Matching and network effects*

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

This version: December 29, 2008
First version: October 2006

Abstract

The matching of individuals in teams is a key element in the functioning of an economy. The network of social ties can potentially transmit important information on abilities and reputations and also help mitigate matching frictions by facilitating interactions among 'screened' individuals. 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 existing social ties.* The objective of this paper is to empirically test this conjecture.

We examine 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: being at a network distance of 2 instead of 3, for instance, raises the probability of initiating a collaboration by 27 percent.

Research collaboration takes place in an environment where fairly detailed information concerning individual ability and productivity – reflected in publications, employment history, etc. – is publicly available. Our finding that social networks are powerful even in this setting suggests that they must affect matching processes more generally.

*A previous version of this paper was circulated under the title, *Scientific Networks and Coauthorship*. We thank the editors and two anonymous referees for useful comments. We also thank Michèle Belot, Jordi Blanes-i-Vidal, Sebi Buhai, Antoni Calvó-Armengol, Steven Durlauf, Jean Ensminger, Joseph Harrington, Vernon Henderson, Matthew Jackson, Jeff Johnson, Michael Moore, Markus Möbius, Kaivan Munshi, Eleonora Patacchini, Imran Rasul, Tom Snijders, Manuel Trajtenberg, Fernando Vega-Redondo, and a number of seminar participants for useful comments. Marco van der Leij would like to thank the Vereniging Trustfonds Erasmus Universiteit Rotterdam and the Spanish Ministry of Science and Innovation (Project grant SEJ2007-62656 and Juan de la Cierva grant JCI558/07) for financial support.

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

‡Faculty of Economics, University of Cambridge. E-mail: sg472@econ.cam.ac.uk

§Departamento de Fundamentos del Análisis Económico, Universidad de Alicante. E-mail: vanderleij@merlin.fae.ua.es

1 Introduction

The matching of individuals into teams to produce intellectual and physical output is a key element in the functioning of an economy – e.g., job market, business partnerships, agency contracts. The formation of a team has to address various types of information problems: there are many potential team members to choose from, but the ability of individuals is privately known (Akerlof, 1970; Diamond, 1982). In economics the formation of teams has traditionally been studied within a search and matching framework.

In this framework, teammates are anonymous agents and search takes place via random draws from the pool of potential partners; see Rogerson, Shimer and Wright (2005) for a survey of this work. It is reasonable to suppose that, to economize on friction costs, individuals will use information easily available in their circle of friends and acquaintances.[1]

Local network embeddedness of the matching process is not an innocuous artefact of social life; it has significant implications on distribution of matchings and welfare. Montgomery (1991) studies the role of social networks in overcoming adverse selection problems in labor markets. He finds that a reliance on social network based referrals has powerful effects on wage inequality. More recently, Jackson & Rogers (2007) propose a dynamic model of network formation in which agents search for partners randomly as well as locally through the network. It is shown that local network search exacerbates the inequality in the distribution of links and, if utility is concave in number of links, this leads to lower social welfare.[2] These theoretical findings motivate the search for direct empirical evidence for the use of social networks.

This paper studies the formation of new collaborations in academic research. Research collaboration is an environment where much public information is available on individual ability – e.g., publications record, employment history, etc. Consequently we would expect matching frictions to be less prevalent than in other team formation processes. If network proximity affects the formation of new teams even in such a favorable environment, we expect that social networks will also matter for matching processes more generally.

We examine data on coauthorship among economists over a 30 year period, from 1970 to 1999. We show that two economists are more likely to publish together if they are close in the network of all economics coauthors. This result is robust and statistically significant. Network distance coefficients are large in magnitude: being at a network distance of 2 instead of 3, raises the probability of initiating a collaboration by 27 percent. Similarly, the probability of two persons writing their first paper together increases by 18 percent if they are at a network

[1] These ideas have been extensively explored in the literature on social networks in sociology as well as economics; see the seminal work of Granovetter (1995). For an overview of this work, see Goyal (2007).

[2] Moreover, subsequent work of Vigier (2008) shows that this inequality exacerbating feature is specific of local *network* search. For example, if local network search plays a minor role, and if the matching process is mostly locally constrained by geographical distance, then the distribution of matchings turns out to be egalitarian.

distance of 5 instead of 6. From this we conclude that social proximity among researchers facilitates the creation of new scientific collaborations. We develop a number of arguments – based on a variety of controls of time invariant as well as time varying factors – to show that this proximity effect can be interpreted as reflecting flows of information about individuals as well as about the quality of the match.[3]

Research collaboration arises when individuals feel that it is beneficial to work together. So common research interests, the educational background, and other individual characteristics will clearly play a role in determining whether a collaboration arises. Indeed, in sections 3 and 4 below we discuss such evidence in some detail.[4] Our results show that, over and above these standard factors, the emergence of a new collaboration tie is decisively shaped by the existing network of collaboration ties.

We do not have experimental data, so we must be very careful in making inferences on network effects. An important aspect of our paper is the care with which we address the problems which arise in making such inferences. In particular, factors such as a common background, research interests and skill complementarity are likely to be correlated with proximity in the social network of coauthorship. To identify social network effects, we need to convincingly control for these confounding factors. This is an estimation problem common to all empirical studies of peer effects. We deal with this difficulty in three ways.

First, we control for pairwise fixed effects. This takes care of all time-invariant complementarity and social proximity effects, such as similarity in age, place of education, stable research interest etc. With pairwise fixed effects, identification of network effects is achieved solely from the timing of collaboration, i.e., we ask whether, conditional on eventually publishing together, a pair of authors is more likely to initiate a collaboration after they got closer in the network of coauthorship.

Secondly, using the available data we construct control variables for time-varying effects, such as changes in productivity and research interests. This takes care of the most serious time-varying confounding factors.

Third, we remain concerned that results may be biased by unobserved time-varying effects – such as non-measurable changes in research interests – that affect the likelihood of collaboration and are correlated with network proximity. Since these effects capture unobserved forces that induce researchers to work together, they should affect the likelihood of all collaborations,

[3]Links to other researchers may also provide access to precious information about research and collaboration opportunities. Consequently, poorly connected researchers may be at disadvantage. This observation provides a link between our study of scientific networks, and the growth and trade literatures which studies technology and information transfer across economies. See, for example, the recent work by Hidalgo et al (2007). Similar ideas have been discussed in the context of job markets by Topa (2001) and in the context of international trade by Casella and Rauch (2002).

[4]For earlier work on the determinants of scientific collaboration see McDowell and Melvin (1983) and Hudson (1996).

not just the first one. In contrast, network effects should only affect the likelihood of the first collaboration between two authors: social networks carry relevant information and create opportunities for face-to-face interaction that may induce two authors to begin work together; but once two authors have published together, they have a lot of information about match quality and network proximity should no longer matter. Building on this observation, we conduct a placebo-like experiment by contrasting the effect of network proximity on first and subsequent collaborations. Time-varying confounding factors that are correlated with network proximity should have a similar effect on first and subsequent scientific collaborations; network effects should only affect the first collaboration. We find that network proximity is only significantly positive for the first collaboration.

This paper contributes to the empirical study of social networks. Informal institutions have been empirically studied extensively in economics and other subjects; see e.g., Granovetter (1985), Greif (2001), Munshi (2003), Munshi and Rosenzweig (2006), North (2001), Fafchamps and Lund (2003) and Fafchamps (2004). The empirical study of the architecture of large and evolving social networks is relatively new. In recent work, Krishnan and Sciubba (2006), Comola (2008), Mayer and Puller (2008) study the formation of links. They argue that individual level heterogeneity – reflected in differences in wealth and race – plays an important role in the creation of new links. By contrast, we control for individual differences and identify a pure network proximity effect in the creation of new links. Our use of longitudinal date allows us to make this stronger inference on network effects.[5]

This paper is also related to the literature on the economics on social interactions – see Manski (2000), Moffitt (2001) and Brock and Durlauf (2001), for overviews. This body of work argues that a significant part of the variation in behavior across individuals faced with similar incentives is due to their being a member of one group rather than another – e.g., Glaeser, Sacerdote and Scheinkman (1996), Bertrand, Luttmer and Mullainathan (2000), Banerjee and Munshi (2004), and Duflo and Saez (2003). The focus of this literature is on explaining behavioral differences across well defined groups, paying special attention to the difficulty of empirically identifying social interaction effects within groups. As in this literature, we take particular care to tackle difficult endogeneity problems, in particular the problem that the network effect may be spurious as relevant unobserved individual characteristics, such as uncontrolled research interests, are correlated in the local neighborhood. Our point of departure is that we look at differences in social connections *within a group* to understand differences in individual behavior.[6]

[5]The effects of social networks are actively being studied. Conley and Udry (2008) who investigate the effects of social communication networks on the individual decision to adopt new crops such as pineapple and Calvó-Armengol, Patacchini and Zenou (2008) study the effects of location in a network on human capital formation and criminal activity.

[6]Bramoullé, Djebbari and Fortin (2008) show that, in contrast to the use of group affiliation data, the use of detailed network data allows stronger identification of endogenous and contextual network effects.

This paper is organized as follows. In Section 2 we present a conceptual framework and introduce the details of the testing strategy. The data are discussed in Section 3 and the econometric results appear in Section 4.

2 Testing strategy

In this Section we begin by presenting a simple referral model, the sole purpose of which is to motivate our estimating equation as an approximation to an arbitrary information sharing network process. With this equation in hand, we present our testing strategy and discuss a number of econometric issues that arise in the estimation of the model.

2.1 The estimating equation

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. The pattern of coauthorship forms 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.

Consider two authors i and j. We assume that, conditional on knowing each other, researchers collaborate with probability $m_t^{ij} \leq 1$. Many factors are likely to affect i and j's willingness m_t^{ij} to form a collaborative team – e.g., complementary skills, shared research interest, proximity in age and background, etc. Some of this information – e.g., publication and citation record – is publicly available, albeit at a financial or time cost; some relevant information is not – e.g., whether the potential collaborator is reliable, easy to work with, etc.

Suppose that authors i and j share a common coauthor k. It is reasonable to suppose that i and j can get information about each other via k, for instance because k talks to i about j and vice versa. It is also possible that i and j met and became acquainted at a professional event – e.g., a conference – organized by k. Since the data does not enable us to distinguish between these different processes, we regard them as equivalent for the purpose of the model.

