TI 2004-001/1

Tinbergen Institute Discussion Paper

Economics: An Emerging Small World?

Sanjeev Goyal[1,2]

Marco van der Leij[1]

José Luis Moraga-González[1]

[1] *Faculty of Economics, Erasmus Universiteit Rotterdam, and Tinbergen Institute,*

[2] *University of Essex.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Economics: an emerging small world?

Sanjeev Goyal* Marco van der Leij†
José Luis Moraga-González‡

Preliminary version: November 2003

Abstract

This paper examines the small world hypothesis. The first part of the paper presents empirical evidence on the evolution of a particular world: the world of journal publishing economists during the period 1970-2000. We find that in the 1970's the world of economics was a collection of islands, with the largest island having about 15% of the population. Two decades later, in the 1990's the world of economics was much more integrated, with the largest island covering close to half the population. At the same time, the distance between individuals on the largest island had fallen significantly. Thus we believe that economics is an *an emerging small world.*

What is it about the network structure that makes the world small? An exploration of the micro aspects of the network yields three findings: one, the average number of co-authors is very small but increasing; two, the distribution of co-authors is very unequal; and three, there exist a number of 'stars', individuals who have a large number of co-authors (25 times the average number) most of whom do not write with each other. Thus the economics world is a set of *inter-connected stars.*

We take the view that individuals decide on whether to work alone or with others; this means that individual incentives should help us understand why the economics world has the structure it does. The second part of the paper develops a simple theoretical model of co-authorship. The main finding of the model is that in the presence of productivity differentials and a shortage of high productivity individuals, inter-connected stars will arise naturally in equilibrium. Falling costs of communication and increasing credit for joint research leads to greater co-authorship and this is consistent with the growth in the size of the giant component.

*University of Essex and Tinbergen Institute. E-mail: sgoyal@essex.ac.uk
†Tinbergen Institute Rotterdam. E-mail: mvanderleij@few.eur.nl
‡Erasmus University Rotterdam, Tinbergen Institute and CESifo. E-mail: moraga@few.eur.nl
This paper has been presented at Amsterdam, Essex, Marseille, UCL, SAET 2003 and the VI Summer School of San Sebastián. We thank Venkataraman Bhaskar, Ken Burdett, Alberto Bisin, Francis Bloch, Alessandra Casella, Andrea Galeotti, Christian Ghiglino, Guillaume Haeringer, Alan Kirman, Otto Swank and participants at these presentations for useful comments.

1 Introduction

The idea of a *small world* is a very simple one. Tom and Sara are said to be at a distance of 1 if they know each other and they are at a distance 2 if they do not know each other but have a mutual acquaintance and so on. The world is said to be small if the average distance between people is very small relative to the size of the population. The order of numbers in this claim is important: for instance, in a world of 5 billion people the average distance between people may well be around 6 (leading to the expression 'six degrees of separation'). The idea that almost all people in the world are socially very close to each other is a very intriguing one. It is also potentially a very important one as the structure of information flows and interaction between individuals has profound implications for a wide range of economic and social phenomena.[1] In this paper we first examine the empirical content of the small world idea in the world of economists and then develop a simple individual incentives based model to help understand the factors that make the economics world small.

Our empirical work looks at the world of economists who publish in journals and we examine the evolution of this world over a thirty year period, from 1970 to 2000. We split this period into three ten year intervals, 1970-1979, 1980-1989, and 1990-1999. For each ten year period we consider all papers published in journals listed by EconLit. This data set contains a very large amount of information on a variety of variables. In this paper our interest is on three sorts of variables: the names of economists who have published at least one paper in one of the journals in the list, whether these papers are co-authored or single authored and the journal where the paper was published. We use this information to map a network of collaboration for each ten year period: every publishing author is a node in the network, and two nodes are linked if they have published a paper or more together in the period under study. Our interest is in the properties of this network of collaboration and the changes in these properties over time.

We start with an empirical examination of the question: is the world of economics small? We identify three features of the network that together provide an answer to this question. The *first* feature is the growth of the network: we find that the world of economists has grown sharply in numbers and has more than doubled in the period from 1970 to 2000. This finding is consistent with the growth in the number of fields/specializations and in the corresponding set of field journals during this period and it leads us to expect that the world has probably become 'larger'. The *second* finding is about the interconnectedness

[1] In recent years a vast literature has appeared on the role of interaction structures. A variety of terms such as local interaction, network effects, peer group effects, have been used in this literature. See e.g., Bala and Goyal (1998, 2001), Ellison and Fudenberg (1993) on social learning, Goyal (1996) and Morris (2000) on norms of coordination, Eshel, Samuelson and Shaked (1998) on norms of cooperation, Allen and Gale (2000) on financial fragility of different inter-bank loan networks, Glaeser, Sacerdote and Scheinkman (1996) on local interaction and crime, Hagerstrand (1969) and Coleman (1966) on technological diffusion, Watkins (1991) on spread of norms in fertility and marriage and Young (1998) on spread of norms on driving.

of the publishing world. In the 1970's the largest group of interconnected economists comprised only about 15% of the population, and there was a large number of small groups. In contrast, in the 1990's there was one huge group of interconnected economists with about 40% of the total population and all the other groups had shrunk in size sharply. The numbers are worth mentioning here: in the 1970's the largest group of economists contained about 5,200 economists while in the 1990's the largest group contained more than 33,000 members. The *third* finding is about the structure of this giant group of economists. We find that the average distance in this giant group is small and has fallen significantly over time (despite the growth of the group), while the level of clustering remains high. We summarize these findings as follows: the economics world is *an emerging small world*.

We then turn to an empirical examination of the structural reasons for the emerging small world properties. What is it about the number and distribution of links in the network that makes this world small? Our analysis of the microstructure of the collaboration network yields several interesting findings. The *first* finding is about the average number of links: we find that the average number of co-authors is very low but has been increasing sharply over the period from the 1970's to the 1990's; this increase seems to be in excess of 75%. The *second* set of findings is about the distribution of links. We find that the distribution is very skewed and unequal; in particular, there exists a fat tail in the distribution, with a significant number of authors having a very large number of links. To get an idea of the figures here we note that in the 1990's the per capita number of collaborators was 1.672 while the maximally connected economist had more than 50 collaborators. Moreover, the clustering in the network in the 1990's was .157 while the clustering for the most connected economist was only .02. This shows that the most connected player had many more links than his cohorts and also that he had very low overlaps among their co-authors as compared to the average person in the network. These numbers lead us to use the term 'stars' for the most connected economists. We next note two striking features about the stars: the output of the most connected economists was very large as compared to the average and most of this output was co-authored. So, for instance, in the 1990's, we find that the 100 most linked economists produced an average of 38 papers (approximately) and almost 85 percent were co-authored; by contrast, the average number of papers for all economists was 2.8 and 40 percent of these were co-authored. Our *third* finding is about the role of these well connected nodes in integrating the network. One way to study this is to examine the impact of deleting the most connected nodes. In the 1990's over 40% of the nodes were in the giant component but a deletion of the 5% most connected nodes leaves less than 1% of the nodes in the giant component, thus completely destroying the network.

Our *final* finding is that there exists a hierarchical structure in the collaboration network: nodes can be more or less partitioned into three levels, and we find that nodes in the lowest level connect only to nodes in the middle level, which in turn connect only to nodes in the top level. Moreover, there are almost no connections across nodes at the

same level and between the top and bottom levels. These findings put together lead us to the view that the *economics network is spanned by a hierarchy of inter-connected stars.*

These empirical findings are fascinating and we would like to develop a theory to account for them. There is a large literature on the small world phenomenon in physics and mathematics. This work takes as a given that the world is small; our empirical work, however, shows that average distances and size of giant component in our network change greatly over time. We therefore need a theoretical approach which can explain the stable architectural features of the network (the existence of stars, short average distance, and high clustering) as well as the changes in the network (such as growing giant component, falling average distance and increasing degree).[2]

In the second part of this paper we develop an incentives based model with the following features. Research papers contain ideas and involve routine work; the quality of a paper depends on the quality of ideas contained in it. Individuals have ideas and can do routine work; however, some people are better at generating high quality ideas than others. Institutions reward individuals on their research output; this reward specifies a threshold level of output quality that is considered for evaluation and also specifies a certain credit to single authored work and co-authored work. There are costs to writing papers which increase in the number of papers written in a research area. Similarly there are costs to meeting and working with others which are increasing in the number of co-authors and which are also sensitive to the (network) distance between the authors. Analysis of this model tells us that stars – which embody unequal distribution of links and links between well connected and poorly connected players – arise naturally in a world with productivity differentials and shortage of high productivity players. We show that equilibrium networks will contain inter-linked stars and hence will exhibit short average distances. We also study the role of two economic developments: a decline in costs of communication and an increase in credit to co-authored papers. We find that they both lead to an increase in the number of collaborators and therefore an equilibrium network with a higher degree, something which helps explain the growth in the size of the giant component.

In recent years there has been a virtual explosion of research in the field of networks. We briefly discuss two related bodies of work in economics and in physics, respectively.[3] The theory of network formation is a very active field of research in economics; recent work includes Bala and Goyal (2000), Jackson and Wolinsky (1996), Kranton and Minehart (2001), among others. Dutta and Jackson (2003) provides a collection of some of the main papers and Goyal (2003), Jackson (2003) and Van den Nouweland (2003) provide

[2]We discuss the work in physics in detail in Section 5.

[3]The idea of a 'small world' has a long history in sociology. Theoretical work on it was initiated in the 1960's by Pool and Kochen (this work was published in 1978) and experimental work was initiated by Milgram (1967). Systematic empirical work on this subject has been hampered by the great difficulties in collecting reliable data sets on large social networks. Theoretical work also seems to have progressed slowly. For an interesting general account of the developments in this field, see Watts (1999) and for a collection of theoretical papers until the late 1980's, see Kochen (1989).

surveys on the developing field. The most distinctive aspect of this work is the use of individual incentives to derive predictions on network architectures. This work has been almost entirely theoretical and our paper's contribution is to introduce the empirical study of large evolving networks in economics.

The empirical properties of large social and economic networks have been investigated extensively by physicists in recent years; see e.g., Barabási and Albert (1999), Barabási, Albert and Jeong (1999), Newman (2001a,2001b,2001c, 2003) and Watts and Strogatz (1998). For comprehensive surveys of the work in physics see Albert and Barabási (2002) and Dorogovtsev and Mendes (2002). This work focuses on the statistical properties of large networks, and uses a variety of techniques ranging from random graph theory to mean field analysis to elaborate on different features of observed networks. This work is discussed in detail in section 5. Here we briefly discuss two ways in which our paper contributes to this body of work. The first contribution is the incentives based approach we develop. We believe that networks of scientific collaboration are an outcome of deliberate decisions by individual economists. This means that the observed networks reflect the technology of production of knowledge and the incentives faced by individuals. We are thus interested in developing a model where technology and incentive schemes are modelled explicitly and we can study their effects on collaboration networks systematically. The second contribution of our paper is our empirical work; several aspects of it are novel. To the best of our knowledge our paper is the first to study economics collaboration networks; existing work concerns the natural sciences, medical sciences and mathematics. This is interesting since these networks seem to exhibit different properties. For example, the relative size of the giant component and the number of co-authors seem to be very different in economics as compared to physics or medical sciences. Moreover, we look at the properties of the network over an extended period of time (thirty years) and this allows us to study the impact of different technological and economic factors on the network, something which has not been done before since earlier studies have looked at the network over relatively short periods of time (the maximum period of time covered seems to be 8 years, see Albert and Barabási (2002)). This difference in time horizon allows us to show the small world property is not a 'constant' of social networks; we find that in the past the economics world was fragmented into many groups and was not small, but that it is slowly getting integrated and is becoming a small world.

Our paper is also a contribution to the literature on different aspects of economics research; recent work includes Ellison (2002a, 2000b) and Laband and Tollison (2000) among others. In particular, the increase in co-authorship has been noted and the reasons for it have been explored in Hudson (1996), while the role of substantial and increasing informal intellectual collaboration is explored in Laband and Tollison (2000). Hamermesh and Oster (2002) present evidence which suggests that collaboration among distant authors has increased over the years. A variety of arguments – such as increasing specialization and the falling costs of communication among others – have been proposed to explain increasing co-authorship among economists.

The rest of this paper is organized as follows. Section 2 presents basic notation and definitions. Section 3 presents the empirical trends. Section 4 develops an incentives based model to explain these network patterns while section 5 discusses alternative explanations. Section 6 contains concluding remarks.

2 Networks

We start by setting down some basic notation which is useful to discuss network features precisely. Let $N = \{1, 2, .., n\}$ be the set of nodes in a network. We shall refer to n as the *order* of the network. We shall be looking at undirected links in this paper, and for two persons/nodes $i, j \in N$, we shall define $g_{i,j} \in \{0, 1\}$ as a link between them, with $g_{i,j} = 1$ signifying a link and $g_{i,j} = 0$ signifying the absence of a link. If two persons have published a paper together then they are said to have a link between them; if they have published no papers together then they have no link. Thus the information on authors and papers allows us to construct a network of collaboration. We shall say that there is a path between i and j either if $g_{i,j} = 1$ or if there is a set of distinct intermediate co-authors $j_1, j_2...j_n$, such that $g_{i,j_1} = g_{j_1,j_2} = ... = g_{j_n,j} = 1$. The collection of all links will be denoted by g. The set of nodes and the links between them will be referred to as a network and denoted by $G(n, g)$. We shall define $\mathcal{N}_i(G) = \{j \in N : g_{i,j} = 1\}$, as the set of people with whom i has a link in network G and $\eta_i(G) = |\mathcal{N}_i(G)|$.

Two persons belong to the same component if and only if there exists a path between them. The path relation therefore defines a partition of the network into components. For a network G the partition will be denoted as $P(G) = \{C_1, ..., C_m\}$ with $m \geq 1$. In case $m = 1$ we have a connected network and in case $m = n$ we have the empty network. The components can be ordered in terms of their size, and we shall say that the network has a giant component if the largest component fills a relatively large part of the graph and all other components are small, typically of order $\mathcal{O}(\log n)$, where n is the total number of vertices.

The geodesic distance between two nodes i and j in network G is the length of the shortest path between them, and will be denoted by $d(i, j; G)$. If there is no path between i and j in a network G then we shall set $d(i, j; G) = \infty$. In case G is connected, the average distance between nodes of a network G is given by

$$d(G) = \frac{\sum_{i \in N} \sum_{j \in N} d(i, j; G)}{n(n-1)}$$

If G is not connected then the average distance is formally speaking infinite. In our data the network is not connected and so to study distances we shall use the average distance in the giant component as a proxy for the average distance in the network. The maximum

distance between any pair of nodes in a network G is referred to as the diameter of the network and it is given by

$$D(G) = \max_{i,j \in N} d(i,j;G)$$

Another measure of short distances, which is defined both for connected and unconnected networks, is the average size of the k-neighborhood. The k-neighborhood for an individual i is defined as

$$\mathcal{N}_i^k(G) = \{j \in N \setminus \{i\} : d(i,j;G) \leq k\}$$

while the size of i's k-neighborhood is given by $\psi_i^k(G) = |\mathcal{N}_i^k(G)|/(n-1)$. Hence $\psi_i^k(G)$ is the fraction of people in the network who are at k or less degrees of separation from individual i. The average size of the k-neighborhood is then $\psi^k(G) = \sum_i \psi_i^k(G)/n$. This statistic gives an estimate of the probability that two random nodes are connected within k degrees of separation.

