Collaborative Knowledge Production in China. Regional Evidence from a Gravity Model Approach

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Cris: for this paper the German and Mandarin abstracts are included

by the authors!

Collaborative Knowledge Production in China

Regional Evidence from a Gravity Model Approach

Thomas Scherngell* and Yuanjia Hu**

*Corresponding author,

Foresight & Policy Development Department, AIT Austrian Institute of Technology

Donau-City-Strasse 1, A-1220 Vienna, Austria

Email: thomas.scherngell@ait.ac.at

**Institute of Chinese Medical Sciences, University of Macau

Av. Padre Tomás Pereira Taipa, Macau, PR China

Email: yuanjiahu@umac.mo

Abstract. In this study we investigate collaborative knowledge production in China from a regional perspective. The objective is to illustrate spatial patterns of research collaborations between 31 Chinese regions, and to estimate the impact of geographical, technological and economic factors on the variation of cross-region collaboration activities within a Negative Binomial gravity model framework. We use data on Chinese scientific publications from 2007 with multiple author addresses coming from the China-National-Knowledge-Infrastructure (CNKI) database. The results provide evidence that geographical space impedes cross-region research collaborations in China. Technological proximity matters more than geography, while economic effects only play a minor role.

JEL Classification: O38, L14, R15

Keywords: Co-publications, collaborative knowledge production, Negative Binomial regional gravity model, Chinese regions

(Abstract in German)

Zusammenfassung. Die vorliegende Studie untersucht kollaborative Wissensproduktion in China aus einer regionalen Perspektive. Zielsetzung ist es, räumliche Muster kollaborativer Wissensproduktion zwischen 31 chinesischen Regionen zu beschreiben und den Einfluss von geographischen, technologischen und ökonomischen Determinanten auf die Variation interregionaler Kollaborationsaktiviäten zu messen. Die Studie verwendet neue Daten aus der China-National-Knowledge-Infrastructure (CNKI) Datenbank über chinesische Ko-Publikationen mit mindestens zwei Autoren aus dem Jahr 2007. Die Ergebnisse zeigen, das die Kollaborationswahrscheinlichkeit signifikant mit zunehmender geographischer Distanz abnimmt. Der Einfluss von technologischer Nähe ist jedoch wichtiger als geographische Distanzeffekte, während ökonomische Unterschiede eine geringere Rolle spielen.

(Abstract in Chinese)

摘要：本研究从区域角度探讨了中国合作知识生产。研究目的为描述中国的

31 个区域在空间上的研究合作关系，并进一步运用负二项重力模型分析地理

因素、技术因素和经济因素对跨区域合作行为的影响。我们使用了中国国家

知识基础设施工程 (CNKI) 数据库中 2007 年具有多作者地址信息的学术期刊

论文数据。结果表明地理距离阻碍了中国跨区域研究合作。技术因素比地理

因素更加重要，而经济因素仅仅具有微弱的影响。

1   Introduction

The literature on the geography of innovation argues that knowledge flows between agents are geographically bounded since knowledge is in part tacit. Though the cost of transmitting codified knowledge may be invariant to distance, presumably the cost of transmitting non-codified knowledge across geographic space rises with geographic distance (see AUDRETSCH and FELDMAN, 1996). Also in scientific research collaborations spatial proximity may occur as an important determinant of the intensity of collaboration between two actors (see KATZ and MARTIN, 1997). Time and money required to engage in collaborations is limited which forces collaborating actors to be highly selective in choosing a collaboration partner. Thus, collaboration intensity between any two actors will not only be affected by learning opportunities, but also by the time and the coordination costs incurred by participating in collaborations (see HOEKMAN et al., 2009). CUMMINGS and KIESLER (2007) illustrate the importance of coordination costs in research collaborations. These costs may increase with geographical distance, but also with the existence of institutional barriers or language borders, since long-distance collaborations are more time-consuming and yield additional travel costs (see, for instance, HINDS and BAILEY, 2002).

From its beginning, the empirical literature on the geographic localisation of knowledge diffusion has faced numerous problems, particularly concerning the measurement of knowledge production and diffusion. KRUGMAN (1991, p. 153) states that "knowledge flows are invisible, they leave no paper trail by which they may be measured and tracked", pointing to difficulties in finding data on knowledge flows. However, during the 1990s the empirical research on the spatial diffusion of knowledge has significantly improved by using new indicators and introducing new (spatial) econometric methods. The pioneering study of JAFFE et al. (1993) uses patent citations as a paper trail for knowledge flows in the United

States. They provide evidence that geographical proximity promotes knowledge diffusion processes. BRESCHI and LISSONI (2001) re-examine the geographical localisation hypothesis, and highlight the importance of additional factors, such as labour mobility, or institutional and cultural barriers. In this context, more recent empirical studies try to control for these additional factors in an advanced econometric framework. Notable examples are the studies of MAURSETH and VERSPAGEN (2002), and FISCHER et al. (2006) analysing knowledge flows across European regions by using patent citation data of the European patent office. SCHERNGELL and BARBER (2009 and 2010) make use of joint research projects funded by the European Framework Programmes as a proxy for knowledge interactions in Europe at a regional level to identify various determinants of knowledge diffusion. In general, these studies provide evidence – as assumed by BRESCHI and LISSONI (2001) – that the localisation of knowledge flows is to a lesser extent related to geographical distance effects, than to institutional or technological conditions, i.e. geographical proximity is not a sufficient condition for knowledge flowing between agents, but still conducive.

Other empirical studies focus on research collaborations to investigate the geographic localisation of knowledge flows. KATZ (1994) investigates the effect of geographical proximity for university collaborations in Canada, Australia and the UK, while LIANG and ZHU (2002), PONDS et al. (2007) and HOEKMAN et al. (2009) use research collaborations as captured by joint scientific publications (co-publications) to analyse localisation effects of knowledge flows. The current study follows the research tradition focusing on research collaborations as captured by co-publications to investigate the geographical dimension of knowledge diffusion. The objective is to provide new empirical insights on the mechanisms of collaborative knowledge production in China by using new information on co-authored scientific publications, and to estimate how spatial, technological and economic characteristics affect the variation of collaboration activities between Chinese regions. We

argue that the regional level is an appropriate choice since "Regions have played and will continue to play a key role in the advancement of S&T in China" (OECD 2007, p. 22). The region is an important unit in a Chinese science and technology policy context since recent policy efforts aim to reduce barriers to cross-region collaborations in science in order to support the development of an harmonised Chinese innovation system.

In using Chinese co-publication data as a proxy for collaborative knowledge production in China, we build on recent work by LIANG and ZHU (2002)[1] but depart from this prior work in at least three major respects: *First*, we follow PONDS et al. (2007) and HOEKMAN et al. (2009) by employing a spatial interaction model of the gravity type to identify and measure separation effects of cross-region collaborative knowledge production, and use a Negative Binomial model specification due to the presence of unobserved heterogeneity in our cross-region co-publication data. This methodological framework enables us – in contrast to the study of LIANG and ZHU (2002) – to control for other determinants than geographical distance affecting cross-region collaboration intensities. *Second*, we make use of new unique data on scientific publications in China of the year 2007 extracted from the China National Knowledge Infrastructure (CNKI) database. *Third*, we investigate differences of spatial patterns of collaborative knowledge production and their determinants across six different scientific fields. Such differences may occur due to different collaboration behaviour across scientific fields, related to different epistemic and cognitive regimes (see, for instance, THAGARD, 1997; WRAY, 2002). Differences in the geographical localisation of scientific disciplines may be related to the localisation of important research equipment, in particular in natural sciences, and to the use of ICTs. PONDS et al. (2007) provide evidence that geographical distance effects to collaboration differ across scientific subfields using Dutch co-publication data.

The remainder of the study is organised as follows. The next section highlights some general aspects of scientific collaborations in China, sheds some light on the Chinese (regional) policy context, and provides a short literature review. *Section 3* describes the empirical setting of the study and discusses in some detail the co-publication data used accompanied by some descriptive statistics. In *Section 4* we analyse spatial patterns of collaborative knowledge production across Chinese regions. Our territorial breakdown is composed of 31 regions, including the 22 provinces of China, five autonomous regions and four municipalities. We use a slightly modified Jaccard similarity coefficient to identify the relative strongest collaboration links in terms of attractiveness. *Section 5* continues to describe the regional gravity model that we use to estimate separation effects of cross-region collaboration activities. The dependent variable are observed cross-region collaborative intensities, the independent variables include origin, destination and separation variables, such as the geographical and technological distance as well as the economic gap between regions. We use a Negative Binomial model specification to account for the non-negative, integer nature of our dependent variable and to allow for the overdispersion in the data by letting each pair of regions parameter have a random distribution of its own. *Section 6* presents the estimation results for all co-publications as well as for six different scientific fields, before we conclude with a summary of the main results, some policy implications and a short outlook in the final section.