Let $b < 1$ denote the probability that author k "refers" i and j to each other, that is, facilitates in one way or another the circulation of information that makes it easier for i and j to assess the potential benefit from a collaboration. To facilitate exposition, assume for a moment that $m_t^{ij} = 1$, that is, conditional on meeting each other, i and j wish to collaborate. In this case, the probability P_t^{ij} of observing a collaboration between i and j at time t is given by:

$$P_t^{ij} = 1 - (1 - b)^c \tag{1}$$

where c is the number of common coauthors between i and j – and thus the number of paths between them: the more common coauthors i and j have, the more likely it is that they get to know each other.

Now let us consider longer paths. The information that j gets about i from k may be passed on to others whom j knows. Alternatively, when i organizes an event afterwards, she may invite some of her past collaborators along with j and k, facilitating contact among collaborators of i, j and k. The quality of information that is being conveyed in this fashion will decay as it passes – indirectly – through more members of the group. The match facilitating effect of network proximity is thus likely to fall with network distance. These considerations lead us to conjecture that the probability of i and j engaging in a collaboration is falling in the distance between i and j in the network of social ties.

To formalize this idea is a simple way, suppose that the probability that a node j transmits information along a given path is independent from the probability that the same node j transmits the same information along another path. With this assumption, the probability of receiving the information over distance k when there are c_k paths of length k linking i to j becomes:

$$P^{ij} = 1 - \prod_{k=2}^{\infty} (1 - b^{k-1})^{c_k} \tag{2}$$

where we have dropped time subscripts on P^{ij} and c_k to improve readability. Let d^{ij} denote the length of the shortest path between i and j and let c^{ij} denote the number of shortest paths between i and j. Rewriting (2) in terms of $1 - P^{ij}$ and taking logs on both sides, we get:

$$\begin{aligned} \log(1 - P^{ij}) &= \sum_{k=2}^{\infty} c_k \log(1 - b^{k-1}) \\ &\approx -\sum_{k=2}^{\infty} c_k b^{k-1} \qquad (3) \\ &\approx -c^{ij} b^{d^{ij}-1} \qquad (4) \end{aligned}$$

The first approximation relies on $\log(1 + a) \approx a$ for a small, while the second approximation relies on b being small. For the last approximation to be reasonable, it must be that c_k does not increase rapidly with distance.

We now use approximation (4) to derive an estimable model of collaboration. Let us assume that P_t^{ij} follows a logit distribution, i.e.:

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

The dependent variable takes value 1 if i and j collaborate, and 0 otherwise. Approximation (4) suggests a reasonable way of writing $X_t'\beta$. Dropping time subscripts to improve readability, we have:

$$\begin{aligned} 1 - P^{ij} &= e^{-c^{ij}b^{d^{ij}-1}} = \frac{1}{1+e^{X'\beta}} \\ e^{c^{ij}b^{d^{ij}-1}} &= 1 + e^{X'\beta} \\ c^{ij}b^{d^{ij}-1} &= \log(1+e^{X'\beta}) \\ &\approx e^{X'\beta} \end{aligned} \tag{5}$$

where we use $\log(1+a) \approx a$ for a small. Approximation (5) is admittedly crude, but since its sole purpose is to motivate the estimation regression, this is not too serious a concern.

We now reintroduce the probability $m_t^{ij} < 1$ of collaborating, conditional on knowing each other. The unconditional probability of collaborating is equal to the probability of being "referred" to each other times the conditional probability of collaborating m_t^{ij}. Let $m_t^{ij} = e^{Z_t'\gamma}$ where Z_t is a vector of variables representing match quality. The probability of i and j collaborating is:

$$P_t^{ij} \approx m^{ij}e^{X'\beta} = e^{X_t'\beta + Z_t'\gamma}$$

Combining the above with (5), our estimated model takes the form:

$$X_t'\beta + Z_t'\gamma \approx -\log b + \log c_{t-1}^{ij} + \log b(d_{t-1}^{ij}) + Z_t'\gamma \tag{6}$$

We thus need to estimate a logit model in which the dependent variable is whether i and j collaborate at time t, and the regressors are the length of the shortest path d_{t-1}^{ij}, the number of shortest paths c_{t-1}^{ij}, and determinants of match quality Z_t. Network variables d_{t-1}^{ij} and c_{t-1}^{ij} are lagged to avoid simultaneity bias. The coefficient of d_{t-1}^{ij} measures the log of unknown probability b (a negative number since $b < 1$) and the coefficient of $\log c_{t-1}^{ij}$ should be approximately 1.

2.2 The acquaintance network

So far, we have assumed that information about author ability and personal attributes travels via coauthor ties only. In practice, information about coauthor ability and other attributes is likely to circulate more broadly among the acquaintances of i and j. To investigate how this may affect inference, let us define the (unobserved) network of personal acquaintance such that a link exists between i and j exists in this network if i and j know each other well enough to transmit accurate and trustworthy information about other researchers' type. The acquaintance 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.[7]

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

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

The problem is that 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 a subset of the acquaintance network, we must 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.

This is illustrated with a simple computer experiment, in which we simulate an 'acquaintance network' and corresponding 'co-author network' by following the procedure of Jackson & Rogers (2007) for a 1000 nodes.[8] Figure (1a) shows a histogram of the simulated acquaintance network, obtained by following the Jackson-Rogers procedure with $m = 8.4$, $r = 4.7$ and $p = 1$. Note that almost all acquaintances tend to be within only 3 degrees of separation. Next, we randomly select 10 percent of the links from the 'acquaintance network' to obtain a corresponding 'coauthor network'.[9] 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 these simulated networks. Figure (1b) shows the results.

As predicted, we 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). To summarize, if we regress P^{ij} on d_c^{ij} and find a significant relationship,

[7] 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 large.

[8] The procedure of Jackson & Rogers (2007) generates networks that mimic the stylized facts of real social networks, namely: a fat-tail degree distribution, short network distances, high clustering, a positive assortativity and a negative relation between degree and clustering.

[9] The model of Jackson & Rogers (2007) only requires the estimation of three parameters: the average in-degree, m, the ratio of random vs. local links, r, and the probability that a searched node is linked to, p. In the case of the co-author network of economists Jackson & Rogers (2007) find that for the 1990s, the best fit is obtained with parameters $m = .84$, $r = 4.7$ and $p = .10$ (Jackson & Rogers, 2007:p. 902, Table 1).

If we consider acquaintances as nodes that are searched, and co-authors as the searched nodes that with probability $p = .10$ receive a link, then we may simulate the 'acquaintance network' by simulating the Jackson & Rogers model with parameters $m = 8.4$, $r = 4.7$ and $p = 1$, and the corresponding 'co-author network' by randomly taking 10 percent of the links of the simulated 'acquaintance network'.

this means that network proximity 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 lower bound of $d_c = 1$ we know that $d_a = 1$ as well). In contrast, 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 can be investigated by regressing P^{ij} on a series of dummy variables, one for each value of d_c. We expect dummy coefficients to be strongest and most significant at low values of d_c while coefficients should be negligible and non-significant for values of d_c above a certain threshold.

Turning to the number of paths, c_c also constitutes an imperfect measure of c_a. To see this, note that if $d_c = d_a$ then $c_a \geq c_c$: if the coauthorship distance is the same as acquaintance distance, then the number of paths between i and j in the coauthorship network provides a lower bound for the number of paths in the acquaintance network. We have 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 becomes weaker 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 should be non-significant.[10]

2.3 Econometric issues

Our testing strategy is to estimate equation (6) and test whether network variables d_t^{ij} and c_t^{ij} are significant with the correct sign. For estimation 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

[10] While d_c and c_c can serve as proxies for $E[d_a]$ and $E[c_a]$, the same cannot be said of the number of links c_k at longer network distances. This is another reason why our estimation is based on equation (4) even though, in principle, we could have used the more general (3).

many factors that are not observed by us. For example, researchers choose to work together because they share common interests or complementary abilities. These determinants of match quality m_t^{ij} are likely to be positively correlated with network distance and may lead to spurious 'network effects', unless they are appropriately controlled for.

We do so by decomposing match quality determinants $Z'\gamma$ into two parts: a pairwise fixed effect μ^{ij} and time-varying controls for match quality Z_t^{ij}. The pairwise fixed effect controls for all time-invariant characteristics of both authors i and j, including their time and place of birth, gender, ethnicity, mother tongue, where they received their education, and where they started their career. Pairwise fixed effects also control for any time-invariant determinants of match quality, such as mutual empathy, complementarity of skills, and commonality of interest and outlook.

Pairwise fixed effects do not, however, control for time variation in individual characteristics and match quality. To this effect, we introduce time-varying controls for productivity and overlap in research interests. How we construct these variables is discussed in details below.

Because experimental data is unavailable, there remains the possibility that time-varying unobservable determinants of match quality may be correlated with network variables d_t^{ij} and c_t^{ij}. To deal with this issue, we note that, after i and j have collaborated once, they know each other and their likelihood of collaborating only depends on match quality m_t^{ij}. It follows that the probability of observing i and j collaborating, conditional on them having collaborated in the past should *not* depend on network variables d_t^{ij} and c_t^{ij}. This suggests a placebo-like test of whether network variables d_t^{ij} and c_t^{ij} are significant in equation (6) only because they are correlated with an unobserved dimension of match quality m_t^{ij}: if this were the case, then d_t^{ij} and c_t^{ij} would have the same effect on first and subsequent collaborations. On the other hand, if we observe an effect in the regression on first collaboration, but *not* such an effect in the regression on subsequent collaboration, then this is strong evidence that our results are not driven by correlated effects due to unobserved or uncontrolled dimensions of the matching quality m_t^{ij}, but are in fact due to pure network effects that alleviate informational frictions.

To clarify how an unobserved dimension of m_t^{ij} creates an spurious network effect on first *and* subsequent collaboration, we first point out that we measure the network distance d_t^{ij} between two economists i and j *exluding* the link between i and j itself. Hence, the distance between i and j is always at least 2, and also after the first collaboration the network distance between i and j will be fluctuating over time. This allows the network distance to be correlated to unobserved dimensions of match quality m_t^{ij} even after the first collaboration. For example, suppose that common research interests is the sole driving force behind collaboration, and that research interests of i and j converge further after the first collaboration. In that case, i becomes more likely to collaborate to one of the co-authors of j, as j's co-authors are likely to have the

same research interests of j and therefore get closer in research interests to i. It is therefore likely that the network distance between i and j (excluding the link i and j itself) becomes shorter even after the first collaboration. If i and j are more likely to strengthen their collaboration due to closer research interests, then we would observe a (spurious) network effect in the regression on subsequent collaboration.

