Matching and network effects*

Marcel Fafchamps† Sanjeev Goyal‡ Marco J. van der Leij§

September 2006

Abstract

Individuals form teams to produce an output. The quality and quantity of a team's output depends on the ability and the effort level of its members. The ability of individuals as well as their work ethic is however imperfectly known. In such an environment, individuals wishing to form a new team can get information about each other from the network of past collaborators. The paper carries out an empirical examination of the existence and magnitude of this network effect.

This empirical work looks at the formation of coauthor relations among economists over a twenty year period. Our principal finding is that a new collaboration emerges faster among two researchers if they are "closer" in the existing coauthor network among economists. This proximity effect on collaboration is strong and robust but only affects new collaboration: it has no effect on subsequent collaboration among the two individuals. Thus the coauthor network acts as a "referral" mechanism among economists.

*A previous version of this paper was circulated under the title, *Scientific Networks and Coauthorship*. We thank Michele Belot, Jordi Blanes-i-Vidal, Sebi Buhai, Jean Ensminger, Joseph Harrington, Vernon Henderson, Matthew Jackson, Jeff Johnson, Markus Möbius, Tom Snijders, Manuel Trajtenberg, Fernando Vega-Redondo, and seminar participants at Alicante, Bristol, Harvard, Oxford, Utrecht and Saint-Etienne for useful comments. Marco would like to thank the Vereniging Trustfonds Erasmus Universiteit Rotterdam for supporting a visit to Oxford University in May 2005.

†Department of Economics, University of Oxford. Email: marcel.fafchamps@economics.ox.ac.uk.

‡Faculty of Economics, University of Cambridge. E-mail: sgoyal@essex.ac.uk

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

1 Introduction

We consider the following matching environment: there is a group of people who collaborate in teams to produce an output. The quality and quantity of a team's output depends on the knowledge, the ability and the effort level of its members. There is some public information available on the attributes of group members, but this information is incomplete. In such a setting it is reasonable to expect that the structure of personal relations among the members of group will play an important role in conveying useful individual information.

To make this idea more precise, suppose that two individuals i and j are linked if they have been a member of the same team at some point. We can define a network of links based on team membership. We will say that i and j are at distance 2 from each other if they have not worked with each other but have worked with someone in common, k. It is reasonable to suppose that i and j can both get information about each other via k. This information may be conveyed through a number of possible channels. One possibility is that k personally introduces i and j or k talks to i about j and to j about i. Another possibility is that i organizes a professional event, such as a conference, in which he invites a selected group of people who then get an opportunity to spend an extended period of time together (and thereby learn about each other). Information about team members will also be conveyed indirectly; for instance k may discuss i's attributes with j who may then discuss them with one of her team partners l, and so on. Similarly, i and j may meet at a conference and when i organizes an event, she may invite some of her past team members along with j whom she met at an event organized by k. This will facilitate the meeting of team mates of i with team mates of k, and so on. Thus the existing pattern of team memberships can convey valuable information directly as well as indirectly about individuals.

It is reasonable to expect that the quality of information that is being conveyed will decay as it passes – indirectly – through more members of the group. In other words, individuals i and j are likely to know more about each other if they are closer in the network of ties. Since this information is critical to their decision on forming a team, we conjecture that *the probability of i and j forming a team is falling in the distance between i and j in the network of social ties.* The main objective of this paper is to test this conjecture.

We develop this argument in a specific empirical context: the formation of new coauthor teams in economics. We examine data on co-authorship over a 30 year period from 1970 to 1999. Using this data we can construct, for any year, the network of coauthor relations.[1] We then

[1] In particular, for any year, we can map a network based on the stock of coauthored publications in an interval of time prior to that year.

examine how the likelihood of the formation of a *new* coauthor tie between any two individuals is related to their location in this network.

Our principal finding concerns strong proximity effects. We find robust evidence that a new collaboration emerges faster if the two group members are "closer" in the existing network. Moreover, we show that this proximity effect extends quite far in the network; there is a statistically significant effect of proximity until distances up to 11 degrees of separation. The effect of proximity is quite powerful: being at a network distance of 2 instead of 3, raises the probability of initiating a collaboration by 29 percent. Similarly, the probability of two persons forming a link increases by 11% if they are at a network distance of 8 instead of 10.

Perhaps the simplest way to understand the effects of proximity at large distances is in terms of a model of social acquaintances. Information about coauthors is conveyed along via the co-author network; but this information also flows along other social connections which are not co-author ties. For small co-author distance values the co-author distance is a good approximation of the distance between two researchers in the overall social network. As distance in the co-author network grows alternative shorter paths between researchers will emerge and so the co-author network distance becomes a noisier signal of the true social distance. This in turn means that at high distance levels a change in distance has a smaller effect on true social proximity, and hence a smaller effect on the probability of forming new ties. Figure 1 (below) relates co-author network distance to distances in an acquaintance network to illustrate this idea.

The estimation of network effects on the formation of new ties is difficult since the link formation process is driven by a number of individual and pairwise variables which are not observed by us. Unobserved factors such as common language, geographical proximity or the help of an advisor creates effects that might be misunderstood as network referral effects. In our testing strategy, we therefore control for pair-wise fixed effects and rely solely on time-varying changes in the network to identify factors that affect the likelihood of co-authorship. Although this methodology controls for factors such as culture, language, gender and graduate school education, the changes in the network may still reflect time-varying attributes of individuals such as changing productivity, changing research interests, or a change in affiliation. We, therefore, include control variates in our analysis, and we find that the network proximity effect remains significant even after controlling for these time-varying factors.

We interpret the above evidence for network proximity effect as saying that the coauthor network acts as a conduit for valuable information concerning authors skills and abilities. To

substantiate this interpretation, we carry out a regression for the effects of proximity on probability of repeat co-authorship *after* the first coauthored publication. The idea is that if proximity influences first collaboration because of a "referral" effect, then there should be no positive effect of network proximity on subsequent or repeated collaboration. We estimate the probability of repeat probability: our finding is that network proximity does not have a positive effect.

This paper is a study of the role of social networks in economic activity. Sociologists and anthropologists have long incorporated networks in their conceptual toolbox.[2] In recent years a growing and influential body of empirical work in economics argues that individual behavior is shaped by the patterns of social interaction. This body of work makes the following substantive point: a significant part of the variation in behavior across individuals who are faced with similar incentives is due to their being a member of one group rather than another. For example, Glaeser, Sacerdote and Scheinkman (1996) argue that social interaction effects within a geographical neighborhood help explain variations in criminal activity. Bertrand, Luttmer and Mullainathan (2000) show that ethnicity is important in understanding differences in participation in welfare programmes. Banerjee and Munshi (2004) argue that membership of a community influences rates of capital investment, while Duflo and Saez (2003) argue that social interaction affects the choice of pensions policy.[3] The focus of this body of work is therefore on explaining differences across well defined groups.

Our point of departure is the idea that, within a group, individuals are likely to have different patterns of interaction and this may affect their behavior. This motivates the principal innovation of the present paper: we look at differences in social connections *within a group* to understand differences in individual behavior. Moreover, we identify the effects of social networks by explicitly relying on the changes in the network over time. This is the second point of departure from existing work, which looks at contexts in which group membership of individuals does not change.

Traditionally in economics, the formation of production teams has been studied within a search and matching framework. In this framework potential teams mates are anonymous and optimizing actors who are searching for the best match. The essential trade-off here is that greater search is costly in terms of effort and time spent while it generates benefits in terms of a better quality match. This approach has been studied extensively and has generated a number of insights; for recent surveys of this line of research see e.g., Mortenson (2003) and

[2]See, for example, Wasserman and Faust (1994) and the references contained.

[3]More generally, the role of informal institutions, such as social networks, in the functioning of an economy has been the subject of extensive study; see e.g., Granovetter (1985), Greif (2001), Munshi (2003), Munshi and Rosenzweig (2006), North (2001), Fafchamps and Lund (2003) and Fafchamps (2004).

Rogerson, Shimer and Wright (2005). Asymmetric information with regard to skills and effort greatly reinforces the traditional matching and search frictions. It is natural to ask if informal institutions, such as social networks, play an important role in resolving these information problems. Figure 2 (below) illustrates the ideas of this paper in a simple way. In a world where network proximity did not gave a significant effect on the formation of new ties, the probability distribution of distance among new matches should roughly mirror the probability distribution of distances among the authors. However, as Figure 2(b) shows the former distribution places much more probability on close by others as compared the latter distribution. This proximity effect may be due to a number of factors, including unobserved heterogeneity across persons. Our econometric work shows that the pictures in Figure 2 reflect a robust proximity effect on the formation of new ties, after we control for such unobserved heterogeneities. Moreover, this proximity effect is due to information transmission in the informal social network of the coauthors.

Recent papers which examine the information sharing aspects of social networks include Conley and Udry (2000) and Calvó-Armengol et al. (2005). Conley and Udry (2000) study the effects of social communication networks on the individual decision to adopt new crops such as pineapple. Calvó-Armengol et al. (2005) study the effects of location in a network on levels of human capital formation and criminal activity. To the best of our knowledge, the present paper is the first attempt in economics at empirically assessing the importance of information sharing in social networks for the formation of new ties.[4]

Our paper is also related to the emerging body of work on the theory of network formation.[5] In a recent paper Jackson and Rogers (2006) (building on the work of Vázquez, 2003), propose a dynamic model of network formation with the following feature: new players form links at random with a set of existing players and then form a set of links with a subset of the neighbors of their initial connections. Our findings may be seen as providing an empirical foundation to the assumption of local linking which is implicit in their model.[6]

[4]For a study of the formation of inter-firm networks in organization theory and sociology, which shares some common features with our work, see Gulati and Gargiulo (1999). There are a number of methodological differences between the two papers; they use random logit model while we estimate a fixed effects logit model. The findings are also different: they highlight the role of centrality in the network, while our results emphasize the role of network proximity in shaping the probability of new tie creation.
Krishnan and Sciubba (2006) study the formation and productivity of labor sharing teams using data from rural Ethiopia. Their interest is in the role of endowment heterogeneity in shaping the architecture of labor sharing networks. By contrast, the focus of our paper is on the role of information transmission in existing social networks in shaping the formation of new ties.

[5]Early work in this area includes Aumann and Myerson (1988), Bala and Goyal (2000), Jackson and Wolinsky (1996), and Kranton and Minehart (2000).

