The Development of Computer Science Research in the People's Republic of China 2000-2009: A Bibliometric Study

Abstract

This paper reports a bibliometric study of the development of computer science research in the People's Republic of China in the 21st century, using data from the Web of Science, Journal Citation Reports and CORE databases. Focusing on the areas of data mining, operating systems and web design, it is shown that whilst the productivity of Chinese research has risen dramatically over the period under review, its impact is still low when compared with established scientific nations such as the USA, the UK and Japan. The publication and citation data for China are compared with corresponding data for the other three BRIC nations (Brazil, Russian and India). It is shown that China dominates the BRIC nations in terms of both publications and citations, but that Indian publications often have a greater individual impact.

Keywords

Bibliometrics; Citation analysis; Computer science research; People's Republic of China; Research impact; Research productivity

Introduction

Information and communications technology (ICT) is arguably the most important technology for supporting the economic development of a nation and thus for enhancing its population's quality of life. It is hence hardly surprising that the People's Republic of China (hereafter China), as the world's largest developing country, has made very extensive efforts to develop its R&D capabilities in ICT. Most notably, the Ministry of Science and Technology identified this as one of six technologies for funding in the National High Technology Programme (the other areas were biotechnology and agriculture, materials, manufacturing and automation, energy, and resources and the environment (Ministry of Science and Technology of the People's Republic of China, no date)). The programme (which is also called the 863 programme) started in 1986 with the aim of upgrading national industrial competitiveness so

that China could compete successfully with established players such as the USA, Japan and the European Union. ICT receives the largest fraction of the extensive government funding that has been made available to the programme (Springut *et al*., 2011), and a range of statistics attest to the success of this initiative: for example, China is now the world's largest mobile communications market and the largest producer of ICT products (Simon, 2011); and it has 75 of the world's 500 most powerful supercomputers, including the second and the fourth fastest (Top500 Computer Sites, 2011).

Computer science is one of the key basic sciences underlying ICT developments (other important areas include electronics, informatics and telecommunications), and Guan and Ma (2004) have noted that China has shown in this respect a "low level of beginning and high speed of developing" in computer science. The "beginning" has been described by Jiuchun and Baichun (2007), who state that computer science originated in the People's Republic of China with the founding of the Chinese Academy of Sciences' Institute of Computing Technology in 1956. Since then it has grown, first steadily (Maier, 1988), and then much more rapidly since the liberalization of the economy. In this paper, we discuss the growth in academic computer science research in China using the methods of bibliometrics (Bar-Ilan, 2008a; Borgman and Furner, 2002; Nicolaisen, 2007; Wilson, 1999).

Previous bibliometric studies have demonstrated the contributions that China is making to research knowledge, not just in general (Royal Society, 2011; Zhou and Leydesdorf, 2006; Zhou *et al*., 2009) but also in particular scientific disciplines such as bioinformatics (Guan and Gao, 2008), digital libraries (Zhao and Zhang, 2011), chemistry (Li and Willett, 2010), liquid crystals (Sangam *et al*., 2010), nanotechnology (Tang and Shapira, 2010), oncology (Yu et al., 2011) and superconductivity (Zhu and Willett, 2011) *inter alia*. However, we have been able to identify only five previous reports focusing specifically on the development of Chinese computer science. In a 2004 study, Guan and Ma (2004) demonstrated a marked growth in publications during the period 1993 to 2002, although they noted that these publications were principally in domestic journals and conference proceedings, with the result that many of their

findings had only a low level of international visibility. In the following year, Kumar and Garg (2005) reported a more extended comparison of Indian and Chinese computer science publications, covering the period 1971 to 2000, and again noted the Chinese preference for publication in domestic journals. More recently, He and Guan (2008) analyzed Chinese papers published in the period 1997 to 2005 in the conference proceedings series *Lecture Notes in Computer Science*. They found that the proportion of Chinese contributions had increased rapidly over this period, but noted that these contributions were not heavily cited and that the increase in conference publication had not been mirrored by a corresponding increase in publications in top computer science journals. A similar focus on quantity was noted by Ma *et al*. (2008) when comparing the competitiveness of world universities in computer science (and also by Calvert and Zengzhi (2001) when reviewing Chinese journals in information and library science, and by Li *et al*. (2012) when comparing the competitiveness of universities in mainland China, Hong Kong and Taiwan across the full range of disciplines). Finally, Li and Ke (2009) discussed research on data mining published in 64 Chinese social science journals in the period 1998-2007.

China is one of the four nations – Brazil, Russia, India and China – that together comprise the BRIC group of large nations and that have rapidly developing economies and levels of research activity (Kumar and Asheulova, 2011). Directories of ICT activity in the BRIC group have been carried out by Simon (2011) for the European Commission and by Sathya (2010) for the European Union, and there have also been several articles considering specific members of the group in more detail. Wainer *et al*. (2009) discussed Brazilian publications in computer science for the period 2001-2005, focusing principally on journal articles and on those conferences included in the *Lecture Notes in Computer Science* series, and compared their results with those for several other Latin countries and the other BRIC countries. In their comparison, Wainer *et al*. noted that Russia had a large fraction of its research in low-IF journals and that it had a very different subject profile (focusing strongly on cybernetics and theoretical computer science) when compared with the other BRIC countries during the period 2001-2005. A subsequent paper by this research group reported on the regional and gender

characteristics of Brazilian computer scientists (Arruda *et al*., 2009). There have been several bibliometric studies of Indian computer science research. Kumar and Garg (2005) compared Indian and Chinese computer science research for the period 1971-2000. India was notably more productive during this period, although China was noted as catching up rapidly, but there was no significant difference in the impacts of the nations' research. An analogous comparison of Indian and Chinese computer science research by Guan and Ma (2004) considered the period 1993-2002, i.e., principally the last part of the period studied by Kumar and Garg, and found that Indian research still had a greater international impact despite a much lower level of overall productivity. Gupta *et al*. (2010) provided extensive productivity data (using the SciVerse Scopus database) for Indian computer science research in a range of areas. In a subsequent paper Gupta *et al*. (2011) compared these results with those for China and Brazil (and also for South Korea and Taiwan), stating that India was far behind the others in terms of publications, but provided no corresponding comparison of the impact of the various nations' research.

