Specifying Social Weight Matrices of Researcher Networks: The Case of Academic Economists

Katarina Zigova

Working Paper Series
2017-10
This paper shows how collaboration and citation networks can be used to specify social weight matrices for a community of researchers. I use two competing theories of social influence on individual behavior, namely communication and comparison. I argue that in research networks collaboration captures communication, while citation captures comparison. I further argue that the comparison principle is likely to be the main social driver of individual research productivity and suggest a benchmark social weight matrix based on this principle. I test the benchmark matrix against several alternatives using a Bayesian model comparison procedure and conclude that the benchmark matrix outperforms alternative specifications. This result lends support to socio-economic theories underlying the benchmark specification of the social weight matrix.

Keywords: collaboration and citation networks, social weight matrix, network autocorrelation, research productivity

JEL Classification: C21, D83, I23

1 Introduction

Networking is a natural phenomenon in any social community. Social network analysis acknowledges that the behavior and the beliefs of an individual are influenced by the behavior and the beliefs of other individuals, especially by those with whom he is directly connected. This influence effect is referred to as social contagion (e.g. Leenders 2002), social effect (e.g. Manski 1993), peer and neighborhood effects (e.g. Moffit 2001), and many other alternative terms, all referring to the effect of social interactions on individual behavior. In this paper, I focus on matrix representations of research networks.

In the research community, networks naturally arise since researchers often work together on a common research topic. As a matter of fact, joint investigation has become more frequent in the recent decades (e.g. Laband and Tollison 2000, Bosquet and Combes 2013). Research collaboration can start with an informal exchange of ideas, followed by structured discussions about common research interests, which later may lead to collaboration. This kind of interaction ends, ideally, with a joint publication. The total of all collaboration activities in a research community constitute a network in which the nodes are the
researchers and the links between them are coauthorships. Joint publications provide a precise definition of the network ties (Newman 2001a) as compared to the rather vague concept of traditional friendship ties or friendship relationships in virtual social media.

In research, not only bi- and multilateral interactions play a role, but also unilateral interactions, namely digesting and adopting the literature. This kind of interaction can (and usually does) exclude personal contact and is, again, precisely defined via referencing. Citations are explicit linkages between studies that have some important content in common (Hummon and Doreian 1989). As compared to the mutual linkages among coauthors, citation links are ‘one-way’, and no reciprocity in referring to each other’s publications needs to take place.

Publications indicate individual productivity. Research productivity is therefore measurable. Econometric network effects models (e.g. Marsden and Friedkin 1993) offer a way to analyze the effect of network interactions on observable outcome variables such as research productivity. The crucial issue in these econometric models is the definition of the so-called weight matrix $W$, which operationalizes the influence working through the network (Leenders 2002, Corrado and Fingleton 2012).

In this study, I show how the specification of a social weight matrix can be derived from a given social network structure by relying on sociological theories of communication and comparison. I illustrate this approach with the documented collaboration and citation behavior of all economists affiliated with a university or research institute in Austria, Germany and German-speaking Switzerland. For that purpose, I use the economists’ journal publications in the 2000-2010 period.

The structure of collaboration networks of researchers has been studied in many disciplines, starting with Newman (2001a, 2001b, 2004) who analyzed collaboration patterns in mathematics, physics and biology. Other examples include philosophy and psychology (Cronin, Shaw, and la Barre 2003), sociology (Moody 2004), economics (Goyal, van der Leij, and Moraga-González 2006), and management science (Acedo et al. 2006). In contrast to studies on collaboration networks, studies on citation networks do not focus on structural aspects. They originate in information science and are targeted at identifying the most influential studies. Hence, most of the studies analyzing citation networks link articles (or journals), rather than individuals. An early well-known example is the study by Derek J. de Solla Price (1965) who analyzes the world citation network of scientific papers using Garfield’s 1961 Science Index Citation data (Garfield 1963). Hummon and Doreian (1989) use the citation network of publications to investigate the chronological sequence of studies which led to the discovery of the DNA. The more recent investigation by White, Wellman, and Nazer (2004) is closer to my approach. It analyzes the evolution of the citation pattern between 16 international researchers (from seven disciplines) all working on a common project.

In my study, the network is viewed as an intermediary input for operationalizing the social weight matrix. I build on theories of social influence which distinguish between contagion via communication and contagion via comparison (Leenders 2002). In the research community, contagion by communication takes place when two or more researchers collaborate. The idea of communication assumes a direct contact between individuals to achieve a common objective. Moreover, contact is always mutual. In contrast, the idea of social contagion via comparison does not rely on direct contacts between individuals, it rather assumes that individuals mimic the behavior of those individuals who are similarly embedded in the network. I use the citation network as a representation of contagion via comparison and I argue that researchers use citations as a means to align their research with the research of the cited peers. Burt (1987, 2010) shows that in competitive environments, comparison is the driving force of network contagion. This finding relates directly to academia since there is abundant evidence that scientists are highly competitive (Stephan 2010). I therefore explore the collaboration and citation networks for the
same community of scientists in order to be able to compare the relative importance of the two types of network effects. Moreover, these two networks that represent the communication and comparison conduits of research productivity allow me to construct the underlying matrix representation.

The matrices based on these networks I call social weight matrices since, technically speaking, they correspond to the spatial weight matrix. Following the tradition, I use the letter $W$ to indicate this weight matrix. $W$ is a square matrix and in my case it is of the size of the network. The $i, j$-th cell of this matrix measures the proximity between the $i$-th and the $j$-th node. The econometric analysis of the network effects uses the same mathematical apparatus as spatial econometrics; the units are, however, individuals and the $W$ matrix mirrors their underlying network structure. One of the issues in solving a network autocorrelation problem (or, for that matter, a spatial econometric model) is to select from a set of weight matrices that are compatible with theoretical considerations the matrix that actually drives the observed social autocorrelation process.

The standard spatial econometric approach to selecting the weight matrix is to test several reasonable matrices using statistical tests (e.g. LeSage and Pace 2009, ch. 6). To address the current critique that calls for a more substantive approach in specifying $W$ (Corrado and Fingleton 2012), I followed a reverse approach. I specify the benchmark social weight matrix $W$ that is compatible with the concept of comparison, i.e. with the competitive nature of research production. In operationalizing the benchmark matrix $W$, I moreover incorporate the sociological view on the sociometric distance and account for the strength of bilateral connections between individuals. To assess the statistical performance of the benchmark $W$, I then specify five alternative matrices, each of them relaxing one criteria imposed on the benchmark matrix $W$. Finally, I use the Bayesian model comparison approach (LeSage and Pace 2009) to test the performance of the benchmark matrix $W$ against these alternatives. The benchmark matrix outperforms all alternative specifications in 4 types of spatial econometric models. This result supports the arguments that comparison, and for that matter competition, is the prevailing social driver of research productivity.

The structure of this paper is as follows. In the next section, I present and visualize properties of the collaboration and citation networks of the considered set of researchers. In Section 3, I justify the choice of the benchmark weight matrix $W$ and determine the alternatives. I then test the performance of the benchmark weight matrix against the alternatives in Section 4. The last section concludes.