2 Scientific research collaborations in China

In the current era of the knowledge-based economy, it is widely agreed that research collaborations, both in the industry and the science sector, but also between the science and the industry sector, are essential for the successful production of new knowledge (see, for example, OECD, 1992, among many others). In the science sector, co-publications accounted

for less than 10% of all publications at the beginning of the last century. In the 1990s, the

share of co-publications has increased significantly to more than 60% of all publications (see

WAGNER-DOEBLER, 2001). Over the last few years, several governments have taken series

of measures to stimulate different kinds of research collaborations between firms, universities

and research organisations. For instance, the European Union provides substantially

increasing financial support to collaborative research activities in the Framework Programmes

for Research and Development.

China follows this trend in the strategic orientation of its science and technology policy.

China faces two major barriers for research collaborations. *First*, there exist enormous

regional disparities, both in the education as well as in the science sector. A considerable

number of empirical studies on this issue confirm the essential inter-regional gap of education

and research systems in China (see TU and YI, 2008; ZHANG and KANBUR, 2005; among

many others). LI (2005) further notes that science-technology talent resources are unequally

distributed between provinces in China, while this kind of imbalanced distribution is

consistent with disparity patterns in the economic development of regions. Statistically, R&D

personnel in the first five regions accounts for 46% of all regions in 2007, whereas the last

five regions only hold 1.11%. Similarly, R&D expenditures in the highest five regions occupy

53% of total expenditures in 2007, whereas the last five only account for 0.66% (NBS/MOST,

2008). Moreover, LIU and JIA (2008), and XU et al. (2005) report considerable disparities

between the coastal, the central and the western regions of China. In this context, OECD

(2007) points out that these regional patterns of R&D and innovation activities in China are

not optimal from a social equity perspective, as innovation systems in lagging regions are

underdeveloped. Thus, OECD (2007) also concludes that the national innovation system is

not fully developed and still imperfectly integrated due to various barriers between the

specific regional innovation systems.

Second, regional protectionism is a widely documented phenomenon in China (see, for instance, BAI et al., 2004). CHEN and WANG (2003) disclose that regional policy makers follow a protectionism perspective in the orientation of their science policy. Also ZHANG (2003) considers regional protectionism as one of the major problems of cross-region collaborations in science and technology. One prominent instrument used by local governments to exert creeping protectionism is China's so-called *hukou* (household registration) system (see CHEN and WANG, 2003). This system has imposed strict limitations on ordinary Chinese citizens changing their permanent place of residence. Employment, housing, and social benefits are commonly linked to *hukou* identification. Rural migrants to urban areas are often unable to obtain equal access to public services such as health care and education. Also the mobility of researchers for scientific knowledge production is affected by the *hukou* system. For instance, it is more difficult for researchers to move to another university located in a city for which they have no *hukou*. Furthermore, regional authorities at the province level play an important role to allocate S&T resources in China. 93% of public research institutions in China are mainly funded by local governments (NBS, 2007). Thus, research collaborations between research institutions located in different regions face significant barriers coming from regional protectionism (see also OECD, 2007).

The Chinese government has recently established a series of policy measures to remove barriers to research collaborations in view of the importance of inter-regional research collaborations. For instance, *The State Science and Technology Development Plan for "the Eleventh Five-Year"* (2006-2010), issued by the Ministry of Science and Technology (MOST) in 2006, notes that inter-regional research collaborations and innovation alliances should be encouraged. The plan emphasises the importance of key S&T projects for impelling the development of inter-regional collaborations and innovation alliances, and the necessity to

improve the central supervision of the allocation of regional S&T resources in order to curb

barriers coming from regional protectionism. Furthermore, *The Grand Western Development*

Program for "the Eleventh Five-Year" (2006-2010), issued by the National Development and

Reform Commission in 2007, points to the crucial importance to foster collaborations and

between East, Central and West China. Further measures in this context include the removal

of administrative barriers to inter-regional collaborations, and the establishment of specific

funding and coordinating organizations aimed to impel such inter-regional collaborative

activities.

There exists a significant body of literature related to the analysis of scientific research

collaborations in China. These studies can roughly be divided into three categories:

Conceptual and theoretical studies, empirical analyses of individual regions, and empirical

research on interactions between regions. The *first* category generally employs qualitative

methods to describe the current situation, including existing problems and policy

recommendations, of scientific research collaborations in China (see, for instance, WEI and

WU, 2004). Briefly speaking, these studies point out the importance and significance of

scientific research collaborations and encourage fostering inter-regional cooperation in

science and technology. The *second* category of studies investigates individual regions with

respect to their collaboration ability by measuring their knowledge absorption capabilities (see

WANG et al., 2005; SU, 2007, among others). The *third* category expands the horizon of the

second by focusing on cross-region interactions in research and science. As mentioned above,

the most notable contribution in this context is the study of LIANG and ZHU (2002). They

investigate factors affecting China´s cross-region research collaborations as captured by co-

publications and conclude that impact of geographical proximity is essential. Further studies

that can be mentioned are WANG et al. (2004) and HONG (2008). The first uses the location

of keywords in scientific publications to investigate the geographical and temporal dimension

of cross-region knowledge spillovers, while the second explores knowledge transfers from Chinese universities to the industry sector at a regional level.

3   The co-publication data and some descriptive statistics

This section discusses in some detail the co-publication data used in the study at hand and describes the empirical setting. Our study area is composed of $i, j = 1, …, n = 31$ regions, including the 22 provinces of China, five autonomous regions and four municipalities. The detailed list of regions is given in Appendix A, accompanied by a map on the spatial configuration of the study area. We use data on scientific publications from the China National Knowledge Infrastructure (CNKI) database, which is the most comprehensive and largest source of China-based information resources and publications. On a basis of random sampling, we use scientific articles published in 2007 covering six thematic fields that are *Agriculture*, *Economics*, *Information Technologies*, *Medicine*, *Natural Sciences & Engineering*, and *Social Sciences*. By using data on publications of the year 2007, the dataset reflects the current sketch of scientific knowledge production in China. Note that the year 2007 is not a special year with respect to differences in publication behaviour or changing patterns in regional protectionism[2]. Our random sample covers about 50% of all publications in 2007, leading to a sample size of 142,548 publications generating 758,682 co-publication flows. We use two-digit zip codes of the author address to trace the specific region of an author.

It is worth emphasising that we only consider publications with financial support, as recorded in the CNKI database, provided by diverse funding sources, including the central government, local governments, high education institutions, research institutions, enterprises, social organizations, military organizations, international bodies, etc. By this, we take into account

policy initiatives aimed to support research collaborations[3]. However, our data are not completely without limitations. *First*, different scientific journals have different ways to deal with zip code information of authors. For instance, some journals only show the zip code information for the corresponding author, while some even do not provide any zip code information. *Second*, we found that some publications are inappropriately classified according to the six thematic fields given in the CNKI database. Further, more general limitations to the use of co-publication data are provided in HOEKMAN et al. (2009). Despite these shortcomings, we still consider our data appropriate and very useful given our research objectives, in particular due to the large sample size used.

By nature of our research question, we need to construct region-by-region co-publication matrices based on a two-step procedure: *First*, we produce a general region-by-region co-publication matrix including all publications (i.e. irrespective of the thematic field), that we label Y, by aggregating the number of individual co-publications to the regional level[4]. y_{ij} represents one element of Y denoting the number of observed co-publications between two region i and j. Thus, Y contains the co-publication intensities between all (i, j)-region pairs, given the $i = 1, \ldots, n = 31$ regions in the rows and the $j = 1, \ldots, n = 31$ regions in the columns[5]. *Second*, the sector specific collaboration matrices for our six thematic fields are extracted from Y, by excluding all publications not belonging to the respective thematic field. We end up with six further n-by-n collaboration matrices for each thematic field, labelled $Y^{(agr)}$ for *Agriculture*, $Y^{(eco)}$ for *Economics*, $Y^{(it)}$ for *Information Technologies*, $Y^{(med)}$ for *Medicine*, $Y^{(nse)}$ for *Natural Sciences & Engineering* and $Y^{(soc)}$ for *Social Sciences*.