This paper reports an analysis of Chinese computer science publications in the first ten years of the 21st century, hence providing a detailed update of the earlier studies of Guan and Ma (2004) and of Kumar and Garg (2005) described above. Moreover, we extend their work in two ways. First, rather than considering computer science at the disciplinary, macro level, we consider China's contribution at the micro level by focusing on three specific aspects of the discipline, *viz* data mining (as an example of a topic that has come very much to the fore over the last ten years with the massive rise that has taken place in the scope and the scale of modern information systems); operating systems (as an example of a topic that has been at the core of computer science ever since it emerged as an academic discipline in the second half of the last century); and web design (as an example of an application that can be expected to grow still further in importance with the increasingly rapid digitization of many aspects of modern society). Second, we place our findings in context by comparing the Chinese research performance both with the other members of the BRIC group, and with three established, productive nations for computer science research, i.e., the USA, the UK and Japan.

Methods

There are now three major systems available for carrying out bibliometric analyses (Bar-Ilan, 2008b; Jacso, 2005, 2008): Web of Science, SciVerse Scopus and Google Scholar.

The Web of Science system from Thomson-Reuters is the longest established of these, and comprises a total of five citation databases: Science Citation Index Expanded, Social Science Citation Index, Arts and Humanities Citation Index, Conference Proceedings Citation Index - Science, and Conference Proceedings Citation Index - Social Science & Humanities. In addition to the extensive, carefully curated publication and citation data, this subscription-based service provides analysis tools that enable sophisticated data mining to be carried out on search outputs, thus permitting the detailed bibliometric analyses reported below. The SciVerse Scopus system from Elsevier Inc. is similar in scope to the Web of Science, and like it is available to users only on a subscription basis. Of these two, Web of Science (hereafter WoS) was used for the work reported here, principally because of its better coverage of the conference proceedings literature. Journal articles provide the principal communication medium for most scientific disciplines but this is not the case for computer science, where conference proceedings are of greater importance (Freyne *et al*., 2010; Rahm, 2008; Sanderson, 2008). The two conference proceedings databases in WoS mean that its coverage of conferences is superior to that of Scopus; indeed, the latter's poor coverage has resulted in criticism of its use for the evaluation of computer science research (Bailes, 2011).

The most widely used citation system now is probably Google Scholar, a free service that generates citation data automatically from publications available via the Google search system. Early versions of Google Scholar were notably error-rich and laborious to use (Meho and Yang, 2007); although this is now much improved (Chen, 2010), it still lacks the data mining tools available in WoS, thus precluding its use for detailed studies of the sort reported here. Similar comments apply to data that can be harvested from the CiteSeer website, as described very recently by Fiala (2012).

In addition to its data mining tools (the Analyze Results and Create Citation Report functions in the system), WoS provides a range of filters that can be applied to search outputs, these including the specification of a broad subject area and of both the year and the country of publication. Searches were hence carried out for the strings "data* mining", "operating system*" and "web design" with the retrieved records satisfying all of the following three search criteria: the string had to occur in either the Title or the Topic fields of a record (the latter including the abstract and both author and database keywords); the record had been assigned one of the seven Computer Science subject categories in the WoS Categories field; and the record had at least one author with an address in the People's Republic of China (for which the searches included both Hong Kong and mainland China). All of the searches reported here were carried out in late 2011 and early 2012. It should be noted that a search string such as "data* mining" cannot possibly provide full recall of all of the articles in the WoS database that pertain to this topic (since they may be indexed under a multitude of words or phrases), but it suffices for a longitudinal comparative study such as this.

The WoS database was used to obtain all of the publication and citation counts reported below, and it was also the source of the h-index scores (Hirsch, 2005) , with two other databases being used to provide external views of the quality of the journals and conference proceedings in which Chinese computer science research is published. The first database was Journal Citation Reports (hereafter JCR), which is also produced by Thomson-Reuters and which contains journal impact factor (IF) data for most of the journals covered by WoS. The IF measures how frequently an "average" article from a specific journal has been cited, so that a journal's IF is calculated by dividing the total number of citations to the journal in a specific year by the number of articles in that journal published in a previous timeframe. For example, if X is the number of citations in 2010 to the Y articles published in a journal 2008-2009, then the two-year IF (which is that used here and which is available in the JCR database back to 2004) is X/Y. The application of IFs to the evaluation of research performance has been questioned, but they continue to be widely used for this purpose (Archambault and Larivière,

2009; Cameron, 2005; Garfield, 2006; Pendlebury, 2009). The second database was the CORE listing of computer science conferences, where CORE is the Computing Research and Education Association of Australasia, an association of university departments of computer science in Australia and New Zealand. Since 2006, CORE has graded the world's major conferences in computer science in order of importance. The 2010 rankings (CORE, 2010) used a three-point scale (A, B or C) to grade a total of 1501 conferences, of which 235 were graded A, 388 graded B and 878 graded C.

Results

The results of the study are presented in Tables 1-10.

Table 1 lists the total numbers of publications in each of the ten years for 2000-2009 for the seven countries (China, USA, UK, Japan, Brazil, India and Russia) where we include all types of publications, i.e., journal articles, conference papers, reviews, editorials etc. Table 2 provides the corresponding data for publications on data mining, with Table 3 listing the numbers of citations to the publications detailed in Table 2. Table 4 lists the mean number of citations per publication, not just for data mining but also for the operating system and web design searches. Table 5 lists the h-index values for publications on data mining, while Table 6 lists the percentages (rounded to the nearest integer) of journal articles (the columns headed A) and conference proceedings papers (the columns headed B) for the publications in all three subject areas.

Table 7 summarizes the IF analyses. The JCR database provides IFs since 2004, and we have chosen to include here the results for publications in 2005 and 2009 to illustrate the changes, if any, that have taken place in publication practices during this time period. Each of the three subject areas has three columns in Table 7: the number of articles in that year (column A); the percentage of those articles in journals with an IF for that year (column B); and the mean IF for those articles in journals with an IF for that year (column C). The mean IF values here are calculated as

$$\frac{\sum n_i IF_i}{\sum n_i},$$

where n_i is the number of articles in the current year in a journal i that has an impact factor IF_i for that year, and where the summation is over all of the journals for which an IF is available. Table 8 then lists the CORE grades (A, B, C or U (for ungraded)) of the conference proceeding publications in the three subject areas. Finally, Tables 9 and 10 list the numbers of publications and citations, respectively, for operating systems and web design in 2000, 2004/5 and 2009 (the 'Total' columns in each case are the total numbers summed over all of the ten years 2000-2009).