[6]An early paper by Montgomery (1991) presents a model of the role of referrals in shaping equilibrium wages and inequality in the presence of asymmetric information on worker ability. The present paper provides robust evidence for the existence of such referral effects and we also show that even indirect referrals have significant effects.

The rest of the paper is organized as follows. Section 2 presents our conceptual framework and discusses our testing strategy. Section 3 defines the relevant variables and presents the descriptive statistics, and Section 4 presents the econometric results. Section 5 concludes.

2 Conceptual framework and testing strategy

We wish to understand how scientific collaborations are formed. We start with the observation that researchers come up with ideas, which they want to work out and develop into articles which are published in due course. When researchers have an idea, they can either work it out individually or they can collaborate with someone else and work out the idea jointly. Researchers have some information about each other, based on their knowledge of each other's work and general reputation (which is partly captured in the curriculum vitae). However, there remains significant residual uncertainty about the ability, work ethic, and personal style of individual researchers. 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. A number of mechanisms – formal as well as informal – have been developed to mitigate the effects of asymmetric information. In this paper the interest is in the role of interpersonal relationships as a conduit for valuable information. The present section develops a simple model of information transmission via social networks.

Let us start by noting that it is always possible for a researcher to publish alone. Since collaboration is voluntary, parties to a scientific collaboration must expect more from working together than what they could achieve in isolation. Thus researchers collaborate only if it is in their mutual interest at the time. We postulate that B^{ij}, the benefit to an author i of collaborating with j relative to working alone, is a function of the effort $\{e_i, e_j\}$ and ability $\{q_i, q_j\}$ of i and j, and an idiosyncratic stochastic term ϵ^{ij} that is unknown to i and j. Naturally, given the authors' abilities both authors strategically choose an effort level that maximizes their benefit.

Let $m^{ij} = \Pr[B^{ij} > 0 \text{ and } B^{ji} > 0 | q_i, q_j]$ denote the probability that this benefit is positive for both authors in case i and j have perfect knowledge of each other. The relation of m^{ij} to the ability of i and j is ambiguous. If effort is irrelevant, for example because B^{ij} does not depend on e_i and e_j, then we expect researchers of similar ability to work together. In other words, we expect m^{ij} to decrease with the ability differential $\Delta q^{ij} = |q^i - q^j|$. This is because high ability researchers only tend to collaborate if the partner is herself of high ability. Otherwise a high ability research could as well work on her own. On the other hand, if effort matters

as well, dissimilar matching can arise whereby a researcher with high ability teams up with a less able researcher who provides much of the effort. In that case it is possible that ability differentials Δq^{ij} may increase incentives to collaborate. These relationships are developed in detail in Appendix A.[7]

In this environment researchers have an incentive to overstate their ability in order to attract either high ability or hard working collaborators. Collaborating with someone reveals valuable information about their ability and motivation. A person i who has coauthored with j is thus in a privileged position vis-a-vis person j: she can convey information in a number of different ways. *One*, i may organize an event – such as a research conference – in which she invites j. Such an invitation is an implicit signal about i's assessment of j. Participation in such a research conference will lead j to interact with a number of other researchers, k and l, who have in turn been similarly screened by i (and some other organizers). These meetings in turn may lead to j and k or l to invite each other at other conferences where they meet a different set of (similarly screened) individuals. *Two*, i may personally introduce j to other researchers x and y, and such an introduction will typically involve the communication of this privileged information. *Three*, person i may refer j to some other persons who are looking for someone of high ability. To summarize, there are different routes through which i may convey nuanced and credible information about j to others such as k and l who may then use it to decide whether to collaborate with j. Moreover, k and l may use this information in turn to facilitate a match between j and some other researchers and so on. In what follows, we will think of these different routes of information transmission as mechanisms of 'referral'. In this paper our interest in assessing the affects of referral on the formation of new coauthor ties.

We formalize our ideas of referral in terms of the following simple model. 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 have not. We describe the pattern of coauthorship as a network in which each author is a node and each mutual acquaintance is a link between two nodes. The set of all $i \in S_t$ and coauthor ties l_t^{ij} forms the network G_t. Because authors enter and exit and links are added as a result of joint publication, the network changes over time.

[7]It is important to clarify that in this model, once they have been referred, researchers only care about the quality of the match in deciding whether they will work together or not. Therefore, the only network effect involved is in facilitating referral, and once two authors have been introduced, they do not care about the additional network effects this collaboration will engender. There are two potential types of additional network effects involved here: one, if i works with j, i gets close to a set of people with whom j is connected and this may help in facilitating future link formation. Two, the neighbors of i and j are brought closer via this link between i and j. The first effect pertains to i and j only, while the second effects is a form of positive externality that coauthor links create. We have assumed that individuals ignore such network proximity creation effects when making decisions on new coauthor ties for reasons of tractability.

Consider two authors i and j. Suppose that authors i and j share a common coauthor k. 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:

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

This idea can be generalized with respect to distances and multiple paths. The total probability of observing a collaboration between i and j at t can be written as:

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

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

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 time consuming process for a network as large as the one we are studying. In what follows we therefore consider only the shortest paths between two researchers in the network. Applying this idea we can approximate P_t^{ij} as:

$$P_t^{ij} \approx m_t^{ij} c_t^{ij} b^{d_t^{ij}-1} \tag{3}$$

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.

If coauthorship networks serve to introduce potential coauthors to each other – directly or indirectly – we should observe a relationship of the form depicted by (3). Equation (3) is estimated using logit. It is then useful to derive the logit functional form that best corresponds to (3). The logit regression takes the form:

$$P_t^{ij} = \frac{e^{\beta X_t^{ij}}}{1 + e^{\beta X_t^{ij}}} \tag{4}$$

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 (4) is approximatively equal to:

$$P_t^{ij} \approx e^{\beta X_t^{ij}} = m_t^{ij} c_t^{ij} b^{d_t^{ij}-1} \tag{5}$$

Taking logs, we obtain:

$$\beta X_t^{ij} = -\log b + \log b(d_t^{ij}) + \log c_t^{ij} + \log m_t^{ij} \tag{6}$$

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 (6) predicts that the coefficient of d_t^{ij} is the log of unknown probability b (a negative number since $b < 0$) and the coefficient of $\log c_t^{ij}$ should be 1.

2.1 The acquaintance network

So far, in the formal model, we have assumed that information about author ability and personal attributes travels via coauthor ties only. As the discussion preceding the model suggests, information about coauthor ability and other attributes is likely to circulate more broadly within the personal acquaintance network of researchers. 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 coauthor network but, and this is the important point, the acquaintance network includes the coauthor network since people who have coauthored a paper together know each other.[8]

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 $P \approx m c_a b^{d_a-1}$ 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: $d_a \leq d_c$. It follows that $E[d_a|d_c]$ increases with d_c. In other words, d_c provides information about unknown d_a since the average value of unobserved d_a increases monotonically with observed d_c.

[8]This is a reasonable assumption in economics, where most coauthored papers have 2 or three authors. This may not be a reasonable assumption in other sciences where the number of authors on a single paper can be very large.

This is illustrated with a simple computer experiment, in which we first generate a random 'acquaintance network' of 1000 nodes and 2500 links between randomly chosen pairs of nodes. Figure 1a shows a histogram of this simulated acquaintance network. Next, we randomly select 1000 links from the 'acquaintance network' to obtain a simulation of the 'coauthor network'. As the coauthor network is a subgraph of the acquaintance network, the distance in the acquaintance network between two nodes is bounded from above by the distance in the coauthor network. We then analyze the relation between d_a and d_c in the simulated networks. Figure 1b shows the results.

We indeed observe that $E[d_a|d_c]$ increases monotonically with d_c. Given that there is a monotonic relation between d_c and d_a, we can therefore regard d_c as a valid proxy variable for d_a (Wooldridge, 2002). The requirement is that d_c is not so much above the distribution of d_a that $\partial E[d_a|d_c]/\partial d_c \to 0$. 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.

It is important to 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 (at the lowerbound of $d_c = 1$ we know that $d_a = 1$ as well). 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.

This idea is 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 lowerbound for the number of paths in the acquaintance network. We have already argued 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. This is also confirmed in our simulation. Figure 1c shows the coefficient of a standard linear regression of c_a on c_c for different levels of coauthor distance d_c. Clearly, the relation between c_a and c_c is accurate for low d_c as the coefficient is close to 1, but the relation quickly becomes more obscure when d_c increases.

If, however, referrals only circulate via the coauthorship network, then equation (6) is the correct model and there is no 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 $d_c \times \log c_c$ to equation (6). If the coauthorship network is embedded inside a denser acquaintance network, attenuation bias implies that the coefficient of the interaction term is negative: $\log c_c$ becomes a worse proxy for $\log c_a$ as d_c increases. If referral circulates only in the coauthorship network, then the interaction term is non-significant.

2.2 Testing strategy

Having presented the theoretical framework of our analysis, we now turn to our strategy to test our predictions. Our strategy fully exploits our availability of a panel data set of research collaborations. In short, we pursue the following strategy. *First*, we split up the data in order to perform two regressions. On the one hand, we analyze the probability of initiating a new collaboration. On the other hand, we analyze the probability of continuing an existing collaboration. The purpose of this is explained below. *Second*, we perform fixed effects logit regressions to control for pairwise unobserved heterogeneity that does not vary over time. *Third*, we include important control variates in the regression to take care of pairwise unobserved effects that do change over time.

For the purpose of estimation, we consider a researcher 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. We split this panel data to perform two regressions, one on the probability to initiate a collaboration, and the other on the probability to continue a collaboration.

This strategy is motivated by our interpretation of the proximity effects as driven by 'referrals'. Before the first collaboration authors do not have complete knowledge of each other's abilities. However, two authors that are closer in the coauthorship network will know more about each other and this is the reason for the greater likelihood in the formation of a new tie. On the other hand, 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 information about each other.

Formally, for first collaborations we want to test whether, conditional $y_{t-s}^{ij} = 0$ for all s, the likelihood that $y_t^{ij} = 1$ decreases in d_{t-1}^{ij} and c_{t-1}^{ij}, i.e., whether for first collaborations:

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 0 \text{ for all } s \geq 1) = f(d_{t-1}^{ij}, c_{t-1}^{ij}, m_{t-1}^{ij}) \tag{7}$$

with $\partial f/\partial d < 0$ and $\partial f/\partial c < 0$. If the coefficients have these signs, then we will say that there exist proximity effects in the formation of new coauthor ties.

