Using Lorenz Curve and Gini Coefficient to Reflect the Inequality Degree of S&T Publications: An Examination of the Institutional Distribution of Publications in China and other Countries

Ma Zheng[1]   Yuan Junpeng   Su Cheng   Hu Zhiyu   Yu Zhenglu   Pan Yuntao   Wu Yishan

03 June 2008

Abstract

As is known to us, Lorenz curve and Gini Coefficient are classic indicators in the field of economics. They have been used to analyze income inequality for about one hundred years since their were designed. Economists or sociologists generally draw a Lorenz curve and calculate the Gini coefficient based on incomes data of a group, a city or a country. The value of Gini coefficient (from 0 to 1) reveals the degree of income inequality (from complete inequality to complete equality). There is tremendous amount of research on the relationship between the degree of income inequality on one hand, and social development and economical growth on the other hand. Later Lorenz curve and Gini Coefficient have been used, beyond economics field, in general quantitative analysis and research. This paper tries to apply these two concepts to explore the institutional distribution of publications. The inequality degree of institutional S&T output will be measured with the Lorenz curve and Gini coefficient, using publications as a proper proxy for S & T output. To compare the data among different countries and analyze the time series of data on each country, recent 10 years of SCIE data is collected. China and other 10 countries (including USA, Russia, Japan, France, UK, Germany, Korea, India, Brazil, and Finland) are selected as samples in this research. These countries are either innovative developed countries or fast-growing developing countries. In addition, we make use of the data from CSTPCD(Chinese S&T Papers and Citations Database),which is produced by Institute of Scientific and Technical Information of China(ISTIC) and covers more than 1,700 core S&T journals published in China. We also discussed, after relevant comparison and analysis, whether the value of Gini coefficient here could be defined as an indicator to judge the S&T development stage of a country. In economics, certain Gini coefficient is used as a warning signal that social inequality seems too sharp that social disruptions are close. In the mean way, we ask whether it is possible to determine certain key value of Gini coefficient here, and use this value to detect or describe the potential characteristics of a country's S&T policy.

1   Introduction

The Lorenz curve is a graphical representation of the proportionality of a distribution among a set of sources (Lorenz, 1905). These sources can be persons (as in the original use of the Lorenz curve), actors (a terminology often used in social network analysis), performers, authors, articles, and so on (Egghe, 2005). Economists or sociologists generally draw a Lorenz curve and calculate the Gini coefficient based on incomes data of a group, a city or a country. The value of Gini coefficient (from 0 to 1) reveals the degree of income inequality (from complete inequality to complete equality). There is tremendous amount of research on the relationship between the degree of income inequality on one hand, and

---
[1]Institiute of Scientific and Technical Information of China (ISTIC) Beijing P.R.China,
mazheng@istic.ac.cn

social development and economical growth on the other hand. Later Lorenz curve and Gini Coefficient have been used, beyond economics field, in general quantitative analysis and research. Examples include income distributions (Lambert, 2001; Kleiber & Kotz, 2003), poverty study (Jenkins & Lambert, 1997; Zheng, 2000), plant size inequality (Weiner, 1985), evenness studies in ecology (Nijssen et al., 1998), vegetation studies based on satellite images (Bogaert et al., 2002), hierarchies (Egghe, 2002), and research evaluation (Rousseau, 1998; Egghe& Rousseau, 2007).

This paper tries to apply the Lorenz curve and the Gini coefficient to explore the institutional distribution of publications. The inequality degree of institutional S&T output will be measured with the Lorenz curve and Gini coefficient, using publications as a proper proxy for S & T output. To compare the data among different countries and analyze the time series of data on each country, recent 10 years of SCIE data is collected. China and other 10 countries (including USA, Russia, Japan, France, UK, Germany, Korea, India, Brazil, and Finland) are selected as samples in this research. These countries are either innovative developed countries or fast-growing developing countries. In addition, we make use of the data from CSTPCD(Chinese S&T Papers and Citations Database),which is produced by Institute of Scientific and Technical Information of China(ISTIC) and covers more than 1,700 core S&T journals published in China. We also discussed, after relevant comparison and analysis, whether there are fuzzy relationships between the Inequality degree of S&T output and the phase or type of S&T improvement in a given country or not. If there is observed regularity according to the value of Gini coefficient and S&T development style of different countries, then the value of Gini coefficient here could be defined as an indicator to judge the S&T development stage of a country. In economics, certain Gini coefficient is used as a warning signal that social inequality seems too sharp that social disruptions are close. In the mean way, we ask whether it is possible to determine certain key value of Gini coefficient here, and use this value to detect or describe the potential characteristics of a country's S&T policy.