Table 1 about here

As a prelude to the analysis the follows, Table 1 presents some descriptive statistics on cross-region co-publications. The mean number of collaborations between any two regions is 789.39, with a high standard deviation of 4,496.22. The statistics for skewness and kurtosis point to an extremely right-skewed distribution, indicating that there are relatively few region pairs with a high number of collaborations, while the majority of the region pairs show relatively low collaboration intensities. The same holds true for the sector-specific collaboration matrices. The coefficient of variation differs slightly across the different co-publication matrices.

4 Spatial patterns of Chinese research collaborations

This section sheds some light on spatial patterns of Chinese research collaborations as captured by co-publications. Figure 1 presents the cross-region network of knowledge flows across Chinese regions. It can be seen that the region of Beijing is the central hub in this spatial co-publication network. One line represents one co-publication link between two regions i and j, the size of the nodes corresponds to a region´s degree centrality, i.e. the number of links of one region from a projected network analysis point of view. At a first glance, the results indicate that co-publications spread out star-shaped from the region of Beijing to all other regions. This may not only be explained by the fact that most universities and researchers are located there. The total number of co-publication links of Beijing equals to 117,226, followed by the region of Jiangsu (60,072) and the region of Guangdong (58,166). The highest cross-region interaction intensity is observed for the region pairs Beijing and Shandong (2,526), followed by Beijing and Shaanxi (2,341) and Beijing and Guangdong (2,202). Figure 1 also illustrates a triangle of strong collaborative activities between the regions of Beijing, Shanghai and Jiangsu. Western regions show relatively lower participation rates. Appendix B presents a spatial collaboration map for each of the thematic subfields. The

region of Beijing shows up as the central hub with respect to the number of links in all sector-specific spatial networks. However, some sectoral differences are observable (see Figure B.1).

Figure 1 about here

This exploratory spatial analysis reveals cross-region co-publication flows in terms of absolute link size. However, from social network analysis we know that we should consider the relative strength of the links between nodes, i.e. regions in our case. One appropriate measure to capture the relative size of the cross-region collaborative links is the Jaccard index (see, for instance, LEYDESDORFF, 2008). In our study the index is defined as[6]

$$J_{ij} = \left(y_{i\bullet} + y_{\bullet j} - y_{ij} \right)^{-1} y_{ij} \qquad\qquad i, j = 1, \ldots, n \qquad\qquad (1)$$

with $y_{i\bullet} = \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} y_{ij}$, $y_{\bullet j} = \sum\limits_{\substack{i=1 \\ i \neq j}}^{n} y_{ij}$ and y_{ij} is the number of observed links between i and j.

The calculation of the J_{ij} coefficient for our (i, j)-region pairs leads to interesting results concerning the spatial structure of cross-region co-publications in China. Table 2 presents the top 5 collaborative links in terms of the Jaccard index J_{ij} for all co-publications and for co-publications in different scientific fields. *First*, it comes out that the relative strongest links are different from the highest links in absolute numbers. *Second*, by far most of the relative strongest links correspond to collaborations between regions that are direct spatial neighbours. It is notable that the connectivity Beijing does not particularly stand out according to the Jaccard index, though Beijing has a rather central role for scientific knowledge production in China as illustrated by Figure 1. This is mainly the outcome of the high geographical diversification of regions collaborating with Beijing, while, for instance, collaborations of the

region Chongqing are highly focused on the region Sichuan, and vice versa. One may also conclude that different disciplinary specialisation and diversification patterns of regions play a role here. However, we cannot identify considerable differences in the disciplinary specialisation of the regions with respect to their publication intensity in our six scientific fields[7]. Thus, we conclude that there may exist – maybe historically well established – links between specific universities that are located in neighbouring regions and are characterized by a higher researchers mobility between them. The strongest connection according to the Jaccard index appears between the regions Chongqing and Sichuan in the western area of China. This pair belongs to the top 5 in all other thematic fields except from *Social Sciences*. A very strong relative connection is also found for the region pairs Jiangsu and Shanghai as well as for Shanghai and Zhejiang. Beijing and Shandong is the only non-neighbour pair that appears twice (all co-publications and *Natural Sciences & Engineering*). These results may point to the existence of strong geographic localisation effects in Chinese co-publication activity.

Table 2 about here

5 The regional gravity model approach

The exploratory analyses of the previous sections points to some interesting issues and reveals spatial patterns of cross-region collaborative knowledge production in China. The question that now arises is what are the determinants of the observed patterns? In this context we aim to let the variation of cross-region collaboration activities depend on a set of exogenous variables, including geographical, technological and economic characteristics. From this perspective, we are interested in models of the form[8]

$$Y_{ij} = X_{ij} + \varepsilon_{ij} \qquad\qquad\qquad i, j = 1, \dots, n \qquad (2)$$

where Y_{ij} is assumed to be a stochastic dependent variable corresponding to the observed collaboration flow $y_{ij} \geq 0$ between two regions i and j. X_{ij} is a function that captures the stochastic relationship to other random variables sampled from a specified probability distribution dependent upon some mean, say μ_{ij}, with $E[Y_{ij}] = \mu_{ij}$. ε_{ij} is a disturbance term with the property $E[\varepsilon_{ij} \mid y_{ij}] = 0$.

For our empirical analysis, the next step is to develop an appropriate model for the systemic part of the model X_{ij}. Since we are dealing with interactions between regions in a multiregional setting, the spatial interaction model of the gravity type is an appropriate instrument. Thus, $\mu_{ij} = X_{ij}$ is specified as a function of covariates measuring the characteristics of origin regions, destination regions, and their separation, given by

$$X_{ij} = f(O_i \ D_j \ S_{ij}) \qquad\qquad\qquad i, j = 1, \dots, n \qquad (3)$$

where O_i is a function characterising region i of interaction, D_j is a function characterising region j of interaction, while S_{ij} is a function characterising the separation between two regions i and j. We follow classical spatial interaction theory (see, for instance, SEN and SMITH 1995) and specify origin and destination functions by using power functions as

$$O_i = o_i^{\alpha_1} \qquad\qquad\qquad i = 1, \dots, n \qquad (4)$$

$$D_j = d_j^{\alpha_2} \qquad\qquad\qquad j = 1, \dots, n \qquad (5)$$

where o_i is the number of researchers[9] in region i, and d_j the number of researchers in region j. α_1 and α_2 are the respective parameters to be estimated[10]. By nature of our research questions, the focus of interest is on the spatial separation function S_{ij} that is specified as

$$S_{ij} = \exp\left[\sum_{k=1}^{K} \beta_k \ s_{ij}^{(k)}\right] \qquad\qquad i, j = 1, \ldots, n \qquad\qquad (6)$$

with $s_{ij}^{(k)}$ representing a measure of separation between i and j. The β_k are unknown parameters. By theory and our research questions, our interest is focused on $K = 6$ separation measures: $s_{ij}^{(1)}$ denotes geographical distance between two regions i and j as measured in terms of the great circle distance between the capital cities of the regions[11]. $s_{ij}^{(2)}$ is a dummy variable that is set to one if two regions i and j are physical neighbours, and zero otherwise[12]. $s_{ij}^{(3)}$ accounts for the economic gap of two regions i and j and is defined as $s_{ij}^{(3)} = G_i - G_j$ where G_i and G_j is the Gross Regional Product in 2007 for region i and j, respectively[13]. Furthermore, we add two variables that control for the fact that China is characterized by huge disparities between the coastal area and the rest of China[14]. $s_{ij}^{(4)}$ is a dummy variable that is set to one if either or both regions i and j are located in the coastal area of China, and zero otherwise, while $s_{ij}^{(5)}$ is a dummy variable that takes values of ones if either or both regions i or j are located in the central area of China[15], and zero otherwise. Finally, $s_{ij}^{(6)}$ controls for the technological distance between two regions i and j. We use data from the Chinese Patent Office to measure this variable. It is constructed as a vector $t(i)$ that measures region i's share of patenting in specific technological classes of the International Patent Classification (IPC)[16]. We follow MORENO et al. (2005) and use the Pearson correlation coefficient given by $r^2 = corr\left[t(i), t(j)\right]^2$ between the technological vectors of two regions i and j to define how close

they are to each other in technological space. Their technological distance is then given by

$$s_{ij}^{(6)} = 1 - r^2.$$

Integrating Equations (4)-(6) into Equation (2) leads to the empirical model:

$$y_{ij} = o_i^{\alpha_1} \, d_j^{\alpha_2} \, \exp\left[\sum_{k=1}^{K} \beta_k \, s_{ij}^{(k)}\right] + \varepsilon_{ij}. \qquad\qquad i, j = 1, \dots, n \qquad (7)$$

At this point, we are interested in estimating the parameters α_1, α_2 and β_k that are elasticities of cross-region scientific collaborations y_{ij} with respect to the origin variable o_i, the destination variable d_j and the separation variables $s_{ij}^{(k)}$.