For subsequent collaborations, we write:[9]

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 1 \text{ for some } s \geq 1) = g(d_{t-1}^{ij}, c_{t-1}^{ij}, m_{t-1}^{ij}) \tag{8}$$

If network effects capture referral, we expect that $\partial g/\partial d = 0$ and $\partial g/\partial c = 0$ since referral is no longer necessary once two researchers have collaborated. Estimating equations (7) and (8) is the objective of this paper.

For estimation of (7) and (8) to yield meaningful inference about network effects, we must control for factors that could create a spurious correlation between y_t^{ij} and d_t^{ij} or c_t^{ij}. Our biggest concern is unobserved heterogeneity. Collaboration depends on many factors that are not observed by us. For example, researchers choose to work together because they share common interests or complementary abilities. Also, researchers tend to collaborate more with colleagues within a department than outside. Unfortunately, many of these factors may lead to spurious 'network effects', unless they are appropriately controlled for.

A simple example clarifies this problem. Suppose that our conjecture of a network proximity effect is false, and instead the only factor that is relevant for collaborating is affiliation. In particular, suppose that the probability of collaborating with a researcher at the same department is p_H, and the probability of collaborating with a researcher at a different department is $p_L << p_H$. Consider i and j being at the same department. If p_H is very high, then i and j are likely to start a collaboration. But both are also likely to have a collaboration with k who is also at the same department. Therefore, there is a high chance that the distance between i and j is very short. On the other hand, if i and j are at different departments then the probability of starting a collaboration is low. Moreover, i is unlikely to have ties with a collaborator of j as well, and therefore the distance between i and j is likely to be high. Viewing the above, we notice a relation between network distance between i and j and their probability to collaborate. However, we have assumed that there is no network effect. This illustrates that not controlling for unobserved heterogeneity may lead to false network effects.

[9]For the regression of subsequent collaboration, network proximity is defined as the distance between two authors in the coauthorship network that ignores their own joint work.

In the presence of unobserved heterogeneity, meaningful inference requires that we control for a pairwise-specific fixed effect μ^{ij}. The models to be estimated are thus of the form:

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 0 \text{ for all } s \geq 1) = f(d_{t-1}^{ij}, c_{t-1}^{ij}, m_{t-1}^{ij}, \mu^{ij}) \quad (9)$$

$$\Pr(y_t^{ij} = 1 | y_{t-s}^{ij} = 1 \text{ for some } s \geq 1) = g(d_{t-1}^{ij}, c_{t-1}^{ij}, m_{t-1}^{ij}, \mu^{ij}) \quad (10)$$

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

We estimate equations (9) and (10) using a fixed effect logit model. However, splitting the data in order to analyze first and subsequent collaboration separately, raises an estimation problem, most notably in the estimation of initiating a collaboration. 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. Therefore, equation (9) is equivalent to a single-spell duration model with fixed effects, except that they are estimated in discrete time. Doing so raises a well known identification problem. It is well known that in single spell duration models, duration dependence and fixed effects cannot be separately estimated.

Nonetheless, it is still possible to identify and consistently estimate coefficients of regressors in the model of (9), as has been shown by Allison & Christakis (2005). However, in order to do so, one needs to assume that these regressors do not contain a trend. Trending regressors mechanically generate spurious correlations between the dependent variable and any regressor that exhibits a time trend. The nature of the problem is illustrated in Appendix B using a Monte Carlo simulation; see also Allison & Christakis (2005).

Unfortunately, almost all our data display trending behaviour. In particular, network distance between i and j tends to become smaller over the years, as both i and j become better connected. The solution we adopt for this problem is to eliminate any time trend in the regressors by de-trending 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 (9) in lieu of the original regressors.[10] In Appendix B we show that this method yields consistent estimates.

[10] We also apply this procedure to model (10) 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.

Control variates. The fixed effects capture many 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. However, research interests, individual network connections, and professional affiliation, are likely to change over time and these changes may be correlated with changes in network distance. Thus the proximity effect on formation of new coauthor ties may actually reflect these changes rather than changes in the transmission of information. This section proposes a number of control variables as a way to address these concerns.

We start by discussing the *propensity to collaborate*. To the extent that this trait is time-invariant, it is captured in the fixed effect. But a researcher's propensity to collaborate may also vary over time: as authors build up coauthoring 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 control for this effect, we calculate the total number of coauthors an author had in the recent past. A researcher who recently had many collaborators is likely to have a higher propensity to collaborate.

We now turn to the relation between network distance and *research overlap*. The distance between researchers in the coauthorship network may reflect proximity in research interests and so changes in the distance may be a measure of how their interests have changed. We control for this possibility by defining an index ω_t^{ij} of research overlap between any two researchers. Recent work on cognitive distance (Wuyts et al., 2005) suggests that this variable affects the probability to collaborate in two ways. On the one hand, the analysis in Appendix A shows that collaboration is only attractive when the researchers involved have complementary knowledge or skills. This suggest that collaboration is unlikely when the research overlap is too large. On the other hand, one must have some common ground in order to collaborate. Hence, research overlap cannot be too small. This suggests an inverted U-curve relation between collaboration and research overlap, and we therefore include a quadratic term $(\omega_t^{ij})^2$ as well.

Finally, we turn to the role of professional affiliations. The propensity to collaborate may increase with common departmental affiliation. If researchers collaborate primarily with departmental colleagues then network proximity may simply capture common affiliation. It is therefore important to control for affiliation.

Even though we include pairwise fixed effects in the regression model as well as some time-varying control variates, one might still be concerned about spurious network proximity effects, in particular because the control variates are difficult to measure and likely to be inprecise. However, we can convincingly close our testing strategy by referring back to the separated analysis of first and subsequent analysis. We conjectured that our network referral mechanism creates a correlation between network proximity and collaboration in the first collaboration regression, but *not* in the subsequent collaboration regression. On the other hand, suppose that referrals are irrelevant, but that some unobserved heterogeneity creates spurious network proximity effects. In that case we would observe a spurious correlation not only in the regression for first collaboration, but also in the regression for subsequent collaboration. This difference is crucial as it allows us to identify a true network referral effect.

3 A description of the data

The data used for this paper come from the Econlit database, a bibliography of journals in economics compiled by the editors of the *Journal of Economic Literature*. From this database we use information on all articles published between 1970 and 1999. We first define all the variables we will use in our study and also describe how we measure them in the context of our data set and then we present descriptive statistics from our data set.[11]

3.1 Definition of variables

We first turn to the definition of the dependent variable. Consider the analysis of initiating a coauthor tie. For this analysis the coauthorship variable y_t^{ij} is defined as follows. Suppose authors i and j coauthor their first paper together in year t_1^{ij}. We create a variable y_t^{ij} that takes value 1 at $t = t_1^{ij}$ and 0 at $t < t_1^{ij}$. 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. To find the last year that both i and j are active, we look in the database for the latest year that i and j separately have a publication, say t_2^i and t_2^j. We then define $t_2^{ij} = \min\{t_2^i, t_2^j\}$. The dependent variable for subsequent collaboration is then defined as $y_t^{ij} = 1$ if i and j co-authored a publication in year t and $y_t^{ij} = 0$ for the period $t : t_1^{ij} < t \leq t_2^{ij}$.

[11] We realize that publication in economics takes place with significant lags. It is indeed not uncommon for a paper to be published in a journal several years after it was first brought out as a working paper. Since all our data comes from published articles, however, the same publication lags affect all variables.

We next consider the definition of the explanatory variables. We construct *network distance* d_t^{ij} as follows. We start by constructing the coauthorship network G_t using authors as nodes and coauthorship as network links and including all publications from year $t-9$ until t. The reason for combining 10 years of publications is that the relation that is formed by coauthoring a paper does not die off instanteneously. As a consequence, we lose the first 10 years of the sample as starting values. Our analysis therefore only considers articles published between 1980 and 1999.[12] Having obtained the coauthorship network, we compute the shortest network distance d_t^{ij} from i to j in G_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 G_t; it is 0 if i and j are unconnected. When computing the distance and the number of shortest paths from i to j, any direct link between i and j is ignored.

If there is no chain of authors 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 sometimes 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.[13] Variable p_t^{ij} is the distance measure used in the estimation of equation (9) and (10).

We turn next to the measure of *productivity*, q_t^{ij}. We would like a system of assessment which takes into account the quality as well as the quantity of work. A standard method for assessing research work combines quality and quantity of work in the following way: there is a quality parameter which is based on the number of citations and a quantity measure which is based on the number of pages. Our aim is to use a simple scheme which captures these ideas and which uses recent citations ranking of journals. These considerations led to use the quality weighting system developed by the Tinbergen Institute Amsterdam-Rotterdam (hereafter TI list) in what follows.[14] This list of journals is used by the Institute to assess the research output of faculty members at 3 Dutch Universities (University of Amsterdam, Erasmus University Rotterdam and Free University Amsterdam). Tenure decisions taken at the Tinbergen Institute are taken based on the number of points a researcher has accumulated. The Institute currently lists 133 journals

[12] 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.

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

[14] The rankings of journals mentioned in Kalaitzidakis et al. (1999) were used as an input in deriving the TI list.

in economics and related fields (econometrics, accounting, marketing, and operations research), of which 113 are covered by EconLit in 2000. This list of journals is split into 3 categories AA, A and B; Appendix C presents the list of these journals. In this system, each article is given a number of points according to the formula:

$$\text{Points} = \frac{\text{Category value} \times \text{Number of pages}}{\text{Number of authors} + 1}$$

The 'category value' is as follows: a journal in category AA yields four points, while a journal in category A yields 2 points and a journal in category B yields 1 point.[15] 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 in year t.

Research output q_t^i is author-specific. To create a pair-specific variable we compute the average productivity of i and j as

$$\overline{q}_t^{ij} \equiv \frac{q_t^i + q_t^j}{2}$$

and the absolute difference in productivity as

$$\Delta q_t^{ij} \equiv \left| q_t^i - q_t^j \right|.$$

We next turn to a number of control variables that are used. We define the *number of coauthors* n_t^i of author i, computed over the ten years preceding time t, and similarly for author j. 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}.

Next we define the research overlap between the research of i and j. Here we use the JEL classification codes contained in the EconLit database to define an index of overlapping interests ω_t^{ij}. We categorize articles into 19 subfields according to the first digit of the JEL codes.[16] If for an article multiple JEL codes are given, then this article is 'divided' and assigned proportionally to the corresponding fields.[17] We then consider the cosine similarity measure as a measure of field overlap between i and j in year t. This measure is computed 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 $\sum_f x_{t,f}^i = 1$). Then

[15] In 1999, 113 out of 687 journals listed in EconLit were classified by the Tinbergen Institute. See `http://www.tinbergen.nl/research/admission.html` for a full description of the TI point system. Journals which are not classified yield zero points.