Formally, 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. For the first collaboration between a pair of authors i and j, we 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:

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

$\partial f/\partial d < 0$ and $\partial f/\partial c > 0$. If the coefficients have the correct signs, this indicates that there exist proximity effects in the formation of new coauthor ties. For subsequent collaborations, we similarly estimate:

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

where network distance d_{t-1}^{ij} is defined ignoring direct coauthorship links between i and j, i.e., d_{t-1}^{ij} is the length of the shortest path between i and j in the coauthorship network that does not include their direct coauthorship link. We expect that $\partial g/\partial d = 0$ and $\partial g/\partial c = 0$ since the authors now know the match quality. Estimating equations (7) and (8) is the objective of the paper.

To control for unobserved heterogeneity in (7) and (8) we include a pairwise fixed effect μ^{ij} for each coauthorship pair:

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

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

Equations (9) and (10) are estimated separately using a fixed effect logit model. Fixed effects μ^{ij} in (10) are allowed to differ from that in (9).[11]

The inclusion of fixed effects in equation (9) raises an estimation problem. To understand the nature of the problem, consider two authors who began their publishing career at time t_0 and coauthor their first paper together at time t_1. By construction, $y_t^{ij} = 0$ for all $t \in [t_0, t_1)$ and $y_t^{ij} = 1$ for $t = t_1$. For each pair ij the time sequence of dependent variables thus takes the

[11] Since equation (10) is estimated using only author pairs that have collaborated at least once, it is potentially subject to selection bias. This is not a cause for concern in this case given that the fixed effect in (10) absorbs any Mills ratio/selection correction term.

form $y^{ij} = \{0, ..., 0, 1\}$, in which the number of 0's varies across author pairs. Equation (9) is equivalent to a single-spell, discrete time duration model with fixed effects.

The estimation of duration models with fixed effects raises a well known identification problem because duration dependence and fixed effects cannot be separately estimated. This lack of identification does not affect us directly, however, because we are not interested in the shape of the hazard function over time, i.e., equations (9) and (10) do not include time as regressor. Our sole focus is the variation of the hazard with time-varying regressors. Given this, it remains possible to identify and consistently estimate coefficients of time-varying regressors, as has been shown for instance by Allison & Christakis (2006). This is unproblematic for regressors that do not trend over time. But regressors that contain a trend mechanically generate a spurious correlation with the dependent variable. This is because, by construction, the dependent variable y_t^{ij} takes the form of a series of 0 followed by a single 1; hence *any* regressor that exhibits a positive or negative trend will automatically help predict y_t^{ij}. We elaborate on the nature of the problem using a Monte Carlo simulation in Appendix A; see also Allison & Christakis (2006). Almost all our regressors display trending behavior. In particular, network distance between i and j becomes smaller over time 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 each of them individually. We accomplish this 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.[12] In Appendix A we show that this method performs well for our purpose. Detrending regressors naturally results in a loss of information and therefore leads to an attenuation bias similar to that which occurs as a result of introducing fixed effects in a linear regression. As a result, our estimates err on the conservative side. If we find significant results under these conservative circumstances, we thus can be confident that they are really significant. The method is nevertheless sensitive to correct specification of the trend. To protect against the possibility of misspecification in detrending, we check the robustness of our results with respect to various detrending methods. This is discussed in detail in subsection 5.4.

Pairwise fixed effects capture many individual or pairwise factors that 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, productivity, research interests, and propensity to collaborate are likely to *change* over a researcher's career, and these changes may be correlated with changes in network distance. Using the available data, we construct a number of control variables to address these concerns. These include proxies for research ability,

[12] 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.

propensity to collaborate, and research overlap. How these variables are constructed is detailed in the data section, to which we now turn.

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. Then we present descriptive statistics from our data set.[13]

3.1 Definition of variables

We first turn to the definition of the dependent variable y_t^{ij}. For the purpose of the econometric analysis, we consider a researcher active from the year of first publication. The set of active authors at time t is denoted S_t. Each researcher $i \in S_t$ can potentially coauthor an article with any other researcher $j \in S_t$. More precisely, 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 for all $t_1^{ij} < t \leq t_2^{ij}$ as $y_t^{ij} = 1$ if i and j coauthored a publication in year t and $y_t^{ij} = 0$ otherwise.

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 instantaneously. 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.[14] Having obtained the coauthorship network, we compute the shortest network distance

[13] 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, on average the same publication lags affect all variables. Nonetheless, publication lags typically vary per article, and this may contaminate our results. To test for robustness we therefore repeat our analysis lagging the explanatory variables by 3 years instead of 1 year, see Subsection 5.2.

[14] As a robustness check we have also considered a 5 year network window for the period 1975-1999, see Section 5.1.

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 (i.e., coauthorship) 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 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. Note that, since we ignore the own link in the computation of d_t^{ij}, p_t^{ij} never takes the value 1, even in the regression on subsequent collaborations, i.e. when i and j already collaborated in the past. Variable p_t^{ij} is the distance measure used in the estimation of equation (9) and (10).

Next we turn to time-varying controls Z_t^{ij} which proxy for changes in match quality and therefore in the conditional collaboration probability m_t^{ij}. We start with *research ability*. Authors who publish more on average have more coauthors and thus are better linked, and thus closer to each other in the coauthorship network. To the extent that authors match on research ability, omitting this variable may lead to incorrect inference. As proxy variable for research ability, we measure individual *productivity* q_t^i using the publication record of each author in the Econlit database. Standard measures of publication quality combine quantity (number and length of published articles) with quality (e.g., journal rank). We use a simple scheme which captures these ideas and relies on recent citations ranking of journals, namely the quality weighting system developed by the Tinbergen Institute, a research center based in Amsterdam and Rotterdam.[15] This list of journals (hereafter the TI list) is used by the Institute to assess the research output of faculty members at 3 leading 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 in economics and related fields (econometrics, accounting, marketing, and operations research), of which 113 are covered by EconLit in 2000. This list of journals, which is reproduced in Appendix B, is split into 3 categories: AA, A and B. Based on this we define a journal quality index as follows: a journal in category AA yields four points, a journal in category A yields 2 points, a journal in category B yields 1 point, and

[15] See `http://www.tinbergen.nl/research/admission.html` for a full description of the TI point system. The rankings of journals mentioned in Kalaitzidakis, Mamuneas and Stengos (2003) were used as an input in deriving the TI list.

a journal in an unlisted journal yields 0 points. For each published article, each author is given a number of points according to the formula:

$$\text{Points} = \frac{\text{Journal quality index} \times \text{Number of pages}}{\text{Number of authors} + 1}$$

Variable q_t^i is the number of points author i has accumulated in the years from $t-9$ to t.

Research output q_t^i is author-specific. Since the estimating equations are dyadic and the underlying network is undirected, Z_t^{ij} regressors must enter the regression in a symmetric way. To this effect, we create two pair-specific variables: average productivity

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

and the absolute difference in productivity

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

High productivity authors may be more likely to collaborate since they are more attractive to each other. If this is the case, we expect m_t^{ij} to increase in $\overline{q}_t^{ij}$. The sign of Δq_t^{ij} depends on whether authors match on productivity, or whether coauthorship is more likely between dissimilar authors, e.g., junior-senior collaborations.[16] Here $\overline{q}_t^{ij}$ and Δq_t^{ij} are simply control variables, so we do not discuss their possible interpretation further.

Our next control variables proxy for *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. More precisely, we compute the *number of coauthors* n_t^i of author i over the ten years preceding time t, and similarly for author j. Since regressors must enter the regression in a symmetric fashion, 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}.

[16] It is possible to show that the relation of the propensity to collaborate to the ability of i and j depends on the forms taken by returns to colaboration. If effort is irrelevant, then we expect researchers of similar ability to work together. 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 may increase incentives to collaborate.

We also wish to control for *research overlap* between authors. A commonality of research interest is probably the single most important factor in determining the likelihood of collaboration. We therefore expect authors to be more likely to collaborate if they share similar research interests. Pairwise fixed effects probably absorb much of the influence of commonality of interests on the likelihood of collaboration. There nevertheless remains the possibility that research interests evolve over time and that this change brings researchers together. Since authors who work on similar topics are more likely to collaborate, changes in network distance are likely to be correlated with changes in research overlap.

To control for this possibility, we construct an index ω_t^{ij} of research overlap between any two researchers. We want this index to capture not just having worked in similar research areas but also overlap in research topics. For instance, if a researcher has worked on, say, development economics and microeconomic theory (2 separate categories in JEL codes), she may be more likely to work with another researcher who has also focused on development and micro. To capture this idea, we construct an index of overlapping interests ω_t^{ij}.

To do this, we use the JEL classification codes contained in the EconLit database. We categorize articles into 121 subfields according to the first two digits of the JEL codes.[17] Articles with multiple JEL codes are 'divided' and assigned proportionally to each of the corresponding fields.[18] 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 $t-9$ to t (such that $\sum_f x_{t,f}^i = 1$). Then

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

The cosine similarity measure is a standard measure used by computer scientists in the development of search engines; 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.

Recent work on cognitive distance (Wuyts et al., 2005) suggests that research overlap affects the probability to collaborate in two ways. On the one hand, collaboration is only attractive when the researchers involved have complementary knowledge or skills. This suggest that collaboration

[17] The JEL classification is the most common field classification system used in Economics, and can be found at `http://www.econlit.org/subject_descriptors.html`. 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. We also experimented with a coarser classification of 9 main fields based on the first digit of the JEL codes, but this did not have any qualitative impact on the results.