Due to the integer non-negative nature of our data and distributional assumptions, OLS estimation procedures are not appropriate (see, for example, LONG and FREESE, 2001). The Poisson distribution is widely considered as a reasonable description for non-negative integer values (see MADDALA, 1983). In our empirical setting, the Poisson density function is given by

$$\Pr\left(Y_{ij} = y_{ij} \mid X_{ij}\right) = e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} \big/ y_{ij}!, \qquad\qquad i, j = 1, \dots, n; \; y_{ij} = 0, 1, 2, \dots \qquad (8)$$

with

$$\mu_{ij} = X_{ij} = \exp\left[\alpha_0 + \alpha_1 \log(o_i) + \alpha_2 \log(d_j) + \sum_{k=1}^{K} \beta_k \, s_{ij}^{(k)}\right] \qquad (9)$$

where α_0 is a constant, and assuming that

$$\mu_{ij} = V\left[Y_{ij}\,\middle|\,X_{ij}\right] = E\left[Y_{ij}\,\middle|\,X_{ij}\right] \tag{10}$$

The parameters are estimated by standard Maximum Likelihood procedures using Newton Raphson iterations (see CAMERON and TRIVEDI, 1998).

Specification (10) implies a particular form of heteroscedasticity due to equality of the conditional mean and the variance of Y_{ij} given X_{ij}. This indicates that the independent variables are assumed to account for all individual deviations. However, in a multiregional setting the existence of unobserved heterogeneity is very likely and may lead to biased estimates (see HAUSMAN et al., 1984). A common strategy to overcome the problem of unobserved heterogeneity is to introduce a stochastic heterogeneity term, $\exp(\xi_{ij})$, in the conditional mean leading to a probability distribution given by

$$\mathrm{Pr}\left(Y_{ij} = y_{ij}\,\middle|\,X_{ij}^*\right) = \exp(-\mu_{ij})^{\exp(\xi_{ij})}\,\mu_{ij}^{\,y_{ij}}\,\middle/\,y_{ij}! \qquad i, j = 1, \dots, n;\ y_{ij} = 0, 1, 2, \dots \tag{11}$$

with $X_{ij}^* = \exp\left[\alpha_0 + \alpha_1 \log(o_i) + \alpha_2 \log(d_j) + \sum_{k=1}^{K} \beta_k\, s_{ij}^{(k)} + \xi_{ij}\right]$ and $\exp(\xi_{ij}) \quad \Gamma(\gamma)$ where γ is an additional parameter allowing for overdispersion[17] and $\Gamma(\cdot)$ denotes the gamma function (see, for example, LONG and FREESE, 2001)[18]. Integrating ξ_{ij} out of Equation (11) produces the unconditional Negative Binomial distribution of y_{ij}:

$$\mathrm{Pr}\left(Y_{ij} = y_{ij}\,\middle|\,X_{ij}^*\right) = \frac{\Gamma(y_{ij} + \gamma^{-1})}{\left[y_{ij}!\,\Gamma(\gamma^{-1})\right]}\,\theta_{ij}^{1/\gamma}\left(1 - \theta_{ij}\right)^{y_{ij}} \qquad i, j = 1, \dots, n;\ y_{ij} = 0, 1, 2, \dots \tag{12}$$

with

$$\theta_{ij} = \gamma^{-1} / (\gamma^{-1} + \mu_{ij}). \qquad\qquad i, j = 1, ..., n \qquad\qquad (13)$$

Model estimation is again done by Maximum Likelihood estimation procedures (see CAMERON and TRIVEDI, 1998 for details on the ML estimation).

6 Estimation results

Table 3 presents the sample estimates of the gravity models for all co-publications (**Y**). The first column gives the results for the basic Poisson model without heterogeneity as specified by Equations (8)-(10), the second column for the Negative Binomial model (Equations (12) and (13)). Standard errors are given in brackets, and parameters are estimated by Maximum Likelihood procedures, using Newton Raphson. The estimated value of the dispersion parameter $\gamma = 0.957$ with $p < 0.001$ in the Negative Binomial specification (second column of Table 3) indicates that the basic Poisson model has to be rejected. Assumption (10) is too restrictive to adequately model our cross-region co-publication data due to unobserved heterogeneity among the (i, j)-region pairs. Thus, we prefer the Negative Binomial model specification leading to a very large increase of the log-likelihood function as compared to the Poisson model without heterogeneity. The Negative Binomial estimates are all significant and robust. However, the estimated standard errors are somewhat larger than for the basic Poisson model due to allowance for an additional source of variance.

In the context of the relevant theoretical and empirical literature as well as within a Chinese policy context, the results provide new valuable insights into the mechanisms of the diffusion of scientific knowledge in China. *First*, geographical distance between authors exerts a significant negative effect on the probability that they collaborate with each other. The parameter estimate of β_1 = -0.354 tells us that for each additional 100 km between two authors, the probability for collaboration decreases by about 29.8%, holding all other variables constant. Since we are investigating knowledge flows within one country, this relatively high negative effect of geographical distance is somewhat surprising as compared to similar studies for regions located in different countries that find similar or even lower distance effects – using different indicators for knowledge flows between European regions, including co-publications (see FISCHER et al., 2006; LESAGE et al.; 2007; PONDS et al. 2007; SCHERNGELL and BARBER, 2009). In an integrated research area collaboration patterns should be based solely on scholarly ground, and not impeded by geographical barriers. Thus, we conclude that China´s regional integration policy faces a big challenge, in particular – but not exclusively – because of negative geographical distance to inter-regional collaborative knowledge production. There seems to be a clear need for further harmonising the Chinese research and education systems.

Table 3 about here

A specific type of localisation is reflected by the estimate for β_2 = 0.254, indicating that collaboration probability increases between regions that are physical neighbours, while suggesting that this effect is slightly smaller in magnitude than the effect of geographical distance. The estimate for the coastal area dummy variable (β_4 = 0.169) indicates that the mean collaboration probability between any two regions increases when at least one region is

located in the coastal area. In contrast, interaction probability between any two regions decreases when at least one region is located in the central area (β_5 = -0.818) pointing out that region pairs involving regions located in the central area have a lower probability for research collaboration than other non-central area pairs.

The estimate for technological distance β_6 = -0.671 suggests that cross-region co-publication activities are more likely to occur between regions that are close to each other in technological space. By this, our econometric analysis confirms the hypothesis for the Chinese case that the variation of cross-region knowledge flows is not only affected significantly by geographical distance, but to a somewhat larger extent by technological distance – as, for instance, proposed by BRESCHI and LISSONI (2001) and confirmed by various empirical studies for the European case (see, for instance, MAURSETH and VERSPAGEN, 2002; FISCHER et al., 2006; LESAGE et al., 2007; HOEKMAN et al., 2009; MAGGIONI and UBERTI, 2009; SCHERNGELL and BARBER, 2009 and 2010). This supports the conclusion that geographical proximity is not a sufficient condition for collaborations to occur between researchers, but still conducive. Regarding the impact of the economic gap between two regions on their collaboration probability (β_3 = -0.048), we find – though significant – a rather small effect. As expected, the estimates for the mass terms ($\alpha_1=\alpha_2=1.001$) are close to one, indicating that a higher number of research staff in a specific region increases the likelihood of collaboration with other regions.

Table 4 presents the estimation results of the Negative Binomial gravity models for different scientific fields. The dispersion parameter γ is significant for all models, indicating that the Negative Binomial specification is the right choice. Note that we have excluded the technological distance variable for the sector specific models. The models highlight important

and significant differences of the results across scientific fields. The most notable outcome is that the elasticity of the collaboration intensity with respect to geographical distance between collaborating researchers is lowest in the field of *Information Technologies*, maybe associated with a higher level of integration by using new information and communication systems. However, for *Information Technologies* we also find the highest neighbouring region effects, i.e. though negative effects of geographical distance are relatively lower, collaborations in this field are subject to a specific type of localisation, namely an increasing collaboration probability between neighbouring regions. The highest elasticity of collaboration intensity with respect to geographical distance is identified for the fields *Economics* and *Medicine*, while for the other fields the magnitude of the geographical distance effect is closer to the average for all publications. Regarding the estimates for the other separation variables in the sector-specific models given in Table 4, the results of the model for all co-publications are generally confirmed, though some estimates are not significant anymore.