[16] The JEL classification changed in 1991. For articles before 1991 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.

[17] 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.

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

The cosine similarity measure is a standard measure used by computer scientists to match text documents in, for example, a search engine; see Salton and McGill (1983). It 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.

Finally, we define *common professional affiliations*. The JEL database contains information about author affiliation, but only after 1989 and occasionally in 1988. Moreover the data is spotty and incomplete. It is nevertheless informative to test whether our results are robust to the inclusion of affiliation data.[18]

We construct a common affiliation variable as follows. Let A_t^i be the set of all affiliations of author i that are mentioned in i's articles published in year t. Note that A_t^i is empty if i did not publish in year t or if no affiliations were mentioned. To fill these empty gaps, we define

$$\tilde{A}_t^i = \begin{cases} A_t^i & \text{if } A_t^i \neq \emptyset \\ \tilde{A}_{t-1}^i & \text{if } A_t^i = \emptyset. \end{cases}$$

This definition implicitly assumes that an author's affiliation remains unchanged until information is given to the contrary. Although not fully accurate, this is the best we can do with the available data. The common affiliation variable, a_t^{ij}, is then defined as follows. If both $\tilde{A}_t^i$ and $\tilde{A}_t^j$ are non-empty, then

$$a_t^{ij} = \begin{cases} 1 & \text{if } \tilde{A}_t^i \cap \tilde{A}_t^j \neq \emptyset \\ 0 & \text{if } \tilde{A}_t^i \cap \tilde{A}_t^j = \emptyset. \end{cases}$$

If either $\tilde{A}_t^i$ or $\tilde{A}_t^j$ is missing, then a_t^{ij} is missing as well.

3.2 Descriptive statistics

We have over 25,000 co-author pairs ij in the database. For each of them, we construct a sequence of y_t^{ij} from the time they first publish independently until their first collaboration. This results in over 160,000 observations. For each of these observations, we compute the various explanatory variables as described in the previous section. Summary statistics on these

[18] Affiliation data is recorded as strings and considerable care is needed in 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.

variables are shown in Table 1. The time elapsed from the first publication until ij publish their first joint article is 6 years on average.

Before the first collaboration, the probability that i and j are directly or indirectly connected via the coauthorship network is 42%. If connected, the average network distance d_t^{ij} between them is around 7. This is a remarkably short distance. For instance, Goyal et al. (2006) find, over the same time period, that the average degree of separation between all the connected pairs of authors in the Econlit database is between 9 and 12. Network distance is thus smaller for pairs that eventually start a collaboration, suggesting that collaboration is associated with 'closeness' in the network.

To illustrate this further, we plot in Figure 2 the histogram of network distances in the entire author network and compare it with that of network distances for collaborating pairs.[19] It is clear that coauthors are on average much closer to each other than pairs of authors taken at random. Dividing one set of frequencies by the other yields a non-parametric measure of the probability of an ij tie conditional on network distance. The result of this calculation, displayed in the last panel of Figure 2, shows a clear monotonic decline with distance.

Productivity and connectedness variables are shown next in Table 1. We see that the productivity differential is quite large relative to average productivity. This indicates that, at the time of first collaboration, authors differ widely in terms of productivity. This can be interpreted as prima facie evidence of dissimilar matching.

Statistics on field overlap and common affiliation appear next in Table 1. The field overlap index ω_t^{ij} is around 50%, indicating that economists typically collaborate with someone in their field. Close to 40% of coauthor pairs had a common affiliation prior to their first collaboration.

Similar statistics are reported in Table 2 for subsequent collaborations. We have close to 15,000 coauthor pairs who published together in more than one year. For each of these pairs we construct a sequence of y_t^{ij} from the year following their first joint publication until the year of the last publication. This gives a little over 105,000 observations. We see from Table 2 that once a collaboration has been successfully initiated, it tends to be repeated: conditional on publishing more than once together, on average a pair of authors publishes jointly in one year out of four. If we compare Table 2 with Table 1 we see that authors who continue collaborating tend to get closer in the author network. Field overlap increases as well.

[19] To ensure comparability, we use the author network from 1980 to 1989 and define collaborating pairs as those who start a collaboration in 1990.

4 Econometric results

4.1 The role of network proximity

We now present the econometric estimation of the models presented in Section 2. We begin with equation (9) which analyzes the determinants of the first collaboration between a pair of researchers. The basic regression model is of the form:

$$\Pr(y_t^{ij} = 1) = \lambda(\alpha + \beta p_{t-1}^{ij} + \gamma_1 \log c_{t-1}^{ij} + \theta_1 \overline{q}_{t-1}^{ij} + \theta_2 \Delta q_{t-1}^{ij} + \mu^{ij}). \tag{11}$$

where $\lambda(.)$ denotes the logit function. Equation (11) is estimated using conditional logit to eliminate the fixed effect μ^{ij}. As detailed in Appendix B, all regressors are detrended to eliminate spurious correlation with the dependent variable.[20]

Results are reported in column (1) of Table 3. The results show a strong positive effect of network proximity p_t^{ij}: the magnitude of the coefficient is large and the z-statistic is highly significant.[21] This suggests that network proximity plays an important role in research collaborations.

We wish to ascertain whether this result is driven by a local effect over very short network distances, or whether it is a more diffuse effect extending to long network distances, as predicted by our model. To investigate this idea, we replace p_t^{ij} with network distance dummies and re-estimate model (11). Coefficient estimates for distance dummies are presented in Figure 3; the dashed lines delineate the 95% confidence interval.

Results shows that network effects are not limited to short distances: distance dummies have a significantly negative coefficient for distances up to 11 degrees of separation. They also indicate that the quantitative impact impact of proximity on the probability of coauthorship is large. Using the approximation of the logit equation given in (5), we see that being at a network distance of 2 instead of 3, raises the probability of initiating a collaboration by 29 percent. The effect remains noticeable at larger distances. For example, being connected at a network distance of 8 instead of 10 implies that the probability of forming a link is 11 percent higher.

At first glance these striking findings appear too good to be true: the likelihood that two authors be introduced to each other by a chain of 11 co-authors appears fairly remote. Perhaps the best way to make sense of these findings is in terms of the social acquaintance network model presented in section 2.1. Information about coauthors is conveyed via the coauthorship

[20] It is essential to detrend regressors using only observations entering in the estimation of (13). So detrending is redone each time the inclusion of a new regressor, such as common affiliation, results in a loss of valid observations.

[21] At short distances, the implied value of b is around 0.71.

network. But information also flows along other social connections which are not coauthor ties – and hence are not observed. At short distances, distance in the coauthor network d_c is a good approximation of distance in the acquaintance network d_a. But as d_c increases, the likelihood rises that a shorter paths exists in the acquaintance network. This implies that, as d_c rises, it becomes a noisier measure of true social distance.

This point is illustrated in Figure 1b where we plotted $E[d_a|d_c]$ as a function of d_c. The Figure highlights two main points. First, d_c reveals information about d_a even when d_c gets large, but beyond a certain level, d_c no longer has any effect on $E[d_a|d_c]$. This is also what we observe in the data: beyond distance 11, d_c no longer has a significant effect on the likelihood of coauthorship. This suggests that the distribution of d_a is concentrated on values from 2 to 11.

Secondly, $E[d_a|d_c]$ can remain relatively small even for large significant values of d_c. This means that a significant coefficient on the distance 11 dummy does *not* imply that unbroken chains of 11 coauthors were customarily used to introduce two authors to each other. In the acquaintance network, distance between the two authors was in all likelihood much shorter than $d_c = 11$. This second point is crucial because it explains why distance dummy 11 can be significant without implying that unrealistically long chains of referral are used to bring authors together.

To confirm this interpretation, we look for indirect evidence of the existence of an acquaintance network. In Section 2.1, we have argued that one way to test for this is to introduce an interaction term $p \times \log c$, i.e., proximity times the number of shortest paths. If there is no acquaintance network, the number of shortest paths in the coauthorship network is measured accurately even at large network distances. But if there is an acquaintance network, $\log c$ becomes an increasingly inaccurate proxy for the number of shortest paths in the acquaintance network. This leads us to estimate the following regression.

$$\begin{aligned} \Pr(y_t^{ij} &= 1) = \lambda(\alpha + \beta p_{t-1}^{ij} + \gamma_1 \log c_{t-1}^{ij} + \gamma_2 p_{t-1}^{ij} \log c_{t-1}^{ij} \\ &\quad + \theta_1 \bar{q}_{t-1}^{ij} + \theta_2 \Delta q_{t-1}^{ij} + \mu^{ij}) \end{aligned} \tag{12}$$

If referral takes place through an acquaintance network, then $\gamma_2 > 0$. Results are shown in column (2) of table 3. We see that the coefficient γ_2 of the interaction term is positive and significant. This suggests that coauthorship referrals circulate in an unobserved acquaintance network that is denser than the observed coauthorship network.

4.2 Changing research interests and affiliation

We have seen that, controlling for pairwise fixed effects and for changes in research productivity, the probability that two researchers i and j start publishing together is related to their proximity in the existing coauthorship network. We have interpreted this as implying that the coauthorship network acts a conduit for valuable information concerning researchers attributes. It is conceivable, however, that changes in network distance in fact proxy for other unobservable factors such as changes in research interest, affiliation, or propensity to coauthor. To control for this possibility, we estimate an expanded regression model:

$$\Pr(y_t^{ij} = 1) = f(\beta p_{t-1}^{ij} + \gamma_1 \log c_{t-1}^{ij} + \gamma_2 p_{t-1}^{ij} \log c_{t-1}^{ij} + \theta_1 \overline{q}_{t-1}^{ij} + \theta_2 \Delta q_{t-1}^{ij} + \lambda z_{t-1}^{ij} + \mu^{ij}) \quad (13)$$

where z_t^{ij} stands for additional controls such as $\overline{n}_t^{ij}$, Δn_t^{ij}, ω_t^{ij} and $(\omega_t^{ij})^2$. Affiliation is introduced later. Estimation results are presented in column (3) of Table 3. They show that our main results are unaffected by the inclusion of these additional controls: network proximity remains significant, and so is the cross term $p \times \log c$.