[18] To give an example, if for one article the JEL codes A10, A11 and B31 are given, then 2/3 of the article is assigned to field A1, while 1/3 of the article is assigned to field B3.

is unlikely when there is too much overlap in skills and thus when research overlap is too strong. 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. To allow for this possibility, we include a quadratic term $(\omega_t^{ij})^2$ in the regression as well.[19]

3.2 Descriptive statistics

In Table 1 we provide summary statistics of the various variables used in the analysis. Column 1 provides a sample of pairs that *never* collaborated, column 2 provides statistics on a sample of collaborating pairs *before* their first collaboration, and column 3 provides data on a sample of collaborating pairs *after* their first collaboration. These data are obtained as follows. In the period from 1980 to 1999 in total 73,873 economists in the dataset collaborated at least once. This allows for about $73873 \times 73872/2 \approx 2.7$ billion potential collaborating pairs, but we observe that only 93,223 of those pairs have collaborated. Since it is computationally impossible to analyze 2.7 billion non-collaborating pairs, we draw a random sample of 162,166 author pairs, with a corresponding number of 921,392 pair-years.[20] The provided statistics, reported in column 1, serve as a benchmark. We observe that economists on average have 2 coauthors, and that two economists are connected via a path with 20 % probability and that it they are connected the average distance is 10. These figures correspond to those given in Goyal, van der Leij and Moraga (2006).

For the estimation of the first collaboration regression (9), we only consider pairs of authors who have a jointly published paper within the 1980-1999 time frame.[21] The fixed effect logit model requires a dependent variable that, for each pair of authors, varies over time. Pairs of authors who never collaborated have $y^{ij} = \{0, ...0\}$ and thus drop out of the analysis.[22] This leaves us with 26,922 collaborating pairs. 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. The time elapsed from the first publication until ij publish their first joint article is 6 years on average.

[19] We also considered the use of affiliation data. However, this turned out to be highly problematic. The JEL database contains information about author affiliation, but only after 1989 and occasionally in 1988. Moreover the data is spotty and incomplete. As a result, the inclusion of this affiliation data reduced the power of any test on significance dramatically. On top of that, the affiliation mentioned on an article is typically the affiliation at the time of publication, and very often this does not correspond to the affiliations at the time of the decision to collaborate. We therefore decided not to include these results in the paper.

[20] In drawing the sample, we have rejected all pairs for which there is no time overlap between the authors, i.e., we ensure that for each randomly selected pair of potential coauthors there is some overlap in the years that the two economists are actively publishing.

[21] Due to lack of data, we cannot take into account pairs of authors who attempted joint work but failed to publish jointly.

[22] Pairs that directly collaborated in the first possible year also drop out since $y^{ij} = \{1\}$.

Before the first collaboration, the probability that i and j are directly or indirectly connected via the coauthorship network is 43%. If connected, the average network distance d_t^{ij} between them is around 7. This is a much shorter distance in comparison to the network distance of those pairs that never collaborated. 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.[23] In a world of random matching, the probability distribution of distance among new matches should roughly mirror the probability distribution of distances in the existing network among the authors. However, as Figure 2(b) shows, 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. Network distance is thus associated with a fall in the likelihood of collaborating. While this constitutes preliminary evidence of network effects, for this evidence to be convincing we need to control for possible confounding factors. To this we now turn.

Control variables described in Section 3 are presented in Table 1. We observe that productivity q_t^i and number of coauthors n_t^i in the second column are higher than in the first column. This mainly reflects the fact that economists with more links are by definition sampled more often in the second dataset.[24] Statistics on field overlap appear next in Table 1. The field overlap index ω_t^{ij} is around .30, much higher than the .05 among non-collaborators, indicating that economists typically collaborate with someone in their field.

Similar statistics are reported in column 3 for subsequent collaborations. As for (9), estimation of regression (10) with pairwise fixed effects requires variation in the dependent variable. This implies that we only sample pairs who have at least one subsequent collaboration, that is, pairs who have collaborated twice in at least two different years. This leaves 14,558 coauthor pairs. For each of these pairs we construct a sequence of y_t^{ij} from the year following first joint publication until the year of last publication. This gives a little over 105,000 observations. We see from Table 1 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 column 3 with column 2 we see that authors who continue collaborating tend to be closer in the author network. Field overlap is higher as well.

[23] For the purpose of this Figure, we use the author network from 1980 to 1989 and define collaborating pairs as those who start a collaboration in 1990.

[24] This fact raises concerns of sample selectivity and clustered correlations. To address these concerns we provide a robustness check in Subsection 5.3, in which we ensure that only one collaboration per author is sampled.

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:

$$\Pr(y_t^{ij} = 1) = \lambda(\beta p_{t-1}^{ij} + \gamma_1 \log c_{t-1}^{ij} + \mu^{ij}). \tag{11}$$

where $\lambda(.)$ denotes the logit function.

We first estimate naive logit regressions on the probability to start and continue a collaboration without controlling for fixed effects, that is, assuming that $\mu^{ij} = \mu$. As explained before, these results are likely to be biased upwards due to unobserved variables that are related to both collaboration and social network distance. However, they serve as a useful benchmark for our later estimations. The results are presented in Table 2.

The first two columns of Table 2 present the results of the logit regression on the probability to initiate a collaboration. This data set contains collaborating pairs as well as pairs that never collaborated.[25] The results in the first two columns confirm the preliminary analysis in Figure 2. In particular, we observe that proximity has a large coefficient of 13.9. Using the approximation of (5), this coefficient implies that the probability to start a collaboration is approximately 10 times larger for pairs that are at distance 2 than for pairs that are at distance 3. Including control variates such as productivity, field overlap and number of coauthors shows that, as expected, standard economic factors have a strongly significant effect on the probability to collaborate. In particular field overlap is strongly significant, more significant than the network proximity variable. These control variables are all correlated with network proximity, and the inclusion reduces the magnitude of the proximity effect somewhat, but note that the reduction of the proximity effect is not very large.

We find a much bigger difference when we compare the results of the first two columns to the results of the last two columns, which contain the results of *continuing* a collaboration. Although the effect of proximity on subsequent collaboration remains significantly positive, the effect is much smaller than in the case of first collaboration. Remember, though, that this estimate is derived from a regression that does not control for pairwise fixed effects.

We now turn to the results of the main regression that *does* control for pairwise fixed effects. Equation (11) is estimated using conditional logit to eliminate the fixed effect μ^{ij}. As detailed in

[25] As explained in Subsection 3.2, there are too many non-collaborating pairs to include them all in the dataset. We therefore only include a random subsample of them, as described in Table 1. Together with the pairs that do collaborate we have about 200,000 pairs and more than a million observations. Reported standard errors are adjusted to correct for the undersampling of non-collaborating pairs.

Appendix A, all regressors are detrended to eliminate spurious correlation with the dependent variable. 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. When we include additional control variates (column 2), the proximity coefficient falls by 28%, but remains strongly significant. In fact, the significance is stronger than any of the control variates. This suggests that network proximity plays an important role in the formation of new research collaboration ties.

We wish to ascertain whether this result is driven by a local effect over short network distances, or whether it is a more diffuse effect extending to long network distances. To investigate this idea, we replace p_t^{ij} with network distance dummies and re-estimate model (11) with control variables, without and with pairwise fixed effects. The coefficient of distance d dummy measures the effect on the probability of tie formation of i and j being at distance $d_t^{ij} = d$ relative to i and j being unconnected. Coefficient estimates for distance dummies are presented in Figure 3 for the simple logit regression and in Figure 4 for the pairwise fixed effects regression. In both figures the dashed lines depicts the 95% confidence interval.

Results indicate that network effects are not limited to short distances: distance dummies remain significant up to 9 degrees of separation. The same pattern is observed irrespective of whether or not we control for pairwise fixed effects. Estimated coefficients are larger in the standard logit regression but remain significant when we control for pairwise fixed effects. Even with fixed effects, the quantitative impact of proximity on the probability of coauthorship is large: being at a network distance of 2 instead of 3 raises the probability of initiating a collaboration by approximately 27 percent. The effect remains noticeable at larger distances. For example, being connected at a network distance of 5 instead of 6 implies that the probability of forming a link is 18 percent higher.

We continue the analysis by turning to subsequent collaborations, conditional on having collaborated once. If network proximity increases the likelihood of collaboration because of some kind of "referral" or mutual introduction 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 to be introduced or given a referral about each other. This suggests a kind of placebo experiment: if we regress repeat co-authorship on network proximity, we should observe no effect. In contrast, if network proximity is correlated with time-varying unobserved match quality, then it should remain significant for subsequent collaborations as well.

To investigate these ideas, we estimate equations (11) using data on subsequent collaborations. Results are summarized in columns (3) and (4) of Table 3. The key finding is the following: *network proximity no longer has a positive effect on coauthorship.* This finding is consistent with the network interpretation and provides 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.

In Table 3 network proximity has a significant but *negative* coefficient. This is unexpected and needs to be explained. To investigate why this is the case, we re-estimate model (11) with distance dummies instead of p_t^{ij}. Results, presented in Figure 5, 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.

The most likely explanation for this finding is in terms of time and capacity constraints.[26] Recall that, in our framework, proximity effects are identified by changes in the levels of proximity. Prior to first collaboration, as two authors move closer to a distance of 2, there are two effects at work: they are likely to get better information about each other; but also one of them has started a new collaboration and has less time available for initiating a new collaboration. The first effect has a positive influence on coauthorship, but the latter has a negative effect on the probability of forming a new tie. For first-time collaborations, our results indicated that the positive effect dominates. But 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. This explanation is consistent with the absence of significant effects at longer distances: as seen in Figure 5, a reduction in network distance from 4 to 3 does not involve an additional link by either i or j and so there is no negative effect of fall in distance.

4.2 Interpretation of control variable coefficients

The coefficients of control variables are interesting in their own right. Consider the regression results for first collaboration given in column (2) of Table 3. The coefficients for average productivity $\overline{q}_t^{ij}$ and average degree $\overline{n}_t^{ij}$ are significantly positive, suggesting that the likelihood of collaboration increases when authors become more productive or gather more coauthors. Estimated coefficients for Δq_t^{ij} and Δn_t^{ij} are negative, indicating that the likelihood of first collaboration falls when authors are more dissimilar in terms of productivity and number of coauthors.

We also note that, in line with our discussion in section 3.1, the effect of field overlap on first coauthorship follows an inverted-U curve: 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 .660, which is much higher than the average field overlap of .054 for random author pairs. Field overlap is thus associated with a higher likelihood of initiating a collaboration.

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

4.3 Interpretation of the results in terms of acquaintance network

Figure 4 shows that network effects remain significant up to distance 9. At first glance this appears too good to be true: the likelihood that two authors be introduced to each other by a chain of 9 coauthors appears remote. Perhaps the best way to make sense of these findings is in terms of the acquaintance network discussed in section 2.2. Information about coauthors is conveyed via the coauthorship network. But information also flows along other social links that are not coauthor ties – and hence are not observed. At short distances, the 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.