This paper is structured as follows. In Section 2 and 3, Method and data are introduced to describe the data resource and how to calculate the Gini coefficient use these publications. In Section 4, it consists two parts. One is 11 Countries Gini Coefficient calculated with top 500 (SCI data). Another is China's Gini Coefficient calculated with CSTPCD.

2   Method

This paper tries to apply the Lorenz curve and the Gini coefficient to explore the institutional distribution of publications.

The Lorenz curve is a graphical representation of the proportionality of a distribution (the cumulative percentage of the values). To build the Lorenz curve, all the elements of a distribution must be ordered from the most important to the least important. Then, each element is plotted according to their cumulative percentage of X(number of institutes) and Y(number of publications), X being the cumulative percentage of elements and Y being their cumulative importance. For instance, out of a distribution of 10 elements (N), the first element would represent 10% of X and whatever percentage of Y it represents (this percentage must be the highest in the distribution). The second element would cumulatively represent 20% of X (its 10% plus the 10% of the first element) and its percentage of Y plus the percentage of Y of the first element.
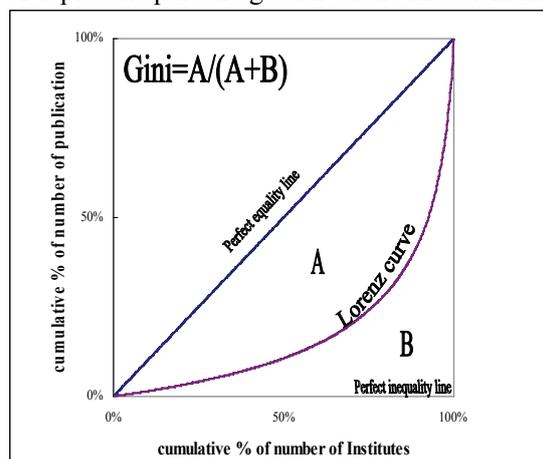

Figure 1: The Lorenz curve and Gini coefficient

The Gini coefficient as is called today was, according to Dalton (1920), named after the fact that "a remarkable relation has been established

between this measure of inequality and the relative mean deference, the former measure being always equal to half the latter." This remarkable relation was first given by Gini in 1912. Dalton (1920) therefore called this mean deference as "Professor Gini's mean deference."

The Gini coefficient can be, as in Figure 1, defined geometrically as the ratio of two geometrical areas in the unit box: (a) the area between the line of perfect equality (45 degree line in the unit box) and the Lorenz curve, which is called Area A and (b) the area under the 45 degree line, or Areas A + B. Because Areas A + B represents the half of the unit box, that is, A+B = 1/2 , the Gini Coefficient, G, can be written as

$$G = \frac{A}{A+B} = 2A = 1 - 2B \quad (1)$$

From our search, we can compute X_i's and Y_i's and then the area below the Lorenz curve

$$B = \frac{1}{2} \sum_{i=0}^{n-1} (X_{i+1} - X_i)(Y_{i+1} + Y_i) \quad (2)$$

Substituting equation (2) into equation (1) yields the Gini Coefficient G（Yao, 1999；）：

$$G = 1 - \sum_{i=0}^{n-1} (X_{i+1} - X_i)(Y_{i+1} + Y_i) \quad (3)$$

Several alternative formulations in fact follow the same tradition, for example, Rao(1969) showed that the Gini Coefficient can be defined as

$$G = \sum_{i=1}^{n-1} (X_i Y_{i+1} - X_{i+1} Y_i) \quad (4)$$

Therefore a Gini coefficient is a number between zero and one that measures the degree of inequality in the distribution of income for a given area. The coefficient would register zero (0.0 = minimum inequality) for an area in which each member received exactly the same output and it would register a coefficient of one (1.0 =maximum inequality) if one member got all the output and the rest got nothing.

