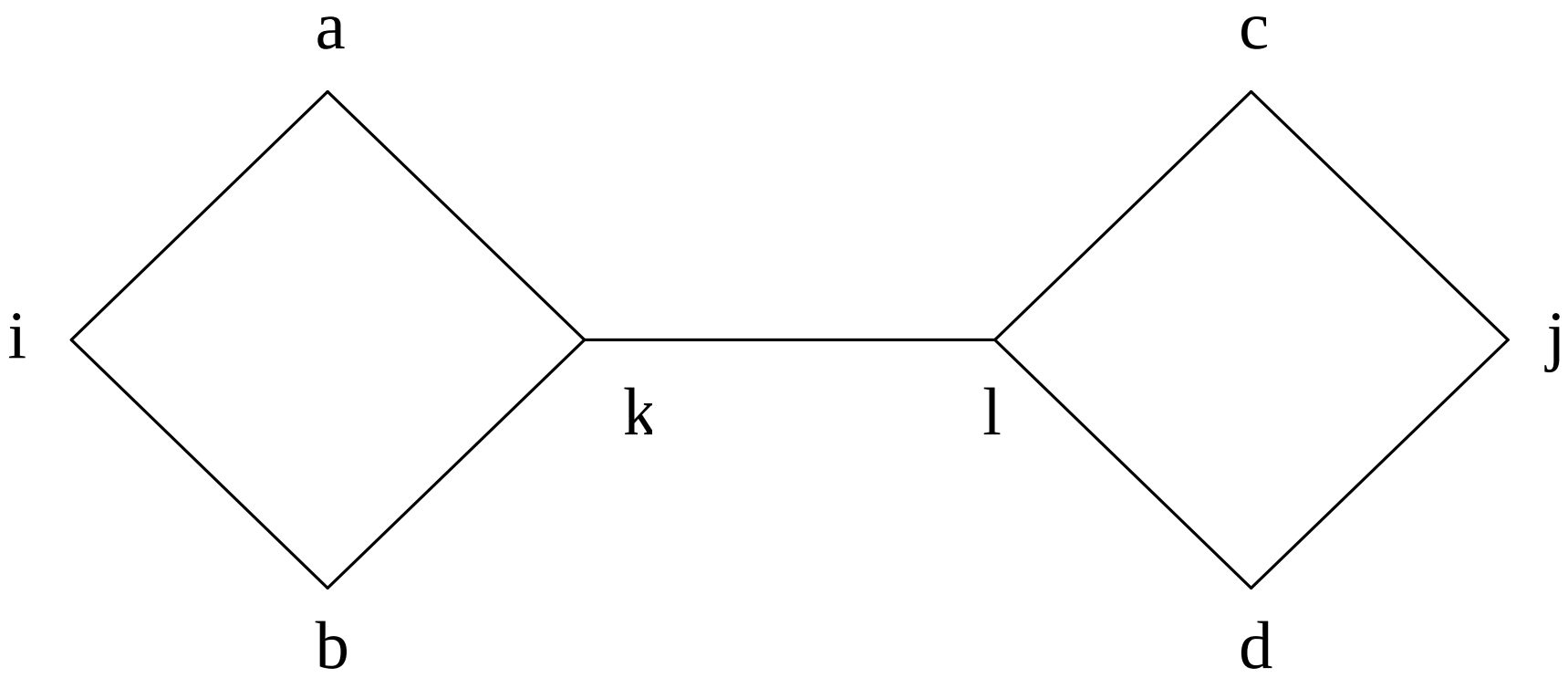


Figure 2. Distance in acquaintance and coauthor networks

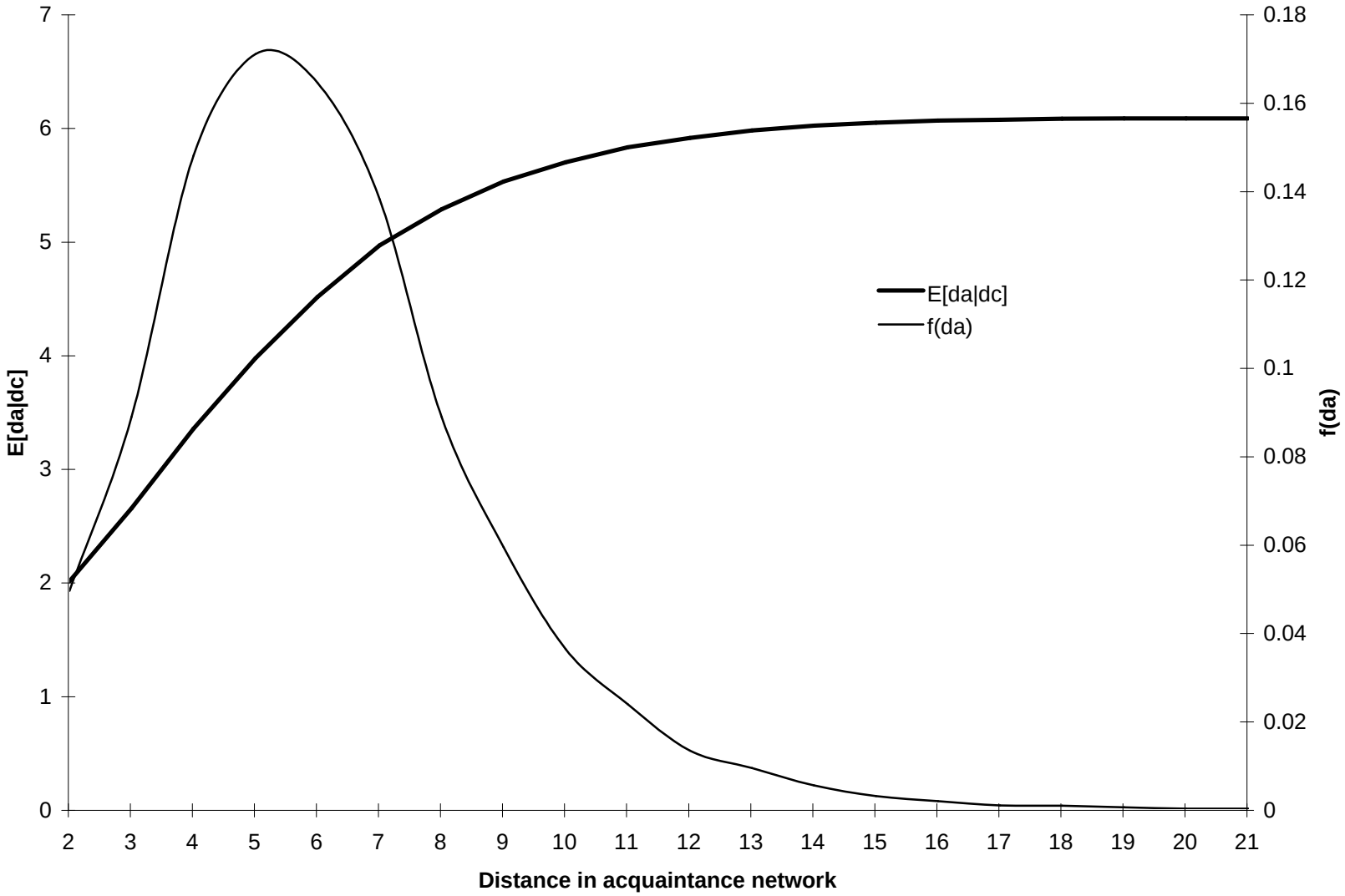