2 Researcher Networks: Descriptive Analysis

In academia, a variety of social interactions take place, both formal and informal. Already in early 1970s, (Crane 1972, ch. 3) delineated four main types of social interactions in the research community: informal discussions of research, joint publications, teacher-student relationships, and influence by publications of other scientists on the choice of research projects and techniques. All of these interactions, if properly documented, could constitute a (social) network in which the nodes represent researchers and links between them the portrayed relationships. Informal relationships are harder to record than formal ones because the researchers, even if they responded to a questionnaire, may not recall everybody with whom they discussed their research. Crane (1972), for example, observed in her sample of mathematicians and

---

1 The spatial weight matrix represents the spatial distribution of units being analyzed and is a crucial input in spatial econometric models (LeSage and Pace 2009).
sociologists that when asked to reveal who influenced their work, 70% of the names provided referred to their citations. Similarly, Melin and Persson (1996) studied research collaboration at the Umea University in Sweden. They found that only 5% of research collaboration does not result in any publication output. These facts allow me to concentrate on social interactions by way of collaboration on joint publications and by way of citations, because these interactions capture the overwhelming majority of social ties in research.

The network based on joint publications is referred to as collaboration network. The nodes are researchers and the links between them represent joint publications. Probably the most prolific author studying scientific collaboration networks is Newman (e.g. 2001a). He finds that collaboration links are consistent in the sense that each link is well documented and all links are of the same type. Citation networks are also consistent in this sense: references are easily observable and of the same type. Collaboration and citation networks are therefore easy to handle because they can be represented by a so called adjacency matrix \( A = a_{ij}, \forall i, j \). If matrix \( A \) represents a collaboration network, \( a_{ij} = a_{ji} = 1 \) if there is a collaboration link between \( i \) and \( j \) and zero otherwise. The adjacency matrix \( A \) of a citation network is in general not symmetric, because here the element \( a_{ij} = 1 \) if \( i \) cites \( j \), but \( a_{ji} = 1 \) would imply that \( j \) also cites \( i \).

This study is based on the publication records of all academic economists affiliated with universities or research institutes in Austria, Germany and German-speaking Switzerland. I retrieved the data from the research monitoring database. This database comprises, in addition to academic economists, also researchers in business administration and finance. It records, on the one hand, the scientists’ relevant individual characteristics such as age, gender, Ph.D. cohort, Ph.D. granting institution, and research field. On the other hand, it provides information on the scientists’ journal publications such as journal name, publication year, title, and (co)authors. The database is under the auspices of German Economic Association. The database is used for the rankings published by the German business magazine *Handelsblatt*. Its main purpose is, however, to support bibliometric research.

For network studies, it is important that all individuals belonging to the scrutinized network are included in the empirical representation. Because all units are included, it is not a random sample. I therefore use the term set rather than sample. My data set comprises all 1572 academic economists beyond the PhD level who worked at the end of 2010 in the German speaking countries. I compiled the social networks of these economists using 13242 journal articles which they published in the 2000-2010 period. In what follows, I will use the adjective ‘relevant’ to indicate these individuals or their publications.

### 2.1 Collaboration Network

For the collaboration network, I gathered all collaborations among the relevant economists as documented in their publications. The average number of publications per individual is about 10.3 with standard deviation 12.9. About 35% of these publications are single-authored, 45% are coauthored with someone outside the set of relevant economists, and 20% of the articles contain relevant coauthorship links. Only these relevant coauthorships are used to construct the collaboration network. Even if the share of 20% is small, it still connects 1036 (66%) individuals into a network via almost 3500 collaboration links.

---

2 In other social networks, the links might not be observed with 100% accuracy, because people might not recall all their acquaintances. Furthermore, acquaintances are of different types: family, friends, colleagues, etc.
3 [www.forschungsmonitoring.org](http://www.forschungsmonitoring.org).
5 e.g. Rauber and Ursprung (2008); Fabel, Hein, and Hofmeister (2008).
linkages. Around 2000 of the linkages connect the same two researchers; altogether the collaboration network then consists of 1544 unique undirected links. The network consists of several components, the largest one, also called the giant component, comprises 54% of the relevant economists, and 75 smaller components. Additionally, there are 536 isolated vertices. These represent researchers who have only coauthors from outside the set of relevant economists (54%), researchers who only published single authored studies (18%), and researchers who did not publish any journal article at all (11%), or did not publish in the considered period (17%).

The nodes of a graph could be distinguished by the number of distinct collaborative links of the researcher, i.e. number of edges of a node, also called degree in network analysis. The top 6% of individuals with high degree (eight and more), are more centrally located. This role could be demonstrated by excluding these highly connected individuals from the graph. In this case the giant component breaks down to over 90 smaller components.

2.2 Citation Network

For the citation network, I collected the citations that link the publications of the relevant economists for the 2000-2010 period, using the citation data from the RePec citation database. Approximately 4500 of the relevant publications cite another paper by a network member. Around 1200 citations connect the same two authors; altogether the citation network hence consists of about 3300 unique directed links. As compared to the collaboration network, the citation network is thus much denser. This is not surprising, as the incidence of citation is much higher than the incidence of collaboration. As a result, there are only 9 smaller components in addition to the giant component. As compared to the collaboration network, there are more isolated vertices (668) which contain economists who did not publish (22%), economists whose publications are not indexed in the Citec database (30%), economists who do not cite or are not cited by any network member (38%), and economists who have only self-citations (10%). In the citation network, the links are directed due to the nature of citations. Each directed link (also called arc in graph theory) from vertex $i$ to vertex $j$ arises if economist $i$ cites one of economist $j$’s publications in one of its publications. To reveal some citation clustering, it is interesting to look at the citation network from the perspective of research fields. The clustering coefficient may grasps this. But it turns out that only the clustering coefficient of the microeconomic field is significantly higher than the overall clustering in the citation network. Excluding the top 6% most prestigious economists, i.e the most cited net of citing, does not cause a breakdown of the giant component as in the case of collaboration. Moreover, keeping only the high prestigious individuals, one would still arrive at a well interconnected network.

2.3 Small-world Properties of the Networks

In network analysis and in network effects models, it is important to cover the whole universe of the investigated topic. In my study, this universe, i.e all academic economics in Austria, Germany and German-

---

6 Economists with ‘only outside collaborators’ (289) could be part of the giant component, if we included outside collaborators. It turns however out that considering outside collaborators integrates only less then half of them (125) into the giant component. The reason for being an isolated economist is a poor publication record: median journal publications of the isolated economists is 4, and the mode is 1, compared to 11 and 4 for the non-isolated economists.

7 Citec = Citation in Economics: http://citec.repec.org/. The Citec database contains citations to about one third of RePec articles (July 2013). The consequence for the relevant economists is that only around 7.000 of the 13.242 journal articles are part of the Citec database.

8 Self-citations would make these economists connected to some member of the citation network, but only via a publication which they self-coauthored. Therefore these links are excluded, not only for them, but for all researchers.
speaking Switzerland, share a homogeneous academic culture. Even though the academic profession in
the German speaking countries is part of the world economic network, these economists are still ‘closer’
to each other than to the rest because they share similar academic institutions. This regional coherence
is reflected by the fact that in the period 2000-2010 almost every third joint publication is coauthored
with another member of the network, and every second one is coauthored with someone with a strong
link to the German-speaking region. Moreover, both networks constitute so-called ‘small-worlds’ (Watts
and Strogatz 1998). I rely here on the definition of the ‘small world’ properties by Goyal, van der Leij, and
Moraga-Gonzáles (2006) who studied the evolution of collaboration networks among economists:

i. $N$ is very large as compared to the average degree
ii. The giant component exists and covers a large share of $N$
iii. The average distance in the giant component is small, i.e. of order $\ln(N)$
iv. The clustering is high as compared to random graphs with the same $N$ and the same average
degree

Table 1 reports the descriptive statistics that confirm the ‘small-world’ properties of my collaboration and
citation networks: $N = 1572$ is very large as compared to the average degree of 2, the giant component
exists and covers in both networks more than half of the nodes.