The clustering coefficient of a network G is a measure of the correlation between links of different individuals. The level of clustering in an individual i's neighborhood is given by

$$C_i(G) = \frac{\sum_{l \in N_i(G)} \sum_{k \in N_i(G)} g_{l,k}}{\eta_i(\eta_i - 1)}$$

for all $i \in N' \equiv \{i \in N : \ \eta_i \geq 2\}$, This ratio tells us what percentage of a person's neighbors are neighbors of each other. The clustering coefficient for the network G can be obtained by averaging across all persons in a network. We shall use an averaging scheme which weights different persons by the size of their neighborhood. This leads us to the following definition for the clustering coefficient:

$$C(G) = \frac{\sum_{i \in N'} \sum_{l \in N_i} \sum_{k \in N_i} g_{l,k}}{\sum_{i \in N'} \eta_i(\eta_i - 1)}$$

We will sometimes refer to η_i as the degree of node i. In case $\eta_1 = \ldots = \eta_n = \eta$ we will refer to η as the degree of the network. In general the degree is not constant across nodes/individuals and there will be inequality in the number of links across individuals. To measure this inequality we will compute Lorenz curves and Gini coefficients for the degree distribution. Suppose the set of nodes $S \subset N$ is ordered, such that $i < j$ if and only if $\eta_i < \eta_j$ for $i, j \in S$, and denote $n_s = |S|$ as the number of nodes in S and $L_S(h) = \sum_{i=1}^{h} \eta_i$ as the number of links in possession of the h least linked nodes. Then the Lorenz curve for S is given by connecting the points

$$(h/n_s, L_S(h)/L_S(n_s)) \in [0,1]^2.$$

for $h = 0, \ldots, n_s$. The Lorenz curve measures the fraction of links that are in possession of the $x\%$ least linked nodes. Note that perfect equality, that is a constant degree across nodes in S, implies that the Lorenz curve follows the 45 degree diagonal.

The Gini coefficient G_S measures the area between the Lorenz curve and the 45 degree diagonal. That is

$$G_S = 1 - \frac{1}{n_s} \sum_{h=1}^{n_s} \frac{L_S(h) + L_S(h-1)}{L_S(n_s)}.$$

An alternative way to write this is in terms of the relative mean difference:

$$G_S = \frac{\sum_{i=1}^{n_s} \sum_{j=1}^{i-1} (\eta_i - \eta_j)}{(n_s - 1) L_S(n_s)}.$$

We note that the Gini coefficient equals 0 in case of perfect equality and 1 in case of perfect inequality, that is when only one individual forms all the links in the network (an impossibility in graph with two-sided links). A higher value of the Gini coefficient is interpreted as greater inequality in the degree distribution.

A star network is a network where one node, referred to as the center node, is linked to all other nodes in the graph while all these other nodes are only linked to the center node. In some networks, there is no single center of the network, but there are a small number of extremely well connected nodes each of whose partners are almost solely connected to them. We shall, somewhat informally, refer to these well connected nodes as 'stars'.

3 Empirical Patterns

We study the world of economists who published in journals which are included in the list of EconLit. We cover all journal papers that appear in a 10 year window and we look at three such windows: 1970-1979, 1980-1989 and 1990-1999. The list of journal articles includes all papers in conference proceedings, as well as short papers and notes. We do not cover working papers and work published in books. The main reason for not covering working papers is that this can potentially lead us into double counting. The main reason for restricting attention to journal articles is that the EconLit database starts covering books from the 1980's only and this would sharply restrict the time frame of our study. Table 1 provides an overview of the coverage of our data. Tables 2 and 3 give us data on the number of EconLit journals and the number of articles published in these journals over this period. The number of journals has grown from 196 in 1970 to 687 in 1999 while the number of journal articles in EconLit has grown from 62,569 in the 1970's to 156,454 in the 1990's. In Table 3 we can also derive that the number of pages per article has increased from 12.85 to 16.49 and that less and less papers have a single author. This trend was highlighted in Ellison (2002a).

The coverage of the Econlit data set is clearly partial and to check the robustness of our findings, we also consider an alternative set of data. We use the list of the Tinbergen Institute Amsterdam-Rotterdam (hereafter TI list) to do this. This list of journals is used by the Tinbergen Institute to assess the research output of faculty members at 3

Dutch Universities (University of Amsterdam, Erasmus University Rotterdam and Free University Amsterdam). 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. The appendix presents the list of these journals and Table 2 shows the growth of this set over the 1970-2000 period. We observe that out of the 113 journals in 2000, only 46 were covered by EconLit in 1970! While some of the new journals are general interest journals, it is fair to say that most of the increase comes from the expansion in the number of field journals. We interpret this as evidence of a broadening as well as a deepening in the subject matter that is covered by economics. Table 4 shows summary statistics for the TI list data set. Not surprisingly we see an increase in the number of papers, the number of pages per paper and the number of coauthored papers.

We thus have six data sets: 3 for the set of all journals covered in Econlit and 3 for the set of TI journals, and we construct a network for each data set. We first find the nodes in the network by extracting the different author names that appear in the data. As in Newman (2001c) we distinguish different authors by their last name and the initials of all their first names. Consequently, authors with the same last name and different initials are considered different nodes. We note that a single author may sometimes be represented by two nodes because of misspellings in the data or because of non-consistent use of first or last names. Further, for papers with more than three authors EconLit reports only the first author and the extension *et alia*. Therefore we do not consider articles with four or more authors when constructing the co-authorship network. We then construct the whole co-authorship network by adding links between those authors that have coauthored a paper. We note that we do not weight the links, that is, we do not distinguish between more or less prolific relationships.[4]

3.1 Aggregate patterns

Our analysis of the collaboration network starts with an examination of the order of the network, i.e., the number of publishing economists. Table 5 tells us that the profession has grown substantially in this period: the number of authors has grown from 33,770 in the 1970's to 81,217 in the 1990's. In Table 3 we saw that the the number of published papers has grown from 62,569 in the 1970's to 156,454 in the 1990's. Thus the number of papers grew by almost 150% over the 30 year period, while the growth in the number of authors is a little lower. This difference is reflected in a slight increase in the per capita number of articles from 2.35 in the 1970's to 2.83 in the 1990's. The data based on the TI list in Tables 4 and 6 is consistent with this trend: the number of authors has increased from 14,051 in the 1970's to 28,736 in the 1990's, the total number of papers

[4]We also analyzed weighted networks, see Newman (2001c). Results are only slightly different. We also considered networks in which separate authors are distinguished by their last name and the first initial only. The conclusions are not affected by this extraction rule.

has increased from 26,802 in the 1970's to 52,469 in the 1990's, and so the per capita number of papers has increased slightly from 2.45 in the 1970's to 2.85 in the 1990's. Our first finding is therefore the following: *the number of journal publishing economists has grown substantially – more than doubling – over the period 1970 to 2000.*

Thus the number of nodes of the network has increased significantly. What about the patterns of connections between these nodes? We shall focus on four macroscopic statistics relating to the pattern of connections: the existence and size of a giant component, the average distance between the nodes in the giant component, the average size of the 10-neighborhood and the clustering coefficient in the giant component as well as in the network as a whole.

We start with a discussion on the existence of a giant component. Table 5 presents these statistics for the periods under study. In the 1970's the largest component contained 5,253 nodes which constituted about 15.6% of the population. This largest component has expanded substantially over time and in the 1990's it contains 33,027 nodes which is roughly 40% of all nodes. Correspondingly, there has been a sharp fall in the proportion of isolated nodes from almost 50% in the 1970's to about 30% in the 1990's. At the same time the second largest component has also declined in size: it had 122 members in the 1970's and only 30 members in the 1990's. This trend is consistent with evidence in the data on the TI list. Table 6 shows that the number of persons in the giant component has increased from 2,775 in the 1970's to 14,368 in the 1990's. In the 1970's the giant component constituted about approximately 20% of the total population, while in the 1990's it comprises 50% of the total population of economists. The proportion of isolated authors has declined from about 42% in the 1970's to about 21% in the 1990's. Finally we note that here too the size of the the second largest component has declined sharply from 74 to 31. These observations lead to our next finding: *The largest component in the collaboration network has grown substantially in absolute numbers as well as in relative size while the size of the second largest component has declined sharply in absolute size. Moreover, the proportion of non-collaborating economists has declined sharply.*

We now turn to the distance between the nodes in the network. As is the norm we set the distance between nodes in the different components to infinity and we use the average distance between nodes in the giant component as a proxy for our measure of average distance in the network. We find that this average distance has declined from 12.86 in the 1970's to 9.47 in the 1990's. We also note that this fall in average distance has been accompanied with a significant fall in the standard deviation in the distances between nodes from 4.03 in the 1970's to 2.23 in the 1990's. This trend is consistent with the trends observed in the data on journals in the TI list: Table 6 tells us that the average distance in the giant component has declined from 11.99 in the 1970's to 9.69 in the 1990's while the standard deviation has declined from 4.02 in the 1970's to 2.35 in the 1990's. This leads to our next finding: *The giant component has grown substantially in numbers but at the same time it has become significantly smaller in terms of distances.*

The above findings suggest that nowadays it is more likely that there exists a path between two random economists and that on average this path is much shorter than it was in the 1970s. This is also confirmed by the trend in the average size of the k-neighborhood. Tables 5 and 6 show that the size of 10-neighborhood has increased sharply – from .7% to 11.8% for the full EconLit data set and from 1.5% to 16.5% for the data set based on the TI list of journals. This trend is also clear if we look at the size of 10-neighborhood within the giant component. For the set of all journals, we find that the mean size of the 10-neighborhood increased from 28.8% in the 1970's to 71.2% in the 1990's and a similar pattern arises for the TI list. Hence the probability that there are less than 10 degrees of separation between two random nodes has increased sharply.

We turn next to the level of overlap between co-authorship which is measured by the clustering coefficient in the network. We present clustering coefficients for the network as a whole as well as the figures for the giant component. Table 5 shows that clustering coefficient for the network as a whole was .193 in the 1970's, .182 in the 1980's and .157 in the 1990's.[5] Table 5 shows that clustering coefficient for the giant component was .136 in the 1970's as well as in the 1990's. This tells us that the clustering coefficient is high and declining through the period under study. We also note that the clustering coefficient for the network as a whole is larger than the clustering coefficient of the giant component in each of the three periods under study. Table 6 presents the figures for clustering coefficient for the journals in the TI list. We find that clustering coefficient for the network as a whole is high but exhibits a declining trend – it was .188 in the 1970's, .180 in the 1980's and .167 in the 1990's. The clustering coefficient of the giant component is high and exhibits an increasing trend. We note that the clustering coefficients of the giant component and of the whole network are converging. This is consistent with the tremendous growth of the giant component: in the limit as the giant component covers the whole network the two clustering coefficients must be the same. These observations lead us to our next finding: *the clustering coefficient for the whole network is high but exhibits a declining trend in the 1970-2000 period.*

The discussion on the data allows us to make *three* general points about the aggregate characteristics of the network of collaboration among economists during the period 1970-2000: the *first* point is about the number of nodes of the network: they have more than doubled in this period. The *second* point is that distinct groups of co-authors seem to have formed links with each other and this has led to the emergence of a giant group of interlinked economists which now covers about half of the total population. The *third* point is about the structure of this giant component: the average distance in this giant group has steadily been falling over time while the level of clustering has remained high. Taken together, these three remarks suggest that the world of economists is expanding but at the same time becoming a small world where most of economists find themselves just a few degrees of separation away from any other economist.

[5]When we say 'high' we mean that it is substantially higher than the clustering coefficient of a random network. For a discussion on this issue, see Section 5

3.2 Link patterns

What are the mechanisms underlying the emergence of the small world? Our approach to this question is founded on the idea that individual economists have a choice between writing papers by themselves or in collaboration with others, and that the network of collaboration we observe arises out of the decisions they make in this regard. Individual decisions on research strategies will thus depend on a variety of factors such as the costs of doing individual and joint work and relative rewards of alternative modes of research. It is also likely that differences in individual productivity have a bearing on collaboration. We would like to develop a model of co-authorship in which these costs and the reward schemes are exogenous parameters while the link decisions and the networks of collaboration are equilibrium outcomes. Thus the crucial micro level data in this approach are the number of collaboration links that an individual forms and the patterns of linking across other economists. We look next at these micro statistics in the data that we have assembled.

We start with the average number of links per capita. In all the data we have assembled we observe the following trend: the average number of links/collaborators per capita has grown significantly in the period 1970 to 2000. For the set of all journals in EconLit, Table 5 tells us that there is almost a doubling in the per capita number of links/collaborators from .894 in the 1970's to 1.672 in the 1990's in the giant group. This figure covers all publishing economists and it is useful to also examine the per capita number of collaborators among people who are in the giant component. Table 5 shows us that the per capita number of collaborators increased from 2.48 in the 1970's to 3.06 in the 1990's. This trend is also visible and clear cut in the TI list of Journals. Table 6 tells us that mean number of links has increased from 1.058 in the 1970's to 1.896 in the 1990's. Similarly, we find that the mean links per author has increased in this data set if we restrict attention to authors in the giant component. This trend of increasing number of collaborators is related to the trend in the proportion of co-authored papers. Table 3 shows that 25% of all articles in EconLit were coauthored in the 1970's, while 42% of the articles were coauthored in the 1990's. When we consider the TI list only, then the proportion of coauthored papers has increased from 28% in the 1970's to 50% in the 1990's (see Table 4). This trend of increasing collaboration is a well-known fact in the research literature, see e.g. McDowell and Melvin (1983) and Eisenhauer (1997).

Although the average number of collaborators per author has increased over time, the above aggregate statistics do not provide information on whether this increase has been equally distributed in the population of economists. We explore this issue by examining the distribution of links in the collaboration network more closely. Figure 1 shows the Pareto plot for distribution of links: this plot shows on a log-log scale the degree of links per author k on the x-axis and the tail distribution, i.e., the fraction of authors for which $\eta_i \geq k$, on the y-axis. This graph shows that there is a first-order stochastic dominance relationship over time. The distribution for the 1980's first order stochastically dominates the distribution of the 1970's, while the 1990's degree distribution first order stochastically dominates the degree distribution of the 1980's. This yields us our first finding on the

micro statistics of the network: *the number of collaborators has been increasing consistently through the 1970-2000 period for all quantiles.*

Another remarkable aspect of Figure 1 is that the Pareto plot seems to converge to a straight line for high degree k. This suggests that at high quantiles the distribution converges to a Pareto or power-law distribution.[6] An important characteristic of such a distribution is the existence of a fat tail. Indeed, extreme degree values appear more frequently in the real data than in a binomial distribution fitted on the 1990's data set. While under the fitted binomial distribution it is unlikely that any author has more than 10 links, in reality we see that more than 1% of the authors have more than 10 links and some of them have 40 to 50 links.