Secondly, $E[d_a|d_c]$ can remain relatively small even for large significant values of d_c. In other words, a significant coefficient on the distance 9 dummy does *not* imply that an unbroken chain of 9 coauthors was used to introduce two authors to each other. The distance between the two authors in the acquaintance network was in all likelihood much shorter than 9. In fact, the simulations in Section 2.2 suggest that the distance between two authors in the acquaintance network is likely to be around 2 or 3. This second point is crucial because it explains why distance dummy 9 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 2.2, we argued that one way to test for this is to introduce an interaction term $p \times \log c$, that is, 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, $\log c_{t-1}^{ij}$ becomes an increasingly inaccurate proxy for the number of shortest paths in the acquaintance network. This leads us to estimate the following regression.

$$\Pr(y_t^{ij} = 1) = \lambda(\alpha + \beta p_{t-1}^{ij} + \gamma_1 \log c_{t-1}^{ij} + \gamma_2 p_{t-1}^{ij} \log c_{t-1}^{ij} + \mu^{ij})$$

If referral takes place through an acquaintance network, then $\gamma_2 > 0$. Results are shown in column (2) of table 4. We see that the coefficient γ_2 of the interaction term is positive and significant. This evidence is consistent with the idea that coauthorship referrals circulate in an unobserved acquaintance network that is denser than the observed coauthorship network.

4.4 Information and time

Our findings strongly suggest that social proximity promotes collaboration. The mechanism is a combination of information revelation and referrals; authors that are socially closer are more likely to obtain information about each other skills and practices as well as more likely to be personally introduced to each other in social and professional gatherings. To strengthen our claim that it is information revelation that drives the results, we examine whether the role of network proximity has changed over time.

The period since the 1980's has witnessed the large scale adoption of new information technologies such as fax and electronic messaging, telephone charges have fallen, air travel has become significantly cheaper, and the world wide web has developed. These developments facilitate the access of information concerning others and expand the pool of potential collaborators. This suggests that the role of social networks in conveying information about others may be less important now than before.

To gain some understanding of this important question, we re-estimate the first collaboration regression separately for the periods before and after 1989. Estimation results, presented in Table 5, indeed show that the estimated coefficient of network proximity is smaller for the 1990s than for the 1980s, suggesting a somewhat reduced role for social proximity in recent times. The coefficient, however, remains significant, indicating that network proximity retains a role in the formation of new coauthor relations even in the present internet age.

5 Robustness

To summarize, the results reported so far support the hypothesis that the likelihood of starting a collaboration increases with network proximity, and that this is probably due to information effects. We now present several robustness checks.

5.1 Shorter duration of network link

In our main regressions, network proximity between i and j in year t is measured as the inverse distance between i and j in the network G_t. The network G_t contains a link whenever i and j have coauthored an article published between year $t-9$ and t. This implicitly assumes that a network link remains active for 10 years. To check that our results do not depend on this assumption, we repeat the analysis using network measures in which G_t has a link if i and j publish between $t-4$ and t, that is, a link persists for exactly 5 years.

Table 6 shows the results for the pairwise fixed effect regressions. It seems that the control variables are able to capture more of the variation in this case, because the proximity effect decreases by half in Table 6 when control variables are included. Nonetheless, we also observe

that our main conclusion remains: network proximity remains significantly positive in the first collaboration regression, whereas it remains significantly negative for subsequent collaborations. Our results are therefore robust to the use of different assumptions on the length of link activity.[27]

5.2 Variation in publication lags

Since our data come from a bibliographic database, we only observe a collaboration project at the time of publication, that is, at its end. A single project typically takes several months if not years to complete. On top of that, there are considerable delays between the moment a paper is submitted to a journal and the moment it is published.[28] There is therefore a considerable difference between the time at which observe a collaboration and the time it actually started. This may be cause for concern.

If project completion time and publication lags were the same for each paper, the dependent variable and all regressors would be lagged by the same number of years. In that case, the estimation results would not be affected. Unfortunately this is not the case: publication lags vary over time and across publications, and this variation perturbs the order and timing of the events. Although we regress the observations of the dependent variable, y_t^{ij}, on observations of *lagged* explanatory variables, it is therefore conceivable that some observations of the dependent variable correspond to decisions made *after* the decision corresponding to the explanatory variables.

To investigate whether this variation in publication lags affects our results, we repeat the analysis lagging explanatory variables by 3 years instead of 1 year. This additional 2 years should mitigate most of the problems of publication lags, since the variation in publication lags rarely exceeds 3 years. By lagging explanatory variables by 3 years, however, we lose a considerable amount of observations. Moreover, we expect the proximity effect to be less pronounced as time passes. The results presented in Table 7, nevertheless show that there is little difference lagging explanatory variables 1 year or 3 year. We therefore conclude that the variation of publication lags probably does not distort the conclusions of our analysis.

5.3 Non-independence across observations

Our estimation approach implicitly assumes that, conditional on a pairwise fixed effects, contemporaneous observations are independent across pairs. Unfortunately, this assumption is

[27] It should in principle be possible to estimate the strength of the ties between all pairs of authors and to measure network proximity in this *weighted* network, for instance to allow links to decay over time. Doing so would require iterating between the estimation of the regressions and the construction of the network G_t and the calculation of the proximity variables p_t^{ij}. Given how long it takes for the computer to calculate a single G_t and p_t^{ij}, such an iterative procedure would represent a massive time investment which we feel is unjustified given the relatively small potential payoff.

[28] Ellison (2002) reports a delay that has been increasing over time in top journals: on average well less than a year in the 1970s, to more than a year and up to 2 years in the 1990s.

unlikely to be entirely appropriate, given that an author who has several collaborators appears in different pairs in the dataset.

Although the pairwise fixed effect controls for individual characteristics that do not change over time, we cannot reasonably treat the pairs in which the same author appears as contemporaneously independent. In particular, time constraints on researchers should imply a negative contemporaneous correlation between pairs with the same author: if i decides to collaborate with j, then i has less time to collaborate with k.

Contemporaneous correlation in residuals does not affect the consistency of the coefficients, but it can bias the standard errors. To the best of our knowledge, there is no known method to correct for this bias within the framework of the fixed effects logit model. However, we can investigate if this contemporaneous bias is problematic by doing the following exercise. We repeat the regressions on a sub-sample of the data set in which it is assured that *each author appears in only one pair*. That is, for every two observations y_t^{ij} and y_t^{kl} we ensure that $\{i,j\} \bigcap \{k,l\} = \emptyset$. This eliminates the possibility that contemporaneous residuals are correlated simply because they have an author in common. But it results in a massive reduction in the number of usable observations since each author only appears once in the data. We therefore expect a loss in significance as a mechanical consequence of the reduction in sample size.

To investigate the magnitude of this loss in significance, we draw from the full dataset a random sample of pairs that has the same size as the sample without duplicate authors. Estimates from this regression gives a sense of the magnitude of the reduction in significance simply due to the reduction in sample size. Results, which are not presented here to save space, show that, as anticipated, significance falls somewhat in both the random sample with duplicate authors and the sample without duplicate authors. But our main results remain by and large unchanged.

5.4 Detrending method

As we mentioned in Section 2 and discussed thoroughly in Appendix A, the estimation of the first collaboration regression with pairwise fixed effects requires us to detrend all explanatory variables prior to estimating the fixed effect logit regression. In the results reported so far, we do this by using linear detrending. We nevertheless worry that results would be affected if explanatory variables followed a nonlinear trend. Simulations (not reported here) show that this is indeed the case: parameters are significantly biased if the explanatory variables have exponential or logarithmic trends but linear detrending is applied.

To investigate whether our results are robust to alternative assumptions regarding the form of the trend, we repeat our analysis using alternative detrending methods, namely, exponential,

logarithmic, and quadratic trends. We do this for the regression on first collaboration only since results for subsequent collaboration are not affected by this problem.

Table 8 shows the results. We observe that the reported coefficients for network proximity are much larger than in Table 3 when we assume an exponential trend and somewhat smaller when we assume a logarithmic or quadratic trend. Our conclusions on network proximity, however, remain.

6 Conclusions

The matching of individuals in teams to create intellectual and physical output is a key element of economic activity. Yet at any point in time there are many potential matches and the ability and skills of these individuals is only imperfectly known. Therefore finding a suitable match takes time and effort and is subject to a great deal of uncertainty. It is plausible then to expect that individuals will seek to economize on search costs by relying on social networks to access easily available information on ability and match quality. The network of past team members is particularly well suited for this purpose since past team work has revealed valuable information about others. The aim of this paper was to examine the empirical relevance of this network effect.

We studied the formation of coauthor relations among economists over a twenty year period from 1980 to 1999. Our principal finding is that a new collaboration is more likely among two researchers if they are "closer" in the existing coauthor network. This proximity effect is positive and robust and extends to network distances of up to 9 degrees of separation. At first glance the network effect appears too large to be true – referral is unlikely to travel across 9 degrees of separation. Our preferred interpretation is that distance in the co-authorship network is an informative statistic about a possibly much shorter social distance in the denser but unobserved acquaintance network.

Our empirical approach takes care of time-invariant unobserved heterogeneity among collaborating pairs as well as time varying observable heterogeneity. Moreover, we show that network proximity does not have a positive effect on subsequent collaboration between authors. This takes account of unobserved time-varying effects – such as non-measurable changes in research interests. From this evidence we conclude that existing social networks have powerful effects on the formation of new coauthor ties.

New research collaboration ties form in an environment where much public information is available on individual ability – e.g., publications record, employment history. So we would expect matching frictions to be less prevalent here than in other team formation processes. Therefore empirical significance of networks in this context suggests that it would be natural to expect social networks to significantly shape most matching processes at work in the economy.

In particular, we hope that our findings will motivate further empirical study of the role of social networks in the functioning of labor markets. One important avenue for future research is to investigate the impact of collaboration and networks on the quality of research. The evidence of network effects in the formation of collaborative teams is indicative of matching frictions, and the presence of matching frictions makes us expect inefficiency in team formation – the best matches are not achieved – and inequity in access to good collaborators – researchers isolated in a social network sense find it harder to identify suitable collaborators.

Appendix A

In this appendix we illustrate the difficulty inherent in estimating a fixed effect logit model for first collaborations, and show how detrending can be used for inference purposes.