<table>
<thead>
<tr>
<th>whole network:</th>
<th>collaboration</th>
<th>citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of vertices ($N$)</td>
<td>1572</td>
<td>1572</td>
</tr>
<tr>
<td>average degree</td>
<td>1.96</td>
<td>2 (in/out)</td>
</tr>
<tr>
<td>isolated vertices</td>
<td>536 (34%)</td>
<td>668 (42%)</td>
</tr>
<tr>
<td>nr. of components (w/o) isolated</td>
<td>76</td>
<td>10</td>
</tr>
<tr>
<td>giant component</td>
<td>845 (54%)</td>
<td>885 (56%)</td>
</tr>
<tr>
<td>overlap:</td>
<td></td>
<td>overlap: 649</td>
</tr>
<tr>
<td>jointly:</td>
<td></td>
<td>jointly: 1081</td>
</tr>
<tr>
<td>second largest component</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>giant component only:</td>
<td>collaboration</td>
<td>citation</td>
</tr>
<tr>
<td>average degree</td>
<td>3.36</td>
<td>3.54 (in/out)</td>
</tr>
<tr>
<td>diameter</td>
<td>24</td>
<td>13*</td>
</tr>
<tr>
<td>average distance</td>
<td>6.9</td>
<td>4.92*</td>
</tr>
<tr>
<td>clustering (1-neighbor)</td>
<td>0.32</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Notes: I used the Pajek software (de Nooy, Mrvar, and Batagelj 2005) to calculate the above measures of small-world properties
* measures are calculated only for reachable vertices in the directed citation network.

Moreover, the second largest component, as compared to the giant component, is in both cases a ‘dwarf’,
of size 7 and 3, respectively. The average distance of 6.9 is close to the ‘six degrees of separation’ (Milgram

---

9 These are local collaborators outside the field of economics or academic economists who worked in some
German-speaking country before or after the cutoff year 2010. Few are non-academic economists, and some of
them are PhD. students, who are not among the relevant economists by definition of the set.
and also very close to $\ln(1572) = 7.4$. The diameter, i.e. the longest distance between two individuals, is also small for both networks. Finally, the clustering in the collaboration network is about 80 times higher than predicted by random graphs with the same parameters, and more than 40 times higher in the citation network. My collaboration and citation networks thus fulfill the properties expected from large social networks. This feature has an important implication: even though my data cover only a regional segment of world economic research, this segment still reflects the properties of the whole and is, therefore, suitable for analyzing network effects. Newman (2001a) conjectures that the ‘small world’ properties are a crucial feature of a functional scientific community.

### 3 Specifying Social Weight Matrices of Research Networks

Analyzing statistical models that explain the effect of social networks on individual behavior has by now become a standard approach in the social sciences (e.g. Friedkin 1990, Leenders 2002, Burt 2010). This development has overlapped with the rise in popularity of spatial econometric models starting by the seminal handbook on spatial econometrics by Anselin in 1988 and by the retrospection of this development by the same author thirty years later (Anselin 2010). Network models and spatial econometric models have the same structure and use the same estimation techniques even though they have rather different motivations. I therefore present here briefly the general structure of the cross-sectional spatial econometric model and then explain the analogies to structural models of network effects.

Adopting the notation of LeSage and Pace (2009), the general spatial econometric model takes the following form:

$$
\begin{align*}
\mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \alpha + \mathbf{X} \mathbf{\beta} + \mathbf{W} \mathbf{X} \mathbf{\theta} + \mathbf{u} \\
\mathbf{u} &= \lambda \mathbf{W} \mathbf{u} + \mathbf{\varepsilon}, \mathbf{\varepsilon} \sim \text{iid}(0, \sigma^2 \mathbf{I}),
\end{align*}
$$

where $\mathbf{y}$ is an $N \times 1$ vector of endogenous variable, the $N \times k$ matrix $\mathbf{X}$ is the matrix of exogenous variables, while $\mathbf{\beta}$ is the underlying parameter vector and $\alpha$ is an optional constant parameter. The terms multiplied with the square matrix $\mathbf{W}$, are the spatial interaction components. The $N \times N$ matrix $\mathbf{W}$ is the spatial weight matrix and comprises the analyst’s explicit hypothesis about the strength and interlocation connections between the units of the analysis, which are usually countries, regions or towns (Corrado and Fingleton 2012). The scalar parameters $\rho, \lambda$ and the $k \times 1$ vector $\mathbf{\theta}$ measure the average effects of the connected units on the variable of interest $\mathbf{y}$. If $\rho \neq 0$, the model contains endogenous interaction effects, if $\mathbf{\theta} \neq \mathbf{0}$, the model contains exogenous effects. A nonzero $\lambda$ implies that the model comprises correlated effects. Model (1) might include any one, any two or all three of the interaction parameters. For the full taxonomy of the variants of model (1) see e.g. Elhorst (2010).

Models of network influence (e.g. Friedkin 1990; Marsden and Friedkin 1993) differ in two substantial ways from spatial models. First, the spatial weight matrix is replaced by the so called social weight matrix and, second, the units of analysis are individuals rather than geographic entities. The portrayed individuals are embedded in some explicitly known social network and the analyst is interested in the effect of this network on a behavioral variable $\mathbf{y}$. The social weight matrix quantifies the structure of the known social

---

10 The clustering coefficient for random graphs of size $N = 845 (885)$, with average degree 3.36(3.54) is predicted to be $\sim 0.004$. 
network. Clearly, the spatial or, as the case may be, the social weight matrix play a crucial role in the estimation of model (1) because the size of the interaction effects $\rho, \lambda$ and $\theta$ depends on the choice of $W$ (Leenders 2002). Because of that, and also as a response to some recent criticism, spatial econometricians have been urged to motivate the choice of matrix $W$ on substantive grounds (Corrado and Fingleton 2012).

In this paper, I take up this challenge by specifying the social weight matrix $W$ by relying on theories of social influence (Burt 1987, Leenders 2002), by justifying the influential sociometric distance (Burt 2010), and by accounting for the strength of the connections between researchers along the lines suggested by Newman (2001b). These theories help me to decide on who is linked with whom and, secondly, how strong the link is. I elaborate my approach in detail in the following subsections.

3.1 Theories of Social Influence: Communication and Comparison

Theories of social influence explore social processes that are at work when an individual forms his attitudes or adapts his behavior. These theories originate in the work of Homans (e.g. 1961) on elementary forms of social behavior in groups. Based on Homans’s analysis, empirical sociologists studied, for example, the role of social influence in decisions on adoption of medical innovation (Burt 1987) and in decisions on non-profit donations (Galaskiewicz and Burt 1991). To assess the extent of social influence in adoption mechanisms, one needs to know the individual’s social environment. All social contacts of all individuals in a studied community constitute the social network of these individuals. The pertinent question for the representative network individual ‘ego’ is which individuals of this social network, the so-called ‘alters’, influence his behavior. Social network theory distinguishes two groups of social processes: between ‘ego’ and ‘alter’: communication and comparison (e.g. Leenders 2002). I study communication and comparison processes in a research community by way of investigation their collaboration and citation networks. The objective is to correctly specify a weight matrix which measures the influence of the research community, which is interlinked via collaboration and/or citation ties, on individual research productivity.

Influence by way of communication considers direct and indirect social contacts between ‘ego’ and his ‘alters’. This is a natural way of portraying social influence since it assumes that social contacts between ego and alter are likely to change ego’s behavior. Most studies of social influence assume communication to be the dominant process (Leenders 2002). The process of comparison assumes that ego adapts his behavior by comparing himself to individuals who occupy a similar position in the social structure (Burt 1987). Comparison highlights the existence of competition between ego and alter because ego adapts his behavior to maintain or improve his status (Burt 2010). The individuals whose ego compares with are not necessarily those whose he communicates with.