We explore the inequality in the number of links per author further by looking at Lorenz curves and Gini coefficients. Figure 2 shows Lorenz curves in the 1970's, 1980's and 1990's based on the data set that includes all EconLit journals. We see an striking inequality in the distribution of links: the 20% most-linked authors account for about 60% of all the links. The plot also reveals a decreasing trend in inequality over time. This observation is confirmed by the Gini coefficients reported in Table 7. This trend is mainly explained by a decrease in the number of isolated authors — from 50% in the 1970's to 30% in the 1990's — since it implies that more and more authors are involved in co-authoring. Interestingly, if we adjust for this participation effect by ignoring isolated authors or by considering the giant component only, then Table 7 shows that the trend is reversed. The above results lead to the following finding: *The distribution of links in the population is very unequal and exhibits a fat tail. Further, inequality in the number of links has increased within components and it has decreased when we consider the whole network.*

We now examine more closely the role of the individuals who are very well connected in the network of collaboration. Table 8 tells us that in the 1970's the maximally connected person had 25 links and the 100 most linked persons had on average 12 links. Looking more closely at the most connected individual we see three very striking features: one, this person published 44 papers out of which 42 (i.e., 95% of them) were co-authored; two, he had 25 collaborators while the average number of collaborations per capita was less than 1; and three, the clustering coefficient for this person was only .05, which is much smaller than .193, the clustering coefficient of the network at large. Similarly, in the 1990's the most connected individual published 66 papers, of which 64 were co-authored (this is 97% of the total), had 54 collaborators (while the per capita number of collaborators was under 2) and a clustering coefficient of .02 (while the clustering coefficient of the network as a whole was .157). Thus the most connected individuals collaborated extensively and most of their co-authors did not collaborate with each other. The most connected individuals can be viewed as 'stars' from the perspective of the network architecture. A closer inspection of Table 8 reveals that these three patterns are quite general and hold for the average of the 100 most linked individuals in the 1970's, 1980's and 1990's. Table 8

[6] A power-law distribution would take the form $f(k) = \alpha k^{-\beta}$, with $\alpha > 0$ and $\beta > 0$.

also tells us that the average number of links among the top 100 stars has more than doubled, while clustering around the stars has decreased. This leads us to state: *There is a large number of 'stars' in the world of economics and they co-author most of their research output and the number of stars in the world of economics is increasing.*

We next examine the role of the stars in connecting different parts of the network. For this purpose we compared the consequences of randomly deleting 2% or 5% of the nodes on network connectivity and clustering with the consequences of deleting star nodes. We do this for the network based on all EconLit journals. Table 9 shows the results. We can see that a removal of 5% of the authors at random has almost no effect on the network connectivity and clustering. For the 1990's, we find that the size of the giant component goes down from .407 to .389, while the average distance within the giant component increases marginally from 9.47 to 9.68. The effects on size of 10-neighborhood and clustering co-efficient are similarly insignificant. By contrast, a removal of the 2% most connected nodes has a devastating effect on the network. The giant component breaks down and its share of the total network goes down from .407 to .256, while the average distance within the giant component increases from 9.47 to 19.00. The effects on the size of the 10-neighborhood are substantial. We would like to note also that the impact on clustering coefficient is very substantial: it increases from 0.157 to 0.250 with the removal of the 2% most connected individuals and to .344 on the removal of the 5% most connected ones. These observations yield our finding on error tolerance of the economics network : *The stars play the role of connectors and sharply reduce distance between different highly clustered parts of the world of economics.*

We would like to plot the networks for the periods of 1970's, 1980's and 1990's to get an overall picture of the networks. This has proved to be very difficult due to the large numbers of nodes involved. We have therefore tried to plot the local network around some prominent well connected economists (Figures 3-6). These plots are fascinating and suggest a number of ideas; we would like to draw attention in particular to one striking feature of the networks: hierarchy. For instance, in the plot for Joseph Stiglitz (Figure 3) we find that he is linked to several persons who are themselves 'stars' in the sense discussed above. Furthermore, we observe that these star co-authors of Mr. Stiglitz typically do not work with each other and also that the co-authors of these persons typically do not work with each other either. In particular, the co-authors of the stars do not work with Mr. Stiglitz. Thus there seems to be a hierarchy of well connected persons. We find this structure remarkable as this hierarchy is mostly self-organizing. A similar structure can be discerned in the plot for Jean Tirole (see Figure 4).

3.3 Data robustness

We now briefly discuss some aspects of the coverage of the data that we use. A shortcoming of the above data for our purposes is the partial coverage of the EconLit list. We observe

that this list has been growing over time and the data discussed above relate to this expanding world. This pattern creates the following possibility: in the 1970's the world of journals was actually very similar to the one we observe today but the EconLit data set does not capture this as it covered a small subset of journals and therefore excluded a large part of the journal publishing world. If this were true then the data above would be about the world of EconLit authors but would not be a good indicator of the world of journal publishing economists per se.

There are different ways of getting around this problem. We have already done one robustness check by looking at the TI list of journals. We now carry out another robustness check. We study the network of collaboration using only the subset of journals that appear in Econlit for the entire sample period. This is the route taken in Tables 11 and 13. The patterns we observe here are broadly in line with what we have observed above. The total number of papers increased about 17%, from 42,115 in the 1970's to 49,245 in the 1990's. The number of authors has however gone up significantly from 22,960 in the 1970's to 32,773 in the 1990's, about 43%. We now turn to the statistics on the pattern of connections. We note that the largest component has grown from 3,076 nodes in the 1970's, which was about 13% of all nodes, to 10,054 nodes in the 1990's, which is about 30% of all nodes. Likewise, the percentage of isolated authors has fallen from about 50% in the 1970's to about 32% in the 1990's. Thus the order of the network is increasing while the network is becoming more integrated. This is also witnessed by the increase in the average size of the 10-neighborhood, which has grown from .6% to 2.8%. With regard to the micro statistics, Table 13 tells us that mean number of links per author has increased from 0.885 in the 1970's to 1.386 in the 1990's. Thus the trends in this set of journals are broadly consistent with the trends identified above for the entire set of journals and the TI list.

We finally consider a fixed set of five core journals, namely, *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics* and *Review of Economic Studies*. Tables 10 and 12 show the results we obtain. They appear to be broadly consistent with our main findings reported before. The size of the giant component has increased from 7% in the 1970's to 25% in the 1990's and the average size of the 10-neighborhood has increased from .5% in the 1970's to 2.9% in the 1990's. Further, the number of per capita collaborators has increased from .833 in the 1970's to 1.429 in the 1990's and the clustering coefficient has remained high over time. There are some interesting differences though. First, we see no trend in distance and clustering. Second, the number of articles in the top 5 journals has decreased from 5,023 in the 1970's to 3,705 in the 1990's. This appears to be mainly due to an increase in the length of the papers. We also observe a much more significant increase in the proportion of coauthored papers. This is highlighted in Figure 7. The figure shows the proportion of coauthored articles for different classes of journals classified on the basis of a quality criterion (AA (top 5 journals), A (second tier) and B (third tier)) as used by the Tinbergen Institute (see the appendix for the list of journals and its classification as AA, A, or B journals).

We see that in the 1970's these 5 journals had the smallest fraction of coauthored papers, while in the 1990's they have the highest proportion of co-authored papers.

3.4 Economics and other subjects

We would like to compare the collaboration network in economics with the networks in other subjects. Table 14 presents a summary of statistics for economics along with those for physics, medical sciences, computer science and mathematics. There are significant differences in period covered and the status of the publications so this table should be seen as being suggestive only. *First*, we would like to draw attention to the substantial differences in the size of the giant component. We observe that both in physics as well as in medical sciences, the giant component covers almost the whole population, while in economics, computer science and mathematics the giant component covers around one half of the population only. *Second,* we would like to mention the average distances: here again we find that average distances in physics and medical sciences are very small as compared to the average distances observed in economics, computer science and mathematics. These differences appear to arise out of substantial differences in the overall number of persons in the network as well as the per capita number of links: in the medical sciences data set the number of authors was over 2 million and the average number of links was 18.1 (for a five year period, 1995-1999), while in economics the total number of authors was 156,454 while average number of links was only 1.67 (over a ten year period, 1990-1999).

4 An incentives based explanation

In this section we try to explain the observed patterns of co-authorship in terms of a model of individual incentives. We are specially interested in explaining the following features of the co-author network: short average distances and high clustering and growth in size of giant component. Our model has three main aspects: productivity differences across individuals, a technology of knowledge production, and academic reward schemes.[7]

4.1 Model

We suppose that there are n players and that a player can be either of High type or Low type. There are n_h High-type players, and n_l Low-type players, and $n = n_h + n_l$. We shall assume that $1 < n_h << n_l$ and that n_l is sufficiently large. We shall denote the set of players by N. Players make decisions on their research strategy: whether to write alone or with others, and if with others, how many co-authorships to form and with which

[7]For a related model of co-authors, see Jackson and Wolinsky (1996). Their interest is in complementarities in collaboration and their equilibrium networks are characterized by complete components of different sizes.

types of players; they also decide how many papers to write and how much effort to put in each paper that they write.

Let $g_{ij}^x \in \{0,1\}$ model i's decision on whether to participate in a project x with author j, where a value of 1 signifies participation while a value of 0 signifies non-participation. Let e_{ii}^x denote the effort that player i spends on a single-authored paper x, and let e_{ij}^x, refer to the time that he spends on a joint paper x with coauthor j. We assume that for a paper to be written the total effort put in by its authors must be at least 1. A research strategy of a player is then given by a row vector, $s_i = \{(g_{ij}^x, e_{ij}^x)_{x\in\{1,..,m\},j\in N}\}$, where m is the number of projects that a player participates in either individually or with any other single coauthor.[8] A player j is a coauthor of player i if $g_{ij}^x = g_{ji}^x = 1$ for some paper x. Let $\eta_i(\mathbf{s})$ be the number of coauthors of player i in research strategy $\mathbf{s}$.

A paper consists of ideas and technical routine work. The quality of a paper depends only on the quality of the ideas it contains and the ideas of the paper in turn depend on the type of the authors of the paper. A High-type author has high quality ideas, while a Low-type author has low quality ideas; the high and low quality types of ideas are denoted by t_h and t_l, respectively, where $t_h > t_l > 1$. It is natural to assume that a type i author will write single-authored papers of quality t_i only, $i = h, l$; we assume that if two authors i and j jointly work on a paper the quality of the paper is given by $t_i \cdot t_j$. Thus quality of a paper can be $q \in \{t_h^2, t_h t_l, t_h, t_l^2, t_l\} = Q$.

We now elaborate on the costs and benefits of writing papers alone and with others. We first note that writing a paper takes time and effort and this has costs in terms of resources and the leisure sacrificed. We assume that the marginal costs of writing papers increase with the number of papers, reflecting increasing marginal opportunity costs of time. Maintaining a coauthor relationship involves communication and coordination across different projects and possibly different partners and these costs are likely to increase as the number of co-authors increases. This leads us to assume that the marginal cost are increasing in the number of coauthors, $\eta_i(\mathbf{s})$. Given these considerations, we are able to write down the costs of a research strategy s_i for a player i faced with a research strategy profile $\mathbf{s_{-i}}$, as

$$\sum_{j\in N} c\left[\sum_{x\in\{1,..,m\}} e_{ij}^x\right]^2 + f\frac{\eta_i(\mathbf{s})^2}{2} \tag{1}$$

with $f > 0$. This cost specification captures the idea that a collaboration relation between two individuals i and j is a research project and that the costs of coming up with interesting ideas and papers increase as more papers are written within the project. This leads us to suppose that writing m papers with m coauthors is less costly than writing m papers with a single coauthor. This assumption pushes individuals toward diversification

[8]The idea of m is introduced for notational convenience and is different from a capacity constraint. This will become clear when we model the costs of writing papers below.

of collaborators. On the other hand, our assumption that costs of linking with others are convex in the number of links pushes toward fewer collaborators. The optimal number of collaborators trades off these two pressures and depends also on the rewards associated with scientific activity.

The rewards from publishing a paper depend on a variety of factors such as its quality of papers and the number of coauthors. We shall suppose that a person is paid on the basis of quality weighted index of papers he publishes, there is discounting of joint work and that there is a minimum quality requirement such that only papers *above* this quality are accepted for publication. We shall suppose that this threshold is given by $\bar{q}$ where $\bar{q} \in [1, t_h^2]$. One interpretation of this threshold is in terms of different journals: a higher ranked journal can be more choosy in the papers it publishes and so it will have a higher threshold as compared to a lower ranked journal.

We suppose that a single-author paper of quality q gets a reward q, while a 2-author paper of quality q yields a reward rq to each author, where $r \in (0, 1)$ reflects the discounting for joint work in the market. Thus $r = 0$ means that joint work is given no credit, while $r = 1$ means that each coauthor of a joint paper gets credit equal to the credit due to an author of a single author paper.[9]

For a strategy profile $\mathbf{s}$, let $I_{ij}^x(\mathbf{e})$ be an indicator function, which takes on value 1 if $g_{ij}^x = g_{ji}^x = 1$, $e_{ij}^x + e_{ji}^x \geq 1$, and $q_{ij}^x \geq \bar{q}$, and it takes a value of 0, otherwise. Given these considerations, for a strategy s_i and faced with a strategy profile $\mathbf{s_{-i}}$, the payoffs to a player are as follows:

$$\Pi_i(s_i, \mathbf{s_{-i}}) = \sum_{j \neq i} \sum_{x \in \{1,..,m\}} I_{ij}^x r q_{ij}^x + \sum_{x \in \{1,..,m\}} I_{ii}^x q_{ii}^x - \sum_{j \in N} c \left(\sum_{x \in \{1,..,m\}} e_{ij}^x \right)^2 - f \frac{\eta(\mathbf{s})^2}{2}. \quad (2)$$

We study the architecture of networks that are strategically stable. Our notion of strategic stability is a refinement of Nash equilibrium. A strategy profile $s^* = \{s_1^*, s_2^*, ..., s_n^*\}$ is said to be a Nash equilibrium if $\Pi_i(s_i^*, \mathbf{s_{-i}^*}) \geq \Pi_i(s_i, \mathbf{s_{-i}^*})$, for all $s_i \in S_i$, and for all $i \in N$. In our model a coauthoring decision requires that both players wish to participate in the paper. It is then easy to see that an autarchic situation in which no one does any joint work is always a Nash equilibrium. More generally, for any pair i and j, it is always a mutual best response for the players not to participate in any joint projects. To avoid this type of coordination problem we supplement the idea of Nash equilibrium with the requirement of pair-wise stability. An equilibrium network is said to be pair-wise stable if no pair of players has an incentive to initiate one or more new joint papers. We define pair-wise stable equilibrium as follows:

[9] We are assuming here that different types involved in a collaboration get the same reward; our results do not change qualitatively if we assume that Low types get a lower payoff than High types.

Definition 1 *A strategy profile* $\mathbf{s}^*$ *is a pair-wise stable equilibrium if the following conditions hold:*

1. $\mathbf{s}^*$ *constitutes a Nash equilibrium.*
2. *For any pair of players,* i *and* j *there is no strategy pair* (s_i, s_j) *such that* $\Pi_i(s_i, s_j, \mathbf{s}^*_{-\mathbf{i}-\mathbf{j}}) > \Pi_i(s_i^*, s_j^*, \mathbf{s}^*_{-\mathbf{i}-\mathbf{j}})$ *and* $\Pi_j(s_i, s_j, \mathbf{s}^*_{-\mathbf{i}-\mathbf{j}}) > \Pi_j(s_j^*, s_j^*, \mathbf{s}^*_{-\mathbf{i}-\mathbf{j}})$.

In what follows, for expositional simplicity we shall use the short form – pws-equilibrium – while referring to pair-wise stable equilibrium. This notion of equilibrium is taken from Goyal and Joshi (2003). We shall say that a network is *symmetric* if all equal-type players have the same number of links with each of the two types of players. This will allow us to talk of the typical number of collaborations between a typical i and j type of players and use η_{ij} to refer to this number.

4.2 Equilibrium analysis

We start characterizing equilibrium networks under the assumption that in a joint project, each author contributes one half of the time needed for routine work and gets credit r for the joint paper. This may be interpreted as a model with no transfers. We note that the optimal choice of number of papers is independent across pairwise collaboration ties. This is due to the cost specification which is additive across projects with different co-authors and own projects. Our first result derives the optimal number of papers that High type and Low type authors will write on their own and with others.[10]

Proposition 1 *Suppose that threshold quality for publication is lower than* t_l*. A High type player optimally chooses* $m_h^* = t_h/2c$ *single author papers,* $m_{hh}^* = 2rt_h^2/c$ *papers in a HH collaboration, and* $m_{hl}^* = 2rt_ht_l/c$ *papers in HL collaboration. A Low type player optimally chooses* $m_l^* = t_l/2c$ *single author papers,* $m_{lh}^* = 2rt_ht_l/c$ *papers in a LH collaboration, and* $m_{ll}^* = 2rt_l^2/c$ *papers in LL collaboration.*

Proof: For a High type the optimization problem with respect to single author papers is

$$\max_{m_h} \; t_h m_h - c m_h^2 \tag{3}$$

Straightforward calculations yield $m_h^* = t_h/2c$. Similarly, for a High type the optimal number of papers in an HH collaboration is the solution to the following optimization problem:

$$\max_{m_{hh}} \; rt_h^2 m_{hh} - c\left[\frac{m_{hh}}{2}\right]^2 \tag{4}$$

This optimization problem yields us the solution that $m_{hh}^* = 2rt_h^2/c$. Similarly, the optimal number if papers for a H type in a HL collaboration are given by $m_{hl}^* = 2rt_ht_l/c$.