Discussion

The publication data for 200-2009 in Table 1 are summarized in Figure 1, which shows the publications in 2000, 2009, and the mean of 2004 and 2005, i.e., the mid-point for the period. From a very low starting point, it will be seen that Chinese productivity has increased dramatically, year on year, with over 14 times as many publications in 2009 as in 2000. This rate of growth is far larger than for any of the other countries in the table, with India's almost seven-fold increase from 2000 to 2009 showing the second largest rate of growth.

The rapid growth of Chinese research evident from Figure 1 has been reported for a range of subject areas, as noted in the Introduction. However, while publication figures are a measure of the quantity of research carried out they say nothing about the quality of that research. A full evaluation of Chinese research performance hence also requires consideration of quality, which is normally measured in bibliometric terms by the identities of the outlets (typically journals or conferences in the case of scientific research) that are used to publish the research, and by the impact, as measured by the numbers of citations (or some function thereof) to the publications. Moreover, the gross figures in Table 1 and Figure 1 say nothing about the research performance in specific parts of the discipline, and we have hence chosen here to study three topics in detail, i.e., data mining, operating systems, and web design.

Table 2 lists the numbers of data mining publications, where it will be seen that all of the nations here have increased their outputs in this important applications area. China has displayed the most rapid growth, with a more than 22-fold increase in productivity over the period. It overtook the USA in terms of numbers of publications in 2006, and is now the source for the largest volume of research in the field, producing more than twice as many publications as USA and more than ten times as many as do Japan and the UK.

The increasing globalization of science means that individual articles can be attributed to two or more countries, e.g., where a collaboration exists between research groups in different countries or where a scientist is affiliated with groups in more than one country. Such occurrences are exemplified by the Chinese data mining publications for 2005 in Table 2. Of these 461 publications, all of which involved at least one Chinese researcher, 23 also involved researchers from the USA, with the other multi-national publications involving Canada (12 publications), Australia (9), Singapore (5), Japan (4), the Netherlands and the United Kingdom (both 2), and Belgium, Germany, Ireland and South Korea (all 1). Whilst it would be possible to introduce some sort of fractional publication weighting scheme, we have chosen on grounds of simplicity to allocate a unit weighting to each country associate with a publication so that, e.g., the 23 publications involving USA authors in the 461 Chinese publications also provide 23 of the 475 USA data mining publications in 2005. As this example shows, the numbers of such multiple occurrences are still small (but will undoubtedly increase in the future as Chinese researchers collaborate more with researchers in other nations).

As noted above, a nation's bibliometric profile depends not only on its publications but also on the citations to those publications. These are listed in Table 3 which shows, for example, a total of 4159 citations to the American data mining publications that were published in 2000. The figures quoted are total citation counts, including self-citations: both Phelan (1999) and Aksnes (2003) have noted that inclusion of these is appropriate when bibliometric analyses are carried out at a national level, and the effect here is certainly small, e.g., the citations to the 461 Chinese publications in 2005 included only four self-citations.

A comparison of Tables 2 and 3 shows that China and the USA have broadly comparable numbers of publications for 2000-2009 but that the USA has over four times as many citations. In part this is because the USA already had a strong publication record in 2000, and thus had many publications that could attract citations throughout the ten-year period, where as China was publishing only a limited amount of material in 2000; however, this is not the entire story as is clear if one considers the mean number of citations per publication. Table 4 lists these mean numbers for the periods 2000-2004 and 2005-2009 (and also for the publications in operating systems and web design that are discussed later in the paper). It will be seen that the Chinese publications attract considerably fewer citations than do those from the three established research nations (the USA, UK and Japan) as shown in Figure 2 (and this is also true for some of the comparisons with the citation rates for the other BRIC countries). The very large numbers of recent Chinese publications are hence attracting very small numbers of citations. Other points of detail in Table 4 are: that the UK shows the smallest difference in mean citation rates between the two periods (possibly because the Research Assessment Exercise has long encouraged academics to focus on just the best journals when publishing their research); and that India has a very high mean citation rate for 2000-2004 (principally because four of the eight Indian publications in 2002 have extremely high citation counts).

A widely reported recent study by the Royal Society (2011) suggested that China would overtake the USA as the most productive scientific nation some time before 2020 (although this conclusion has since been disputed (Jacso, 2011; Leydesdorff, 2012)). A χ^2 test using Yates correction on the Chinese and USA data mining data in Table 4 shows no significant difference (χ^2=0.21, p> 0.64), i.e., there has been no change in the relative impact between 2000-2004 and 2005-2009. This is in marked contrast to the productivity data from Table 2 for the same two periods, where there is a highly significant difference (χ^2=532.47, p><0.00001), as would be expected from inspection of Figure 1.

The h-index (Hirsch, 2005) has become widely used as a simple, single-number criterion of

research impact. A researcher (or group of researchers) has index *h* if *h* of their *N* publications have each attracted at least *h* citations, and the other *N*-*h* have fewer than *h* citations. Table 5 lists the *h* values, where China would appear to be now competitive with, or superior to, all of the other countries. However, the *h*-index can be rather misleading if, as is the case here, the number of publications, *N*, varies considerably. If we consider the 2009 publications, then inspection of Tables 2 and 5 shows that ten of the 1153 Chinese publications (less than 1% of them) attracted at least ten citations, whereas the USA achieved the same result from less than half the number of publications. The 2009 comparison with the UK is still more stark, since eight of the latter's 113 publications (over 7% of them) attracted at least eight citations.