The coefficients of control variables are interesting in their own right. Average productivity appears with a significant negative sign: researchers are more likely to coauthor when their productivity is flagging. We also see that the coefficient of Δq_t^{ij} is positive and significant, indicating that two authors are more likely to collaborate when their productivity differs. This is suggestive of dissimilar matching.

As anticipated, the likelihood of initiating collaboration increases with the average number of coauthors $\overline{n}_t^{ij}$: the more coauthors researchers have, the more likely they are to collaborate with each other. The difference in the number of collaborators has a negative sign and is significant as well. We expected the effect of field overlap on coauthorship to follow an inverted-U curve: too little overlap and the authors have nothing in common; too much overlap and they do not complement each other. This is indeed what we find: the coefficient of the field overlap index ω_{t-1}^{ij} is significantly positive, whereas the coefficient of the quadratic term $(\omega_{t-1}^{ij})^2$ is significantly negative. The likelihood of forming a collaboration is highest when the field overlap index is .657.

Finally we turn to the effect of common 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 coauthorship. To investigate this possibility, we re-estimate the model with the joint affiliation variable a_t^{ij}. As pointed out

earlier, affiliation information is only available after 1988. This means that including a_t^{ij} in the regression leads to a massive loss of observations – more than half.

Since we lose so many observations, we first repeat the estimation of models (11), (12) and (13) for this smaller sample before including a_t^{ij}. The results, presented in Table 4, show that the reduction in the number of observations weakens the results markedly. Most variables are not significant anymore, except for network proximity which remains positive and significant. Contrary to expectations, common affiliation has a negative effect on the likelihood of initiating a first collaboration. When interpreting this finding, it is important to remember that our estimation controls for pairwise fixed effects. This means that identification is obtained only from pairs of authors who sometimes have a common affiliation, and sometimes not, either because they joined the same university or because they moved away from each other. A negative sign means that researchers are more likely to start collaborating after moving apart. A possible explanation is the following: there are significant lags in economics and so it is quite possible for productive authors to start work when they have a common affiliation but that this works comes out in print after they move on to new affiliations.

4.3 Subsequent collaboration

We have found that proximity in the existing co-authorship network shapes the formation of new coauthor ties. We have interpreted this finding as evidence that the co-authorship network acts as a conduit for valuable information concerning authors skills and abilities. We now test the validity of this interpretation as follows.

As discussed in section 2.2, if proximity influences collaboration because of a "referral" effect, we should expect network proximity *not* to be significant for subsequent collaborations. This is because once two researchers have worked together, they no longer need a referral for each other. It follows that if we regress repeat co-authorship on network proximity, we should observe no effect. In contrast, if proximity influences collaboration because of, say, an inherent in-breeding bias, then it should remain significant for subsequent collaborations as well. The same observation applies if network proximity proxies for unobservable effects that happen to be correlated with network proximity.

To investigate these ideas, we re-estimate equations (11), (12) and (13) using data on subsequent collaborations. Results are summarized in Table 5. We find that network proximity no longer has a positive effect on co-authorship. This finding is consistent with the referral interpretation. It also provides further reassurance that the positive network effect on first collaboration

is unlikely to be the result of omitted variable bias. Indeed, this would require that the omitted variable only affects the likelihood of first collaboration, something we find improbable.

However, Table 5 says that that network proximity now has a significant but *negative* coefficient. To investigate why this is the case, we re-estimate model (13) with distance dummies instead of p_t^{ij}. Results, presented in Figure 4, show that the only negative and significant dummy is for distance 2 – that is, for authors who have one or several common coauthors; other distance dummies are not significant. Our explanation for this finding is in terms of time/capacity constraints of authors.[22]

Recall, that in our framework proximity effects are identified by changes in the levels of proximity. Prior to first collaboration, note that as two authors move closer to a distance of 2, there are two effects at work: one, they are likely to get better information about each other, and two, at least one of them have started a new collaboration and this means that they have less time for an additional new collaboration. The former effect has a positive influence while the latter has a negative effect on the probability of formation of a new tie. Our results on first collaboration suggest that the positive effect dominates. Once i and j have collaborated, there are no informational advantages to be gained from proximity and so the negative effect prevails dampening the probability of repeat collaboration.[23]

5 Conclusions

We have examined an environment in which individuals form teams to produce an output. The quality and quantity of a team's output depends on the ability and the effort level of its members, but individual ability and work ethic are imperfectly known. In such an environment, individuals wishing to form a new team can get information about each other from the network of past collaborators. The paper investigated the existence and magnitude of this network effect.

This investigation was carried out within a specific empirical context: the formation of coauthor relations among economists over a twenty year period from 1980 to 2000. Our principal finding is that a new collaboration emerges faster among two researchers if they are "closer" in the existing coauthor network. This proximity effect is positive and robust and extends quite far in the network. Moreover, we found that network proximity does not have such a positive effect on prospects of repeat collaboration.

[22] For a model of co-author network formation in which capacity constraints play an important role, see Jackson and Wolinsky (1996).

[23] This explanation is consistent with the absence of significant effects at longer distances. A fall in distance to 3 or 4 does not necessarily involve an additional link by either i or j and so the capacity constraint effects are indirect and weaker.

Our interpretation of these findings is that, in an environment characterized by pervasive asymmetric information on individual ability and working styles, informal institutions, such as social networks, convey valuable information which facilitates the formation of new collaboration ties. The coauthor network among economists provides us with information on the actual social network and this explains why on the one hand, coauthor network proximity effects extend quite far into the network and on the other hand, it also explains why this effect tapers off at high network distances.

In recent years, the effects of social networks on economic outcomes has been the subject of a large and influential literature. This literature has focussed on explaining how differences in outcomes *across* groups may be an outcome of group level social interaction effects. The present paper departs from this tradition and the concern here has been in showing how network differences *within* a group help explain differences in individual behavior. The formation of production teams is a central aspect of the functioning of an economy and we hope that our findings will motivate further empirical study of the role of referral networks in labor markets.

Appendix A

This appendix develops a simple model of scientific collaboration to explore the relative role of ability and effort sharing in determining incentives for 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 a strictly continuous increasing function in all its arguments and concave in e_i and e_j.

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 a joint project with j to solve:

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

which yields a first order condition of the form:[24]

$$k\frac{\partial r(e_i, e_j, a_i, a_j)}{\partial e_i} = \frac{\partial \bar{R}(E - e_i, a_i)}{\partial e_i} \tag{14}$$

Equation (14) 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^*.

[24]We assume that a solution exists and that it is unique and interior.

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 $r = (a_i + a_j)$ and $\overline{R} = a_i$. 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 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 competence (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 $(a_i + a_{ij})\bar{R}(e_i)$ is the return from allocating effort $e_i \in [0, E]$ to research with others, with $\bar{R}'' < 0$. 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:

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

from which we obtain:

$$\frac{\bar{R}'(E - e_i)}{\bar{R}'(E - e_j)} = \frac{a_{ij}}{a_i + a_{ij}} \tag{15}$$

Equation (15) shows that the marginal return to allocating effort to alternative projects is lower for i than for j. This implies that $E - e_i > E - e_j$, and hence $e_i < e_j$ since $\bar{R}'' < 0$ 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 there exists an effort level $e_i \in [0, E]$, such that

$$\begin{aligned} k(a_i + a_{ij})r(e_i + e_j) + (a_i + a_{ij})\bar{R}(E - e_i) &> (a_i + a_{ij})\bar{R}(E) \\ \Leftrightarrow kr(e_i + e_j) + \bar{R}(E - e_i) &> \bar{R}(E) \end{aligned} \tag{16}$$

The low ability author prefers to collaborate so long as there exists an effort level $e_j \in [0, E]$, such that

$$k(a_i + a_{ij})r(e_i + e_j) + (a_{ij})\bar{R}(E - e_j) > (a_{ij})\bar{R}(E) \tag{17}$$

It is immediately clear that as long as $e_j > 0$ condition (16) 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_{ij}. Furthermore, from the first order condition (15) we see that e_j increases in a_i, and so from equations (16)-(17) 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.

We summarize our findings as follows. 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.

Appendix B

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 $u_i \sim N(0, 5)$.[25] 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:

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

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

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:

$$y_{it}^{*} = -2 + x_{it} + u_i + \varepsilon_{it} \tag{18}$$

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 (18), 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 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 6. 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_{it}^{a} 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_{it}^{a} 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} should get rid of the problem. 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 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, partialling out the effect of time is a valid solution to our inconsistent estimation problem.

To test that this is indeed the case, we estimate the following regressions:

$$\begin{aligned} x_{it} &= \gamma_x t + v_i^x + e_{it}^x \\ z_{it} &= \gamma_z t + v_i^z + e_{it}^z \end{aligned}$$

and obtain $x_{it}^d = x_{it} - \widehat{\gamma}_x t$ and $x_{it}^d = x_{it} - \widehat{\gamma}_z t$. We then regress y_{it} on x_{it}^d and z_{it}^d. If detrending solves the spurious correlation problem, coefficient estimates and inference should be similar to the results obtained in the first panel of Table 6. For the sake of comparison, we also regress y_{it}^a on x_{it}^d and z_{it}^d.

Results are presented in Table 7. 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_{it}^a – i.e., repeated collaboration – regression. There is of course a large loss of precision between the y_{it}^a 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_{it}^a after the first 1 realization. There is also 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.

Appendix C

The Tinbergen Institute List of Journals:

Journals (AA): 1. American Economic Review 2. Econometrica 3. Journal of Political Economy 4. Quarterly Journal of Economics 5. Review of Economic Studies

Journals (A): 1. Accounting Review 2. Econometric Theory 3. Economic Journal 4. European Economic Review 5. Games and Economic Behavior 6. International Economic Review 7. Journal of Accounting and Economics 8. Journal of Business and Economic Statistics 9. Journal of Econometrics 10. Journal of Economic Literature 11. Journal of Economic Perspectives 12. Journal of Economic Theory 13. Journal of Environmental Economics and Management 14. Journal of Finance 15. Journal of Financial Economics 16. Journal of Health Economics 17. Journal of Human Resources 18. Journal of International Economics 19. Journal of Labor Economics 20. Journal of Marketing Research 21. Journal of Monetary Economics 22. Journal of Public Economics 23. Management Science(*) 24. Mathematics of Operations Research (*) 25. Operations Research (*) 26. Rand Journal of Economics / Bell Journal of Economics 27. Review of Economics and Statistics 28. Review of Financial Studies 29. World Bank Economic Review.