Generally speaking, social influence is operationalized with the $N \times N$ matrix $W$, whose elements $w_{ij}$ equal normalized proximity measures (Burt 1987)

$$w_{ij} = \frac{\text{proximity } i \text{ to } j}{\sum_k (\text{proximity } i \text{ to } k)}, \forall k \neq i. \quad (2)$$

The used proximity measure, usually indicated by $d_{ij}$ as a distance between individuals $i$ and $j$, depends on the kind of social influence that is to be portrayed.

To operationalize social influence via communication, researchers applied a proximity measure that is based on a measure of closeness (cohesion) which includes information on the number, length, and strength of the links between network members (Leenders 2002). As cohesive proximity measure
between \(i\) and \(j\), Burt (1987) uses the inverse path distance required to reach individual \(j\) from individual \(i\).\(^{11}\) The proximity is one if \(i\) and \(j\) are direct neighbors, it is less than one if there is larger distance between them, and it is zero if there is no direct path between \(i\) and \(j\). This measure of proximity is nonzero for each network pair regardless of how many intermediaries lies in between them, and does not depend on the strength of the links between the individuals. The cohesion measure advocated by Leenders (2002) restricts the cohesion proximity to be non-zero only for neighboring pairs of individuals, and zero for all other pairs, but his measure does allow for the strength of the connections. In this study, I use a proximity measure that lies somewhere between those two extremes. I follow Leenders (2002) and use the strength of the links, but I include not only the first but also the second-order neighborhood of ego and normalize the measure as suggested by Burt (1987).

The original idea of social contagion via comparison presumes that an individual \(i\) behaves like other individuals who are similarly embedded in the network. Using the same network that was also used for operationalizing communication, comparison is operationalized by defining proximity from (2) as the degree to which \(i\) and \(j\) structurally equivalent. For instance Burt (2010) uses the Euclidean distance measure

\[
\begin{align*}
  d_{ij} &= \left[ \sum_k (z_{ik} - z_{jk})^2 + \sum_k (z_{ki} - z_{kj})^2 \right]^{1/2},
\end{align*}
\]

where \(z_{ij}\) is the normalized path distance between individuals \(i\) and \(j\), and the sums are across all \(k\) other than \(i\) and \(j\). Note that \(d_{ij}\) is zero for perfectly equivalent individuals, i.e. for individuals whose distances to and from all other members of the network are equal. The larger the proximity \(d_{ij}\), the less equivalent is the position of the \(i\) and \(j\) in the network. To arrive at a measure that increases with increasing proximity, Burt (2010) represents the \(i,j\)-th cell of the social weight matrix by the term

\[
\begin{align*}
  w_{ij} &= \frac{d_{ij} - d_{ji}}{\sum_{k \neq i} (d_{max} - d_{ki})},
\end{align*}
\]

where \(d_{max}\) is the maximum path distance in the network.

Leenders (2002) suggested an alternative measure of structural equivalence which is also based on the Euclidean distance:

\[
\begin{align*}
  d_{ij} &= \left( \frac{(\hat{i} - \hat{j})(\hat{i} - \hat{j})}{\sqrt{2n}} \right)^{1/2},
\end{align*}
\]

where \(\hat{i}\) is the stacked vector of the \(i\)-th row and the \(i\)-th column of the adjacency matrix \(A\), \(\hat{j} = 1 - \hat{j}\), and \(n\) is the number of network members. \(d_{ij}\) equals 0 for completely nonequivalent individuals, while \(d_{ij}\) equals 1 in the case of perfect structural equivalence.

The comparison concept overlaps to some extent with the concept of communication. This lack of discrimination has led several researchers to arbitrarily discard either communication or comparison as an influencing mechanism. Leenders (2002) illustrates this dilemma with four consecutive studies (Coleman, Katz, and Menzel 1966; Burt 1987; Marsden and Podolny 1990; and Strang and Tuma 1993) that analyze the same network of physicians and its influence on the adoption of a new drug. These papers

---

\(^{11}\) Path distance of two network members is given by the minimal number of links between them. Normalization of the path distance means discounting the path distance by some inverse distance measure. This is an equivalent approach to the one used in spatial econometrics.
arrive at four different answers, namely that the adoption was driven by communication, by comparison, by none of the two, and by both of them.

The blurring of the concepts of comparison and communication arose from the fact that both concepts have usually been operationalized with the help of one and the same social network but using different measures of distance between two individuals. In my study, I take a different approach and utilize the availability of two distinct networks of the same community of researchers. I argue that one of them already captures the communication process, the other one the comparison process. Thus unlike most previous studies on social weight matrices, I apply a cohesion and a structural equivalence measures to both, citation and collaboration networks. I show that rather than the choice of the measure, the particular network itself embodies the comparison, or communication principle.

Moreover, I justify here the ex-ante preference for social influence via comparison as the driving force of research productivity. Communication assumes that two individuals talk to each other directly. Scientists need to communicate directly with their collaborators. Hence, it is natural to assume that the collaboration network reflects communication patterns. Moreover, research collaboration takes time. Communication ties therefore often turn into long term acquaintance and sometimes even friendship. Independent of the depth of the relationship, communication implies reciprocity.

Comparison, on the other hand, simply assumes that individuals observe each other, without the necessity of conversation, acquaintance, and, most importantly, reciprocity. Comparison network can therefore be highly impersonal. In the medical innovation studies described by Burt (1987, 2010) the network of physicians was based on professional ties. Some of the ties were formal, some strategic, and some were perhaps also based on friendship. This network was therefore suited to analyze the effects of both communication and comparison. A pure collaboration network does, however not really fit the concept of comparison, because it captures many ties that clearly transcend simple comparison. The citation network, on the other hand, captures non-personal information on the literature which serves as an input for the author’s own work. Citations are explicit linkages between studies that have some important content in common (Hummon and Doreian 1989). While it is not unheard of that researchers communicate with authors whom they cite, this is rather an exception. Exactly because citation networks are largely anonymous, they are well suited for modeling comparison effects on research productivity.

<table>
<thead>
<tr>
<th>frequencies of pairs</th>
<th>citation</th>
<th>no citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>collaboration</td>
<td>2%</td>
<td>8%</td>
</tr>
<tr>
<td>no collaboration</td>
<td>3%</td>
<td>59%</td>
</tr>
</tbody>
</table>

An important evidence to my argument is the practical non-existence of overlapping in my collaboration and citation networks. Table 2 lists possible types of collaboration and citation patterns. Most of the economists either collaborate but do not cite each other, or they do not collaborate, but cite and/or are
cited by another relevant economist. These ‘clean’ combinations make up more than 90% of the all existing links in the two networks (cf. Table 2). Hence these two networks represent largely distinct ties.

The research question of my analysis is whether the researcher’s productivity is driven by the network of collaborators or by the network of authors they cite. My hypothesis is that the comparison effect is more important than the communication effect. The reason lies in the nature of science. The science system is competitive (Carayol 2008; Stephan 2010). Competition among scientists has monetary aspects, but more important are prestige, esteem, and acknowledgment of one’s work in the academic community (e.g. Merton 1973). Burt (1987) also strongly argues in favor of the comparison effect in the presence of competition.

Homans (1961) surveyed experiments showing that individuals who solve a puzzle in a cooperative setting specialize in particular tasks rather than try to solve the puzzle on their own. The increasing specialization and interdisciplinary nature of research are the main reasons for scientific collaboration and its steady increase over time (Laband and Tollison 2000; Goyal, van der Leij, and Moraga-González 2006).