To illustrate how identification is achieved, consider the following example. Imagine we have observations on collaborator pairs over two periods. Since we restrict our attention to first collaborations, we have $\{y_1, y_2\} = \{0, 1\}$. Let network distance d take only two values, say 2 and 3. There are only two possible data configurations:

$$\begin{bmatrix} y & d \\ 0 & 2 \\ 1 & 3 \end{bmatrix} \text{ and } \begin{bmatrix} y & d \\ 0 & 3 \\ 1 & 2 \end{bmatrix}$$

If there are an equal number of observations of each type, there is no systematic relationship between network distance d and the timing of first collaboration. In contrast, if most observations fall in the second category, the initiation of a collaboration is more likely when coauthors are closer. In this simple example, inference can be achieved by applying a simple t-test to d across the two time periods – or alternatively by using regression analysis with y as dependent variable. It is clear that this example can be generalized to more periods, more values of d, and more regressors.

To analyze this problem, we construct a Monte Carlo simulation that reproduces the kind of data we have. We begin by generating pair-wise fixed effects $u_i \sim N(0, 2000)$.[29] We then create two potential regressors x_{it} and z_{it} indexed over individual (e.g., pair of authors) i and time t. Each regressor is constructed as a trend with noise:

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

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

$$y_{it}^* = -2 + x_{it} + u_i + \varepsilon_{it} \tag{12}$$

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 (12), 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

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

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

This simple observation suggests that removing the trend in x_{it} and z_{it} should get rid of the problem. To investigate whether this is indeed the case, we estimate the following regressions:

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

and obtain $x^d_{it} = x_{it} - \widehat{\gamma}_x t$ and $x^d_{it} = x_{it} - \widehat{\gamma}_z t$. We then regress y_{it} on x^d_{it} and z^d_{it}. If detrending solves the spurious correlation problem, coefficient estimates and inference should be similar to the results obtained in the first panel of Table 9. For the sake of comparison, we also regress y^a_{it} on x^d_{it} and z^d_{it}.

Results are presented in Table 10. They show that in the Monte Carlo simulation detrending eliminates the bias in both coefficients in the y_{it} – i.e., first collaboration – regression while keeping things basically unchanged in the y^a_{it} – i.e., repeated collaboration – regression. There is a large loss of precision between the y^a_{it} regression and the detrended y_{it} regression. But this is a mechanical consequence of the way we generated the data, which leads us to throw away all observations of y^a_{it} after the first 1 realization.

As a cure to our estimation problem, detrending is not without side-effects. This is because detrending reduces the variation in x, and hence the amount of information that can be used to identify its coefficient. Table 10 illustrates what happens in the best of cases. Compare the detrended and un-detrended regressions using repeated collaboration data. For these data, the data generation process is such that the un-detrended regressions yield consistent estimates.

We see that detrending leads to a loss of precision – the Monte Carlo sample variance of the x coefficient increases as a result of detrending.

Detrending can also introduce an attenuation bias which is somewhat analogous to what happens in fixed effect linear regression models. If the data generation process is such that the average duration to first collaboration is very short, or if most of the variation in x is persistent over time, then much of the variation in x_{it} is eliminated after detrending. As a result, the coefficient of x in (12) is biased towards 0. This is confirmed by Monte Carlo simulations. However, the attenuation bias also means that we can nearly never reject the null hypothesis that the coefficient of x is zero. Inference has low power and is thus biased towards failing to reject the null. What this means is that, if we can reject the null hypothesis after detrending, the likelihood of a type I error is small – smaller on average than the reported p-value. This is also confirmed by Monte Carlo simulations: depending on the data generation process, detrending can entail a loss of power and may fail to reject the null hypothesis even when it is false. But it nearly never result in a type I error, that is, rejecting the null when it should not be rejected. In other words, hypothesis testing based on detrending is too conservative.

Appendix B

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] Akerlof, G. (1970), The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, 84, 3, 488-500.

[2] Allison, P.D. and Christakis, N.A. (2006), Fixed effects methods for the analysis of non-repeated events. *Sociological Methodology*, 36, 1, 155-172.

[3] 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.

[4] Bertrand, M., E.F.P. Luttmer, and S. Mullainathan (2000), Network effects and welfare cultures, *Quarterly Journal of Economics*, 115, 3, 1019-1055.

[5] Bramoullé, Y., H. Djebbari, and B. Fortin (2008), Identification of peer effects through social networks, mimeo, Université Laval.

[6] Brock, W. and S. Durlauf (2001), *Interaction-based Models*, in J. Heckman and E. Leamer (eds), *Handbook of Econometrics*, Volume 5, Amsterdam: North Holland.

[7] Casella, A. and J. Rauch (2002), Anonymous market and group ties in international trade, *Journal of International Economics*, 58, 1, 19-47.

[8] Calvó-Armengol, A., E. Patacchini and Y. Zenou (2008), Peer effects and social networks in education, *Review of Economic Studies*, forthcoming.

[9] Comola, M. (2008), The network structure of informal arrangements: Evidence from rural Tanzania. mimeo, Universitat Pompeu Fabra.

[10] Conley, T. and C. Udry (2008), Learning about a new technology: Pineapple in Ghana, Economic Growth Center, Yale University.

[11] Diamond, P. (1982), Aggregate demand management in search equilibrium, *Journal of Political Economy*, 90, 5, 881-894.

[12] 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.

[13] Ellison, G. (2002), The slowdown of the economics publishing process, *Journal of Political Economy*, 110, 5, 947-993.

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

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

[16] Glaeser, E., B. Sacerdote and J. Scheinkman (1996), Crime and Social Interactions, *Quarterly Journal of Economics*, 111, 2, 507-548.

[17] 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.

[18] Goyal, S. (2007), *Connections: an introduction to the economics of networks.* Princeton University Press, Princeton, New Jersey.

[19] Granovetter, M. (1985), Economic Action and Social Structure: The Problem of Embeddedness, *American Journal of Sociology,* 91, 3, 481-510.

[20] Granovetter, M. (1995), *Getting a Job: A Study of Contacts and Careers*, 2nd edition, Chicago: University of Chicago Press.

[21] 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.

[22] Hidalgo, A., Klinger, B., Barabási, A-L., Hausmann, R. (2007), The Product Space Conditions the Development of Nations, *Science*, 317, 5837, 482 - 487.

[23] Hudson, J. (1996), Trends in multi-authored papers in economics, *Journal of Economic Perspectives*, 10, 3, 153-158.

[24] Jackson, M. and B. Rogers (2007), Meeting strangers and friends of friends: How random are social networks?, *American Economic Review*, 97(3), 890-915.

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

[26] Kalaitzidakis, P., T. Mamuneas, and T. Stengos (2003), Rankings of academic journals and institutions in economics, *Journal of European Economic Association*, 1, 6, 1346-1366.

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

[28] Manski, C. (1993), The Identification of endogenous social effects: the reflection problem, *Review of Economic Studies*, 60, 3, 531-542.

[29] Mayer, A. and S.L. Puller (2008), The old boy (and girl) network: Social network formation on university campuses, *Journal of Public Economics*, 92, 1-2, 329-347.

[30] McDowell, J.M. and M. Melvin (1983), The determinants of co-authorship: An analysis of the economics literature, *Review of Economics and Statistics*, 65, 1, 155-160.

[31] Moffitt R. (2001), Policy interventions, low-level equilibria, and social interactions, in: S. Durlauf and P. Young (eds.), *Social Dynamics*, Cambridge: MIT Press.

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

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

[34] Munshi, K. and M. Rosenzweig (2006), Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy, *American Economic Review*, 96, 4, 1225–1252.

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

[36] 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.

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

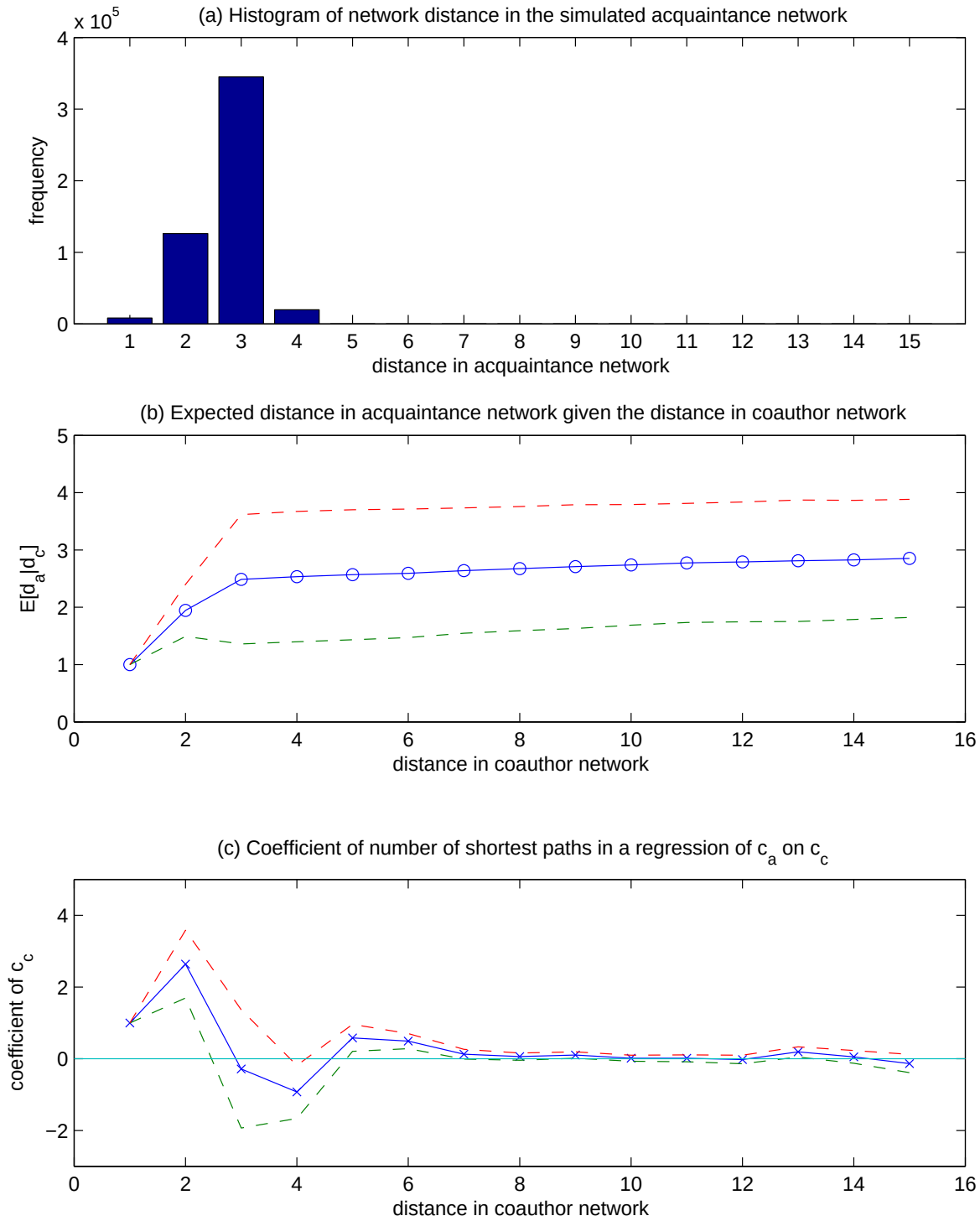
[38] Topa, G. (2001), Social Interactions, Local Spillovers, and Unemployment, *Review of Economic Studies*, 68(2): 261-95

[39] Vigier, A. (2008), Globalization, Education, and the Topology of Social Networks, mimeo, University of Cambridge.