Journals (B): 1. Accounting and Business Research(*) 2. Accounting, Organizations and Society(*) 3. American Journal of Agricultural Economics 4. Applied Economics 5. Cambridge Journal of Economics 6. Canadian Journal of Economics 7. Contemporary Accounting Research(*) 8. Contemporary Economic Policy 9. Ecological Economics 10. Economic Development and Cultural Change 11. Economic Geography 12. Economic History Review 13. Economic Inquiry / Western Economic Journal 14. Economics Letters 15. Economic Policy 16. Economic Record 17. Economic Theory 18. Economica 19. Economics and Philosophy 20. Economist 21. Energy Economics 22. Environment and Planning A 23. Environmental and Resource Economics 24. European Journal of Operational Research(*) 25. Europe-Asia Studies(*) 26. Explorations in Economic History 27. Financial Management 28. Health Economics 29. Industrial and Labor Relations Review 30. Insurance: Mathematics and Economics 31. Interfaces(*) 32. International Journal of Forecasting 33. International Journal of Game Theory 34. International Journal of Industrial Organization 35. International Journal of Research in Marketing(*) 36. International Monetary Fund Staff Papers 37. International Review of Law and Economics 38. International Tax and Public Finance 39. Journal of Accounting Literature(*) 40. Journal of Accounting Research 41. Journal of Applied Econometrics 42. Journal of Applied Economics 43. Journal of Banking and Finance 44. Journal of Business 45. Journal of Comparative Economics 46. Journal of Development Economics 47. Journal of Economic Behavior and Organization 48. Journal of Economic Dynamics and Control 49. Journal of Economic History 50. Journal of Economic Issues 51. Journal of Economic Psychology 52. Journal of Economics and Management Strategy 53. Journal of Evolutionary Economics 54. Journal of Financial and Quantitative Analysis 55. Journal of Financial Intermediation 56. Journal of Forecasting 57. Journal of Industrial Economics 58. Journal of Institutional and Theoretical Economics / Zeitschrift für die gesamte Staatswissenschaft 59. Journal of International Money and Finance 60. Journal of Law and Economics 61. Journal of Law, Economics and Organization 62. Journal of Macroeconomics 63. Journal of Mathematical Economics 64. Journal of Money, Credit and Banking 65. Journal of Population Economics 66. Journal of Post-Keynesian Economics 67. Journal of Risk and Uncertainty 68. Journal of the Operations Research Society(*) 69. Journal of Transport Economics and Policy 70. Journal of Urban Economics 71. Kyklos 72. Land Economics 73. Macroeconomic Dynamics 74. Marketing Science 75. Mathematical Finance 76. National Tax Journal 77. Operations Research Letters(*) 78. Organizational Behavior and Human Decision Processes(*) 79. Oxford Bulletin of Economics and Statistics / Bulletin of the Institute of Economics and Statistics 80. Oxford Economic Papers 81. Oxford Review of Economic Policy 82. Probability in the Engineering and Informational Sciences(*) 83. Public Choice 84. Queuing Systems(*) 85. Regional Science and Urban Economics 86 Reliability Engineering & System Safety(*) 87. Resource and Energy Economics / Resource and Energy 88. Review of

Income and Wealth 89. Scandanavian Journal of Economics / Swedish Journal of Economics 90. Scottish Journal of Political Economy 91. Small Business Economics 92. Social Choice and Welfare 93. Southern Economic Journal 94. Theory and Decision 95. Transportation Research B - Methodological 96. Transportation Science(*) 97. Weltwirtschaftliches Archiv / Review of World Economics 98. World Development 99. World Economy

(*) Journal not covered by EconLit

References

[1] Allison, P. and Christakis, A. (2005), Fixed effects methods for the analysis of non-repeated events, *Sociological Methodology*, forthcoming.

[2] Aumann, R. and R. Myerson (1988), Endogenous formation of links between players and coalitions: an application of the Shapley Value, in A.Roth, (ed), *The Shapley Value*, Cambridge University Press.

[3] Bala, V. and S. Goyal (2000), A non-cooperative model of network formation, *Econometrica*, 68, 1181-1230.

[4] Banerjee, A. and K. Munshi (2004), How efficiently is capital allocated? evidence from the knitted garment industry in Tirupur. *Review of Economic Studies*, 71, 1, 19-42.

[5] Bertrand, M., E. Luttmer, and S. Mullainathan (2000), Network effects and welfare cultures, *Quarterly Journal of Economics*, 115, 1019-1055.

[6] Calvó-Armengol, A., E. Patacchini and Y. Zenou (2005), Peer effects and social networks in education and crime, *mimeo*, Universitat Autònoma de Barcelona.

[7] Chamberlain, G. (1985), Heterogeneity, omitted variable bias, and duration dependence. In J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Cambridge.

[8] Conley, T. and C. Udry (2000), Learning about a new technology: Pineapple in Ghana, *Working Paper 817*, Economic Growth Center, Yale University.

[9] Duflo, E. and E. Saez (2003), The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment, *Quarterly Journal of Economics*, 118, 3, 815-842.

[10] Fafchamps, M. (2004), *Market institutions in Sub-Saharan Africa*. MIT Press, Cambridge, Mass.

[11] Fafchamps, M. and S. Lund (2003), Risk sharing networks in rural Philippines, *Journal of Development Economics*, 71, 261-287.

[12] Glaeser, E., B. Sacerdote and J. Scheinkman (1996), Crime and Social Interactions, *Quarterly Journal of Economics*, 111, 507-548.

[13] Goyal, S., M.J. van der Leij and J.L. Moraga (2006), Economics: an emerging small world, *Journal of Political Economy*, 114, 2, 403-412.

[14] Granovetter, M. (1985), Economic Action and Social Structure: The Problem of Embeddedness, *American Journal of Sociology,* 3, 481-510.

[15] Greif, A. (2001), Impersonal exchange and the origin of markets: From the community responsibility system to individual legal responsibility in pre-modern Europe. In M. Aoki and Y. Hayami (eds). *Communities and Markets in Economic Development.* Oxford University Press. Oxford.

[16] Gulati, R. and M. Gargiulo (1999), Where do interorganizational networks come from, *American Journal of Sociology*, 104, 5, 1439-1493.

[17] Jackson, M. and B. Rogers (2006), How random are social networks?, *mimeo*, Caltech.

[18] Jackson, M. and A. Wolinsky (1996), A Strategic Model of Economic and Social Networks, *Journal of Economic Theory*, 71, 1, 44-74.

[19] Kalaitzidakis, P., T. Mamuneas, and T. Stengos (2003), Rankings of academic journals and institutions in economics, *Journal of European Economic Association*, 1346-1366.

[20] Kranton, R. and D. Minehart (2001), A theory of buyer-seller networks, *American Economic Review*, 91, 3, 485-508.

[21] Krishnan, P., and Sciubba, E. (2006), Links and Architecture in Village Networks, *Working Paper*, University of Cambridge and Birkbeck College, London.

[22] Montgomery, J. (1991), Social networks and labor-market outcomes: toward an economic analysis, *American Economic Review*, 81, 5, 1408-1418.

[23] Mortenson, D. (2003), *Wage Dispersion: Why are similar workers paid differently?* MIT Press. Cambridge. MA.

[24] Munshi, K. (2003), Networks in the modern economy: Mexican migrants in the U. S. labor market, *Quarterly Journal of Economics*, 118, 2, 549-599.

[25] Munshi, K. and M. Rosenzweig (2006), Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy, *American Economic Review*, forthcoming.

[26] North, D. (2001), Comments. In *Communities and markets in economic development*, edited by M. Aoki and Y. Hayami,. Oxford University Press. Oxford.

[27] Rogerson, R., R. Shimer, and R. Wright (2005), Search-Theoretic Models of the Labor Market: A Survey, *Journal of Economic Literature*, 43, 4, 959-988.

[28] Salton, G. and M. McGill (1983), *Introduction to modern information retrieval.* McGraw-Hill.

[29] Vázquez, A. (2003), Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67, 056104

[30] Wasserman, S. and K. Faust (1994), *Social network analysis: Methods and applications.* Cambridge U Press. Cambridge, UK.

[31] Wooldridge, J. (2002), *Econometric analysis of cross section and panel data.* MIT Press. Cambridge Mass.

[32] Wuyts, S., M.G. Colombo, S. Dutta, and B. Nooteboom (2005), Empirical tests of optimal cognitive distance, *Journal of Economic Behaviour & Organization*, forthcoming.

Table 1: Summary statistics of the data for the sample of the first collaboration regression

Variable	Description	Mean	Std.Dev.	Max	Correlation w. Proximity
	Number of pairs	26922			
	Number of observations	160339			
$t_1^{ij} - t_0^{ij}$	Duration to first collab.	5.96	3.80	20	p_{t-1}^{ij}
p_{t-1}^{ij}	Proximity	.086	.133	.5	
$\Pr(p_{t-1}^{ij} > 0)$	Connected	.428	.495	1	
$d_{t-1}^{ij} \vert p_t^{ij} > 0$	Distance if connected	7.06	3.67	35	
c_{t-1}^{ij}	Number of shortest paths	.902	1.80	69	.282
$\overline{q}_{t-1}^{ij}$	Avg. productivity	6.30	10.62	187.25	.175
Δq_{t-1}^{ij}	Dif. in productivity	8.98	15.66	282.33	.130
$\overline{n}_{t-1}^{ij}$	Avg. number of coauthors	3.19	2.63	30	.487
Δn_{t-1}^{ij}	Dif. in number of coauthors	3.32	3.70	47	.156
ω_{t-1}^{ij}	Field overlap	.489	.352	1	.193
a_{t-1}^{ij}	Common affiliation	.384	.486	1	.037

Table 2: Summary statistics of the data for the sample of the subsequent collaborations regression

Variable	Description	Mean	Std.Dev.	Max	Correlation w. Proximity
	Number of pairs	14558			
	Number of observations	105854			
y_t^{ij}	Subsequent collaboration	.239	.427	1	p_{t-1}^{ij}
p_{t-1}^{ij}	Proximity	.276	.227	.5	
$\Pr(p_{t-1}^{ij} > 0)$	Connected	.686	.464	1	
$d_{t-1}^{ij} \vert p_t^{ij} > 0$	Distance if connected	3.50	2.82	30	
c_{t-1}^{ij}	Number of shortest paths	1.11	1.44	42	.258
$\overline{q}_{t-1}^{ij}$	Avg. productivity	7.48	12.75	209.33	.077
Δq_{t-1}^{ij}	Dif. in productivity	8.38	15.83	282.33	.063
$\overline{n}_{t-1}^{ij}$	Avg. number of coauthors	5.53	3.61	39	.404
Δn_{t-1}^{ij}	Dif. in number of coauthors	4.10	4.37	46	.095
ω_{t-1}^{ij}	Field overlap	.755	.238	1	.075