[10] In what follows we treat the number of papers and the number of co-authors as continuous variables.

Given that the publication threshold is below t_l, L types will also write papers on their own. It follows from above calculations for the H-type players that the optimal number of single author papers for an L-type are $m_l^* = t_l/2c$. Using arguments analogous to the above we can also conclude that the optimal number of papers for the Low type in a LH collaboration is $m_{lh}^* = 2rt_lt_h/c$, while the optimal number of papers in a LL collaboration is $m_{ll}^* = 2rt_l^2/c$.

■

This proposition tells us that H-types will write more single authored paper than L-types. Moreover, the optimal number of papers in a HH relationship is greater than the number of papers in a LL co-author relation. These results follow directly from the initial productivity differences across players. We also note that the number of optimal papers varies negatively with the costs of writing papers, while they vary positively with the individual credit given in co-authored papers.

One implication of the above result is that different type of players get very different aggregate returns from working alone and working with others. Moreover, players of different types also get very different payoffs from co-authorship relations. Let π_i refer to the payoff that a i type player gets from working alone, and π_{ij} refer to the reward that a type i player gets from working with a type j player. Then the above proposition allows us to write down the following payoffs for different type players.

$$\pi_h^* = \frac{t_h^2}{4c};\ \pi_{hh}^* = \frac{r^2t_h^4}{c};\ \pi_{hl}^* = \frac{r^2t_h^2t_l^2}{c}; \tag{5}$$

$$\pi_l^* = \frac{t_l^2}{4c};\ \pi_{lh}^* = \frac{r^2t_l^2t_h^2}{c};\ \pi_{ll}^* = \frac{r^2t_l^4}{c}. \tag{6}$$

Equipped with these returns from solo research and collaborative research we can now study the incentives for collaboration. The following result characterizes the architecture of symmetric co-author networks in the case where there are no constraints on finding suitable partners. In what follows, our interest is primarily in the nature of co-author networks that arise and we shall omit mention of single author papers throughout the discussion.

Proposition 2 *Suppose that $n_h - 1 \geq r^2t_h^4/cf$ and there are an even number of H-types and L-types in the population and $\bar{q} = t_l$. A symmetric equilibrium network exists and it has the following properties.*

1. *If $f > 2r^2t_h^4/c$ then it is empty.*
2. *If $2r^2t_l^4/c < f < 2r^2t_h^4/c$, then $\eta_{hh}^* = \frac{r^2t_h^4}{cf}$, $\eta_{lh}^* = 0$ and $\eta_{ll}^* = 0$.*
3. *If $f < 2r^2t_l^4/c$, then $\eta_{hh}^* = \frac{r^2t_h^4}{cf}$, $\eta_{hl}^* = 0$ and $\eta_{ll}^* = \frac{r^2t_l^4}{cf}$.*

Proof: We first characterize the incentives to collaborate. Part (1) follows directly from noting that $\pi^*_{hh} < f/2$ implies that there is no incentive for two H-types to collaborate. Since this is the highest possible return from co-authorship no links can arise in equilibrium. We now prove parts (2) and (3). First, we note that an H-type and an L-type will not collaborate in a symmetric equilibrium. The returns to an H type from an HH link, $\pi^*_{hh} > \pi^*_{hl}$, the returns from an HL link and so a H type will not link up with a L type if there is an H type available. The assumptions that $n_h - 1 \geq r^2 t_h^4/cf$ and that n_h is an even number guarantee that this will be the case (the critical number of high types is derived below). Second, we note that an L type would only be willing to collaborate with H-types if $f/2 < \pi^*_{lh}$; similarly, he will be willing to collaborate with L-types only if $f/2 < \pi^*_{ll}$.

We now turn to optimal choice of partners. If $f/2 < \pi^*_{hh}$ then the optimal number of links for an H type, η_{hh}, can be computed by solving:

$$\max_{\eta_{hh}} \quad \eta_{hh}\pi^*_{hh} - f\frac{\eta^2_{hh}}{2} \tag{7}$$

The solution is given by $\eta^*_{hh} = \frac{r^2 t_h^4}{cf}$. Thus if $n_h - 1 > \frac{r^2 t_h^4}{cf}$, then there are enough H-types around and an H-type will not collaborate with an L-type.

Similarly, an L type collaborates with another L type if and only if $f/2 < \pi^*_{ll}$. In this case the optimal number of links for an L type, η_{ll}, can be computed by solving:

$$\max_{\eta_{ll}} \quad \eta_{ll}\pi^*_{ll} - f\frac{\eta^2_{ll}}{2} \tag{8}$$

The solution is given by $\eta^*_{ll} = \frac{r^2 t_l^4}{cf}$. Thus if $n^*_{ll} < n_l - 1$, then there are enough L-types around and the proof is complete, which follows from our earlier assumptions $n_l >> n_h$ and $n_h - 1 \geq r^2 t_h^4/cf > \frac{r^2 t_l^4}{cf}$.

The existence of symmetric equilibrium follows directly from the fact that an optimal number of papers and co-authors exist and there are enough players of each type to make the optimal number of co-authors feasible. ■

Figure 8 illustrates equilibrium networks with a large number of H and L types. Proposition 2 tells us that if two persons involved in a collaboration equally share the effort required to write a paper, then only links between same type players will form in a symmetric equilibrium. Moreover, H-types will have more co-authors than L-types. The difference in the number of papers and the number of co-authors is however directly a function of the difference between t_h and t_l, which means that if the two types are similar in productivity then the equilibrium outcomes and payoffs will also be similar.

We now comment on the role of the two institutional reward variables: the threshold level for publication, $\bar{q}$, and the credit for joint work r. The threshold $\bar{q}$ is critical in defining

the level and types of co-authorship: for instance, if the threshold is t_h^2 then we will only see HH co-author relations and t_h^2 quality papers. On the other hand, if the threshold is t_l^2 then two L types will collaborate as well. This leads us to ask: does an increase in $\bar{q}$ always raise the proportion of co-authored papers? The answer to this depends on the relative value of t_h and t_l. If $t_h < t_l^2$ then the proportion of co-authored papers is increasing in $\bar{q}$. If on the other hand, $t_h > t_l^2$ then there is a non-monotonicity: as $\bar{q}$ crosses t_l the proportion increases and as it increases beyond t_l^2 it falls before rising again to a value of 1 as $\bar{q}$ crosses t_h. We also note that the number of joint papers as well as the number of co-authors is increasing in r, the level of individual credit for co-authored work. This also implies that the proportion of co-authored papers is increasing in r.

Proposition 2 implies that there are no connections between Low and High type players. Moreover, in equilibrium, links only exist between players with the same number of links. This seems to be at variance with one of the crucial aspects of empirically observed networks: the existence of a large number of stars (which arise when highly connected players connect with very poorly connected players). This difference between observed patterns and equilibrium predictions leads us to explore two aspects of the model more closely: the number of H-types available and the possibility of transfers between High and Low types.

One reason for the 'same-type collaboration only' result is that there are enough players of each type. What happens if an H-type wants to collaborate with 10 H-types but there are only 5 H-types around? In this case, High type players will not be able to reach their global optima and collaboration between unequal types could emerge. This observation leads us to the following result.

Proposition 3 *Suppose that $n_h - 1 < r^2 t_h^4/cf$ and the threshold for publication is $\bar{q} = t_l$. Then a symmetric equilibrium has the following features.*

1. *If $f > 2r^2 t_h^4/c$ then it is empty.*
2. *If $2r^2 t_l^4/c < f < 2r^2 t_h^4/c$, every H-type has $n_h - 1$ H-type co-authors, and also has $\eta_{hl} = \max\{0, \frac{r^2 t_h^2 t_l^2}{cf} - n_h + 1\}$ L-type co-authors. L-types do not work with each other.*
3. *If $f < 2r^2 t_l^4/c$ an H-type has exactly the same co-author pattern as in (2), while each L-type has $\eta_{lh} \in (1, n_h)$ H-type co-authors and $\max\{0, \frac{r^2 t_l^4}{cf} - \eta_{lh}\}$ L-type co-authors.*

Proof: Part 1 follows as in Proposition 2. Part 2 is now proved. Since $n_h - 1 < r^2 t_h^4/cf = n_{hh}^*$, it follows that there are not enough High-type players around so that a High-type may find it worthwhile to form collaborations with Low-types. Since $\pi_{lh}^* > \pi_{ll}^*$ a Low-type always prefers to collaborate with a High-type rather than with another Low-type. Thus, the payoff to a High-type may be written as

$$\left(t_h m_h - c m_h^2\right) + (n_h - 1)\pi^*_{hh} + \eta_{hl}\pi^*_{hl} - f\frac{(n_h - 1 + \eta_{hl})^2}{2}. \tag{9}$$

It is now easy to see that the optimal number of HL collaborations is given by $\eta_{hl} = r^2 t_h^2 t_l^2/cf - n_h + 1$. We now consider the incentives of L types. First note that since $f > 2r^2 t_l^4/c$ there will be no LL co-author papers. It then follows that an L-type player will have $\eta_{lh} \in \{1, n_h\}$ H-type co-authors in a symmetric equilibrium. This completes the proof of part 2.

Consider part 3 next. Note first that incentives of H types are as in part 2. Also note that L types prefer to form a relation with an H type over a relation with an L type. Denoting by η_{lh} the number of H-type co-authors for a Low-type player, the payoff to the Low-type player is given by:

$$\eta_{lh}\pi^*_{lh} + \eta_{ll}\pi^*_{ll} - f\frac{(\eta_{lh} + \eta_{ll})^2}{2}. \tag{10}$$

The first order conditions yield $\eta^*_{ll} = \frac{r^2 t_l^4}{cf} - \eta_{lh}$. ■

Figure 9 presents equilibrium networks, when the number of H types is small.

The above proposition says that a small number of H-types has a number of implications. The first implication is that there is a wide range of parameters for which HL collaborations arise in equilibrium. Related to this we note that H types and L types will have an unequal number of co-authors. In particular, in part 2, for $n_h << n_l$, the networks will have the core-periphery structure: all H types will co-author with each other while each of them will co-author with a number of L-types, who do not co-author with each other.

We now examine the scope of 'a sharing of scarce resources' motivation for collaboration between an H-type and an L-type. We start by examining a case in which L-types offer 'time' for routine work and in return get High quality ideas from H-types. An important issue here is how the exchange of ideas and time takes place. We start by looking at the case where an L-type only shares in the routine work and does not share the costs of communication and maintaining links f. To keep matters simple we shall suppose that an H-type makes a take-it-or-leave-it offer $\alpha \in (1/2, 1]$ to an L-type, where α measures the amount of time the Low-type must contribute per paper the two parties write together.[11] What will be the optimal level for α and how many co-authorships between H and L types will arise in this case? The following proposition provides a complete answer to this question.

Proposition 4 *Assume that $n_h - 1 \geq r^2 t_h^4/cf$ and $\bar{q} = t_l$. Suppose that an H-type makes a take-it-or-leave-it offer of $\alpha \in (1/2, 1]$ to an L-type with regard to sharing routine work in a joint paper. Then in equilibrium there will be no HL co-authorships and therefore networks will have the same structure as in Proposition 2.*

[11] We assume that the α-contract is enforceable.

Proof: Given an $\alpha \in (1/2, 1]$, an L-type faces a trade-off: work with other L-types and share routine work equally or work with an H-type and put in a fraction of time $\alpha > 1/2$ per paper in exchange for high quality ideas. From 6 we know that if an L-type works with another L-type then he receives a payoff given by $\pi_{ll} = r^2 t_l^4/c$. If a Low-type works with an H-type instead, then he receives a payoff given by

$$rt_h t_l m_{lh} - c(\alpha m_{lh})^2,$$

where α denotes the fraction of time an L-type must put in to be able to work with an H-type. From the first order conditions it follows that $m_{lh} = rt_h t_l/2c\alpha^2$, which yields a payoff given by $\pi_{lh}(\alpha) = r^2 t_h^2 t_l^2/4c\alpha^2$ to the Low-type. We can compare π_{ll}^* and π_{lh} and we find that $\pi_{lh} > \pi_{ll}$ if and only if $t_h^2 > 4\alpha^2 t_l^2$. This gives us the set of α that an L-type will accept.

Consider now the decision of an H-type. An H-type faces a trade-off: work with High-types and share writing costs or work with Low-types and exchange quality of ideas for working time. From (5) we know that the payoff from an HH link to an H-type player are $\pi_{hh}^* = r^2 t_h^4/c$. Consider now the payoffs to an H-type from a HL link, where the L-type puts in α share of the routine work, while the H-type puts in $(1-\alpha)$ fraction of the routine work.

$$rt_h t_l m_{hl} - c((1-\alpha)m_{hl})^2$$

From the first order condition, it follows that $m_{hl} = rt_h t_l/2c(1 - 2\alpha + \alpha^2)$. Since $\alpha \geq 1/2$ it follows that $m_{hl} \geq m_{lh}$. We assume that for a given α, the number of papers written in a HL relation is $\min\{m_{hl}, m_{lh}\}$. Thus the payoff to an H-type from a link with an L-type is

$$\bar{\pi}_{hl}(\alpha) = \frac{r^2 t_h^2 t_l^2}{2c\alpha^2} \frac{\alpha(\alpha+2)-1}{2\alpha^2}$$

We now note that $\bar{\pi}_{hl}(\alpha) \geq \pi_{hh}^*$ if and only if $t_h^2 < (\alpha(\alpha+2)-1)/4\alpha^4) t_l^2$. This gives us the set of parameters for which an H-type would be willing to co-author with an L-type for a given α.
Putting together the restrictions for the H and L types we get that for a fixed $\alpha \in (1/2, 1]$, an HL pair will co-author only if

$$4\alpha^2 t_l^2 < t_h^2 < (\alpha(\alpha+2)-1)/4\alpha^4) t_l^2$$

It is easy to verify that $4\alpha^2 > (\alpha(\alpha+2)-1)/4\alpha^4)$ for all $\alpha \in (1/2, 1]$. Thus there is no sharing of routine work between H and L types that can make a HL relation mutually incentive-compatible. ■.

The above proposition says that if transfers are restricted to the sharing of routine work then there will be no HL co-author relation in equilibrium. This result leads us to ask: are there other richer transfer schemes which would allow mutually profitable HL co-authorships. An obvious candidate is an arrangement by which an L type does all the

routine work and most of the costs of maintaining the relation. In that extreme case, an H type incurs no costs in writing papers with an L-type, while a L-type has to compare the relative returns of entering into such an unequal relation versus working on equal terms with another L-type. From above calculations we can deduce that the payoffs to an L-type from such a HL relation are: $\pi_{lh}(\alpha = 1) = r^2 t_h^2 t_l^2 / 4c$. The payoffs to an L-type from an LL relation are $\pi_{ll}^* = r^2 t_l^4 / c$. Now it is easy to see that for fixed f we can always find t_h and t_l with $t_h >> t_l$ such that an L type would prefer to link with an H type rather than link with another L type. Thus if sharing of routine work as well as the costs of maintaining the relation are possible then an HL relation can arise in equilibrium.