On the other hand, Homans (1961) also argues that people who compete are more likely to become more similar in their expertise than people who cooperate. In economic theory there is a similar concept called ‘minimum differentiation’ (Hotelling 1929), showing that competition between two sellers makes them similar in what they offer to their customers. The ‘minimum differentiation’ principle holds also for \( n \) sellers, if the consumers’ tastes are sufficiently heterogeneous (Palma de et al. 1985). For competitive scientists this implies that their research productivity becomes similar over time. This lends additional support to my hypothesis that comparison plays a major role in science.

The comparison effect is likely to be stronger for more productive individuals because their reward is higher. Elhorst and Zigova (2014) show that competition among top economic institutes is twice as high as competition among institutes in general. Figure 1 shows the academic networks of the 50 top-ranked economists from the German-speaking countries according to the Handelsblatt ranking in 2011. The two networks are completely different. Collaboration is extremely sparse among these individuals, while the citation network shows a dense net of interconnections. This evidence again suggests that comparison is the primary social driver of scientific productivity. I therefore contend that the benchmark social weight matrix should be based on the citation network.

3.2 Network Horizon

In the previous section, I have argued that influence by comparison, represented by the citation network, is the main social process governing research behavior. Now, two additional issues in specifying the best social weight matrix need to be addressed: how far reaches the horizon of the relevant network and what is the magnitude of influence of the peers who are within this horizon? The network horizon is the maximal sociometric distance between two individuals who still exert influence on each other. The magnitude of the influence comprises network factors that strengthen or weaken the link between two individuals. For the \( W \) matrix, the horizon of influence determines which cells \( w_{ij} \) are zero and which are not, while the magnitude of the influence determines the value of the nonzero cells \( w_{ij} \).

---

12 Recall that the self-citations are excluded from the citation network.
14 The citation network is approximately 3 times denser than the collaboration network of the top economists, even if taking into account the higher incidence of citation links in general.
Figure 1 Collaboration (a) and citation (b) networks of the top-50 ranked economists. The filled circles represent economists who are connected to at least one other top economist via a common publication (a) or an in- or out-citation (b). The selection of the top economists is based on the Handelsblatt 2011 ranking, based on the lifetime research productivity. See http://tool.handelsblatt.com/tabelle/index.php?id=79&pc=250
The sociometric distance between two individuals $i$ and $j$ in a network is defined as minimum number of links that are necessary to reach $j$ from $i$. Direct neighbors have a sociometric distance of one. Those pairs that are not direct neighbors and have at least one common neighbor have a sociometric distance of two. The average sociometric distance in social networks is surprisingly small given the size of the whole network. Milgram (1967) has called this the ‘small world’ feature. In my study the average distance is 7 in the collaboration network, and 5 in the citation network (see Table 1).

The neighborhood circle of an individual is clearly smaller than the average distance of a network, because otherwise nearly everybody would matter for everybody else. On the other hand, as observed by Milgram (1967), ‘five intermediaries means an enormous psychological distance’ between two individuals. If the network matters at all for social relationships, the influential circle must include at least the individual’s immediate neighbors and be clearly smaller than the circle defined by the average sociometric distance; the question is how much smaller. Leenders (2002) advocates to limit the influential distance to two or three. Burt (2010) devotes an entire book to the analysis of so-called neighbors’ networks and discusses numerous examples of professional networks in business, politics, and medicine, focusing on the importance of neighborhood ties for the individual’s decisions in his profession, and so advocates the relevant sociometric distance setting to two.

Crane (1972, Ch.3), reports evidence from mathematics and sociology. In these disciplines over two thirds of the researchers are linked with a prolific scholar by not more than one intermediary. In the citation network analyzed in my study, the average number of researchers reachable via two links is 25, via three links it is already 92.\(^{15}\) It seems reasonable to limit in this network the influential circle by the distance two because, given that the social driver of research productivity is comparison, it is plausible to assume that there are about 20 peers a researcher wants to compare with. In the $W$ matrix, I therefore set the sociometric distance which matters for a researcher to two.

### 3.3 Benchmark Social Weight Matrix

In the previous sections I established the choice of the citation network and sociometric distance two as the crucial building stones for the benchmark weight matrix $W$. The remaining issue is the choice of the proximity measure, i.e. cohesion or structural equivalence. Based on previous literature I suggest here one cohesion measure and one structural equivalence measure for the citation network. I show later that the citation network is statistically always preferred to the collaboration network, irrespectively which measure is applied.

Any proximity measure is quantified for all network pairs with help of the adjacency matrix $A$. In the case of the citation network, the $i,j$-th element of matrix $A$ is 1 if $i$ cited $j$ at least once, and zero otherwise. The $A$ matrix is asymmetric, because the $i$-th row of $A$ indicates which researchers $i$ cites and the $i$-th column indicates which researchers are citing $i$.

So far, I have assumed that all citation links are equally strong. An author can, however, cite another author more than once, and the citing paper can be coauthored by several authors. Newman (2001b), dealing with this issue in collaboration networks, defined the mutual link between two collaborators $i$ and $j$ by the following term

\(^{15}\) These calculations are done for scientists belonging to the giant component (cf. Table 1).
\[ c_{ij} = \sum_{k} \frac{\delta_{ik}^k \delta_{jk}^k}{n_k} \]

(6)

where the sum runs over all publications coauthored by \( i \) or \( j \), and the indicator \( \delta_{ik}^k \) is one if researcher \( i \) has collaborated on the \( k \)th publication written by \( n_k \) researchers. The matrix \( \{ c_{ij}, \forall i, j \} \) is called the weighted adjacency matrix. The rationale for specification (6) is to increase the strength of the relationship by the multiplicity of collaboration and decrease it by the term \( 1/n_k - 1 \), the latter implying that \( i \) divides the time spent on study \( k \) interacting with \( n_k - 1 \) other researchers. Newman’s (2001b) specification (6) can be adapted to citation networks as follows:

\[ c_{ij} = \sum_{k} \frac{\delta_{ij}^k}{n_k} \]

(7)

where the indicator \( \delta_{ij}^k \) is one if \( i \) cited \( j \) in the \( k \)th publication written by \( n_k \) authors. As in (6), the directed link between \( i \) and \( j \) is again increased by the multiplicity of citations and decreased by the term \( 1/n_k \) which measures the probability that \( i \) was responsible for citing \( j \) in publication \( k \).

To make multiplicity effective in the benchmark matrix, I multiplied the respective adjacency matrix \( A \) with the term \( c_{ij} \) in (7). Using the weighted adjacency matrix of citation network, the cells of the \( W \) are defined as follows:

\[ w_{ij} = \begin{cases} c_{ij} & \text{if } i \text{ cites } j \\ \frac{1}{2} \max_{\forall k} \{ c_{ik} + \frac{c_{kj}}{2} \} & \text{if } i \text{ cites } k \text{ and } k \text{ cites } j, \end{cases} \]

(8)

where the \( k = (1, 2, \ldots, K) \) is the set of researchers who \( i \) cite. The measure in (8) discounts indirect citation by one half. This is in line with the normalized path distances used by Burt (2010). The idea behind the maximum sum of weighted paths between \( i \) and \( j \) is that \( i \) is familiar with the publications of \( j \) from \( K \) different ‘sources’. Since the information on \( j \) is not cumulated but rather overlapping, \( i \) cannot know more about \( j \) than what he learns about him via the maximal sum of his connections with \( k \) and \( k \)'s connection with \( j \). The matrix \( W \) defined in (8) is the benchmark weight matrix \( W \) prior to normalization.