Table 4 about here

Figure 2 visualizes the geographical distance decay probability on cross-region co-publications in China in different scientific fields along with the distance decay for all co-publications. In general, the curves demonstrate that the probability of cross-region co-publication activities is geographically bounded. Other things being equal, for total co-publications the collaboration probability decreases by about 57.12% within a distance of 300km. Further, it is clearly shown that for *Information Technologies* the distance decay associated with the probability of cross-region collaboration is much lower, pointing to a probability decrease of about 25.5% within a distance of 300km. The strongest decay is observed for *Economics* and *Medicine*.

Figure 2 about here

7   Closing comments and discussion

The empirical analysis of knowledge flows and knowledge diffusion in geographical and technological space is one of the key research areas of contemporary economic geography and economics of innovation. The study at hand adopts a regional gravity model framework to analyse the variation of cross-region collaborative knowledge production in China as captured by scientific co-publications. We focused on the question whether geographical space is a significant determinant of collaborative knowledge activities between Chinese regions, at the same time controlling for economic and technological differences. We use a Negative Binomial model specification of our regional gravity model to allow for overdispersion in the data.

The analysis has produced a number of interesting results in the context of the relevant empirical and theoretical literature. The probability of collaboration between researchers significantly decreases by geographical distance. Furthermore, technological distance between regions shows a somewhat larger negative effect on cross-region collaborative activities, indicating that the collaboration probability increases between regions that are located close to each other in technological space. By this, our study confirms the hypothesis for the Chinese case that the localisation of knowledge flows is to a lesser extent related to geographical distance effects, than to technological conditions, i.e. geographical proximity is not a sufficient condition for collaborations to occur between researchers, but still conducive. However, geographical effects are statistically significant across all scientific disciplines, though its magnitude varies. In the fields of *Information Technologies* and *Natural Sciences and Engineering* we find a lower negative effect of geographical distance. This may pertain to

21

the localisation of important research equipment, in particular in *Natural Sciences and Engineering*, and to the more intensive use of information and communication technologies. Furthermore, the results point to the existence of quite high barriers to collaboration between authors located in the central area and authors located in the coastal area of China. Effects of economic differences between regions are rather small but statistically significant.

The study at hand produces significant policy implications, and shows that there is much space for improvement in a Chinese policy context. Though we are investigating knowledge flows within one country, we find strong evidence for the negative effect of geographical distance on the variation of cross-region collaborative knowledge production in China. Thus, Chinese regional integration policy is claimed to establish further measures to ease barriers between regions, for instance by harmonising research and education systems. Two major factors that hinder inter-regional collaboration need to be addressed by the policy: *First*, regional protectionism still plays a crucial role. Thus, the central Chinese government needs to improve the supervision of the allocation of regional S&T resources by regional governments, and remove barriers for researchers mobility coming from inconsistencies because of different regional household registration systems. *Second*, regional economic and technological differences are significant factors. A series of measures has been taken to encourage cooperation between regions with different comparative advantages, but the strong negative effect of technological proximity to the variation of inter-regional research collaborations in the current study implies that there remains much space for improvement in this direction. Concerning economic effects, Chinese government has launched a series of policies to reduce regional economic disparities in the recent past, particularly *The Grand Western Development Program* starting in 1999. The empirical results imply that these measures may have to some extent achieved a positive feedback, at least in the context of cross-region research collaborations, though the impact of economic disparities is still

22

statistically significant. Thus, further policy initiatives, in particular *The Grand Western Development Program for "the Eleventh Five-Year"* (2006-2010), are intended to further reduce regional economic disparities[19].

Concerning a future research agenda, at least two points come to mind: *First*, a dynamic analysis of collaborative knowledge production as captured by co-publications across Chinese regions would provide valuable insight into the evolution of the estimated effects over time. The influence of the above mentioned policy measures could be more accurately analyzed under the framework of a dynamic model. *Second*, analysing other indicators for cross-region knowledge flows, such as co-patents or patent citations, would be important to enrich our understanding on the mechanisms of knowledge diffusion in China.

Endotes

[1] The study of LIANG and ZHU (2002) uses co-publication data to analyse the effect of geographical proximity by means of some simple correlations. They find that geographical proximity is one the most important factors of cross-region knowledge flows. However, they do not control for other factors that may affect collaboration intensities. For instance, geographical distance may just be a proxy for technological distance (see FISCHER et al., 2006)

[2] Research input, as captured by R&D expenditures, and research output, as captured by publications, in China shows a smooth upward trend over time from 1997 to 2007 (NBS/MOST, 2008).

[3] By focusing on publications with financial support, we delimit the scope of the study to a special form of co-publications that is – to our view – a more suitable proxy for cross-region collaboration intensities since we take into account policy initiatives aimed to support research collaborations. By this our indicator comes more close to recently used indicators for knowledge flows between European regions as captured by joint R&D projects in the European Framework Programmes (see SCHERNGELL and BARBER, 2009 and 2010).

[4] We follow full counting procedure, i.e. a publication with n authors produces $n(n\text{-}1)/2$ links that we count. For instance, for an article with three authors in three different regions, we count three links: from region a to region b, from b to c and from a to c. When all three authors are located in one region we count three intraregional links. Authors from the same institution are included when counting collaborations. They enter our calculations as intra-regional observations. However, bias in this direction does not play a role since the share of 'same institution'-collaborations is similar across all regions (between 27% and 31%).

[5] Note that the $n\text{-}by\text{-}n$ matrix is symmetric by construction ($y_{ij} = y_{ji}$).

[6] Note that Equation (1) is equally applicable to the sector specific networks. We do not distinguish between them in the formal presentation.

[7] The index of specialization calculated in this study is defined as $S_i = (1/2) \sum \left| y_{ik} - \overline{y}_{ik} \right|$, where y_{ik} is the share of publications in the scientific field k in region i, and $\overline{y}_{ik}$ is the respective average of China (see, for instance, HALLET, 2000). The average specialization index is 0.09, i.e. regions are highly diversified with respect to their publication intensity in different disciplines. The Herfindal concentration index shows values between 0.05 and 0.07 for the six fields, i.e. they are relatively equal distributed across Chinese regions.

---

[19]   *The Five-Year Plans of China* are a series of economic development initiatives, for example, *The First Five-Year Plan* (1953-1957), *The Tenth Five-Year Plan* (2001-2005) and *The Eleventh Five-Year Plan* (2006-2010). *The Grand Western Development Program for "the Tenth Five-Year"* (2001-2005) has been completed.

References

AUDRETSCH D.B. and FELDMAN M.P. (1996) R&D spillovers and the geography of innovation and production, The American Economic Review 86, 630-640;

BAI C.E., DU Y.J., TAO Z.G. and TONG S.Y. (2004) Local protectionism and regional specialization Evidence from China's industries, Journal of International Economics 63, 397- 41;

BRESCHI S. and LISSONI F. (2001) Knowledge spillovers and local innovation systems A critical survey, Industrial and Corporate Change 10, 975-1005;

BRÖCKER J. (1989) How to eliminate certain defects of the potential formula, Environment and Planning A 21, 817-830;

CAMERON A.C. and TRIVEDI P.K. (1998) Regression Analysis of Count Data, Cambridge University Press, Cambridge [UK] and New York [US].

CHATTERJEE S., HADI A.S. and PRICE B. (2000) Regression Analysis by Example, John Wiley & Sons, New York [US].

CHEN C.T. and WANG T. (2003) The influence to the local S&T legislation from the view of CAST operation mode forum on science and technology in China, Forum on Science and Technology in China 4, 3-6 (in Chinese);

CUMMINGS J.N. and KIESLER S. (2007) Coordination costs and project outcomes in multi-university collaborations, Research Policy 36, 1620-1632;

FISCHER M.M. and GRIFFITH D. (2008) Modeling spatial autocorrelation in spatial interaction data. An application to patent citation data in the European Union, Journal of Regional Science 48, 969-989;

FISCHER M.M., SCHERNGELL T. and JANSENBERGER E. (2006) The geography of knowledge spillovers between high-technology firms in Europe. Evidence from a spatial interaction modelling perspective, Geographical Analysis 38, 288-309;

HALLET M. (2000) Regional Specialisation and Concentration in the EU. European Commission, Brussels.