Table 3: Results of fixed effects logit regression on first collaboration

Variable	Description	(1)	(2)	(3)
	Number of pairs	26922	26922	26922
	Number of observations	160339	160339	160339
p_{t-1}^{ij}	Proximity	2.002**	1.940**	1.448**
		(.101)	(.102)	(.111)
c_{t-1}^{ij}	Log number of shortest paths	.0684**	-.1046**	-.0708
		(.0177)	(.0382)	(.0383)
$p_{t-1}^{ij} \times c_{t-1}^{ij}$	Interaction term		1.514**	1.055**
			(.294)	(.297)
$\overline{q}_{t-1}^{ij}$	Avg. productivity	-.00693**	-.00702**	-.00890**
		(.00149)	(.00149)	(.00151)
Δq_{t-1}^{ij}	Dif. in productivity	.00269**	.00270**	.00290**
		(.00092)	(.00092)	(.00092)
$\overline{n}_{t-1}^{ij}$	Avg. number of coauthors			.1016**
				(.0102)
Δn_{t-1}^{ij}	Dif. number of coauthors			-.0157**
				(.0059)
ω_{t-1}^{ij}	Field overlap			.777**
				(.136)
$(\omega_{t-1}^{ij})^2$	Squared Field overlap			-.591**
				(.138)

Table 4: Results of fixed effects logit regression on first collaboration for the sample with affiliation data

Variable	Description	(1)	(2)	(3)	(4)
	Number of pairs	11278	11278	11278	11278
	Number of observations	47498	47498	47498	47498
p_{t-1}^{ij}	Proximity	.354*	.361*	.492**	.491**
		(.165)	(.165)	(.179)	(.179)
c_{t-1}^{ij}	Log number of shortest paths	-.0009	.0385	.0258	.0271
		(.0249)	(.0575)	(.0578)	(.0578)
$p_{t-1}^{ij} \times c_{t-1}^{ij}$	Interaction term		-.315	-.183	-.191
			(.415)	(.420)	(.420)
$\overline{q}_{t-1}^{ij}$	Avg. productivity	-.00197	-.00195	-.00162	-.00135
		(.00202)	(.00202)	(.00203)	(.00203)
Δq_{t-1}^{ij}	Dif. in productivity	.00067	.00066	.00060	.00052
		(.00128)	(.00128)	(.00128)	(.00128)
$\overline{n}_{t-1}^{ij}$	Avg. number of coauthors			-.0377*	-.0362*
				(.0171)	(.0171)
Δn_{t-1}^{ij}	Dif. number of coauthors			.0138	.0137
				(.0094)	(.0094)
ω_{t-1}^{ij}	Field overlap			.462	.470
				(.265)	(.0265)
$(\omega_{t-1}^{ij})^2$	Squared Field overlap			-.394	-.397
				(.260)	(.0260)
a_{t-1}^{ij}	Common affiliation				-.2441**
					(.0508)

Table 5: Results of fixed effects logit regression on subsequent collaborations

Variable	Description	(1)	(2)	(3)
	Number of pairs	14558	14558	14558
	Number of observations	105854	105854	105854
p_{t-1}^{ij}	Proximity	-1.443**	-1.238**	-.671**
		(.0710)	(.0723)	(.0775)
c_{t-1}^{ij}	Log number of shortest paths	-.1993**	.2475**	.1542**
		(.0238)	(.0369)	(.0373)
$p_{t-1}^{ij} \times c_{t-1}^{ij}$	Interaction term		-1.972**	-1.093**
			(.131)	(.139)
$\overline{q}_{t-1}^{ij}$	Avg. productivity	-.01262**	-.01232**	-.01197**
		(.00117)	(.00117)	(.00118)
Δq_{t-1}^{ij}	Dif. in productivity	.00560**	.00555**	.00574**
		(.00087)	(.00087)	(.00088)
$\overline{n}_{t-1}^{ij}$	Avg. number of coauthors			-.11007**
				(.00807)
Δn_{t-1}^{ij}	Dif. number of coauthors			.03376**
				(.00505)
ω_{t-1}^{ij}	Field overlap			-1.031**
				(.272)
$(\omega_{t-1}^{ij})^2$	Squared Field overlap			-.409
				(.216)

Table 6: Monte Carlo results without detrending.

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

Notes: *E[coef]* is the mean coefficient value in the sample of 1000 simulations. *σ[coef]* is the standard deviation of the coefficient values. *% significant* is the fraction of coefficients in the sample of 1000 simulations that have an absolute t-value larger than 2.

Table 7: Monte Carlo results with detrending.

	E[coef]	σ[coef]	% significant
A. y_{it}^a is the dependent variable			
coefficient of x_{it}^d	0.085	0.009	100%
coefficient of z_{it}^d	0.000	0.008	5%
Number of observations	2000		
B. y_{it} is the dependent variable			
coefficient of x_{it}^d	0.089	0.025	98%
coefficient of z_{it}^d	0.000	0.021	4%
Average number of usable observations	237		

Notes: *E[coef]* is the mean coefficient value in the sample of 1000 simulations. *σ[coef]* is the standard deviation of the coefficient values. *% significant* is the fraction of coefficients in the sample of 1000 simulations that have an absolute t-value larger than 2.

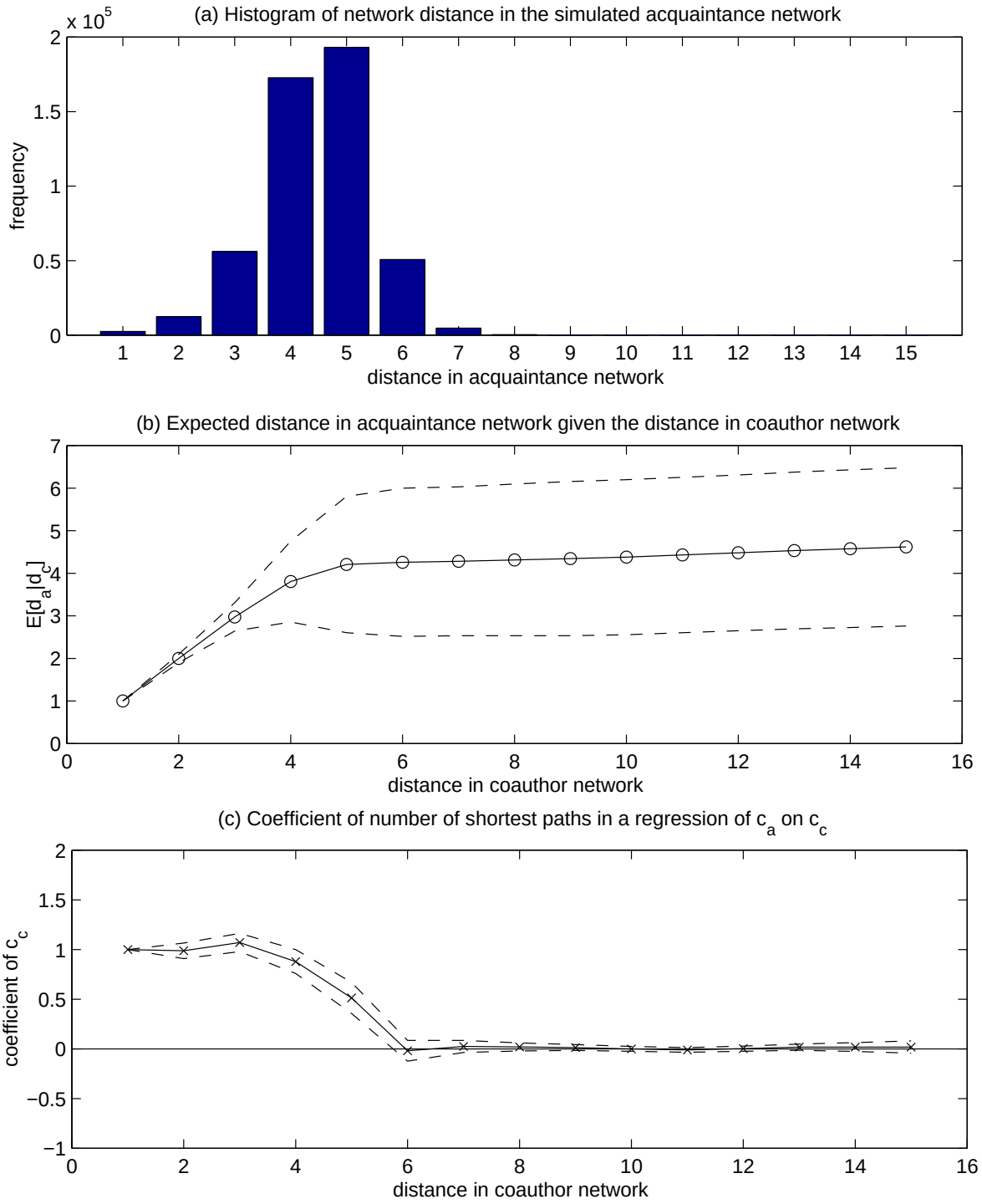

Figure 1: Relation between the distance in the coauthor network and the distance in the acquaintance network.

Note: Acquaintance network is a simulated Erdős-Renyi graph with 1000 nodes and 2500 links. The coauthor network is simulated by taking a random subgraph with only 1000 links of the simulated acquaintance network.