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

[41] Wuyts, S., M.G. Colombo, S. Dutta, and B. Nooteboom (2005), Empirical tests of optimal cognitive distance, *Journal of Economic Behaviour & Organization*, 58, 2, 277-302.

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



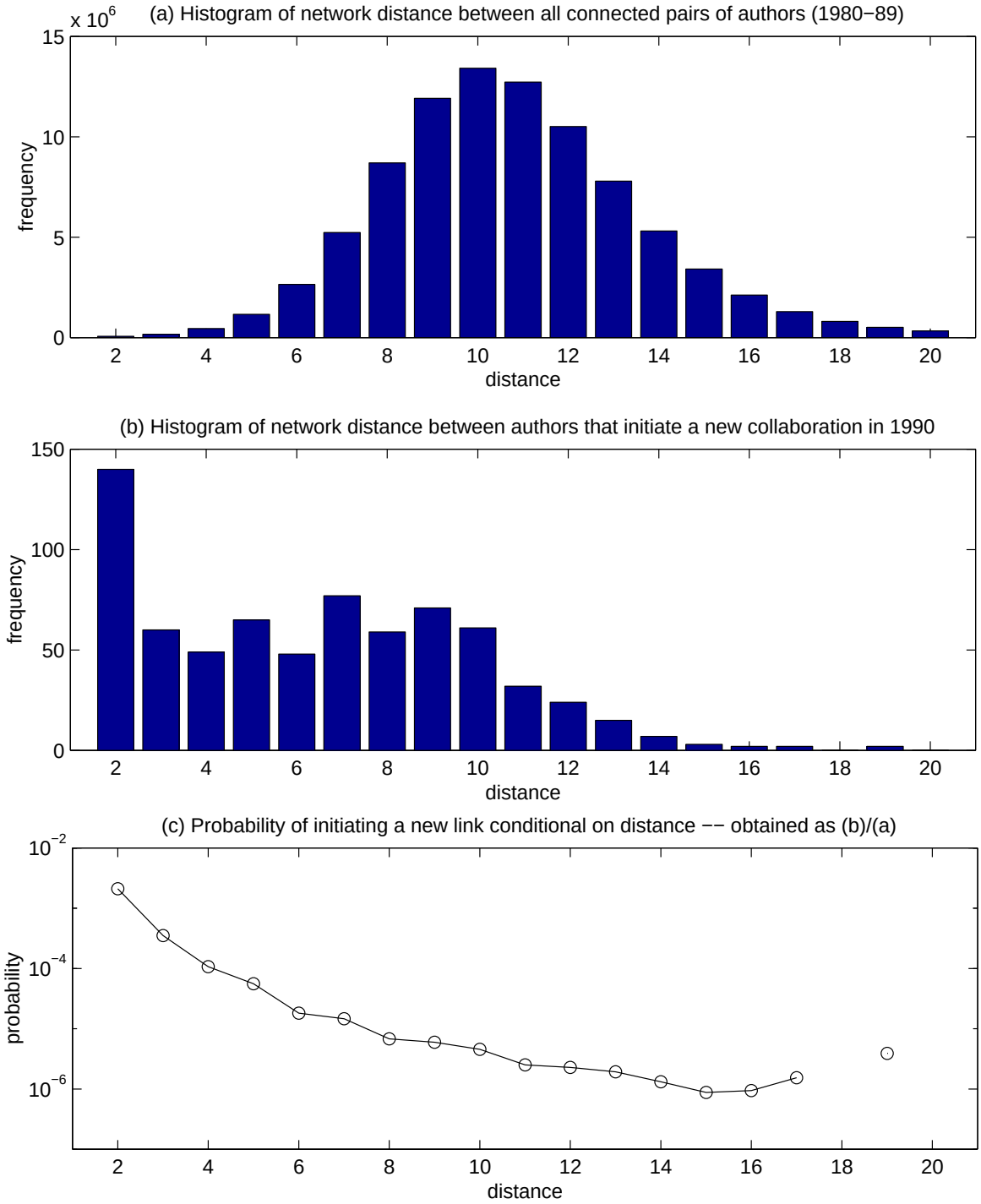


Figure 2: Histogram of distance in the network of the 1980s and the formation of links in 1990.

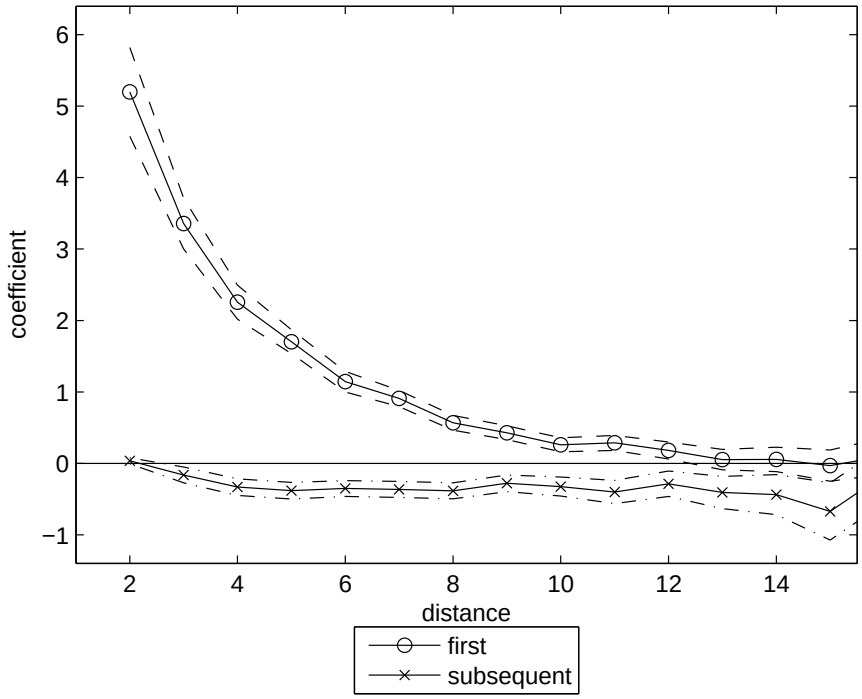


Figure 3: Coefficients of distance dummies in regressions on first collaboration and subsequent collaboration, estimated with a logit estimator without controlling for pairwise fixed effects.

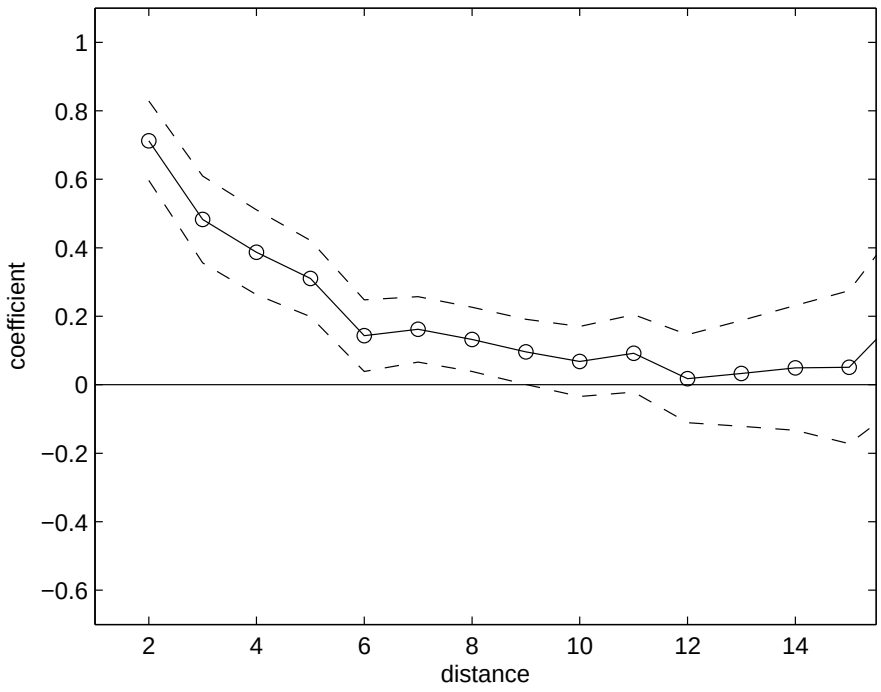


Figure 4: Coefficients of distance dummies fixed-effects logit regression on first collaboration.