HAUSMAN J., HALL B.H. and GRILICHES Z. (1984) Econometric Models for Count Data with an Application to the Patents-R & D Relationship, Econometrica 52, 909-938;

HINDS P. and BAILEY D. (2002) Out of Sight, Out of Sync understanding conflict in distributed teams, Organization Science 14, 615-632;

HOEKMAN J., FRENKEN K., and VAN OORT F. (2009) The geography of collaborative knowledge production in Europe, The Annals of Regional Science 43, 721-738;

HONG W. (2008) Decline of the center The decentralizing process of knowledge transfer of Chinese universities from 1985 to 2004, Research Policy 37, 580-595;

ISMAIL N. and JEMAIN A.A. (2007) Handling overdispersion with Negative Binomial & Generalized Poisson Regression Models. Casualty Actuarial Society Forum, Winter, 103-158;

JAFFE A.B., TRAJTENBERG M. and HENDERSON R. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations, Quarterly Journal of Economics 108(3), 577-598;

KATZ J.S. (1994) Geographical proximity and scientific collaboration, Scientometrics 31, 31-43;

KATZ J.S. and MARTIN B.R. (1997) What is research collaboration?, Research Policy 26, 1-18;

KRUGMAN P. (1991) Geography and Trade, MIT Press, Cambridge [MA].

LESAGE J., FISCHER M.M. and SCHERNGELL T. (2007) Knowledge spillovers across Europe. Evidence from a Poisson spatial interaction model with spatial effects, Papers in Regional Science 86, 393-421;

LEYDESDORFF L. (2008) On the normalization and visualization of author co-citation Data Salton's cosine versus the Jaccard index, Journal of the American Society for Information Science and Technology 59, 77-85;

LI F.Z. (2005) Regional differences and types of science-technology talent resources distribution in China, Resource Development & Market 1, 19-22 (in Chinese);

LIANG L. and ZHU L. (2002) Major factors affecting China´s inter-regional research collaboration Regional scientific productivity and geographical proximity, Scientometrics 55, 287-316;

LIU N.Q. and JIA Y.L. (2008) Regional policy, regional openness and regional disparity in economic development, Contemporary Economy & Management 10, 65-72 (in Chinese);

LONG J.S. and FREESE J. (2001) Regression Models for Categorical Dependent Variables Using Stata. Stata Corporation, College Station [Texas].

MADDALA G. (1983) Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press, New York.

MAGGIONI M. and UBERTI T. (2009) Knowedge networks across Europe which distance matters?, The Annals of Regional Science 32, 691-720;

MAURSETH P. B. and VERSPAGEN B. (2002) Knowledge spillovers in Europe A patent citation analysis, Scandinavian Journal of Economics 104, 531-545;

MORENO R., PACI R. and USAI S. (2005) Spatial spillovers and innovation activity in European regions, Environment and Planning *A* 37, 1793-1812;

NBS (National Bureau of Statistics) (2007) China Statistical Yearbook, Statistics Press, Beijing, China.

NBS/MOST (National Bureau of Statistics/Ministry of Science and Technology) (2008) China Statistical Yearbook on Science and Technology, Statistics Press, Beijing, China.

OECD (1992) Technology and the Economy The Key Relationships. OECD, Paris.

OECD (2007) OECD Reviews of Innovation Policy China. Synthesis Report. OECD, Paris.

PONDS R., VAN OORT F. and FRENKEN K. (2007) The geographical and institutional proximity of research collaboration, Papers in Regional Science 86, 423-443;

SCHERNGELL T. and BARBER M. (2009) Spatial interaction modelling of cross-region R&D collaborations. Empirical evidence from the 5th EU Framework Programme, Papers in Regional Science 88, 531-546;

SCHERNGELL T. and BARBER M. (2010) Distinct spatial characteristics of industrial and public research collaborations Evidence from the 5th EU Framework Programme, The Annals of Regional Science [forthcoming];

SEN A. and SMITH T.E. (1995) Gravity Models of Spatial Interaction Behaviour. Heidelberg, Berlin and New York, Springer.

SU F.L. (2007) An empirical analysis on China's provincial R&D knowledge spillovers on using GWR, The Journal of Quantitative & Technical Economics 2, 145-153 (in Chinese);

THAGARD P. (1997) Collaborative Knowledge, Nous 31, 242-261;

TU C.L. and YI W.H. (2008) Factors and regional disparity of self-innovation capabilities in China, Science and Technology Management Research 1, 39-42 (in Chinese);

WAGNER-DOEBLER R. (2001) Continuity and discontinuity of collaboration behaviour since 1800 – from a bibliometric point of view, Scientometrics 52, 503-517;

WALSH K. (2007) China R&D A high-tech field of dreams, Asia Pacific Business Review 13, 321-335;

WANG Z., MA C.F., WENG G.L. and WANG, Y. (2004) A study on temporal and regional process of knowledge spillover, The Annals of Regional Science 38, 595-606;

WANG Y., YISHAN W., PAN Y., ZHENG M. and ROUSEAU R. (2005) Scientific collaboration in China as reflected in co-authorship, Scientometrics 62, 183.198;

WEI S.H. and WU G.S. (2004) The cause, model and policy implications of cooperation and communication of science and technology among administrative regions in China, China Soft Science 7, 103-108 (in Chinese);

WRAY K.B. (2002) The epistemic significance of collaborative research, Philosophy of Science 69, 150-168;

XU J.H., LU F., SU F.L., and LU Y. (2005) Spatial and temporal scale analysis on the regional economic disparities in China, Geographical Research 24, 57-68 (in Chinese);

ZHANG G.C. (2003) Study on the environment of science and technology innovation, Railway Technical Innovation 2, 7-10 (in Chinese);

ZHANG X.B. and KANBUR R. (2005) Spatial inequality in education and health care in China, China Economic Review 16, 189-204;

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix A

This study disaggregates China's territory into 31 regions composed of 22 provinces, five autonomous regions and four municipalities. Due to data limitations we have to exclude the special administrative regions Hong Kong and Macao.

Table 5 about here

Figure 3 about here

Appendix B

Figure 4 here

Table 1   Some descriptive statistics on collaborative knowledge production among Chinese regions as captured by scientific co-publications

	Sum	Mean	Standard Deviation	Min	Max	Skewness	Kurtosis	CV*
All co-publications Y	758,602	789.39	4,496.22	0	81,640	10.71	145.57	5.70
Agriculture $Y^{(agr)}$	73,156	76.12	370.37	0	5,124	8.10	77.61	4.87
Economics $Y^{(eco)}$	38,985	40.57	228.18	0	4,291	11.36	163.99	5.62
Information technologies $Y^{(it)}$	57,628	59.97	399.34	0	8,544	13.57	238.21	6.66
Medicine $Y^{(med)}$	225,851	235.02	1,571.14	0	26,850	11.31	154.43	6.69
Natural sciences & engineering $Y^{(nse)}$	343,201	357.13	1,954.09	0	38,771	11.70	181.86	5.47
Social sciences $Y^{(soc)}$	19,781	20.58	108.09	0	1,495	8.43	81.74	5.25

*Coefficient of Variation

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 2 Top 5 collaborative links in terms of the Jaccard index J_{ij} for all cross-region co-publications and for cross-region co-publications in different scientific fields.

Region pair	J_{ij}	Spatial neighbours	Region pair	J_{ij}	Spatial neighbours
All co-publications Y			Information Technologies $Y^{(it)}$		
Chongqing and Sichuan	0.087	yes	Chongqing and Sichuan	0.165	yes
Jiangsu and Shanghai	0.077	yes	Beijing and Jilin	0.093	no
Shanghai and Zhejiang	0.062	yes	Beijing and Shanghai	0.092	no
Henan and Shaanxi	0.058	yes	Hebei and Tianjin	0.091	yes
Beijing and Shandong	0.058	no	Henan and Shaanxi	0.080	yes
Agriculture $Y^{(agr)}$			Natural Sciences & Engineering $Y^{(nse)}$		
Jilin and Nei Mongol	0.125	yes	Chongqing and Sichuan	0.070	yes
Shanghai and Zhejiang	0.111	yes	Beijing and Shandong	0.066	no
Shanghai and Jiangsu	0.107	yes	Hebei and Tianjin	0.060	yes
Anhui and Jiangsu	0.083	yes	Henan and Shaanxi	0.058	yes
Chongqing and Sichuan	0.079	yes	Jiangsu and Shanghai	0.058	yes
Economics $Y^{(eco)}$			Social Sciences $Y^{(soc)}$		
Chongqing and Guizhou	0.124	yes	Jilin and Liaoning	0.133	yes
Anhui and Jiangsu	0.115	yes	Guizhou and Sichuan	0.125	yes
Chongqing and Sichuan	0.096	yes	Hubei and Hunan	0.113	yes
Henan and Hubei	0.074	yes	Heilongjiang and Jilin	0.106	yes
Shanghai and Zhejiang	0.072	yes	Gansu and Ningxia	0.102	yes
Medicine $Y^{(med)}$					
Jiangsu and Shanghai	0.128	yes			
Chongqing and Sichuan	0.096	yes			
Guangdong and Hubei	0.082	no			
Jilin and Liaoning	0.068	yes			
Guangdong and Henan	0.059	no			