As a complement to simple numbers of citations, it has been suggested that research quality can be assessed by considering where the research is published. Table 6 details the document types for the publications in Table 2 (and also for publications in operating systems and web design that are discussed later in the paper), and shows that the overwhelming majority of them (in excess of 98%) are conference proceedings papers or journal articles. As expected for computer science topics (Freyne *et al*., 2010; Rahm, 2008; Sanderson, 2008), the former is the more popular type of outlet, with Chinese researchers having the strongest preference for conference papers of all the seven countries considered here (note that there is some degree of overlap in the figures since, for example, *Lecture Notes in Computer Science* is a serial publication that covers conferences). As noted in the Methods section, we have studied the quality of research in journals and conference proceedings using IF and CORE data, respectively; these analyses are described below.

Table 7 summarizes the IF results for data mining research (and also for operating systems and web design as discussed later in the paper) published in 2005 and 2009. Inspection of Table 7 shows that the percentage of Chinese articles in journals with IFs is competitive with all of the other nations, and that the mean IF is notably lower than for the USA and the UK but comparable to all of the other nations. Thus, from the IF perspective, China has managed to increase the quantity of its data mining research whilst at least maintaining the quality of same

for that small fraction of its output that appears in the refereed journal literature. In 2009, for example, the three most popular journals for publishing Chinese data mining research were *Expert Systems and Applications* (IF=2.91, 22 papers), *Information Sciences* (IF=3.29, 8 papers) and *IEEE Transactions on Knowledge and Data Engineering* (IF=2.29, 5 papers).

Table 8 summarizes the CORE results for the nations' 2009 publications in conference proceedings. In this table, the columns labeled A, B and C are the numbers of publications in the proceedings for conferences with that CORE grading, and the column labeled U (for ungraded) is the number of publications in the proceedings for conferences that do not appear in the CORE rankings. Since the homepage for the rankings (at http://core.edu.au/index.php/categories/conference%20rankings/1) states that "CORE has been engaged in an exercise to rank fully refereed conferences in which its members publish", it seems not unreasonable to assume that a U conference is likely to be of lower quality than one that has been graded. If this assumption is accepted then inspection of Table 8 suggests that only a small fraction of WoS Chinese conference papers on data mining (and a still smaller fraction of the corresponding Indian conference papers) are published in the best conference proceedings. The USA has the smallest fraction of publications in U conferences, with the other countries all having comparable fractions (though, as elsewhere in the tables, the numbers for Russia are very small (*vide infra*)).

The only study of which we are aware of Chinese research in data mining is that of Li and Ke (2009), who discussed publications on this topic in the *Chinese Social Science Citation Index* (CSSCI) database. CSSCI covers over 500 scholarly Chinese journals in the humanities and social science, and Li and Ke analyzed 342 CSSCI articles that contained "data mining" in the title or keyword fields and that had been published in a total of 64 Chinese journals (with two of these – the *Journal of Information* and *Statistics and Decision* - accounting for over one-third of all the articles). They reported the most productive authors and institutions, and it is of interest to compare these Chinese-language publications with the (overwhelmingly) English-language publications that form the basis for the data reported here. Rather surprisingly, none of the

seven individuals listed as being the most productive in CSSCI are amongst the ten most productive in WoS, using either journal articles or proceedings papers as the basis for comparison. There is marginally more agreement when institutions are considered: Huazhong University of Science and Technology (ranked 4th in CSSCI) is ranked 8th in WoS for conference papers; Zhejiang University (7th) is ranked 10th in WoS for conference papers; and Tsinghua University (10th) is ranked 2nd and 6th in WoS for conference papers and journal articles, respectively. It would hence appear that there are two near-distinct groups of researchers, one focusing on publication in national outlets and the other in international outlets. It may be that there is a language factor at work since the WoS rankings are dominated by Hong Kong-based institutions, which are predominantly English-speaking, with four occurring in the top ten places for conference papers and five occurring in the top ten places for journal articles; however, these institutions are notably absent from the top of the CSSCI rankings.

We have discussed the data mining results in detail, and hence present more briefly the results that were obtained in analogous sets of searches for operating systems and for web design, with Tables 9 and 10 listing numbers of publications and citations, respectively, for 2000, for the mean of 2004 and 2005, for 2009, and the total number summed over the entire ten-year period. The trends evident in Tables 9 and 10 are very similar to those seen in Tables 2 and 3, respectively. Thus, the productivity of Chinese research in both operating systems and web design has increased very substantially over the decade, so that it now exceeds that of the USA. However, the impact of the two nations' research is still very different. For example, in 2009, China and the USA generated 473 and 389 operating system publications, but these yielded 86 and 625 citations, respectively, and the comparison with the UK's operating system research is even more striking (see also the central portion of Table 4). The reader should note that the UK total of 6811 citations to operating system research in Table 10 is dominated by the 3441 citations to a 2007 article by Larkin *et al*. (2007) describing a technique for aligning multiple biological sequences, a vital tool in molecular biology database systems; and the comparable Russian total is similarly dominated by citations to two individual articles published in 2006 and 2007. The h-index values (data not shown) for the nations' research into operating

systems and web design mirror closely the citation counts in Table 10, in the same way that Table 5 mirrored Table 2 for data mining research.

The central and right-hand portions of Table 6 demonstrate the marked Chinese preference for conference, as against journal, publication that we have already noted for data mining. The impact of the journal and conference publications is summarized in the central and right-hand portions of Tables 7 and 8. Similar comments apply to these two research areas as applied to data mining with the exception of the 2009 IF data for operating systems, which demonstrates a high level of impact for Chinese research in this area.

The discussion above has focused on the results for China alone. We now consider research across the BRIC nations, where China's dominant position is clearly evident from the various tables of results. This dominance is exemplified diagrammatically by Figure 3, which shows the publication and citation counts from Tables 9 and 10 for research on web design. China is by far the most productive nation, and also attracts the most citations; however, the impact of individual papers (i.e., the mean number of citations per publication) is greater for Indian research (although this is not always the case as demonstrated by Table 4). A study by Wainer *et al*. (2009) of Brazilian publications in computer science for the period 2001-2005 found that it was most similar to Russia of the BRIC countries, and inspection of Table 1 shows that they produced comparable numbers of publications at the start of the decade. However, Russia has not subsequently increased its productivity, whereas Brazil, China and India have all increased their output volumes over the decade, China and India to a very considerable extent. Indeed, inspection of the various tables of results highlights the very low level of Russian research activity in both the three specific areas, and in computer science as a whole. This is rather surprising given the long history of computer science there (Klimenk, 1999; Prokhorov, 1999). One reason might be that Russian computer scientists do not publish in the predominantly English-language journals that form the bulk of the input to the WoS database; however, Wagner and Wong have demonstrated recently that high-quality science carried out in the BRIC countries is adequately represented in the *Science Citation Index Expanded*, the largest

component of WoS (Wagner and Wong, 2012), so it is not clear why Russia might be preferentially disadvantaged when compared to Brazil and China, the other non-English speaking BRICs nations.