We conclude this section by noting that in the equilibrium networks discussed so far there is no a priori reason to expect high clustering levels. In the context of part 3 of Proposition 3, we know that an equilibrium network will have the property that L types will co-author with each other. The equilibrium characterization however does not say whether these L-types will be linked to the same H-type or to different H-types. In the former case we will get reasonable levels of clustering and an inverse relation between degree and clustering, while in the latter case aggregate clustering levels will be close to 0 and the H-types will have the highest levels of clustering. The former is consistent with the empirical patterns while the latter is clearly not. A simple way to rule out the latter type of equilibrium is to suppose that the costs of co-authoring between i and j are slightly less if they have a common co-author. This captures the idea that we usually start collaborating with individuals whom we already know and we are more likely to know someone with whom we share a common acquaintance. A slight difference in cost is sufficient to induce two L-types who share a common H-type co-author to form a link and de-link from other more distant L-types.

5 Alternative explanations

We now discuss some other theories that have been used to explain the emergence of a small world. All the work we will discuss in this section has been done by physicists and they have used statistically driven models to explain empirical patterns in large collaboration networks. These models are methodologically very different from the incentives-based approach we use and it is therefore difficult to make a direct comparison. The statistical models are developed to explain salient features of the large networks observed empirically and we will therefore ask how well they achieve this objective. Our main point here will be that the basic statistical models do not perform well in this regard. We will also point out that recent attempts at matching the high clustering levels and the inverse relation between degree and clustering suffer from some shortcomings.

We start with a discussion of the (uniform) Erdős-Rényi random graph model, in which the probability that there is a link between i and j is constant for all i and j. As a first step we set the evidence on mean number of links per capita against the average distance in the giant component. From earlier work in the theory of graphs we know that for a

given degree level, the random graph is very efficient at reducing average distances (see e.g., Bollobás, 1985). In our collaboration network the figures for the average distance compare favorably with the hypothetical figures for a random graph with a similar average degree. This may suggest that random link formation is a good description of the process underlying the above model. However, as we noted above the clustering coefficient in our network is much larger than the clustering coefficient of a corresponding random graph. To make this concrete let us consider the figures from the 1990's. On average a person has 1.672 co-authors, while there are 81,217 authors in all; we can interpret this as saying that there is a probability of approx .000025 of a link being formed. In a random graph, since the probability of link formation is independent, the clustering coefficient should be approximately equal to this probability of a link. However, the actual clustering coefficient is given by .157, which is well over 6000 times the level predicted by the random model of link generation. Thus the random model of link formation is not a good explanation of the process that generates scientific collaboration networks such as the one we are studying.

Perhaps the best known attempt at reconciling low average distance, low average degree and high clustering is the random rewiring model of Watts and Strogatz (1998). To get a flavor of this model suppose people are arranged on a circle and interact with their immediate and next to immediate neighbors. This pattern yields a very high clustering coefficient (of .5). Now suppose that with a small probability links are rewired: they are taken away from immediate neighbors and redrawn at random with the whole world. It can be shown that for low levels of rewiring probability, this process yields a network in which clustering remains high while average distances fall dramatically relative to the original circle. Does the collaboration network among economists look like a rewired network? One of the features of this rewiring in the Watts-Strogatz model is that it is random: the probability of my rewired link being formed with someone is equal across people in the network at large (in fact this is an important aspect of the procedure which reduces distances in the network). This in turn means that the resulting network will have a fairly uniform distribution of links. Formally speaking, the distribution of links will approximately follow a Poisson distribution. However, the empirical distribution of links in our collaboration network is sharply skewed and there is a large number of economists who have a very large number of links (see Figure 1). Thus the rewiring process fails to capture an essential aspect of our collaboration network and we must look further.

In a series of papers, Albert-Lázló Barabási and his collaborators have developed a model of preferential attachment and growing networks to address this problem of a highly skewed distribution of links.[12] In the basic preferential attachment model, we start with a few nodes and then add nodes one at a time. Each entering node forms links with existing nodes in a preferential manner: the probability of a newcomer i linking with some j is increasing in the number of links that j currently has. As time passes the network grows, and it is shown that the limit distribution of degrees is scale free. A scale

[12]For a survey of this work see Albert and Barabási (2002).

free distribution allows for fat tails and for a large number of very highly connected nodes and so the preferential attachment model provides a very simple and elegant explanation for the skewed empirical distribution of degrees. How does this model account for the clustering coefficient? Albert and Barabási (2002) provide simulations which suggest that this clustering is about 5 times higher than that of a random network, and that it is declining in the size of the network n at a rate $n^{-0.75}$. We now recall that the clustering in our data is about 6,000 times the clustering in the random network, while the clustering level in the network as a whole has remained quite stable over the period 1970-2000 while the number of nodes has more than doubled. These differences lead us to conclude that the basic preferential attachment model cannot account for the clustering levels observed in the data. Another feature of the data is that there is a clear and inverse relation between the degree and the clustering coefficient of a node. Figure 10 shows this clearly. By contrast, simulations of the standard preferential attachment model suggest that clustering levels are degree independent (see e.g., Ravasz and Barabási (2003). These two problems have led Barabási and his collaborators to explore modifications of the preferential attachment model.

In recent work Ravasz and Barabási (2003) propose a model of networks which uses modules as building blocks. Start with a completely linked set of 5 nodes, represented as in Figure 11a (to be inserted). Replicate this set 5 times and connect the central player of the initial set with the four outer nodes of each of the other 4 sets of nodes. This gives us a 2-level hierarchy as in Figure 11b, with the central player in the middle set of 5 being considered the top layer of the hierarchy. We can expand this network by replicating this 25 node network 5 times and correspondingly connecting the central node. Figure 11c implements this step. We now have 125 nodes in all. Ravasz and Barabási (2003) show that such modules based networks exhibit the following three properties: scale free distribution of links, clustering coefficients which are scale free and an inverse relation between degree and clustering coefficients of nodes. We briefly comment about the precise nature of the hierarchy. In this structure, the central player in the central set of 5 nodes can be considered as the top player in the hierarchy. This player is linked to 64 level 1 players but is not linked to any level 2 player. How does this compare with the hierarchy that we see in our collaboration network? Let us focus on the local network of Joseph Stiglitz for the 1990's (Figure 3). This network resembles a classical hierarchy, with persons at level 1 interacting mostly with their immediate superior at level 2 and these superiors at level 2 in turn interact with a single person Mr. Stiglitz at level 3. Moreover, persons at the same level do not interact very much with each other. Moreover there are very few instances of a level 1 person interacting with a level 3 person. In our view this is quite different from the hierarchy generated by the method in Ravasz and Barabási, where the top level player is connected to practically all the level 1 persons and none of the level 2 persons. We have examined the effects of a change in the pattern of links: consider a network in which node 1 is linked to all the level 2 players, to none of the level 1 players, and, in addition, to the 20 outer players, in his own replica of 25 players. This network generates a clustering coefficient of 1/9 (approx) for the central

player which seems to be much too high compared to what is needed to satisfy the $1/k$ scaling required for the clustering coefficient by Ravasz and Barabási.

6 Concluding remarks

The idea that the social world is small is an intriguing one. We are led to wonder if the world is indeed small and if so what makes the world small? Moreover, we would like to ask: does the smallness have any important implications for things that we care about?

In this paper, we have focused on the first question. We study the world of journal publishing economists during 1970-2000. The first part of our paper is an empirical investigation of the co-author network over this 30 year period. Our main finding on the issue of whether the world is small may be summarized as follows: In the 1970's the world of economics was a collection of islands, with the largest island having about 15% of the population. Two decades later, by the 1990's the world of economics has become much more integrated, with the largest island covering close to half the population. At the same time, the average distance between individuals on the largest island group has fallen while the clustering remains high ⇒ **an emerging small world.**

We then ask: what is about the number and distribution of co-authorships that accounts for these aggregate patterns? Our main findings are, one, that average number of co-authors is low but increasing sharply; two, the distribution is very unequal (fat tails are present); and three, there are many stars (these are highly connected economists work with economists who have few or no other co-authors)! This leads us to view the world of economists as a set of inter-linked stars.

Existing work on small worlds – in physics and mathematics – takes the smallness of a world as a given and develops statistical explanations. Our empirical work however shows that the smallness of the world is not a constant, and in fact seems not to be true at all in the 1970's. We take the view that the co-author network arises out of individual decisions – on working alone and with others –and their structure must therefore be understood in terms of individual incentives. This perspective leads us to naturally study changes in 'smallness' as an outcome of changes in incentives.

We develop a simple model of production of knowledge in economics with the feature that a paper consists of novel ideas and routine work. The quality of the paper depends on the quality of ideas, We embed this basic technology into a setting where individuals choose how many papers to write, and whether to write them alone or with others. There are three features of the model that play a role: differences in productivity of economists, institutional rewards schemes, and network effects. Our main results are that in a world with relatively few High productivity individuals the distribution of links/co-authors will be very unequal and links between people who have many co-authors and those who

have very few co-authors will arise naturally. Thus stars are an important aspect of equilibrium networks. Moreover, we find that the stars will be connected with each other and so equilibrium networks will look like inter-connected stars. This architecture will display short average distances. If costs of co-authoring are related to network distance then the network will also display high clustering. Finally, we show that lower costs of communication and higher individual credit for co-authored work will both lead to greater co-authoring and can help explain the growth in the size of the giant component.

Appendix: 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

Table 1: Coverage of Econlit: Basic statistics

	1970's	1980's	1990's
Books	5	5302	16156
Book Review	0	0	1029
Collective Volume Articles	0	35422	96307
Dissertation	0	2649	9649
Journal Article	62518	95033	156601
Working Paper	41	12215	23446

Table 2: Number of journals in Econlit: 1970-1999

Years	Number of Journals	Number of Journals in TI List
1970	196	46
1971	198	48
1972	198	47
1973	209	53
1974	203	55
1975	200	56
1976	220	58
1977	227	61
1978	242	64
1979	248	65
1980	256	67
1981	264	67
1982	262	68
1983	285	74
1984	304	79
1985	311	81
1986	318	86
1987	317	87
1988	324	90
1989	340	95
1990	353	98
1991	368	101
1992	425	104
1993	439	106
1994	491	107
1995	535	109
1996	590	110
1997	624	111
1998	656	112
1999	687	113

Table 3: Summary statistics for all articles in EconLit.

period	70's	80's	90's
total papers	62569	95027	156454
mean pages per paper	12.85	14.45	16.49
standard deviation	(9.94)	(10.27)	(10.59)
# Authors per paper: Distribution			
single-authored	.753	.678	.578
two authors	.210	.256	.309
three authors	.031	.055	.090
four or more authors	.005	.011	.023

Table 4: Summary statistics for all articles in 113 selected journals (TI list).

period	70's	80's	90's
total papers	26802	38133	52469
mean pages per paper	12.17	13.76	16.29
standard deviation	(8.69)	(8.40)	(9.08)
# Authors per paper: Distribution			
single-authored	.716	.616	.504
two authors	.244	.311	.371
three authors	.035	.064	.104
four or more authors	.005	.009	.020

Table 5: Network statistics for the network based on all articles in EconLit.

period	70's	80's	90's
total authors	33770	48608	81217
mean papers per author	2.35	2.65	2.83
standard deviation	(3.18)	(3.65)	(4.10)
max papers per author	89	101	129
size of giant component	5253	13808	33027
as percentage	.156	.284	.407
second largest component	122	30	30
isolated authors	16735	19315	24578
as percentage	.496	.397	.303
mean links per author	.894	1.244	1.672
standard deviation	(1.358)	(1.765)	(2.303)
max links per author	25	35	54
mean size 10-neighborhood	.007	.037	.118
clustering coefficient	.193	.182	.157
Giant component			
mean papers per author	4.80	4.63	4.48
standard deviation	(5.91)	(5.53)	(5.67)
percentage coauthored	.410	.476	.541
mean links per author	2.48	2.77	3.06
standard deviation	(2.09)	(2.40)	(2.93)
mean distance	12.86	11.07	9.47
standard deviation	(4.03)	(3.03)	(2.23)
maximum distance (diameter)	40	36	29
mean size 10-neighborhood	.288	.456	.712
clustering coefficient	.136	.151	.136

Table 6: Network statistics for the network based on all articles in 113 selected journals (TI list).

period	70's	80's	90's
total authors	14051	19694	28736
mean papers per author	2.45	2.77	2.85
standard deviation	(3.15)	(3.50)	(3.57)
max papers per author	62	69	49
size of giant component	2775	7283	14368
as percentage	.197	.370	.500
second largest component	74	32	31
isolated authors	5859	5999	6156
as percentage	.417	.305	.214
mean links per author	1.058	1.467	1.896
standard deviation	(1.433)	(1.815)	(2.224)
max links per author	21	27	43
mean size 10-neighborhood	.015	.060	.165
clustering coefficient	.188	.180	.167
Giant component			
mean papers per author	4.79	4.39	4.02
standard deviation	(5.45)	(4.89)	(4.53)
percentage coauthored	.411	.502	.590
mean links per author	2.48	2.70	2.95
standard deviation	(2.05)	(2.25)	(2.61)
mean distance	11.99	11.12	9.69
standard deviation	(4.02)	(3.07)	(2.35)
maximum distance (diameter)	33	31	26
mean size 10-neighborhood	.379	.438	.658
clustering coefficient	.139	.154	.149

Table 7: Gini coefficients computed for several data sets.

period	70s	80s	90s
Whole network			
All journals	.659	.619	.585
TI List	.609	.559	.524
Network excluding isolated authors			
All journals	.324	.367	.404
TI List	.329	.366	.394
Giant component			
All journals	.381	.395	.420
TI List	.378	.385	.402

Table 8: Network statistics for the economists with the highest number of links.

Author	Papers	% Coauthored	Links	Distance 2	Clust.Coeff
		1970s			
tollison rd	44	0.955	25	57	0.053
heady eo	30	0.833	23	13	0.028
feldstein ms	73	0.288	21	40	0.024
schmitz a	23	0.870	20	29	0.042
smith vk	72	0.514	20	26	0.032
Average top 100	23.87	0.724	11.94	25.67	0.062
Average all	2.35	0.243	0.89		0.193
		1980s			
mccarl ba	36	0.889	35	97	0.022
thisse jf	34	0.971	30	80	0.055
lee cf	36	1.000	29	106	0.030
whalley j	52	0.808	29	44	0.022
schmitz a	26	0.846	26	118	0.058
Average top 100	28.42	0.827	16.36	49.80	0.062
Average all	2.65	0.315	1.24		0.182
		1990s			
thisse jf	66	0.970	54	244	0.022
lee j	58	0.586	45	158	0.019
sirmans cf	67	1.000	41	172	0.045
nijkamp p	67	0.940	41	57	0.034
michel p	48	0.938	34	169	0.036
Average top 100	37.69	0.849	25.31	99.40	0.043
Average all	2.82	0.409	1.67		0.157

Table 9: Error and attack tolerance of the network based on all articles in EconLit.

period	70's	80's	90's
Size of the giant component (in perc.)			
Whole network	.156	.284	.407
w/o random 2%	.149	.276	.398
w/o random 5%	.137	.263	.389
w/o top 2%	.002	.067	.256
w/o top 5%	.000	.001	.001
Average distance within giant component			
Whole network	12.86	11.07	9.47
w/o random 2%	12.88	11.17	9.58
w/o random 5%	12.89	11.21	9.68
w/o top 2%	9.26	29.80	19.00
w/o top 5%	2.64	5.71	8.91
Average size of 10-neighborhood			
Whole network	.007	.037	.118
w/o random 2%	.006	.034	.110
w/o random 5%	.005	.030	.102
w/o top 2%	.000	.000	.002
w/o top 5%	.000	.000	.000
Clustering coefficient			
Whole network	.193	.182	.157
w/o random 2%	.193	.183	.158
w/o random 5%	.192	.182	.157
w/o top 2%	.318	.280	.250
w/o top 5%	.440	.380	.344

Table 10: Summary statistics for all articles in *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics* and *Review of Economic Studies*

period	70's	80's	90's
total papers	5023	4565	3705
mean pages per paper	10.65	13.04	17.47
standard deviation	(7.73)	(9.65)	(11.29)
# Authors per paper: Distribution			
single-authored	.743	.604	.458
two authors	.225	.330	.424
three authors	.029	.059	.099
four or more authors	.003	.007	.019