Note Spatial neighbours are defined to share a common border

Table 3   Estimation Results of the regional gravity models for all co-publications [n^2 = 961 observations; asymptotic standard errors given in brackets]

	Basic Poisson model without heterogeneity	Negative Binomial model
Origin and destination variable [$\alpha_1=\alpha_2$]	0.695*** (0.003)	1.001*** (0.040)
Geographical distance [β_1]	-0.354*** (0.001)	-0.351*** (0.030)
Neighbouring region [β_2]	0.096*** (0.006)	0.254** (0.101)
Economic gap [β_3]	-0.027*** (0.001)	-0.048* (0.014)
Coastal area dummy [β_4]	0.119*** (0.003)	0.169*** (0.003)
Central area dummy [β_5]	-0.483*** (0.003)	-0.818*** (0.004)
Technological distance [β_6]	-0.179*** (0.011)	-0.671*** (0.158)
Intercept [α_0]	-7.487*** (0.031)	-6.101*** (0.000)
Dispersion parameter (γ)	–	0.957*** (0.042)
Log-Likelihood	-125,571.12	-5,408.68
AIC	251,160.23	10,837.37
Pseudo *R*-squared	0.723	0.769

Notes The dependent variable is the cross-region co-publication intensity between two regions *i* and *j*. The independent variables are defined as given in the text. Note that due to symmetry of the origin and destination variables we have a special case with [$\alpha_1=\alpha_2$]. The multicollinearity condition number yields a value of 4.023, while the Variance Inflation Factors (VIFs) for the variables range between 1.27 and 3.83, so we infer there are no multicollinearity problems (see CHATTERJEE et al., 2000).We tested the residual vector for the existence of spatial autocorrelation which can also be a problem in the context of interaction data (see FISCHER and GRIFFITH 2008). The respective Moran´s *I* statistic is insignificant, i.e. spatial autocorrelation in the error term does not exist. ***significant at the 0.001 significance level, **significant at the 0.01 significance level, *significant at the 0.05 significance level.

Table 4 Estimation Results of the Negative Binomial regional gravity models for different scientific fields [$n^2 = 961$ observations; asymptotic standard errors given in brackets]

	Negative Binomial regional gravity models					
	Agriculture $Y^{(agr)}$	Economics $Y^{(eco)}$	Information technologies $Y^{(it)}$	Medicine $Y^{(med)}$	Social sciences $Y^{(soc)}$	Natural sciences & engineering $Y^{(nse)}$
$\alpha_1=\alpha_2$	0.830*** (0.060)	1.025*** (0.061)	1.301*** (0.069)	0.873*** (0.058)	0.878*** (0.061)	1.027*** (0.044)
β_1	-0.324*** (0.044)	-0.434*** (0.040)	-0.225*** (0.042)	-0.448*** (0.043)	-0.368*** (0.039)	-0.289*** (0.032)
β_2	0.193 (0.151)	0.345** (0.133)	0.639*** (0.141)	0.063 (0.145)	0.151 (0.131)	0.261* (0.115)
β_3	-0.027 (0.038)	0.033 (0.035)	-0.180*** (0.036)	-0.025 (0.037)	-0.024 (0.034)	-0.057* (0.028)
β_4	0.001 (0.125)	0.434*** (0.003)	0.625*** (0.123)	0.250* (0.122)	0.304** (0.114)	0.224* (0.092)
β_5	-0.665*** (0.115)	-0.781*** (0.003)	-0.624*** (0.110)	-1.152*** (0.112)	-0.606*** (0.000)	-0.767*** (0.000)
α_0	-12.582*** (0.954)	-17.748*** (0.976)	-23.771*** (1.109)	-12.785*** (0.924)	-15.129*** (0.970)	-15.501*** (0.697)
γ	0.440*** (0.022)	0.614*** (0.036)	0.554*** (0.032)	0.467*** (0.024)	0.700*** (0.047)	0.803*** (0.036)
Ps. R-sq.	0.570	0.694	0.714	0.674	0.708	0.715

Notes The dependent variable is the cross-region co-publication intensity between two regions i and j in different scientific fields. The independent variables are defined as given in the text. Note that due to symmetry of the origin and destination variables we have a special case with [$\alpha_1=\alpha_2$]. Technological distance is not included in the sector-specific models. We tested the residual vector for the existence of spatial autocorrelation which could be a problem in the context of interaction data (see FISCHER and GRIFFITH 2008). The respective Moran´s I statistic is insignificant, i.e. spatial autocorrelation in the error term does not exist. ***significant at the 0.001 significance level, **significant at the 0.01 significance level, *significant at the 0.05 significance level.

Table 5 List of regions used in the study

ID	Region	Capital	Type	ID	Region	Capital	Type
1	Anhui	Hefei	Province	17	Jilin	Changchun	Province
2	Beijing	Beijing	Municipality	18	Liaoning	Shenyang	Province
3	Chongqing	Chongqing	Municipality	19	Nei Mongol	Hohhot	Autonomous Region
4	Fujian	Fuzhou	Province	20	Ningxia	Yinchuan	Autonomous Region
5	Gansu	Lanzhou	Province	21	Qinghai	Xining	Province
6	Guangdong	Guangzhou	Province	22	Shaanxi	Xian	Province
7	Guangxi	Nanning	Autonomous Region	23	Shandong	Jinan	Province
8	Guizhou	Guiyang	Province	24	Shanghai	Shanghai	Municipality
9	Hainan	Haikou	Province	25	Shanxi	Taiyuan	Province
10	Hebei	Shijiazhuang	Province	26	Sichuan	Chengdu	Province
11	Heilongjiang	Harbin	Province	27	Tianjin	Tianjin	Municipality
12	Henan	Zhengzhou	Province	28	Xinjiang	Urumqi	Autonomous Region
13	Hubei	Wuhan	Province	29	Xizang	Lhasa	Autonomous Region
14	Hunan	Changsha	Province	30	Yunnan	Kunming	Province
15	Jiangsu	Nanjing	Province	31	Zhejiang	Hangzhou	Province
16	Jiangxi	Nanchang	Province				