Conclusions

In this paper, we have reported a bibliometric study of Chinese research in computer science for the first decade of the 21st century, focusing on the areas of data mining, operating systems and web design, and using data available in the WoS, JCR and CORE databases. The data presented here demonstrates clearly the very substantial increase in the productivity of Chinese research output in our three chosen areas of computer science. For example, by 2009, China had become the largest source of publications on data mining, producing more than the total outputs of the other six countries considered here (the USA, the UK, Japan and Brazil, India and Russia) (see Table 2). The predominance is not quite so strong for operating systems and web design research; even so, China has again become the largest producer nation for both of these subject areas. The impact, however, of many of these publications is quite low when compared with the impact of publications from established research nations such as the USA. For example, across all three subject areas, the mean number of citations per data mining publication in the period 2005-2009 was over three times larger for USA publications than it is for Chinese publications, and the differential was still greater for the other two subject areas (see Table 4) . If impact is quantified in terms of presentation at high-quality conferences, rather than mean numbers of citations per publication, then China again lags far behind the USA: for example, in data mining, only 21.6% of the 1013 Chinese conference presentations in 2009 were at conferences graded A-C in the CORE listings, where as the corresponding figure for the USA was as high as 52.3%, and the Chinese percentages for the other two subject areas were even lower (see Table 8). There is hence a marked disparity between the quantity and the quality of Chinese research: this disparity is a significant problem that needs to be addressed as a matter of some urgency if China is to contribute fully to the future development of computer science.

Acknowledgement

We thank the China Scholarship Council for funding ZX's visit to the University of Sheffield

References

Aksnes, D.W. (2003) A macro study of self-citation. *Scientometrics* 56(2): 235-0246.

Archambault É and Larivière V (2009) History of the journal impact factor: contingencies and consequences. *Scientometrics* 79(3): 635-649.

Arruda D, Bezerra F, Neris VA et al. (2009) Brazilian computer science research: Gender and regional distributions. *Scientometrics* 79(3): 651-665.

Bailes P (2011) ERA challenges for Australian university ICT activity. Available at: http://www.acs.org.au/attachments/EraIssuesforICT.pdf (accessed 2nd May 2012)

Bar-Ilan J (2008a) Informetrics at the start of the 21st century – a review. *Journal of Informetrics* 2(1): 1-52.

Bar-Ilan J (2008b) Which h-index? - A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 74(2): 257-271.

Borgman CL and Furner J (2002) Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology* 36: 3-72.

Calvert PJ and Zengzhi S (2001) Quality versus quantity: contradictions in LIS journal publishing in China. *Library Management* 22(4/5): 205-211.

Cameron BD (2005) Trends in the usage of ISI bibliometric data: uses, abuses, and implications. *Portal: Libraries and the Academy* 5(1): 105-125.

Chen X (2010) Google Scholar's dramatic coverage improvement five years after debut. *Serials Review* 36(4): 221-226.

CORE (2010) Computing conferences sorted by alphabetical title. Available at: http://core.edu.au/cms/images/downloads/conference/08sorttitleERA2010_conference_list.pdf. (accessed 2nd May 2012)

Fiala D (2012) Bibliometric analysis of CiteSeer data for countries. *Information Processing and Management* 48(2): 242-253.

Freyne J, Coyle L, Smyth B et al. (2010) A quantitative evaluation of the relative status of journal and conference publications in computer science. *Communications of the ACM* 53(11): 124-132.

Garfield E (2006) The history and meaning of the journal impact factor. *Journal of the American Medical Association* 295(1): 90-93.

Guan J and Gao X (2008) Comparison and evaluation of Chinese research performance in the field of bioinformatics. *Scientometrics* 75(2): 357-379.

Guan J and Ma N (2004). A comparative study of research performance in computer science. *Scientometrics* 61(3): 339-359.

Gupta BM, Kshitij A and Singh Y (2010). Indian computer science research output during 1999-2008: Qualitative analysis. *DESIDOC Journal of Library & Information Technology* 30(6): 39-54.

Gupta BM, Kshitij A and Verma C (2011) Mapping of Indian computer science research output, 1999-2008. *Scientometrics* 86(2): 261-283.

He Y and Guan J (2008) Contribution of Chinese publications in computer science: A case

study on LNCS. *Scientometrics* 75(3): 519-534.

Hirsch JE (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102(46): 6569-16572.

Jacso P (2005) As we may search - comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science* 89(9): 1537-1547.

Jacso P (2008) The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases. *Online Information Review* 32 (2): 266-283.

Jacso P (2011) Interpretations and misinterpretations of scientometric data in the report of the Royal Society about the scientific landscape in 2011. *Online Information Review* 35(4): 669-682.

Jiuchun Z and Baichun Z (2007) Founding of the Chinese Academy of Sciences' Institute of Computing Technology. *IEEE Annals of the History of Computing* 29(1): 16-33.

Klimenk SV (1999) Computer science in Russia: A personal view. *IEEE Annals of the History of Computing* 21(3): 16-30.

Kumar N and Asheulova N (2011) Comparative analysis of scientific output of BRIC countries. *Annals of Library and Information Studies* 58 (3): 228-236.

Kumar S and Garg KC (2005) Scientometrics of computer science research in India and China. *Scientometrics* 64 (2): 121-132.

Larkin MA, Blackshields G, Brown NP et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21), 2947-2948.

Leydesdorff L (2012) World shares of publications of the USA, EU-27, and China compared and predicted using the new interface of the Web-of-Science versus Scopus. *El Professional de la información* 21(1): 43-49.

Li C and Ke J (2009) Bibliometric analysis of data mining in the Chinese Social Science Circle. In: Second International Workshop on Knowledge Discovery and Data Mining (Moscow, Russia, 2009), pp. 231-234. Chicago: IEEE Computer Society Press.