Table 11: Summary statistics for all articles in journals that have been covered by EconLit since 1970.

period	70's	80's	90's
total papers	42115	46075	49245
mean pages per paper	12.53	13.79	16.12
standard deviation	(9.96)	(10.24)	(11.74)
# Authors per paper: Distribution			
single-authored	.753	.674	.587
two authors	.213	.263	.311
three authors	.030	.055	.085
four or more authors	.005	.008	.018

Table 12: Network statistics for the network based on all articles in *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics* and *Review of Economic Studies*.

period	70's	80's	90's
total authors	3186	3387	3171
mean papers per author	2.02	1.94	1.87
standard deviation	(2.05)	(1.94)	(1.78)
max papers per author	34	33	21
size of giant component	237	608	779
as percentage	.074	.180	.246
second largest component	25	28	22
isolated authors	1507	1143	701
as percentage	.473	.337	.221
mean links per author	.833	1.142	1.429
standard deviation	(1.095)	(1.279)	(1.405)
max links per author	12	17	17
mean size 10-neighborhood	.005	.014	.029
clustering coefficient	.253	.259	.257
Giant component			
mean papers per author	4.08	3.48	3.05
standard deviation	(4.05)	(3.24)	(2.68)
percentage coauthored	.446	.540	.680
mean links per author	2.45	2.45	2.55
standard deviation	(1.80)	(1.77)	(1.82)
mean distance	7.94	11.92	11.02
standard deviation	(3.43)	(5.22)	(4.12)
maximum distance (diameter)	22	33	29
mean size 10-neighborhood	.775	.435	.475
clustering coefficient	.143	.183	.190

Table 13: Network statistics for the network based on all articles in journals that have been covered by EconLit since 1970.

period	70's	80's	90's
total authors	22960	27539	32773
mean papers per author	2.33	2.28	2.20
standard deviation	(3.02)	(2.77)	(2.62)
max papers per author	86	97	122
size of giant component	3076	5899	10054
as percentage	.134	.214	.307
second largest component	82	57	23
isolated authors	11260	11062	10572
as percentage	.490	.402	.323
mean links per author	.885	1.134	1.386
standard deviation	(1.312)	(1.508)	(1.695)
max links per author	23	25	30
mean size 10-neighborhood	.006	.013	.028
clustering coefficient	.198	.218	.216
Giant component			
mean papers per author	4.78	4.00	3.44
standard deviation	(5.60)	(4.31)	(3.83)
percentage coauthored	.413	.518	.585
mean links per author	2.45	2.62	2.70
standard deviation	(2.03)	(2.11)	(2.18)
mean distance	12.15	12.63	12.33
standard deviation	(3.75)	(3.65)	(3.36)
maximum distance (diameter)	29	37	34
mean size 10-neighborhood	.341	.279	.296
clustering coefficient	.134	.168	.174

Table 14: Comparison across subjects: 1995-1999[1]

	Biomedical	Physics	Maths	Computer Sc.	Economics
1. Total papers	2163923	98502	587000	13169	156454
2. Total Authors	1520251	52909	192000	11994	81217
3. Mean papers per author	6.4	5.1	4.97	2.55	2.83
4. Mean authors per paper	3.75	2.53	1.63	2.22	1.56
5. Mean links	18.1	9.7	2.84	3.59	1.672
6. Clustering coefficient	.066	.43	.15	.496	.157
7. Giant component	1395693	44337	n.a.	6396	33027
% of population	92.6	85.4	60*	57	40
8. Mean distance	4.6	5.9	9*	9.7	9.47
9. Max Distance	24	20	12*	31	29

[1] The data for medicine, physics and computer science is taken from Newman (2001c). The data for biomedical sciences comes from the Medline database and covers publications in referreed journals, the data for physics comes from the Los Almos Archive and covers preprints, the data for computer science comes from the database NCSTRL, and covers preprints. The data for mathematics comes from a database on articles in Mathematical Reviews and covers all papers from 1940 onwards; our summary numbers are taken from Grossman (2002). The data for mathematics and economics covers the 1990-1999 period. The data pertaining to mathematics marked with a star relates to the network with all papers dating from 1940.

References

[1] Albert, R. and A-L. Barabási (2002), Statistical mechanics of complex networks, *Review of Modern Physics*, 74, 47–97.

[2] Allen, F. and D. Gale (2000), Financial Contagion, *Journal of Political Economy*, 108, 1-33.

[3] Bala V. and S. Goyal (1998), Learning from neighbours, *Review of Economic Studies* 65, 595-621.

[4] Bala, V. and S. Goyal (2000), A non-cooperative model of network formation, *Econometrica*, 68, 5, 1181-1231.

[5] Bala, V. and S. Goyal (2001), Conformism and diversity under social learning, *Economic Theory*, 17, 101-120.

[6] Barabási, A. and Albert, R. (1999), Emergence of scaling in random networks, *Science*, 286, 509-512.

[7] Barabási, A., Albert, R., and Jeong, H (1999), Mean-field theory for scale-free random networks, *Physica A*, 272, 173–187.

[8] Bollobás B. (1985), *Random Graphs*, Academic Press, London.

[9] Coleman, J. (1966), *Medical Innovation: A Diffusion Study*, Second Edition, Bobbs-Merrill, New York.

[10] Dorogovtsev, S. and J. Mendes (2002), Evolution of networks, *Advances in Physics*, 51, 1079-1187.

[11] Dutta, B. and M. Jackson (eds) (2003), *Networks and Groups: Models of Strategic Formation*, in series: 'Studies in Economic Design'. Springer-Verlag, Berlin.

[12] Eisenhauer, J. (1997), Multi-authored papers in economics, *Journal of Economic Perspectives*, 11, 191-192.

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

[14] Ellison, G. (2000b), Evolving Standards for academic publishing: A q-r theory, *Journal of Political Economy*, 110, 5, 994-1034.

[15] Ellison, G. and D. Fudenberg (1993), Rules of Thumb for Social Learning, *Journal of Political Economy*, 101, 612-644.

[16] Eshel, I., Samuelson, L. and A. Shaked (1998), Altruists, egoists, and hooligans in a local interaction model, *American Economic Review*, 88, 157-179.

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

[18] Goyal, S. (1996), Interaction structure and social change, *Journal of Institutional and Theoretical Economics*, 152, 3, 472-495.

[19] Goyal, S. (2003), Learning in networks: a survey. mimeo.

[20] Grossman, J (2002), The evolution of the mathematical research collaboration graph, *Congressus Numerantium*, 158, 201–212.

[21] Hamermesh, D and Oster, S (2002), Tools or toys: The impact of high technology on scholarly productivity, *Economic Inquiry*, 40, 539–555.

[22] Hagerstrand, T. (1969), *Innovation diffusion as a spatial process.* University of Chicago Press. Chicago.

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

[24] Jackson, M. (2003), A survey of models of network formation: Stability and efficiency. mimeo.

[25] Jackson, M. and A. Wolinsky (1996), A strategic model of economic and social networks, *Journal of Economic Theory*, 71, 44-74.

[26] Kochen, M. (ed) (1989), *The small world*, Ablex, Norwood, NJ.

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

[28] Laband, D. and Tollison, R. (2000), Intellectual collaboration, *Journal of Political Economy*, 108, 3, 632-662.

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

[30] Milgram, S. (1967), The small world problem, *Psychology Today*, 2, 60-67.

[31] Morris, S. (2000), Contagion, *Review of Economic Studies*, 67, 57-79.

[32] Newman, M. (2003), Coauthorship networks and patterns of scientific collaboration, *Proceedings of the National Academy of Sciences*, in press.

[33] Newman, M. (2001a), Scientific collaboration networks: I. Network construction and fundamental results, *Physical Review E*, 64.

[34] Newman, M. (2001b), Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Physical Reviews E*, 64.

[35] Newman, M. (2001c),The structure of scientific collaboration networks, *Proceedings of the National Academy of Sciences*, 98, 404-409.

[36] Pool, I de Sola. and Kochen, M. (1978), Contacts and Influence, *Social Networks*, 1: 1-48.

[37] Ravasz E. and A.-L. Barabási (2003), Hierarchical organization in complex networks, *Physical Review E*, 67, art. 026112.

[38] van den Nouweland, A. (2003), Models of network formation in cooperative games. mimeo.

[39] Watkins, S. (1991), *Provinces into Nations: Demographic Integration in Western Europe, 1870-1960*, Princeton University Press, Princeton, New Jersey.

[40] Watts, D. (1999), *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, New Jersey.

[41] Watts, D. and Strogatz, S. (1998), Collective dynamics of 'small world' networks, *Nature*, 393, 440–442.

[42] Young, P. (1998), *Individual Strategy and Social Structure*, Princeton University Press, NJ.

Figure 1: Pareto plot of links per author for the network based on all articles: 1970s to 1990s.

Figure 2: Lorenz curves of links per author for the network based on all articles: 1970s to 1990s.

Figure 3: Neighborhood of network around J. Stiglitz in the 1990's.

Figure 4: Neighborhood of network around J. Tirole in the 1990's

Figure 5: Neighborhood of network around J.F. Thisse in the 1990's

Figure 6: Neighborhood of network around A. Dixit in the 1990's

Figure 7: Fraction of coauthored articles in AA journals, A journals, B journals and other journals.

Figure 8: Equilibrium networks

Figure 9: Equilibrium networks with size constraints.

Figure 10: Degree versus clustering.

Figure 11: The construction of a hierarchical network. Source: Ravasz and Barabási (2003)

Figure 1

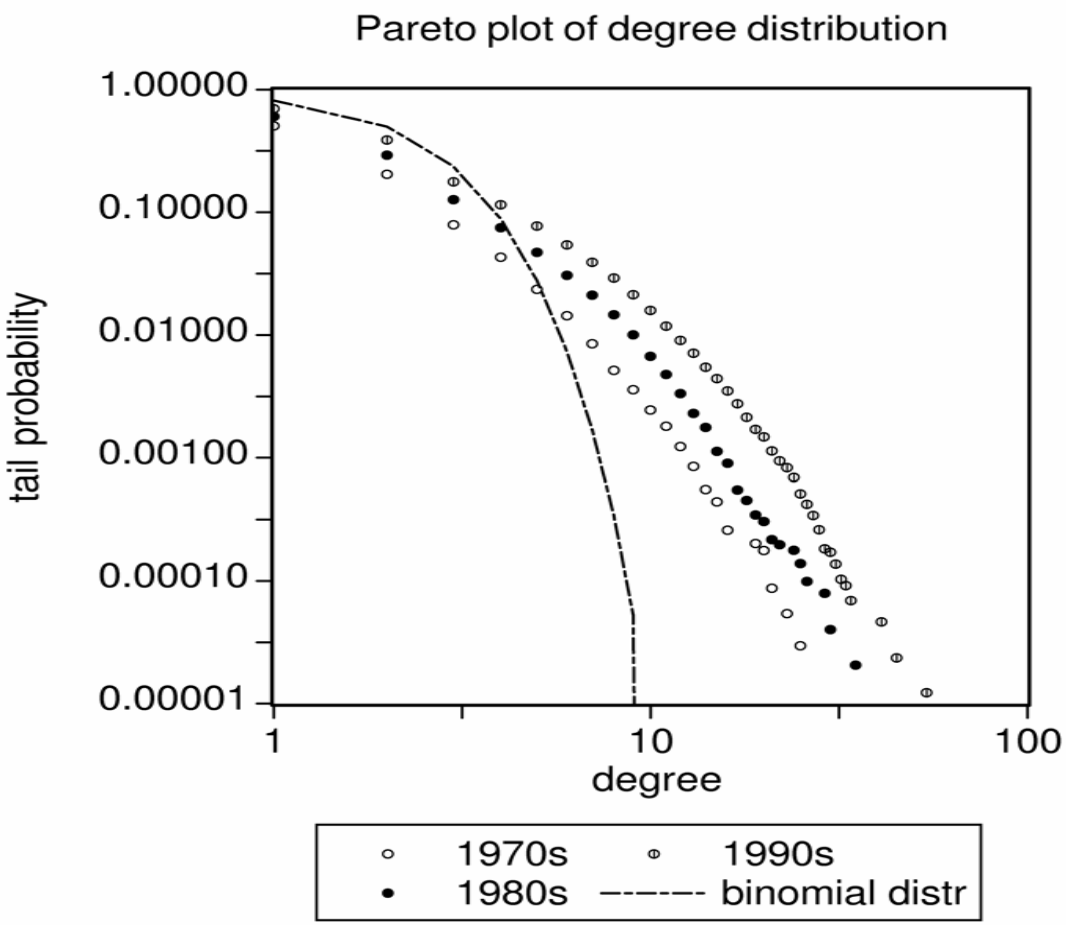

Pareto plot of degree distribution
1.00000
0.10000
0.01000
0.00100
0.00010
0.00001
tail probability
1
10
100
degree
1970s
1980s
1990s
binomial distr