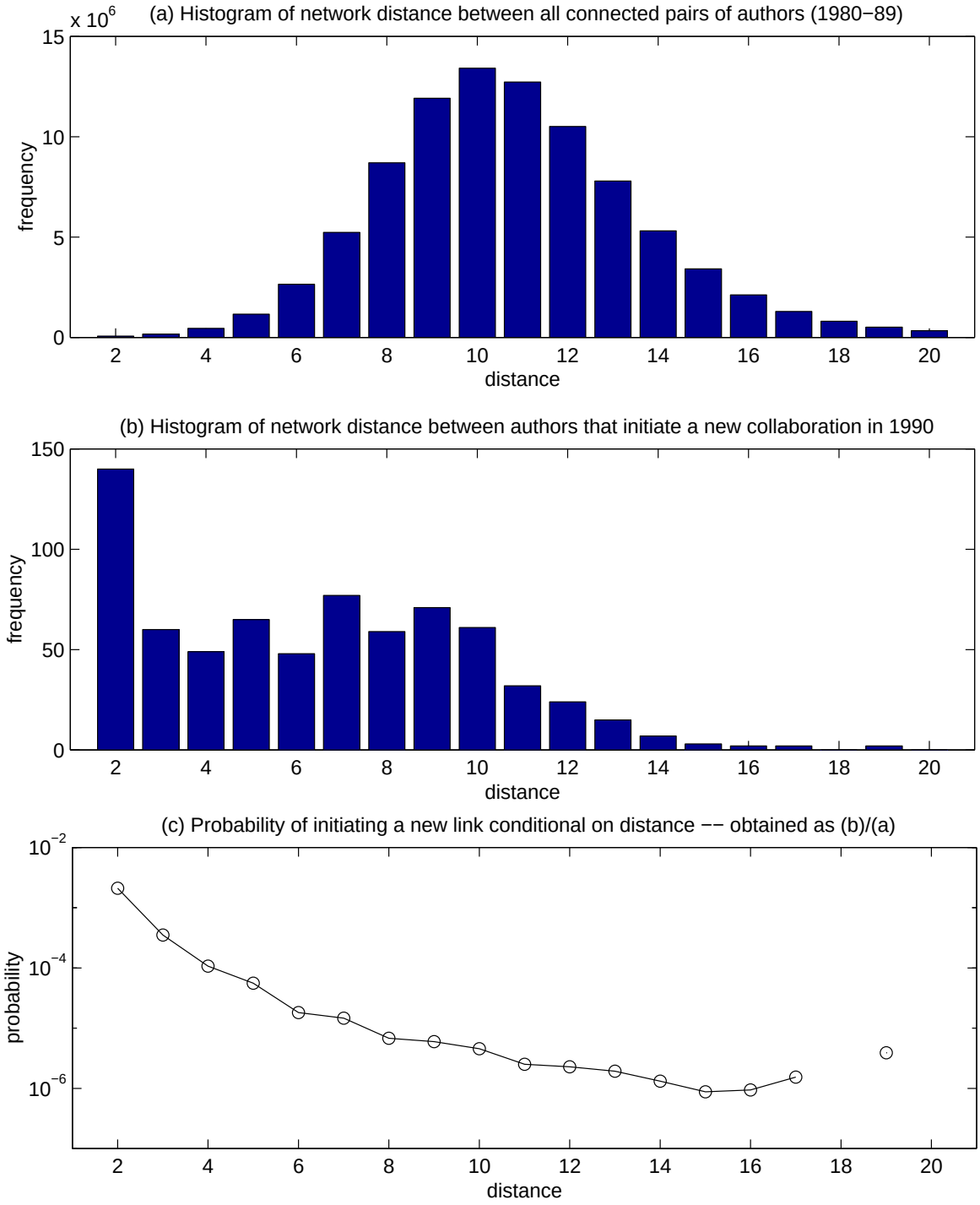

Figure 2: Histogram of distance in the network of the 1980s and the formation of links in 1990.

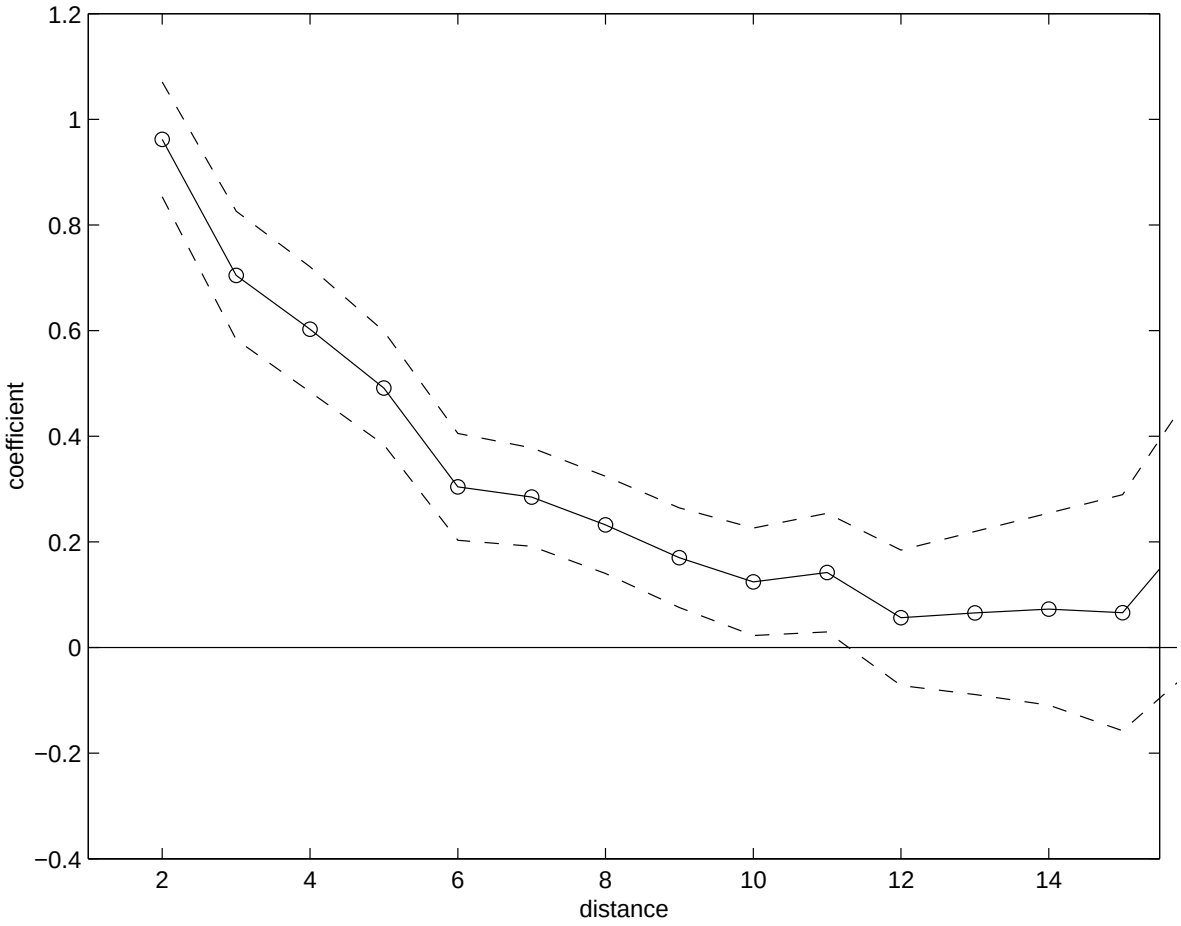

Figure 3: Coefficients of distance dummies in model (11) for first collaboration.

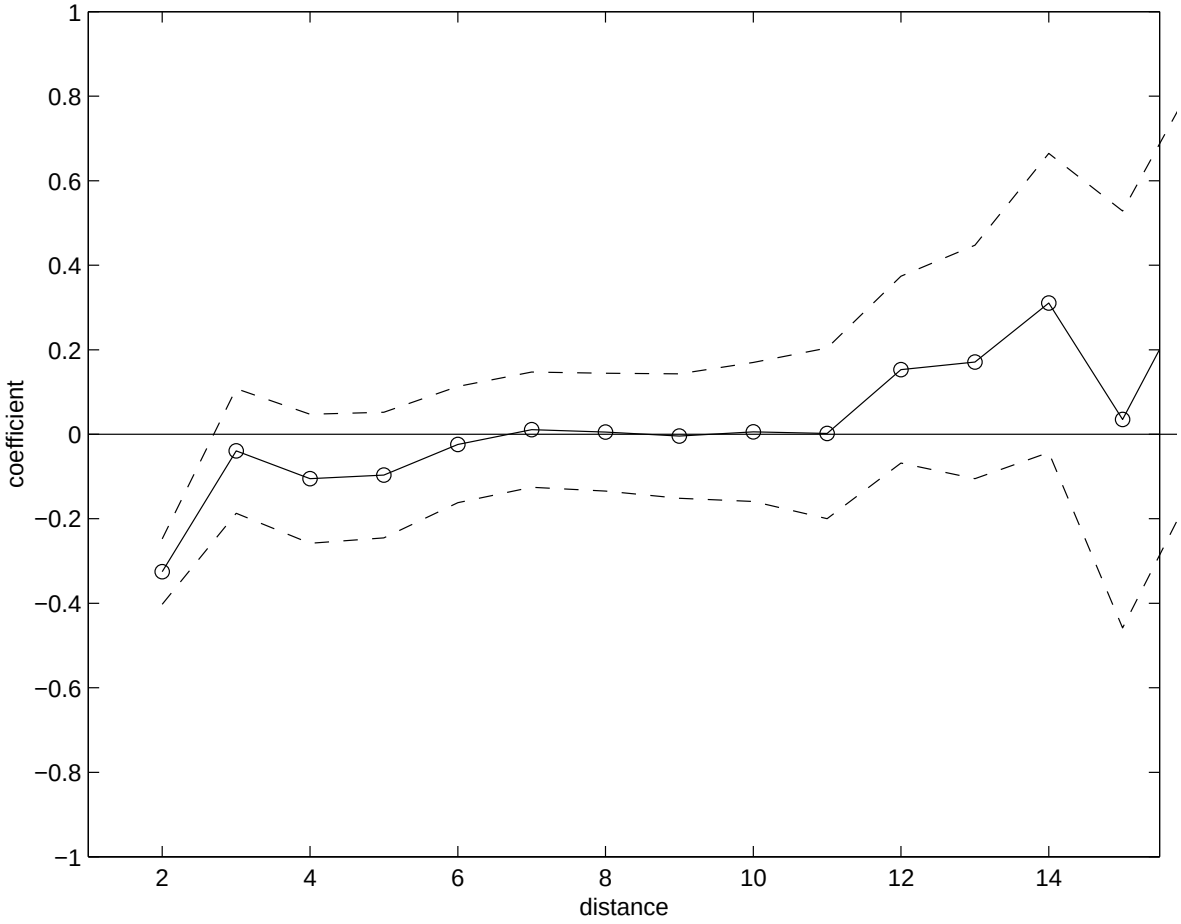

Figure 4: Coefficients of distance dummies in model (13) for subsequent collaborations.