Li F, Yi Y, Guo X et al. (2012) Performance evaluation of research universities in mainland China, Hong Kong and Taiwan: based on a two-dimensional approach. *Scientometrics* 90(2): 531-542.

Li J and Willett P (2010) Bibliometric analyses of Chinese research in cyclisation, Maldi-Tof and antibiotics. *Journal of Chemical Information and Modeling* 50(1): 22-29.

Ma R, Ni C and Qiu J (2008) Scientific research competitiveness of world universities in computer science. *Scientometrics* 76(2): 245-260.

Maier JH (1988) Thirty years of computer science developments in the People's Republic of China: 1956-1985. *IEEE Annals of the History of Computing* 10(1): 19-34.

Meho L and Yang K (2007) Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology* 58(13): 2105-2125.

Ministry of Science and Technology of the People's Republic of China (no date). National High-tech R&D Program (863 Program). Available at: http://www.most.gov.cn/eng/programmes1/200610/t20061009_36225.htm (accessed 2nd May 2012).

Nicolaisen J (2007) Citation analysis. *Annual Review of Information Science and Technology,* 41: 609-641.

Pendlebury DA (2009) The use and misuse of journal metrics and other citation measures. *Archivum Immunologiae et Therapiae Experimentalis* 57(1): 1-11.

Phelan, TJ (1999) A compendium of issues for citation analysis. *Scientometrics 45*(1): 117-136.

Prokhorov SP (1999) Computers in Russia: Science, education, and industry. *IEEE Annals of the History of Computing* 21(3): 4-15.

Rahm E (2008) Comparing the scientific impact of conference and journal publications in computer science. *Information Services & Use* 28(2): 127-128.

Sanderson M (2008) Revisiting h measured on UK LIS and IR academics. *Journal of the American Society for Information Science and Technology* 59(7): 1184-1190.

Sangam SL, Liming L and Ganjihal GA (2010) Modeling the growth of Indian and Chinese liquid crystals literature as reflected in Science Citation Index (1997–2006). *Scientometrics* 84(1): 49-52.

Sathya R (2010) BRIC countries ICT Landscape Report. Available at: http://www.my-fire.eu/documents/11433/37829/BRIC_Country_report_v2.pdf (accessed 2nd May 2012).

Simon JP (2011) The ICT Landscape in BRICS Countries: Brazil, India, China. Available at: http://ftp.jrc.es/EURdoc/JRC66110.pdf (accessed 2nd May 2012)

Royal Society (2011) *Knowledge, Networks and Nations. Global Scientific Collaboration in the 21st Century.* London: Royal Society.

Springut M, Schlaikjer S and Chen D (2011) China's program for science and technology modernization: Implications for American competitiveness. Available at: http://www.uscc.gov/researchpapers/2011/USCC_REPORT_China%27s_Program_forScience_and_Technology_Modernization.pdf (accessed 2nd May 2012)

Tang L and Shapira P (2010) Regional development and interregional collaboration in the growth of nanotechnology research in China. *Scientometrics* 86(2): 299-315.

Top500 Computer Sites (2011) Japan's K computer tops 10 Petaflop/s to stay atop TOP500 list. Available at: http://www.top500.org/lists/2011/11/press-release (accessed 2nd May 2012).

Wagner CS and Wong SK (2012) Unseen science: representation of BRICs in global science. *Scientometrics* 90(3): 1001-1013.

Wainer J, Xavier EC and Bezerra F (2009) Scientific production in computer science: a comparative study of Brazil and other countries. *Scientometrics* 81(2): 535-547.

Wilson CS (1999) Informetrics. *Annual Review of Information Science and Technology* 34: 107-247.

Yu Q, Shao H and Duan Z (2011). Research groups of oncology co-authorship network in China. *Scientometrics* 89(2): 553-567.

Zhao L and Zhang Q (2011) Mapping knowledge domains of Chinese digital library research output, 1994–2010. *Scientometrics* 89(1): 51-87.

Zhou P and Leydesdorf L (2006) The emergence of China as a leading nation in science. *Research Policy, 35*(1), 83-104.

Zhou P, Thijs B and Glanzel W (2009) Is China also becoming a giant in social sciences?

Scientometrics 79(3): 593-621.

Zhu Q and Willett P (2011) Bibliometric analyses of Chinese superconductivity research, 1986-2007. *Aslib Proceedings* 63(1): 101-119.

FIGURE AND TABLE CAPTIONS

Table 1. Numbers of publications on computer science during 2000-2009

Table 2. Numbers of publications on data mining during 2000-2009

Table 3. Numbers of citations to publications on data mining during 2000-2009

Table 4. Mean number of citations per publication on data mining, operating systems and web design.

Table 5. h-index values for publications on data mining during 2000-2009

Table 6. Percentages (rounded to the nearest integer) of journal articles (A) and conference proceedings papers (B) for publications during 2000-2009. The very small remaining percentages in some cases include document types such as editorials, letters, corrections etc.

Table 7. IF analyses for journal article publications on data mining, operating systems and web design in 2005 and 2009: Number of articles in that year (A); Percentage of those articles in journals with an IF for that year (B); Mean IF for those articles in journals with an IF for that year (C).

Table 8. CORE grading (A, B, C or U) of conference proceeding publications on data mining, operating systems and web design in 2009

Table 9. Numbers of publications on operating systems and web design during 2000-2009 (the 'Total' columns are the total numbers summed over all ten years)

Table 10. Numbers of citations to publications on operating systems and web design during 2000-2009 (the 'Total' columns are the total numbers summed over all ten years)

Figure 1. Computer science publications 2000-2009

Figure 2. Mean citations per publication for research on data mining

Figure 3. Web design publication and citation data in 2000 and 2004/5 for the four BRIC countries

Country	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
China	2828	4500	6555	7673	11770	14745	20639	24867	30407	41931	165915
USA	16742	16399	20995	22505	23123	23729	22275	23904	23435	23920	217027
UK	4023	3964	4607	5054	5295	5747	5924	6378	6662	6556	54210
Japan	3868	3836	5695	5572	5858	5481	6286	6359	6335	6714	56004
Brazil	676	766	988	1087	1243	1185	1318	1585	1850	1836	12534
India	540	584	917	971	1511	1648	2381	2774	3032	3768	18126
Russia	681	572	632	721	789	726	663	570	613	664	6631