Figure 2

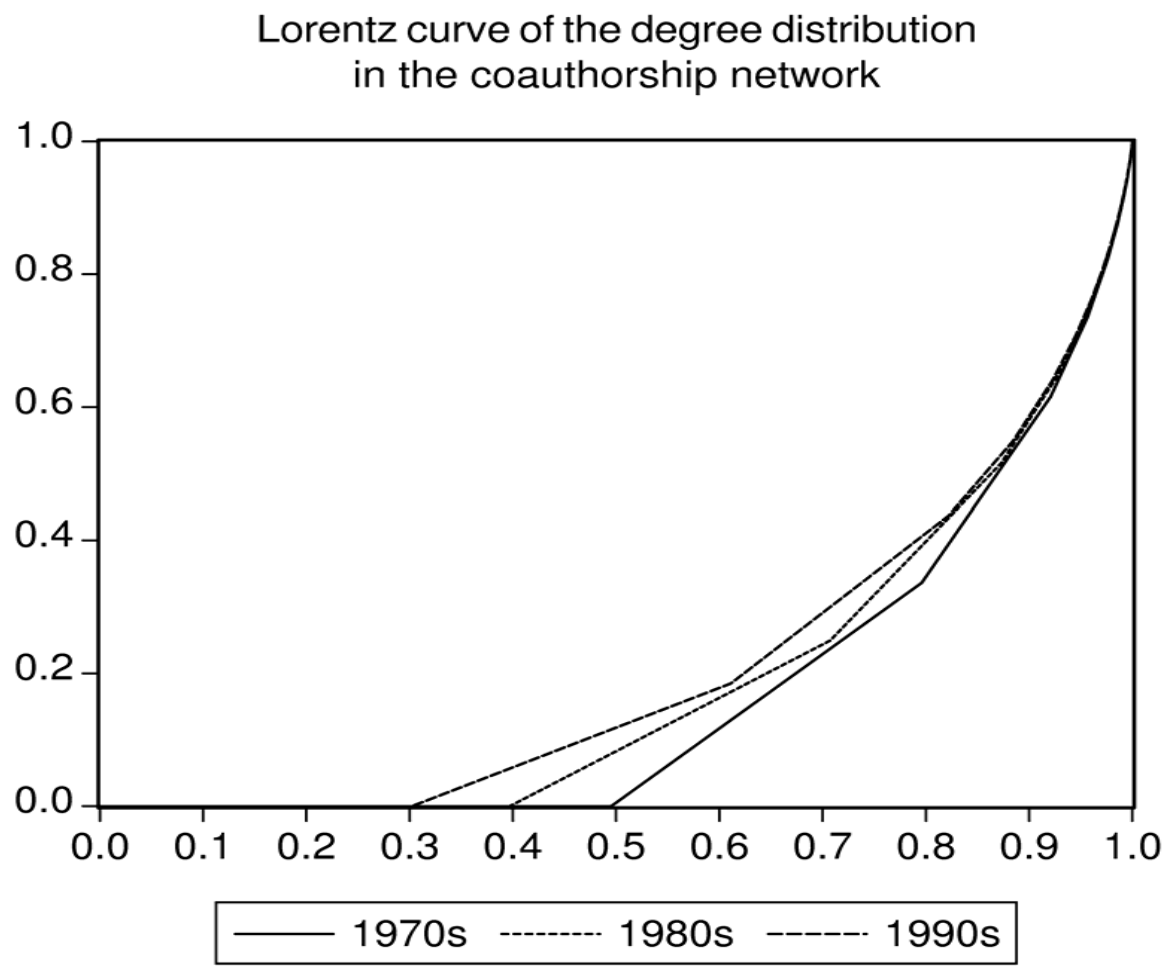
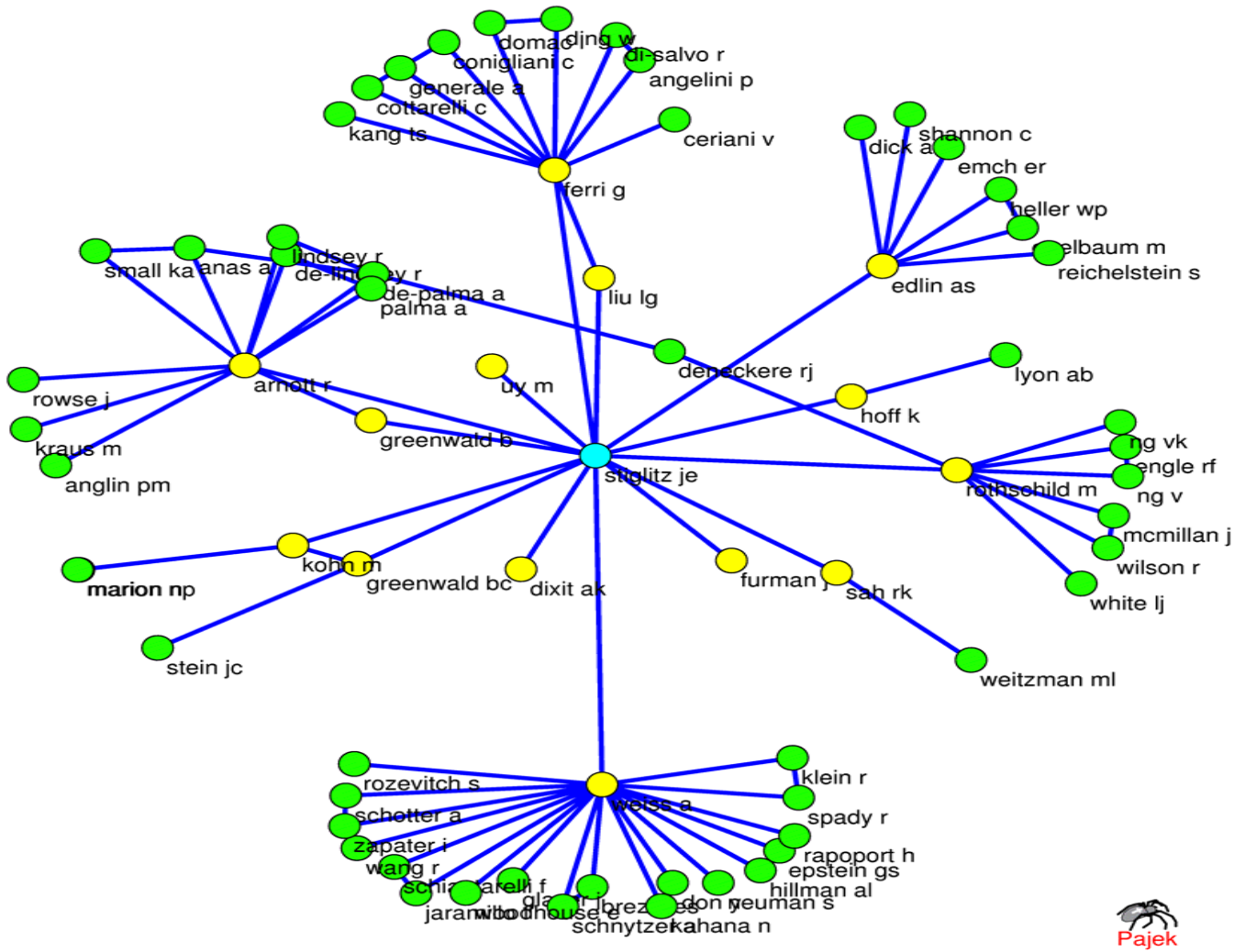
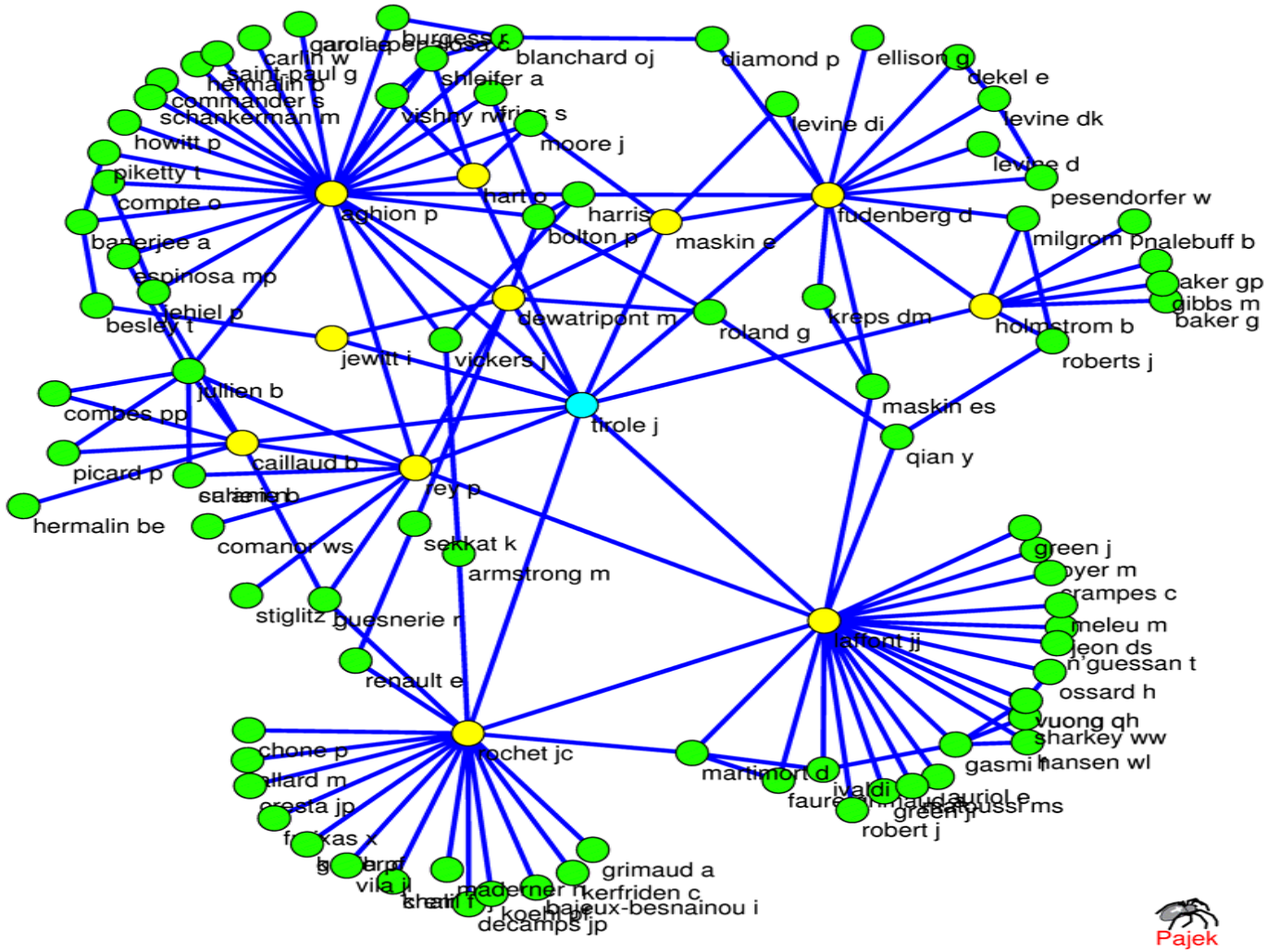

Lorentz curve of the degree distribution
in the coauthorship network
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
1970s
1980s
1990s

Figure 3: Local network of J. Stiglitz in 1990's



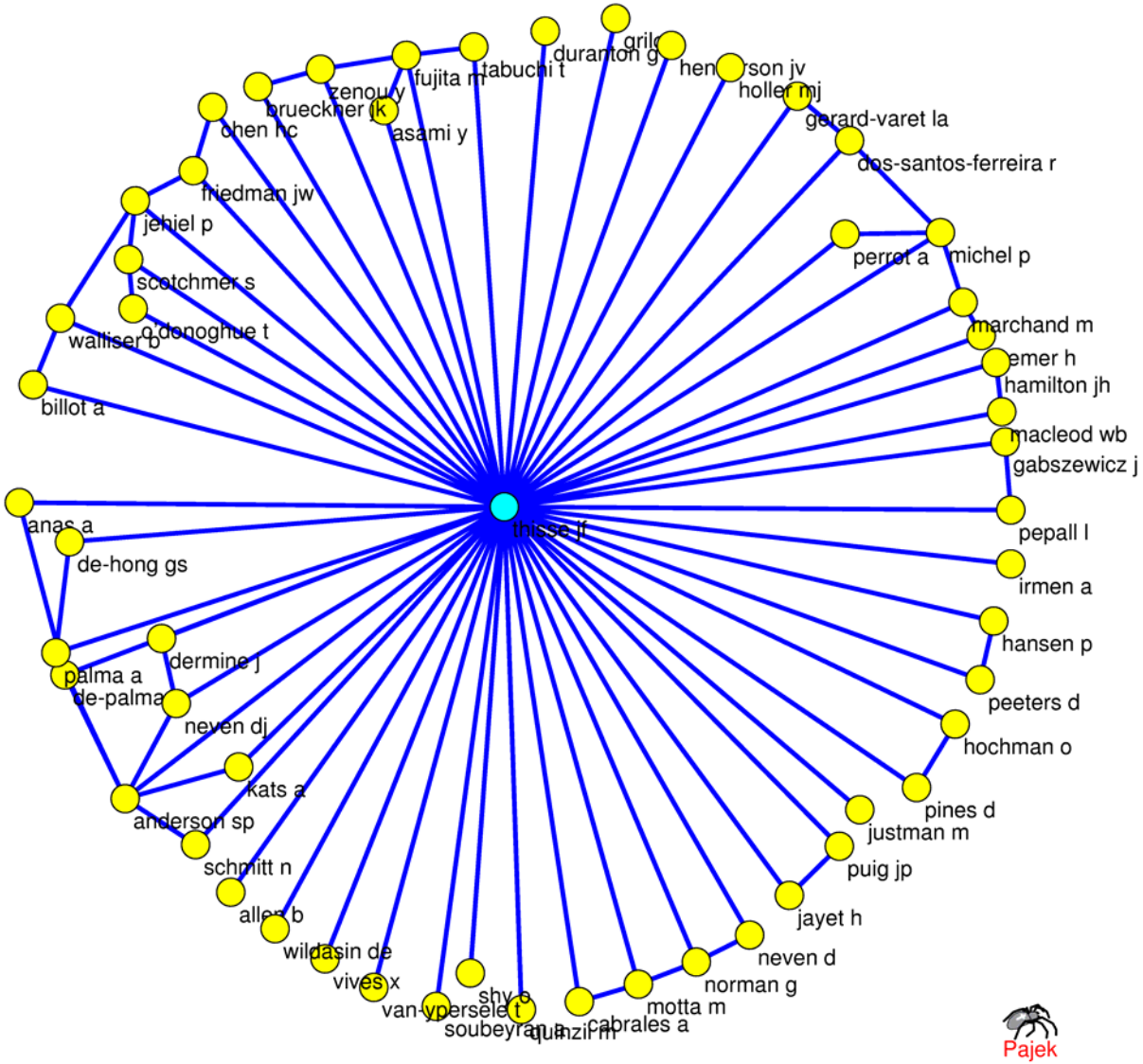
Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Figure 4: Local network of J. Tirole in 1990's



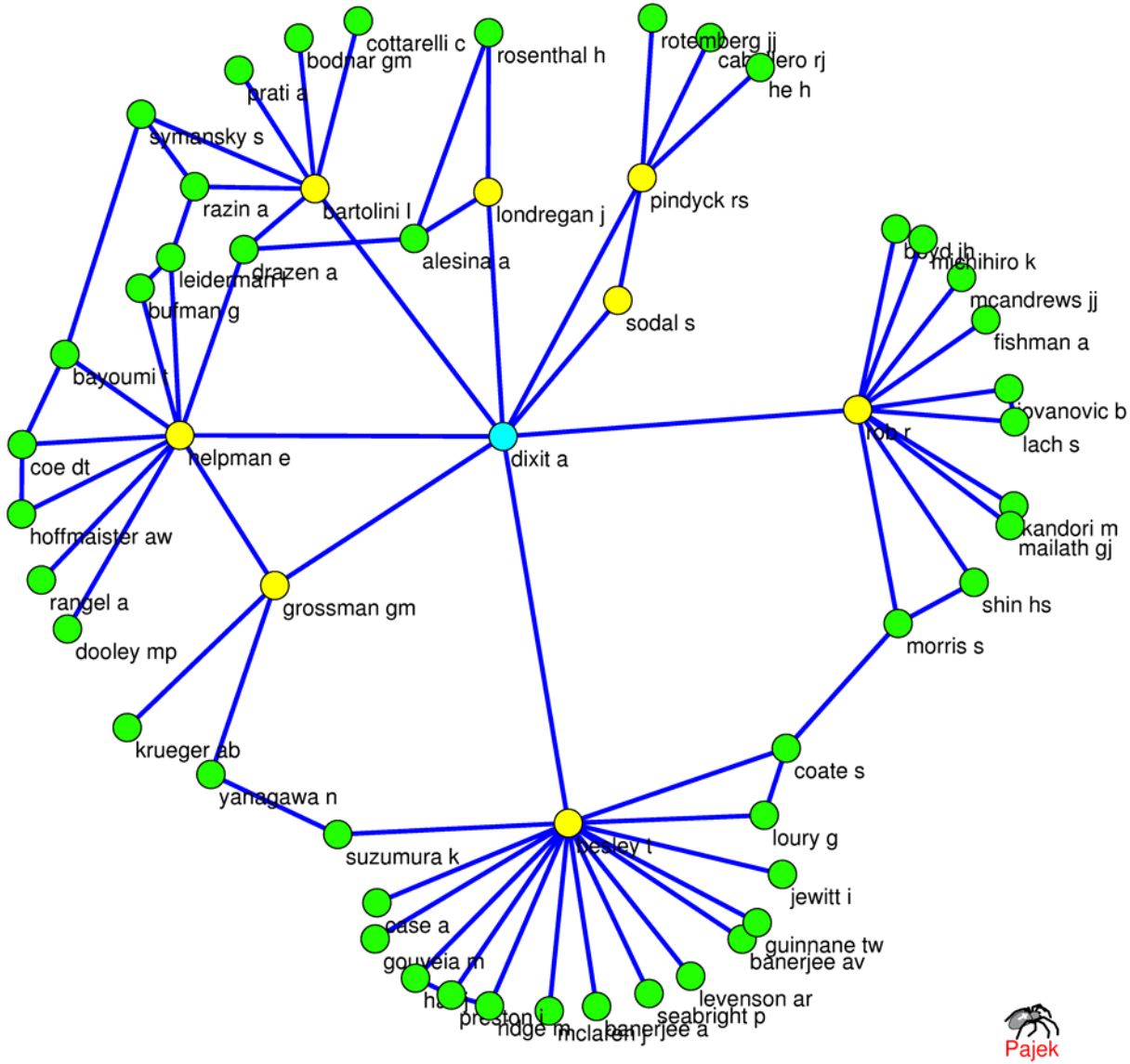
Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Figure 5: Local network of J. Thisse in 1990's



Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Figure 6: Local network of A. Dixit in 1990's



Note: Some economists might appear twice or are missing due to the use of different initials or misspellings in EconLit.