The final step in the operationalization of the social weight matrix is a normalization. I apply row-normalization \( w_{ij} = w_{ij}/w_{i} \), where \( w_{i} \) equals the sum of all elements of the \( i \)-th row of \( W \). The row-normalization rescales the received influence, while the column normalization rescales the exerted influence (Leenders 2002). In our context, the row normalization means that the sum of the influences received by all researchers who exert influence to \( i \) sums to one, but the relative strength of the individual influence remains unchanged.

### 3.4 Alternative Social Weight Matrices

In the previous sections I presented a sequence of arguments that led me to the preferred operationalization of the social weight matrix (8). The standard spatial econometric approach to selecting the weight matrix is to test the performance of several matrices with the help of statistical tests of model comparison (e.g. LeSage and Pace 2009, ch. 6).

To address the recent critique calling for more substantive approaches (Corrado and Fingleton 2012), I first specified the benchmark social weight matrix \( W \) based on socio-economic concepts and now test its performance by comparing it with several alternative matrices. Each of these alternative matrices relaxes one criteria imposed on the benchmark matrix \( W \).
The first alternative matrix, $W_1$, operationalizes the weight matrix via a cohesion measure in the collaboration network. Using the weighted adjacency matrix (6) of collaboration network, the cells of the $W_1$ are defined as follows

$$w_{ij} = \begin{cases} c_{ij} & \text{if } i, j \text{ direct neighbors} \\ \frac{1}{2} \max_k \left\{ c_{ik} + \frac{c_{jl}}{2} \right\} & \text{if } i, j \text{ soc. distance 2 via } k \\ \frac{1}{3} \max_{k,l} \left\{ c_{ik} + c_{jl} + \frac{c_{kl}}{3} \right\} & \text{if } i, j \text{ soc. distance 3 via } k \text{ and } l \end{cases}$$

(9)

similarly to (8) the $k = (1, 2, \ldots, K)$ is the set of researchers with whom both $i$ and $j$ collaborated. The measure (9) discounts indirect collaborations by one half. Testing $W$ versus $W_1$ gives statistical preference for the social process, i.e. comparison or communication, at work in the community of researchers.

The second alternative, $W_2$, relates to the citation network but ignores in the cohesion measure (8) the multiplicity of citations and the number of coauthors (7). Hence, the $W_2$ matrix is created applying (8) to plain adjacency matrix $A$ based on citation network. Rejecting $W_2$ against $W$ gives statistical preference for accounting of the ‘strength’ of the connection from $i$ to $j$.

The next two alternatives, $W_3$ and $W_4$, modify the horizon of influence. $W_3$ expands the maximum sociometric distance to three. I modify the cohesion measure (8) to account for distances up to three as in (10) and apply it to the citation network.

$$w_{ij} = \begin{cases} c_{ij} & \text{if } i, j \text{ direct neighbors} \\ \frac{1}{2} \max_k \left\{ c_{ik} + \frac{c_{jl}}{2} \right\} & \text{if } i, j \text{ soc. distance 2 via } k \\ \frac{1}{3} \max_{k,l} \left\{ c_{ik} + c_{jl} + \frac{c_{kl}}{3} \right\} & \text{if } i, j \text{ soc. distance 3 via } k \text{ and } l \end{cases}$$

(10)

On the contrary, $W_4$ limit the horizon of influence only to direct cites. Rejecting $W_3$ or $W_4$ against $W$ gives statistical preference for the socioeconomic distance 2 among researchers.

Table 3 Descriptive statistics of the benchmark social weight matrix $W$ and the alternative matrices $W_1$ to $W_6$ prior to row-normalization. Means and ranges are based on non-zero values only.

<table>
<thead>
<tr>
<th>matrix</th>
<th>deviation from $W$</th>
<th>non-zero cells</th>
<th>range (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>cohesion in citation network</td>
<td>19'086</td>
<td>0.10 - 15.79 (0.67)</td>
</tr>
<tr>
<td>$W_1$</td>
<td>cohesion in collaboration network</td>
<td>12'670</td>
<td>0.03 - 26.50 (1.33)</td>
</tr>
<tr>
<td>$W_2$</td>
<td>as $W$ but ignore multiplicity and coauthors</td>
<td>19'086</td>
<td>0.75 - 1.75 (0.83)</td>
</tr>
<tr>
<td>$W_3$</td>
<td>as $W$ but sociometric distance up to 3</td>
<td>69'561</td>
<td>0.10 - 15.79 (0.58)</td>
</tr>
<tr>
<td>$W_4$</td>
<td>as $W$ but sociometric distance restricted to 1</td>
<td>3'297</td>
<td>0.10 - 11.00 (0.78)</td>
</tr>
<tr>
<td>$W_5$</td>
<td>comparison in citation network</td>
<td>17'517</td>
<td>0.07 - 16.58 (0.83)</td>
</tr>
<tr>
<td>$W_6$</td>
<td>comparison in collaboration network</td>
<td>11'254</td>
<td>0.01 - 22.05 (0.74)</td>
</tr>
</tbody>
</table>

For matrices $W_5$ and $W_6$, I adapt the structural equivalence measure (5) by Leenders (2002) to citation and collaboration networks. Having citation network in mind, the value $d_{ij}$ in (5) measures the fraction of researchers who $i$ and $j$ both cite (or both not cite), and of researchers who cite both, $i$ and $j$ (or both not cite). Hence, $d_{ij}$ is a measure of similarity in citation, or for that matter of non-citation, patterns. $d_{ij}$ is exactly one, if $i$ and $j$ cite exactly the same set of researchers and are cited by the same set of researchers (possibly distinct from the citing set). $d_{ij}$ is less than one if $i$ and $j$ cite and are cited by sets of authors that only partially overlaps. $d_{ij}$ is zero if $i$ cites exactly those researchers who $j$ does not cite and is cited
by exactly those researchers who do not cite \( j \). The drawback of \( d_{ij} \) in (5) is that it can be non-zero also for quite distant individuals. Their structural equivalence can be large just because they both do not cite many other researchers. Therefore, Leenders (2002) advocates to calculate the measure of structural equivalences (5) only for pairs of individuals who are able to observe each other. Because of that, I restrict the value of structural equivalence to be non-zero only for pairs separated by a sociometric distance of one or two and, moreover, base the structural equivalence measure of such pairs only on a subnetwork consisting of all their direct citers and citees. This strategy is in line with the above arguments in favor of the sociometric distance two. In short, I modify the structural equivalence measure of Leenders (2002), defined in (5), as follows

\[
\tilde{d}_{ij} = \begin{cases} \left( \frac{(\hat{s}_i - \hat{s}_j)^2}{2n_{ij}} \right)^{1/2} & \text{for } i \neq j \text{ and } A_{ij}^2 > 0 \\ 0 & \text{otherwise} \end{cases}
\]

(11)

where the matrix \( A^2 = A \cdot A \), consists of non-zero values for pairs \( ij \) of individuals, separated by a path of length one or two. \( n_{ij} \) is the size of the subnetwork around \( i \) and \( j \) containing citees an citers of \( i \) and \( j \). The vector \( \hat{s}_i \) consists of the selected stacked elements of the \( i \)-th row and \( i \)-th column of the adjacency matrix \( A \) underlying to the subnetwork around \( i \) and \( j \). As in (5), \( \hat{s}_j \) is symmetric and measures, via the Euclidean distance, the ratio of the number of common citees and citers and the number of all citees and citers of any pair. Figure 2 illustrates and discusses the measure \( \tilde{d}_{ij} \) for a hypothetic pair \( ij \).