Table 1

Country	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
China	52	71	144	229	311	461	576	670	863	1153	4530
USA	197	248	321	398	458	475	462	500	444	512	4015
UK	33	33	49	69	95	100	106	110	76	113	784
Japan	47	43	58	64	77	109	98	128	92	110	826
Brazil	14	9	13	17	38	31	33	34	43	49	281
India	6	10	8	25	36	58	53	94	97	131	518
Russia	3	4	3	11	7	12	9	8	14	7	78

Table 2

Country	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
China	158	160	450	943	1177	1188	1048	874	867	563	7428
USA	4159	3193	3385	3975	4391	4050	3659	2201	1626	700	31339
UK	171	408	393	376	659	747	625	477	434	371	4661
Japan	253	163	343	336	501	326	270	232	214	100	2738
Brazil	21	67	252	113	165	66	64	41	29	32	850
India	65	96	448	141	222	147	136	171	92	72	1590
Russia	7	0	9	29	26	41	28	6	36	20	202

Table 3

Country	Data mining		Operating systems		Web design	
	2000-2004	2005-2009	2000-2004	2005-2009	2000-2004	2005-2009
China	3.58	1.39	1.91	0.53	2.91	0.81
USA	11.78	5.11	6.83	3.16	10.51	4.64
UK	7.19	5.26	5.31	13.65	8.34	3.72
Japan	5.52	2.13	2.38	0.88	1.60	0.78
Brazil	6.79	1.22	3.86	1.23	1.72	1.53
India	11.44	1.43	1.78	1.47	4.07	1.34
Russia	2.54	2.62	1.31	12.76	0.85	4.87

Table 4

Country	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
China	6	9	11	16	17	19	16	13	13	10
USA	32	29	28	32	32	31	25	19	18	10
UK	4	12	8	10	16	13	14	11	11	8
Japan	7	7	8	8	10	9	8	7	6	5
Brazil	1	2	4	5	7	5	4	4	3	3
India	3	2	4	6	9	6	6	7	7	5
Russia	1	0	2	3	3	3	2	2	3	2

Table 5

Country	Data mining		Operating systems		Web design	
	A	B	A	B	A	B
China	20	80	12	87	19	81
USA	43	55	37	62	43	56
UK	43	54	42	58	47	51
Japan	32	67	26	74	23	77
Brazil	27	73	28	72	33	67
India	27	72	28	72	30	69
Russia	37	61	55	45	44	56

Table 6

Country	Data mining						Operating systems						Web design					
	2005			2009			2005			2009			2005			2009		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
China	49	80	1.73	132	78	1.99	14	71	0.98	28	89	1.81	25	48	1.62	54	67	1.50
US	150	67	2.20	189	79	2.40	102	65	1.61	114	75	1.59	128	62	1.75	145	70	1.93
UK	22	73	3.31	46	85	2.49	20	55	1.94	38	82	1.79	55	65	3.33	54	57	1.95
Japan	14	71	1.68	28	89	1.98	11	73	1.17	13	100	1.87	10	80	0.88	12	83	3.08
Brazil	5	100	1.69	8	75	1.64	2	50	0.36	4	75	1.41	8	75	1.82	8	63	0.76
India	13	60	2.25	17	71	2.29	9	67	0.97	19	68	1.66	7	57	3.20	5	60	1.09
Russia	1	0	0.0	1	100	2.91	4	100	1.27	1	100	1.96	3	100	0.69	1	100	4.93

Table 7

Country	Data mining				Operating systems				Web design			
	A	B	C	U	A	B	C	U	A	B	C	U
China	46	16	157	794	0	21	31	358	9	55	45	475
US	81	8	71	146	37	9	17	170	39	24	15	179
UK	4	6	9	42	5	4	3	41	7	8	11	58
Japan	13	8	11	48	1	4	4	50	3	8	12	24
Brazil	0	4	12	25	0	2	1	16	0	1	1	10
India	4	1	2	104	1	0	2	43	3	1	3	32
Russia	0	1	0	3	0	0	0	3	0	0	1	1

Table 8

Country	Operating systems				Web design			
	2000	2004/5	2009	Total	2000	2004/5	2009	Total
China	17	120	473	1689	22	217	642	2411
USA	307	367	389	3917	251	388	421	3861
UK	61	53	99	724	50	113	147	1127
Japan	40	64	75	642	19	54	59	559
Brazil	8	16	23	170	9	20	20	171
India	7	31	70	306	5	15	44	194
Russia	3	10	4	66	1	4	3	28

Table 9

Country	Operating systems				Web design			
	2000	2004/5	2009	Total	2000	2004/5	2009	Total
China	87	128	86	1625	114	430	169	3228
USA	3118	2131	625	26008	4372	2717	710	34010
UK	507	255	425	6811	333	533	281	6185
Japan	190	115	17	1163	26	39	37	786
Brazil	178	21	13	373	17	55	13	270
India	24	95	56	722	5	91	8	619
Russia	0	15	19	506	0	17	16	84