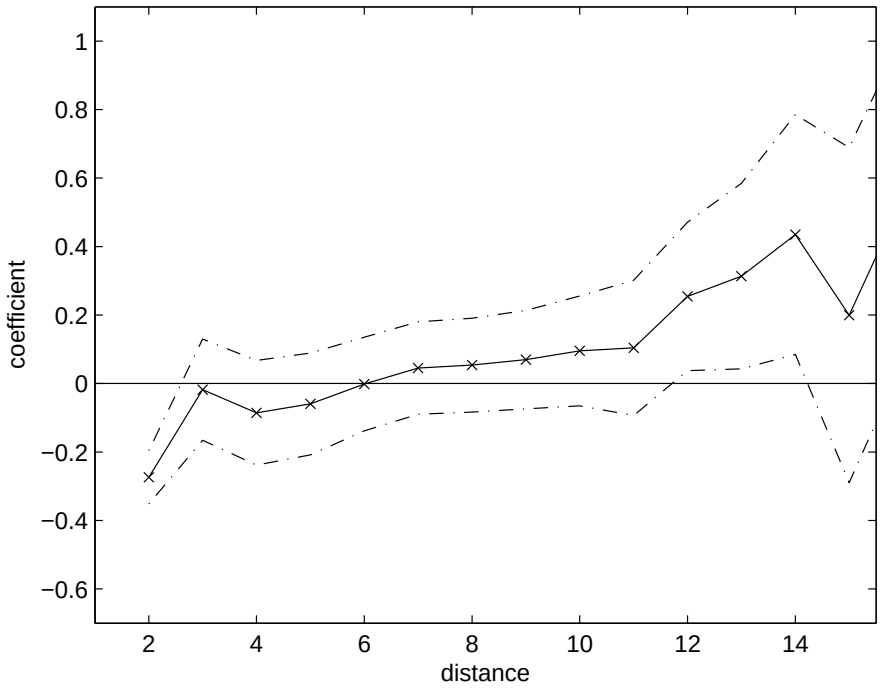


Figure 5: Coefficients of distance dummies in fixed-effects logit regression on subsequent collaborations.

Table 1: Summary statistics of the data

Variable	Random sample	Before collaboration	After collaboration
Number of pairs	162166	26922	14558
Number of observations	921392	160339	105854
Duration to first collab.		5.96	
		(3.80)	
Subsequent collaboration			.239
			(.427)
Proximity	.021	.086	.276
	(.046)	(.133)	(.227)
Connected	.189	.428	.686
Distance if connected	9.91	7.06	3.50
	(2.84)	(3.67)	(2.82)
Number of shortest paths	.545	.902	1.11
	(1.81)	(1.80)	(1.44)
Avg. productivity	20.50	44.70	65.93
	(33.75)	(60.09)	(87.23)
Dif. in productivity	32.06	50.92	57.78
	(59.37)	(80.65)	(94.27)
Avg. number of coauthors	2.01	3.19	5.53
	(1.88)	(2.63)	(3.61)
Dif. in number of coauthors	2.33	3.32	4.10
	(2.77)	(3.70)	(4.37)
Field overlap	.054	.305	.631
	(.143)	(.315)	(.264)

Notes: for each variable and dataset the sample mean and (in parentheses) the standard deviation are shown. *Duration to first collab.*: the duration to the first collaboration of a pair. *Subsequent collaboration*: number of subsequent collaboration after the first collaboration. *Connected*: fraction of pairs that are directly or indirectly connected to each other. Other variables are explained in the text.

Table 2: Results of logit regression on first collaboration and subsequent collaboration without controlling for pairwise fixed effects.

Regression	First collaboration		Subsequent collaboration	
Pairs	195255	195525	40263	40263
Observations	1094861	1094861	242595	242595
Proximity	13.885**	10.387**	.239**	.227**
	(.339)	(.584)	(.037)	(.041)
Log Shortest Paths	-.230**	-.200**	.234**	.249**
	(.020)	(.022)	(.020)	(.020)
Avg. productivity		.0121**		.0020**
		(.0009)		(.0002)
Dif. in productivity		-.0065**		-.0016**
		(.0005)		(.0002)
Avg. number of coauthors		-.072**		-.031**
		(.024)		(.006)
Dif. in number of coauthors		.089**		.029**
		(.011)		(.003)
Field overlap		6.145**		1.636**
		(.230)		(.131)
Sq. Field overlap		-2.856**		-.701**
		(.328)		(.112)
Career time	-.032**	-.052**	-.087**	-.079**
	(.004)	(.004)	(.002)	(.002)
Intercept	-12.751**	-13.371**	-1.591**	-2.311**
	(.018)	(.023)	(.016)	(.039)

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

Regression	First collaboration		Subsequent collaboration	
Pairs	26922	26922	14558	14558
Observations	160339	160339	105854	105854
Proximity	1.970**	1.427**	-1.452**	-.588**
	(.100)	(.111)	(.071)	(.078)
Log Shortest Paths	.067**	.048**	-.204**	-.068**
	(.018)	(.018)	(.024)	(.024)
Avg. productivity		.0013**		-.0041**
		(.0005)		(.0004)
Dif. in productivity		-.0007*		.00174**
		(.0003)		(.0003)
Avg. number of coauthors		.087**		-.097**
		(.011)		(.008)
Dif. in number of coauthors		-.012*		.030**
		(.006)		(.005)
Field overlap		.841**		-1.718**
		(.141)		(.223)
Sq. Field overlap		-.637**		-.011
		(.160)		(.192)

Table 4: Results of fixed effects logit regression on first collaboration: the effect of the interaction of log number of shortest paths and proximity.

Regression	First collaboration	
Pairs	26922	26922
Observations	160339	160339
Proximity	1.908**	1.409**
	(.101)	(.111)
Log Shortest Paths	-.104**	-.068
	(.038)	(.038)
Avg. productivity		-.0013*
		(.0005)
Dif. in productivity		-.0006
		(.0003)
Avg. number of coauthors		.082**
		(.011)
Dif. in number of coauthors		-.011
		(.006)
Field overlap		.848**
		(.141)
Sq. Field overlap		-.646**
		(.160)
Proximity × Log Shortest Paths	1.493**	1.022**
	(.294)	(.297)

Table 5: Results of fixed effects logit regression on first collaboration and subsequent collaboration in the 1980s and 1990s.

Regression	First collaboration		Subsequent collaboration	
	1980s	1990s	1980s	1990s
Pairs	7732	17829	4776	10763
Observations	34651	81653	27661	58775
Proximity	1.014**	.598**	-1.149**	-1.169**
	(.226)	(.144)	(.166)	(.114)
Log Shortest Paths	.018	.014	-.189**	-.070*
	(.042)	(.021)	(.057)	(.031)
Avg. productivity	-.0032**	-.0005	-.0074**	-.0083**
	(.0011)	(.0009)	(.0010)	(.0007)
Dif. in productivity	.0011	-.0011*	.0014*	.0030**
	(.0007)	(.0005)	(.0007)	(.0005)
Avg. number of coauthors	.006	.010	-.107**	-.160**
	(.026)	(.015)	(.022)	(.013)
Dif. in number of coauthors	-.009	.010	.026*	.048**
	(.015)	(.008)	(.013)	(.007)
Field overlap	-.263	.142	-4.921**	-3.970**
	(.288)	(.195)	(.505)	(.354)
Sq. Field overlap	.306	-.035	1.339**	-.955**
	(.336)	(.218)	(.418)	(.295)

Table 6: Results of fixed effects logit regression on first collaboration and subsequent collaboration assuming that links are active for 5 years.

Regression	First collaboration		Subsequent collaboration	
Pairs	25047	25047	14829	14829
Observations	144612	144612	103672	103672
Proximity	2.511**	1.189**	-.736**	-.264**
	(.097)	(.104)	(.056)	(.063)
Log Shortest Paths	.092**	-.005	-.164**	-.062*
	(.024)	(.025)	(.028)	(.029)
Avg. productivity		.0109**		-.0043**
		(.0007)		(.0006)
Dif. in productivity		-.0035**		.0015**
		(.0004)		(.0004)
Avg. number of coauthors		.260**		-.079**
		(.012)		(.009)
Dif. in number of coauthors		-.058**		.023**
		(.007)		(.006)
Field overlap		.951**		-.019
		(.124)		(.158)
Sq. Field overlap		-.640**		-.748**
		(.141)		(.140)

Table 7: Results of fixed effects logit regression on first collaboration and subsequent collaboration, in which all explanatory variables are lagged 3 years relative to the dependent variable.

Regression	First collaboration		Subsequent collaboration	
Pairs	17941	17941	8153	8153
Observations	100969	100969	57538	57538
Proximity	1.963**	1.420**	-1.217**	-.592**
	(.133)	(.147)	(.096)	(.104)
Log Shortest Paths	.056*	.037	-.195**	-.079*
	(.022)	(.023)	(.032)	(.033)
Avg. productivity		.0020**		-.0046**
		(.0007)		(.0005)
Dif. in productivity		-.0004		.0021**
		(.0004)		(.0004)
Avg. number of coauthors		.071**		-.096**
		(.014)		(.011)
Dif. in number of coauthors		-.008		.016*
		(.008)		(.007)
Field overlap		1.306**		-.706*
		(.176)		(.313)
Sq. Field overlap		-1.091**		-.070
		(.201)		(.265)

Table 8: Results of fixed effects logit regression on first collaboration with different detrending methods.

Regression	Exponential trend		Logarithmic trend		Quadratic trend	
Pairs	26922	26922	26922	26922	26922	26922
Observations	160339	160339	160339	160339	160339	160339
Proximity	12.747**	3.736**	1.382**	1.049**	1.344**	1.165**
	(.194)	(.193)	(.098)	(.109)	(.099)	(.110)
Log Shortest Paths	.379**	.174**	-.012**	.002	.025	.020
	(.018)	(.021)	(.018)	(.018)	(.018)	(.018)
Avg. productivity		.0250**		-.0037**		-.0029**
		(.0011)		(.0005)		(.0005)
Dif. in productivity		-.0101**		.0004		.0002
		(.0006)		(.0003)		(.0003)
Avg. number of coauthors		1.469**		.103**		.064**
		(.020)		(.010)		(.011)
Dif. in number of coauthors		-.162**		-.010		-.005
		(.010)		(.006)		(.006)
Field overlap		4.687**		-.114		.206
		(.217)		(.140)		(.140)
Sq. Field overlap		-2.838**		-.175		-.132
		(.233)		(.159)		(.159)

Table 9: Monte Carlo results without detrending.

	E[coef]	σ[coef]	% significant
A. y^a_{it} 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 10: Monte Carlo results with detrending.

	E[coef]	σ[coef]	% significant
A. y^a_{it} is the dependent variable			
coefficient of x^d_{it}	0.085	0.009	100%
coefficient of z^d_{it}	0.000	0.008	5%
Number of observations	2000		
B. y_{it} is the dependent variable			
coefficient of x^d_{it}	0.089	0.025	98%
coefficient of z^d_{it}	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.