To adjust the structural equivalence matrix \( \tilde{D} \) accordingly, I multiply each element of \( \tilde{D} \) by

\[
C_{ij} = \sum_l c_{il}, \forall l: A_{ii} = 1 \text{ and } A_{ji} = 1,
\]

(12)

where the sum runs over all citations \( l \) which \( i \) and \( j \) have in common. The strength of the structural equivalence defined in this way can be interpreted as how much researcher \( i \) desires to align with the research of \( j \). The social weight matrix \( W_5 \) then consists of non-zero elements \( w_{ij} \) for all pairs of researchers who are not separated by more than two links, and \( w_{ij} \) is the product of structural equivalence (11) and the strength of this equivalence (12), i.e.

\[
w_{ij} = C_{ij} \tilde{d}_{ij}.
\]

(13)

The matrix \( W_6 \) applies the structural equivalence measure (13) in the collaboration network, taking into account the strength of collaboration ties \( c_{ij} \) as defined in (6). Testing \( W \) against \( W_5 \) can statistically resolve the question, whether comparison, via the use of simple cohesion measure in citation network, is sufficiently represented or not. Moreover rejecting \( W_6 \) against \( W \) and \( W_5 \) gives clear statistical preference for citation network irrespectively of the measure used.

Table 3 summarizes the notation and presents some descriptive statistics relating to the benchmark matrix \( W \) and the alternative matrices \( W_1 \) to \( W_6 \).

**4 Testing the W in the Network Autocorrelation Model**

In this section I test the alternative matrices \( W_1 \) to \( W_6 \) against the benchmark matrix \( W \) with the help of network effect models. Notice, that rejecting the preferred matrix in these tests does not automatically mean that \( W \) is not properly specified and the underlying theories are wrong. It might also mean that by
Figure 2 Illustration of the structural equivalence measure $\tilde{d}_{ij}$ for a hypothetic pair $ij$ of a network consisting of 9 nodes. The nodes 1 to 5 are citees or citers of $i$ or $j$ and hence belong to the subnetwork of $ij$. The nodes 6 and 7 are not part of the $ij$ relevant subnetwork, even if they are part of the whole network, due some linkages with the nodes 1 to 5 (here it is not relevant what are they like). The nodes 6 and 7 would effect the $d_{ij}$ (5), but not the measure $\tilde{d}_{ij}$ (11).
relying solely on statistical criteria one might fail to select the optimal social weight matrix. Because the tests are based on the size and significance of the spatial effect parameters, the matrix is selected that performs best in purely statistical terms. As pointed by Corrado and Fingleton (2012), the size and significance of the spatial effect parameter can however be misleading, since these values might simply pick up the effect of omitted spatially correlated dependent variables.

The ultimate reason for operationalizing the social weight matrix in line with social influence theories, is to arrive at correct estimates of the effect of professional social network on the researchers’ productivity. In the field of economics of science, the quality and quantity of the publication output is viewed as the most credible measure of a scientist’s productivity. In this section I compare the performance of the benchmark social weight matrix $W$ with its alternatives $W_1$ to $W_6$ with the help of a network autocorrelation model of research productivity and apply a Bayesian testing procedure.

For that purpose, I use versions of the general spatial model (1), where $\rho \neq 0$ or $\lambda \neq 0$. According to the taxonomy of the spatial models (e.g. Elhorst 2010), there are 5 such models:

$$y = \rho Wy + \alpha + X\beta + \varepsilon$$ (SAR)
$$y = \alpha + X\beta + u; u = \lambda Wu + \varepsilon$$ (SEM)
$$y = \rho Wy + \alpha + X\beta + u; u = \lambda Wu + \varepsilon$$ (SAC)
$$y = \rho Wy + \alpha + X\beta + WX\theta + \varepsilon$$ (SDM)
$$y = \alpha + X\beta + WX\theta + u; u = \lambda Wu + \varepsilon$$ (SDEM)

where $\varepsilon \sim iid(0, \sigma^2 I)$. These models assume that the network influences the research productivity directly via the productive efforts of other researchers in the analyzed community (if $\rho \neq 0$) or via unobserved factors (if $\lambda \neq 0$). In the models above I impose an $N \times N$ row-normalized social weight matrix $W$. This can be either the benchmark weight matrix $W$ or one of the six alternative alternatives. $t$ is an $N \times 1$ vector of ones associated with the constant term parameter $\alpha$ and $X$ is an $N \times k$ matrix of research productivity covariates measuring individual and institutional characteristics. Furthermore the $y$ is an $N \times 1$ vector of research productivities. Research productivity is usually measured as the quality adjusted number of journal articles. In this paper I firstly calculate for each individual $i$ his research productivity in each year $t$ of the 2000-2010 period as

$$y_{it} = \sum_{k=1}^{N_{it}} \frac{q_{ik}}{a_k}.$$ (14)

The expression (14) consists of the sum of the ratios of journal quality weights $q_j$ over the number of authors $a_j$ of the $k$th publication of scientist $i$. The sum runs over all publications which economists $i$ has published in the year $t$. For $q \in [0,1]$, I use the journal quality weights proposed by Combes and Linnemer (2010). These authors index about 1200 economics journals. Finally, I calculate for each individual $i$ his average annual research productivity to smooth the volatile annual productivities and to account for

16 The abbreviations used to name these models are standard in the spatial econometric literature. SAR refers to spatial autocorrelation model, SEM means the spatial error model, SDM is the spatial Durbin model, SDEM is the spatial Durbin error model, while SAC has no direct meaning.

17 Referring to my previous research with research productivity data (, ), I use the following set of control variables: academic age, gender, research field, academic position, country, research institute, size of the department, measure of outside collaboration, and share of the non-publishing faculty.
different career age. This average productivity becomes the dependent variable in models (SAR - SDEM) to be explained by the set of control variables focusing on the network effect captured by $\rho$ or by $\lambda$.

Table 4 Estimates of the network effect parameters $\rho$ and $\lambda$ in four types of spatial models for the benchmark matrix $W$ and the six alternative matrices $W_i$.

<table>
<thead>
<tr>
<th></th>
<th>SAR model $\rho$ std. dev.</th>
<th>SEM model $\lambda$ std. dev.</th>
<th>SAC model $\rho$ std. dev.</th>
<th>SDM model $\lambda$ std. dev.</th>
<th>SDEM model $\rho$ std. dev.</th>
<th>SDEM model $\lambda$ std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>0.264</td>
<td>0.027</td>
<td>0.251</td>
<td>0.038</td>
<td>0.238</td>
<td>0.050</td>
</tr>
<tr>
<td>$W_1$</td>
<td>0.226</td>
<td>0.009</td>
<td>0.199</td>
<td>0.036</td>
<td>0.444</td>
<td>-0.343</td>
</tr>
<tr>
<td>$W_2$</td>
<td>0.262</td>
<td>0.028</td>
<td>0.250</td>
<td>0.040</td>
<td>0.248</td>
<td>0.026</td>
</tr>
<tr>
<td>$W_3$</td>
<td>0.290</td>
<td>0.033</td>
<td>0.294</td>
<td>0.051</td>
<td>0.260</td>
<td>0.061</td>
</tr>
<tr>
<td>$W_4$</td>
<td>0.149</td>
<td>0.007</td>
<td>0.127</td>
<td>0.011</td>
<td>0.176</td>
<td>0.034</td>
</tr>
<tr>
<td>$W_5$</td>
<td>0.241</td>
<td>0.017</td>
<td>0.226</td>
<td>0.024</td>
<td>0.245</td>
<td>0.042</td>
</tr>
<tr>
<td>$W_6$</td>
<td>0.116</td>
<td>0.007</td>
<td>0.072</td>
<td>0.011</td>
<td>0.180</td>
<td>0.108</td>
</tr>
</tbody>
</table>

Table 4 reports the estimates of $\rho$s and $\lambda$s for all considered matrices in the five spatial models. The estimated $\rho$s are highly significant for all weight matrices specifications and the network effect they signal is positive ranging from 0.03 to 0.44, but in most cases being above 0.2. The effect of $\rho$ is lower only in the models with network correlated explanatory variables (SDM), because part of the network effect act via covariates. Similar pattern can be observed with estimated $\lambda$s. They are in most cases larger than 0.2 in the ‘clean’ SEM models. In the SDEM models, where the network correlated disturbances are estimated together with network correlated explanatory variables, are the $\lambda$s smaller but in most cases still significant. Interesting view offer the SAC models, where the network autocorrelation is combined with the network correlated disturbances. Here the $\rho$s hardly differ from those estimated in the SAR model, except for the two cases when they compensate the significant $\lambda$s.