Table 10

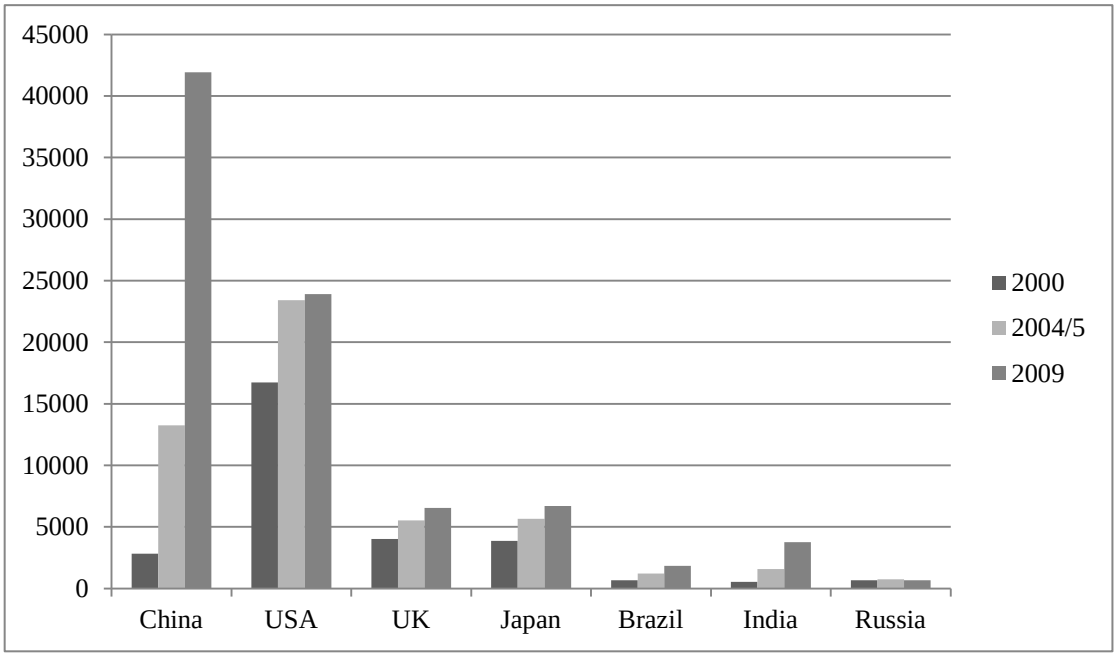

Figure 1

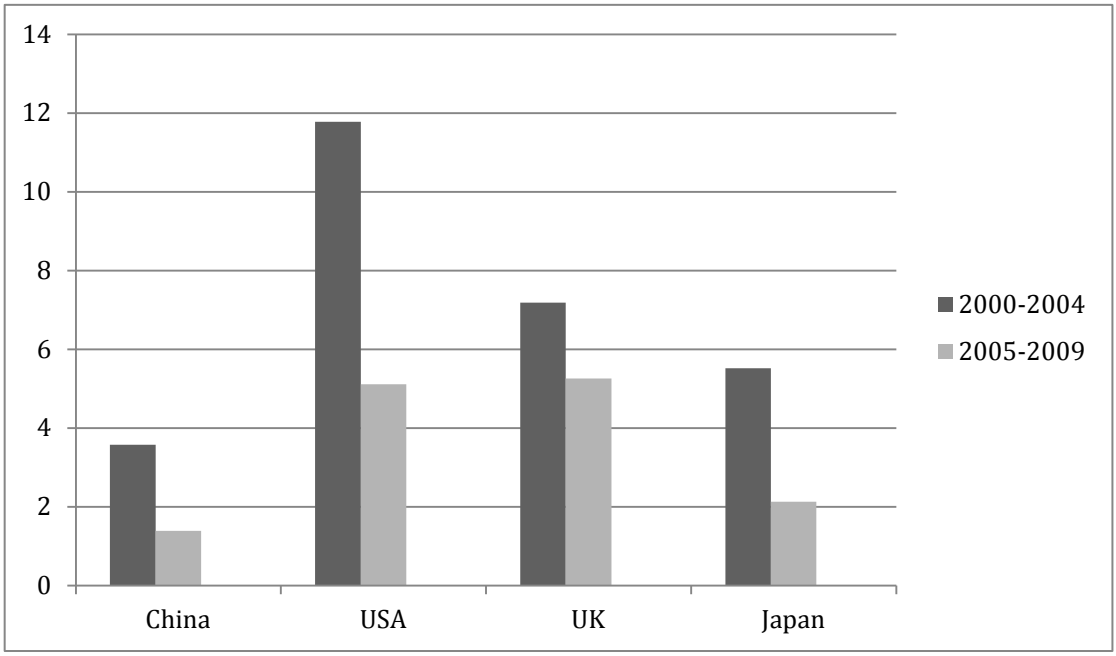

Figure 2

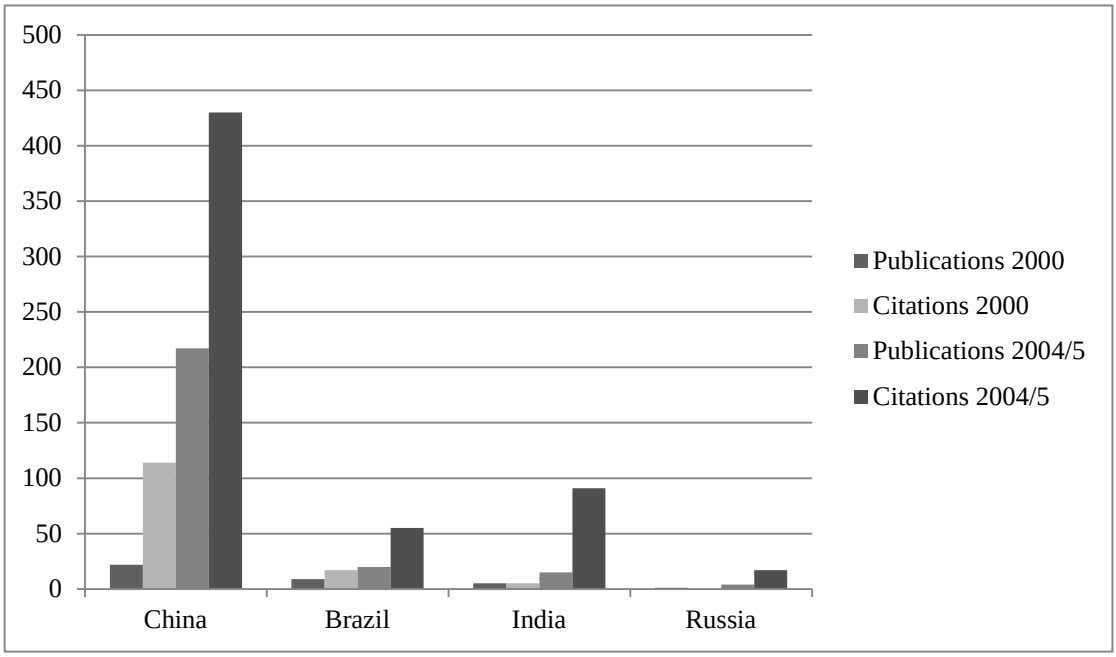

Figure 3