Figure 3. Relative productivity and collaboration

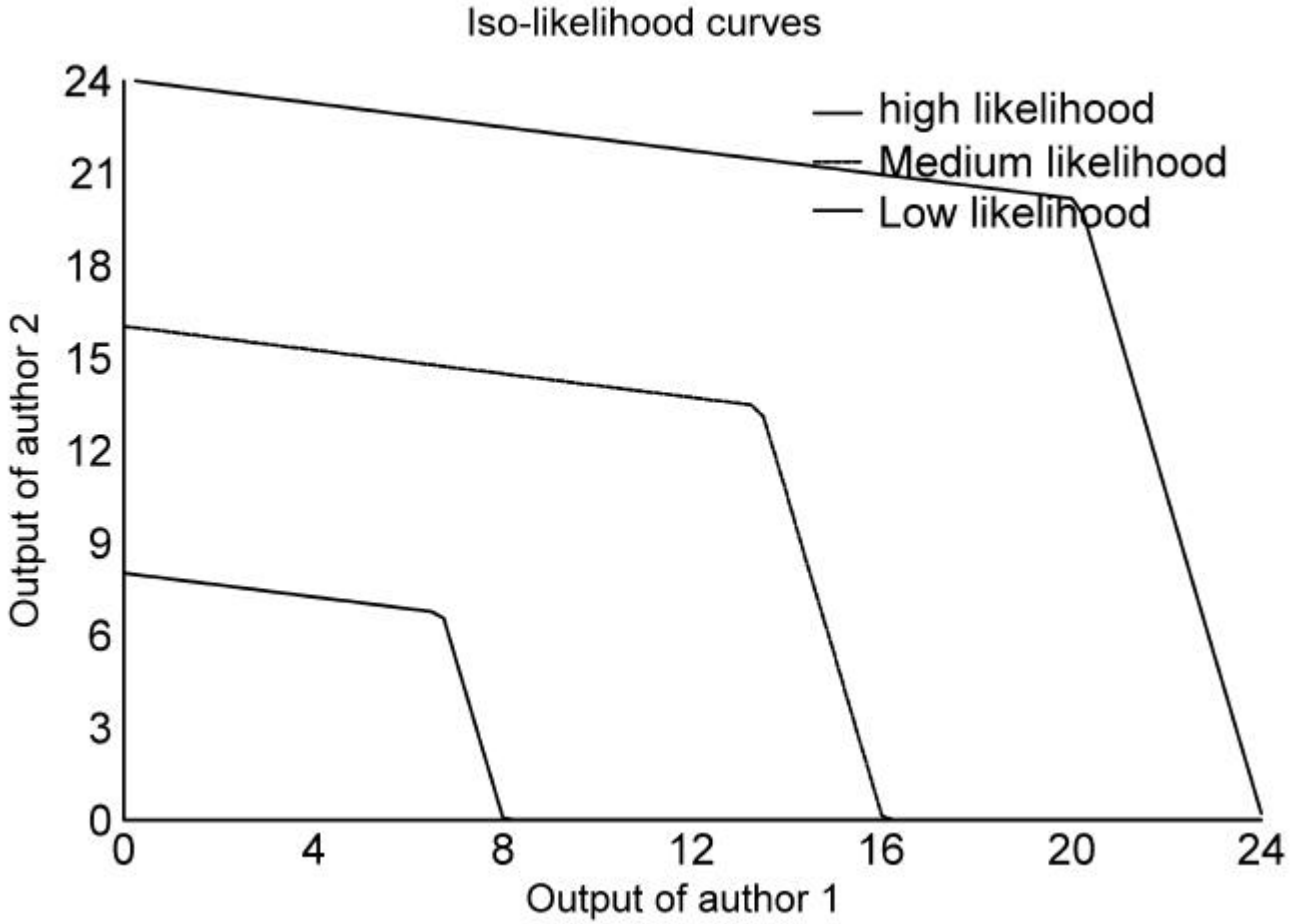
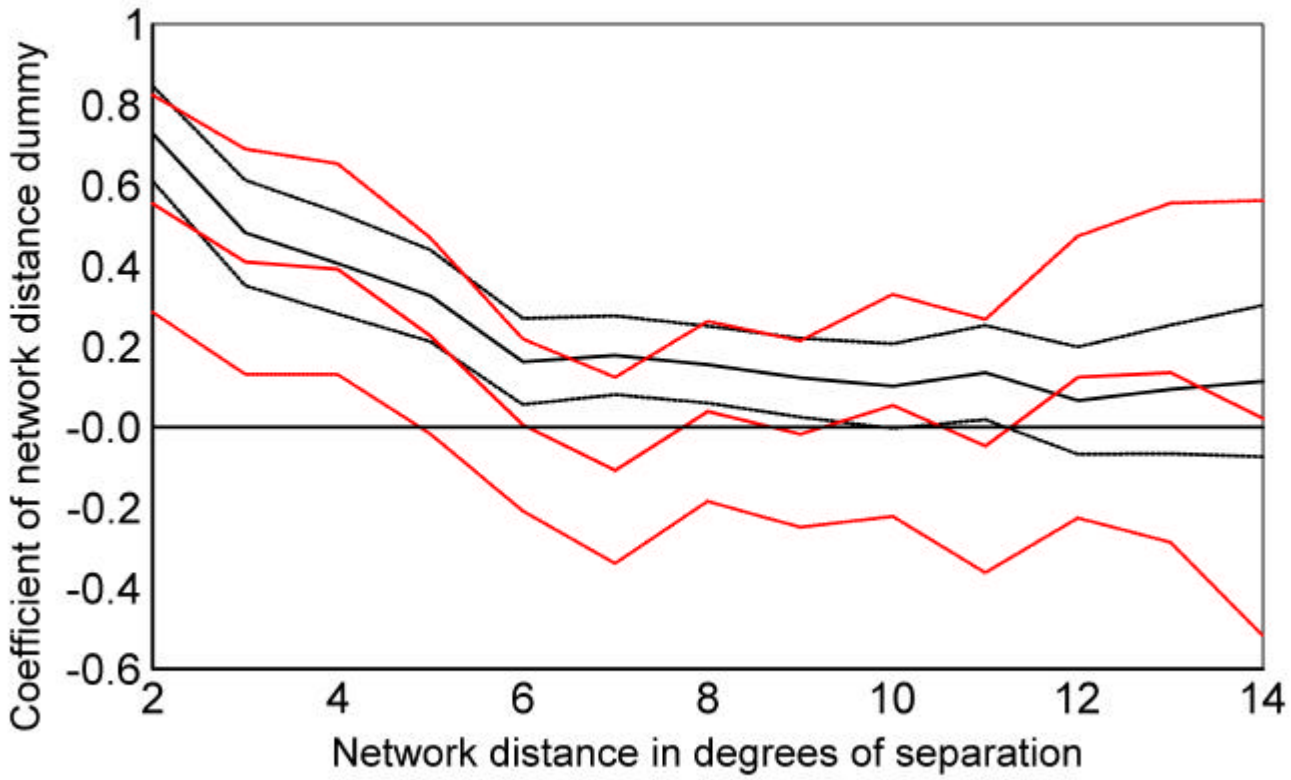


Figure 4. Network distance coefficients

95% confidence interval shown



Full sample black; restricted sample red