Figure 1. Knowledge flows as captured by co-publications across Chinese regions

Node size corresponds to a
region's degree centrality
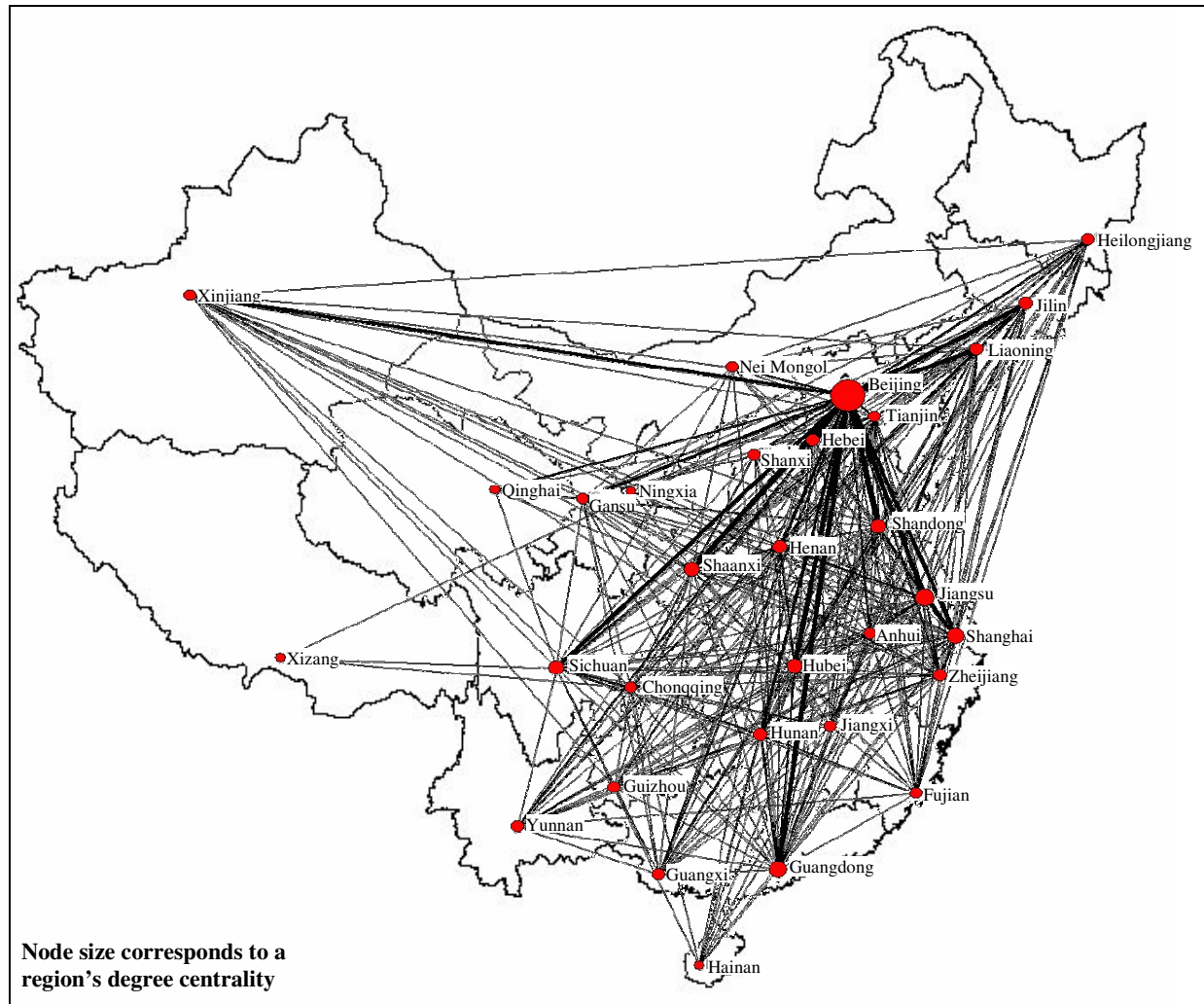
Figure 2   Distance decay probability on co-publications in China in different scientific fields [decay parameter β_1 estimated in Negative Binomial regional gravity models]
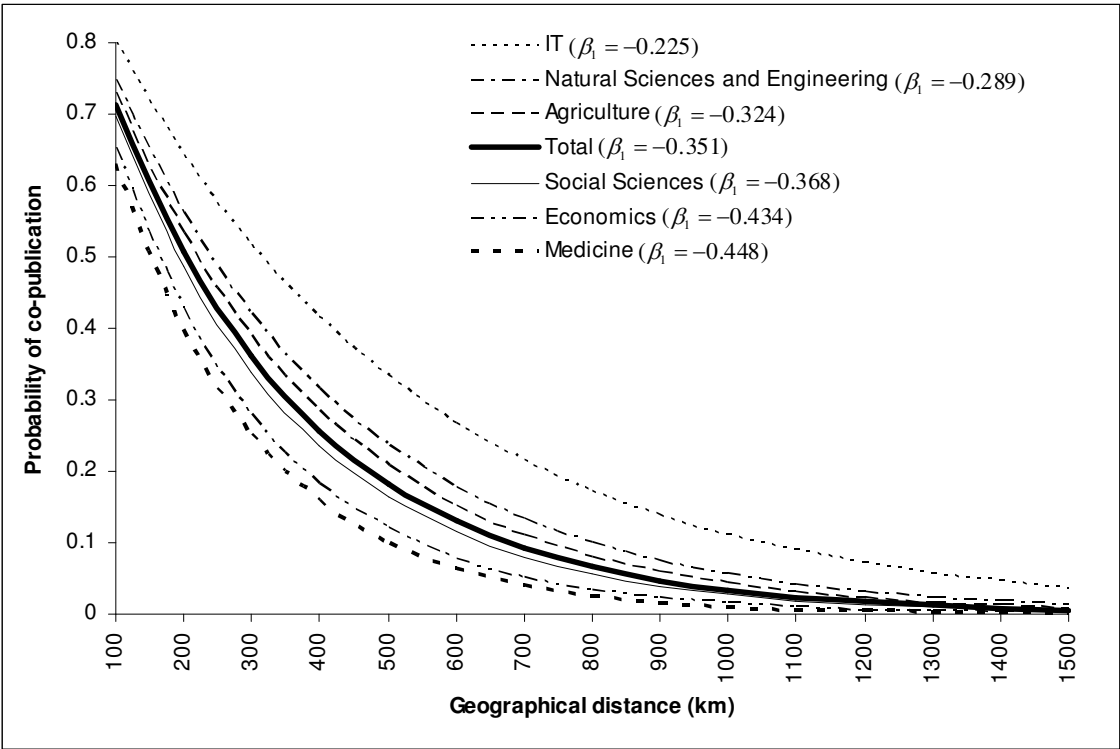
Figure 3 Study area (*n* = 31 regions)

Figure 4   Knowledge flows as captured by scientific co-publications across Chinese regions in six
different scientific fields A. Agriculture, B. Economics, C. Information Technologies,
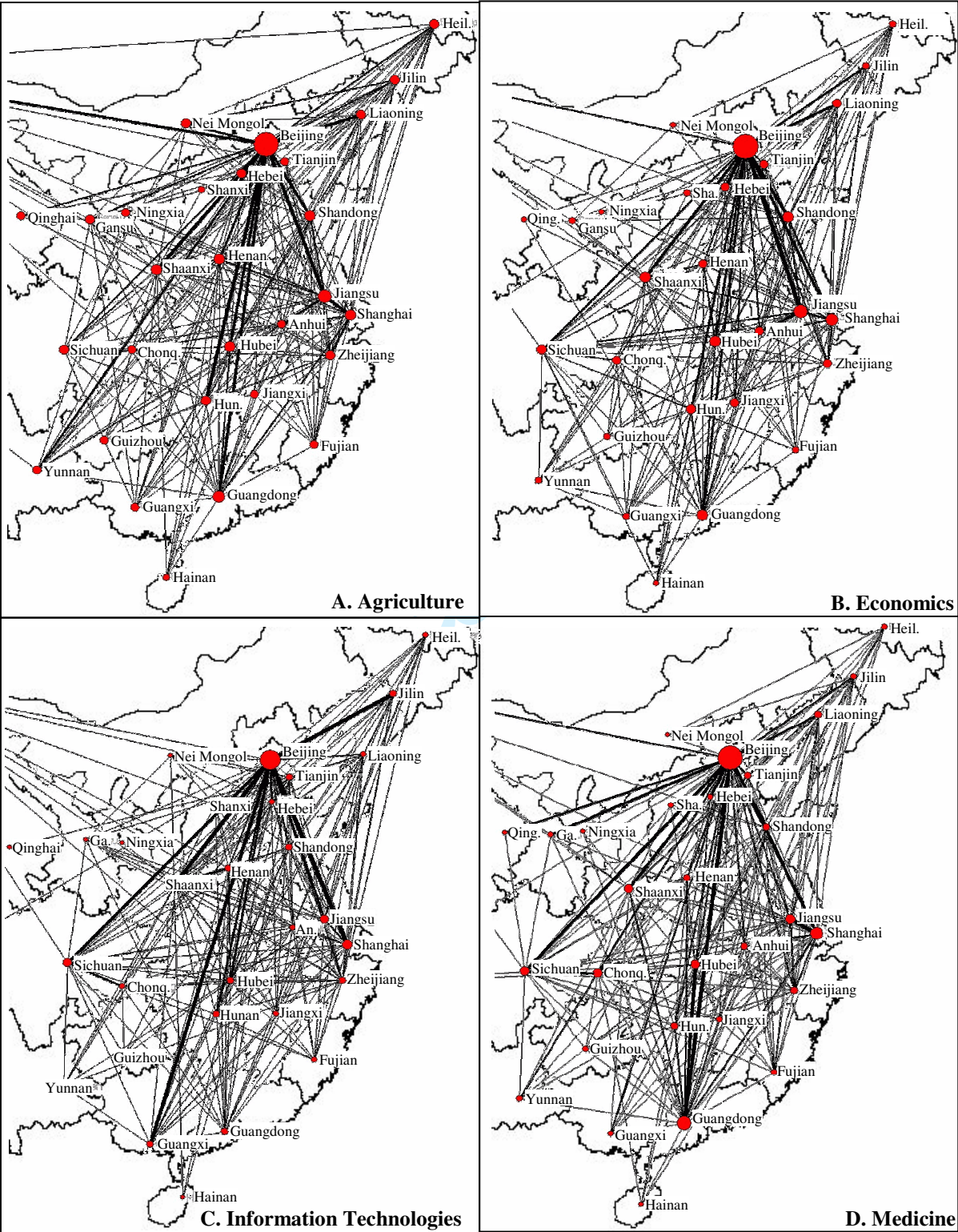D. Medicine, E. Social Sciences, F. Natural Sciences & Engineering

Figure B.1 continued

E. Social Sciences

F. Natural Sciences & Engineering