Figure 7
Collaboration across journals

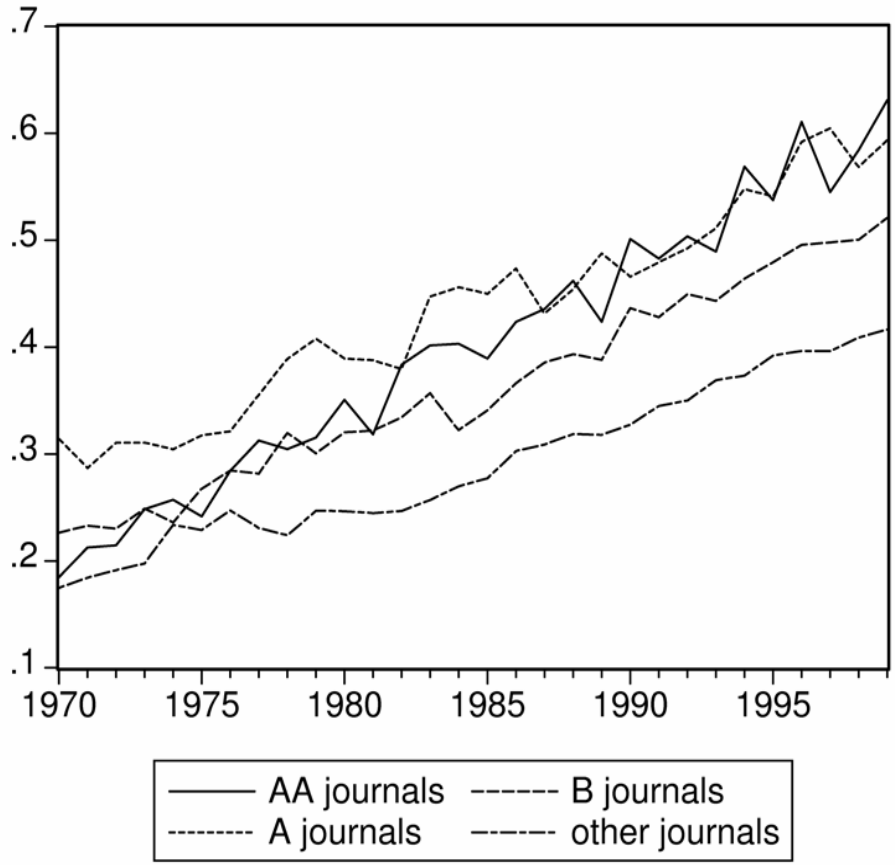
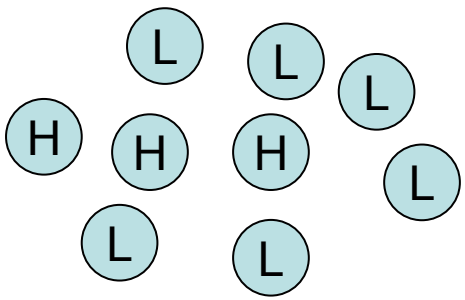


Figure 8: Symmetric equilibrium networks

$n_h=3$ & $n_l=6$

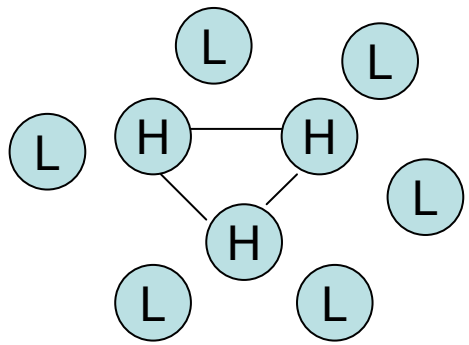
$f>f_{hh}$



Empty network

$f_{ll}<f<f_{hh}$

$\eta_{hh}=2$

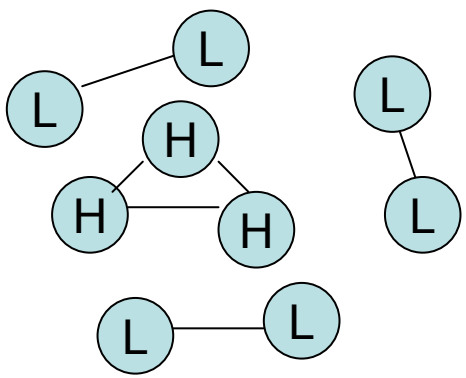


Clique of High Types + isolated low types

$f<f_{ll}$

$\eta_{ll}=1$

$\eta_{hh}=2$



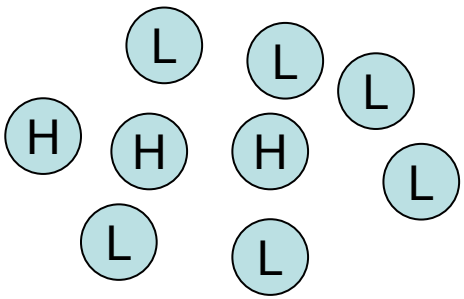
Separate cliques

Figure 9: Symmetric equilibrium networks with size constraints.

$n_h=3$ & $n_l=6$

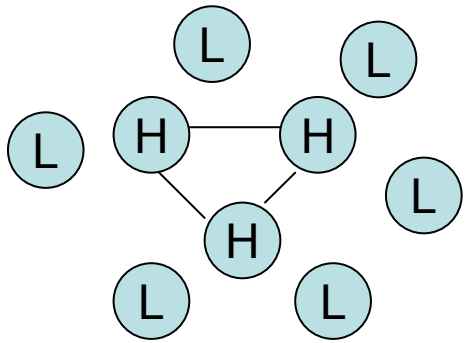
$f>f_{hh}$

$\eta_{hh}>n_h-1$



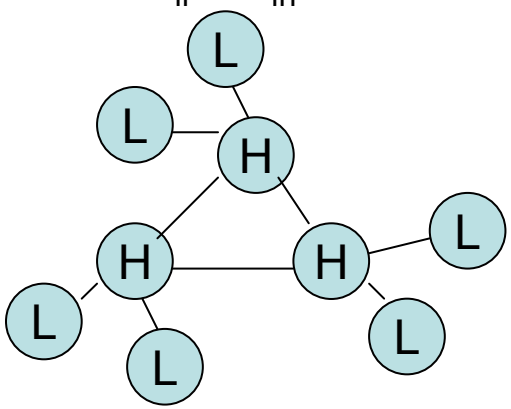
Empty network

$f_{lh}<f<f_{hh}$



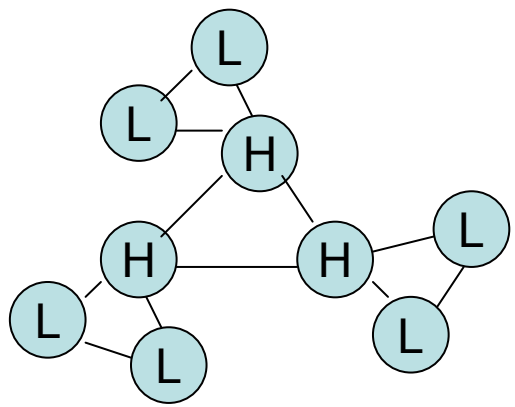
Clique of High Types

$f_{ll}<f<f_{lh}$



Interlinked stars

$f<f_{ll}$



Interlinked stars + high clustering

Figure 10
Degree vs clustering

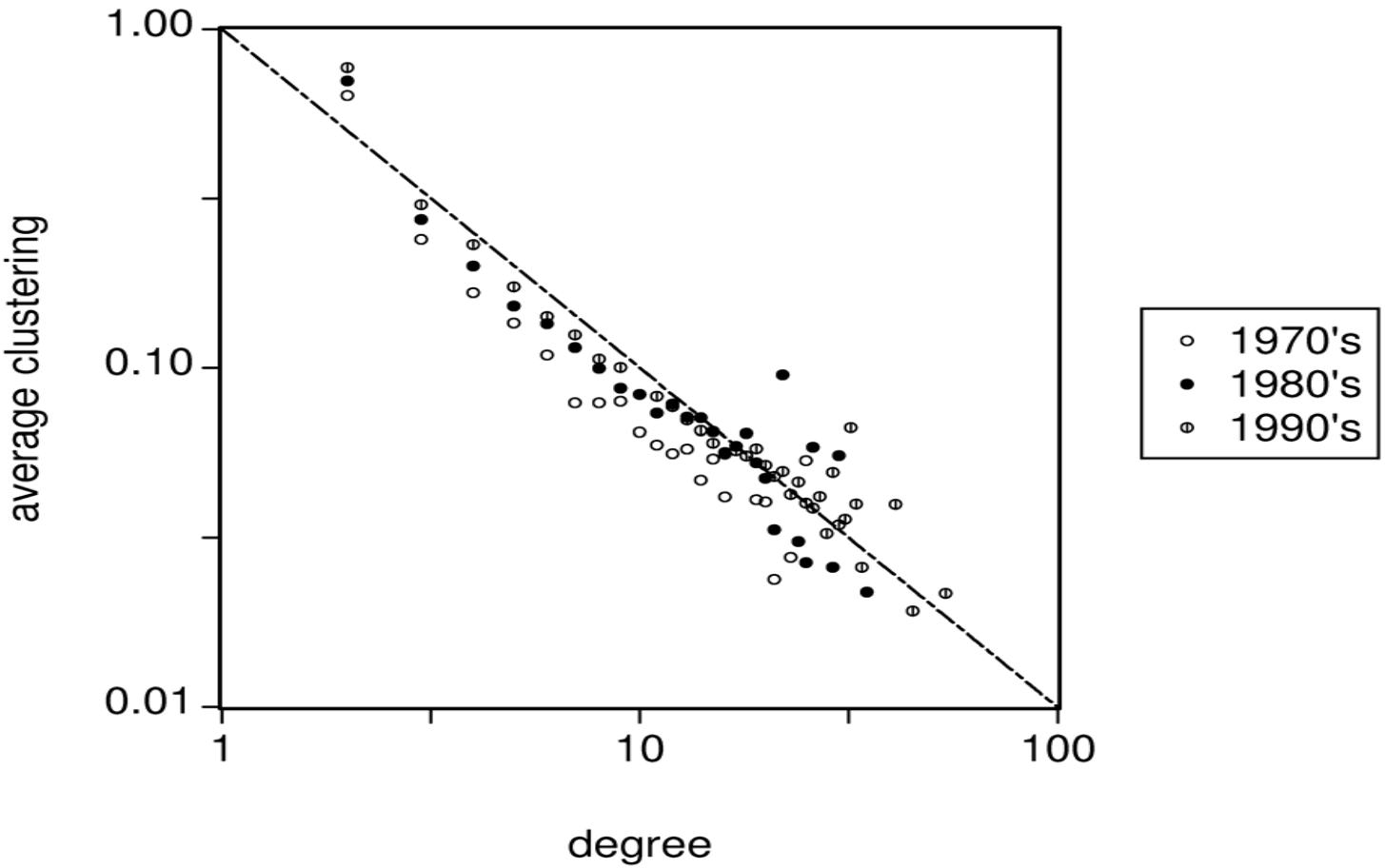
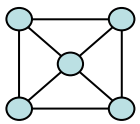
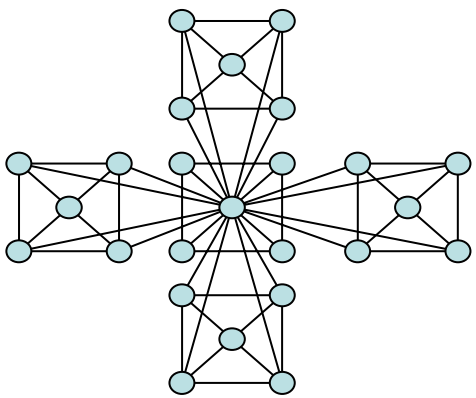


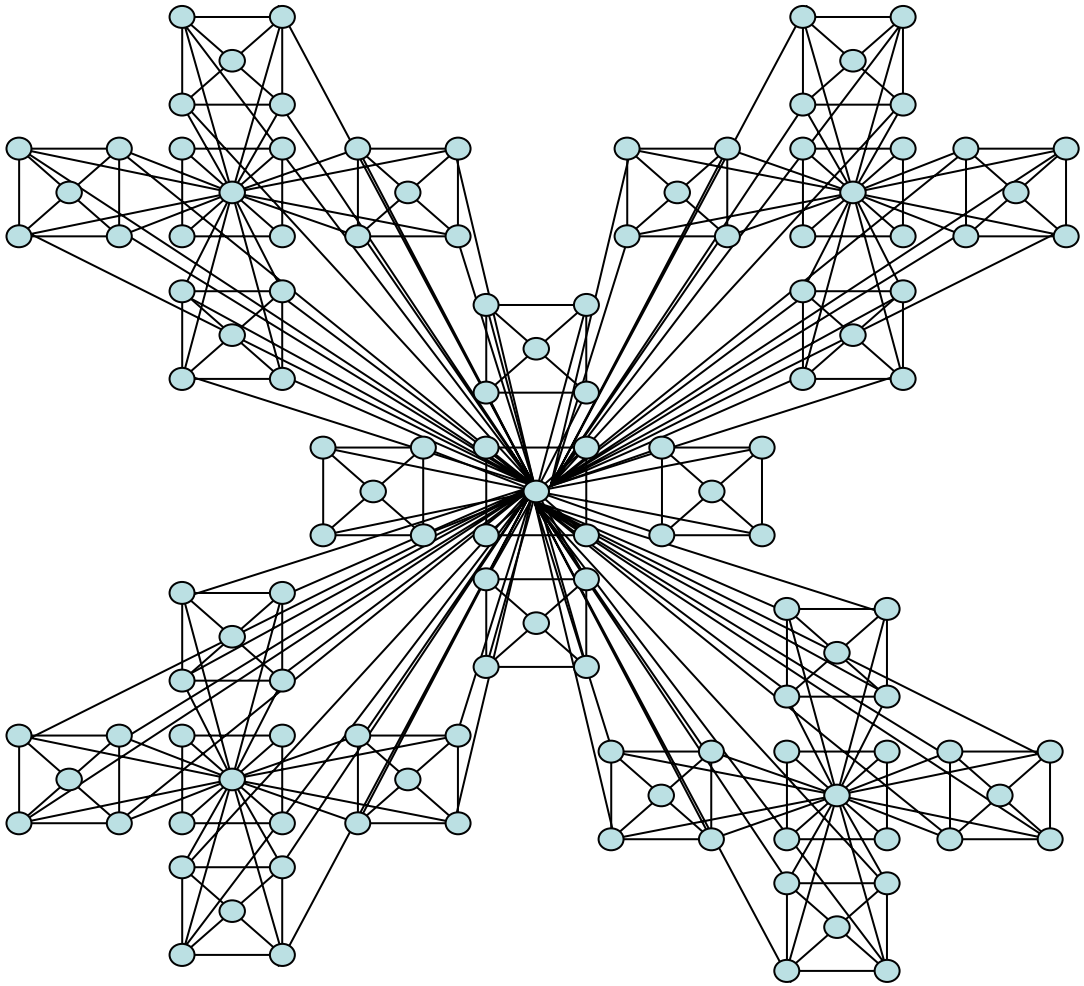
Figure 11
The construction of a hierarchical network



(a) 1 level, n=5



(b) 2 levels, n=25



(c) 3 levels, n=125