However the focus of this study lies in the correct specification of the social weight matrix, there is an interpretation of the estimated network parameter. This can be, in the case of the benchmark matrix $W$ and the SAR model, described as follows. If a researcher from the community publishes an additional journal article of quality $q$, the productivity of researchers who are influenced by this researcher increase on average by $0.16 \times q$ (cf. [res_prod]) journal articles. The coefficient 0.16 is the average row (or column) sum of the non-diagonal elements of the matrix $(I - \rho W)^{-1}$, with $\rho = 0.264$ and $W = W$ (see details in LeSage and Pace 2009).

To test the performance of the benchmark social weight matrix $W$ against the alternatives, I employ Bayesian posterior probabilities. The Bayesian approach does not require a null hypothesis and can test any number of weight matrices simultaneously in order to find the statistically preferred one. This method which builds on Bayesian theory of model comparison is advocated by LeSage and Pace (2009, ch. 6) when there exist uncertainty about the exact specification of the spatial weight matrix.
Consider $M$ alternative weight matrices, $W_i, i = 1, 2, \ldots, M$. Each of them produces in a network autocorrelation model (SAR to SDEM) a value of the likelihood function.\(^{18}\) Assume now a prior distribution of the parameter space $\theta = (\rho, \beta, \lambda, \sigma^2)$. The posterior probabilities for each $W_i$-model is

$$p(W_i|y) = \frac{p(y|W_i)p(W_i)}{p(y)}.$$ \hspace{1cm} (15)

The $p(y|W_i)$ in (15) is the marginal likelihood which is obtained by (16) using the MCMC algorithm.\(^{19}\)

$$p(y|W_i) = \int p(y|\theta^i, W_i)p(\theta^i|W_i)d\theta^i \hspace{1cm} (16)$$

The probabilities $p(y|W_i)$ are used to test whether the benchmark matrix $W$ is statistically superior in explaining the research productivity models. Table 5 reports the Bayesian posterior probabilities of joint tests between the benchmark matrix $W$ and the alternatives. The benchmark matrix statistically outperforms all alternative matrices in each of the four models. Thus, based on these Bayesian tests, I can statistically exclude the alternative $W_1$ and thereby the hypothesis of social influence via communication, i.e. via the network of collaborations. Taking the strength of the links into account when measuring structural equivalence is statistically superior to ignoring strength, since $W$ also outperform $W_2$. Matrices $W_3$ and $W_4$ are rejected, too. Note that this is in spite of the larger network effect of the $W_3$ in all four models. I am thus led to conclude that the sociometric distance two is necessary and sufficient to explain the network effects on research productivity. The benchmark matrix $W$ is statistically preferred to matrix $W_5$ and hence supports my earlier claims, that the comparison effect is sufficiently expressed by the specific matrix of citations. Moreover a pairwise test between $W_5$ and $W_6$ (not shown here) rejects $W_6$, i.e. the citation network is statistically superior to collaboration network also when using a measure of structural equivalence in both networks.

<table>
<thead>
<tr>
<th></th>
<th>SAR model</th>
<th>SEM model</th>
<th>SDM model</th>
<th>SDEM model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>0.966</td>
<td>0.983</td>
<td>0.723</td>
<td>0.793</td>
</tr>
<tr>
<td>$W_1$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$W_2$</td>
<td>0.033</td>
<td>0.016</td>
<td>0.030</td>
<td>0.028</td>
</tr>
<tr>
<td>$W_3$</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$W_4$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.247</td>
<td>0.179</td>
</tr>
<tr>
<td>$W_5$</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$W_6$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: The fifth model here was the SAC model but unfortunately the Bayesian posterior model probability is not yet available.

The statistical argumentation via the Bayesian posterior probabilities helped me to discriminate between the alternative specification. Moreover, the analysis has shown that weight matrices based on the citation

\(^{18}\) As the Bayesian posterior model probability is not available for the SAC model I do not consider it for the Bayesian testing procedure.

\(^{19}\) I use here the Matlab codes sar_g, sdm_g, sem_g, developed by LeSage. It is freely available at http://www.spatial-econometrics.com/
network outperform weight matrices based on the collaboration network irrespectively of the measure used, indicating that the dominant conduit of social network effect on research productivity works through competition incentives.

5 Conclusion

In this paper I propose a structured approach to specifying social weight matrices that reflect the influence of social networks on individual behavior. I illustrate this approach by applying it to investigating the impact of collaboration and citation networks on the research productivity of academic economists.

My approach proceeds in three steps. First I report with descriptive statistics the important properties of the analyzed networks in order to document that they satisfy the ‘small-world’ properties which implies that they are well-behaved social networks. Second, I rely on two socio-economic theories to identify the circle of influential individuals. These two competing theories of social influence on individual behavior describe the role of communication and comparison. In my illustration I then apply these theories on the community of researchers, arguing that mainly due to the competitive forces prevailing in research, the comparison process is the preferred social driver of research productivity. These first two steps lead me to a benchmark specification of the social weight matrix which is based on the citation network. To quantify proximity between pairs of individuals, I use a measure of cohesion and a measure of structural equivalence.

The third step consists of defining six alternative weight matrices which I obtain in each case by relaxing a specific assumption of the benchmark social weight matrix. I then estimate for each alternative matrix five types of network autocorrelation model of research productivity based on the taxonomy of the spatial econometric models. Afterwards, I apply a Bayesian procedure of model selection to measure the performance of the benchmark matrix as compared to the alternative matrices. In my illustrating example, I find that the benchmark social weight matrix outperforms the alternatives in all cases. Moreover, the Bayesian analysis has shown that weight matrices based on the citation network outperform weight matrices based on the collaboration network irrespective of the measure used.

My results suggest that the citation network, which represents the sociological process of comparison, is the preferred portrait of interactions among scientists. This result agrees with Burt’s (1987, 2010) view that comparison is the strongest social process in a group of competing individuals.

Even though I use bibliographic data on German academic economists, the proposed approach can readily be applied to researchers in other disciplines and also to other socio-economic contexts in which network effects on individual behavior are to be investigated.

Acknowledgments

I thank Toke Aidt, Heinrich Ursprung, Eylem Gevrek, Laurent Bach and Paul Elhorst for their comments on the previous drafts of this paper. I am grateful to Simon Heß, Susann Adloff and Sebastian Kopf for research assistance and to José Manuel Barrueco, the maintainer of the citec database, for kindly sharing with me a simplified format of the citec data.
References
DE NOOY, W., A. MRVAR, and V. BATAGELJ (2005), Exploratory network analysis with Pajek, Cambridge University Press.
Administrative Science Quarterly 36, 88–105.

GARFIELD, E. (1963), “Science citation index 1961”, Science Citation Index, v–xvi.