We use equation (3) to calculate the Gini coefficient. In the process of comparing ten major countries, we use the TOP 500 institutes to calculate the Gini coefficient, because several characteristics of classical Lorenz curves make them unsuitable for the study of a group of top-sources. For example, Lorenz curves do not reflect intensity or incidence of the top. They are, moreover, invariant under scale transformations （Egghe& Rousseau, 2007）.

3   Data

The data which to compare different countries were provided by Thomson ISI, which indexes more than 8,000 journals in 36 languages, representing most significant material in science and engineering. The Web of Science provides seamless access to current and retrospective multidisciplinary information from approximately 8,700 of the most prestigious, high impact research journals in the world. Web of Science also provides a unique search method, cited reference searching. With it, users can navigate forward, backward, and through the literature, searching all disciplines and time spans to uncover all the information relevant to their research.

The analysis focuses on the ten major countries (the USA, Russia, Japan, France, UK, Germany, South Korea, India, Brazil, Finland and China). We also included South Korea because this comparison may teach us something about the differences in the dynamics between Asian versus other OECD countries. (Korea has been a member of the OECD since 1996.)

On May 25, 2008, we searched the Web of Science by the title of countries (USA, Russia, Japan, France, UK, Germany, South Korea, India, Brazil, Finland and China) from 1995 to 2007, and then we downloaded each save which limited to 500.

The Web-of-Science installation of the Science Citation Index allows for the measurements including the most recent year (2004), but there are some limitations on the retrieval. The system does not provide an exact number when the recall is larger than 100,000, and the download for each save is limited to 500. In order to solve the first problem, when we search USA's data, we take a sample of the USA's data to less than 100,000.

In addition, we make use of the data from 1991 to 2006 in CSTPCD(Chinese S&T Papers and Citations Database) to calculate the Gini coefficient of China, which is produced by Institute of Scientific and Technical Information of China(ISTIC) and covers more than 1,700 core S&T journals published in China.

4   Results

4.1   The 11 Countries Gini Coefficient calculated with top 500

Before to construct the series Lorenz Curves, the number of publications and number of institute of each country by years are calculated firstly (Figure 2). There are more than 300,000 papers from USA issued by SCIE annually, which is far from that in other countries. To show the distribution of the number of institute against the number of publications from other countries clearly, there are data points from 10 other countries except USA in Figure 2.

From Figure 2, it can be seen that the traditional strong countries in science research, Germany, Japan, UK, France, etc. for example, publish more papers and the number of publications increase fast by the term of year. At the same time, the number of institutes which publication papers covered by SCIE increase fast.

Other countries like India, Brazil, Finland and Korea have less number of publication, but the rate of increasing scale of number of institutes to that of the number of publication is like such strong countries, that can be found from Figure 2, all the 8 countries' data points move along a group similarly parallel lines.

Different to above countries, China's number of publication increase most (from about 20,000 in 1995 to about 100,000 in 2007). However, there is not so big increasing scale in the number of institute.

Figure 2 presents that the number of institute is relate to the number of publications, in other words, more institutes product more publication, but as we know, few top institutes in a country generally share a heave percentage of total publications, so there is difference degree of inequality amount different countries. This paper tries to use Lorenz Curve and Gini Coefficient to reflect this inequality degree of S&T publications.

To construct the series Lorenz Curves and calculate Gini Coefficient of top 500 institutes, the number of publication of each institute is statistics per year. And then for different year, to rank the top 500 institutes, the Lorenz curve is the plot of the cumulative percentage of publications against the cumulative percentage of institutes.
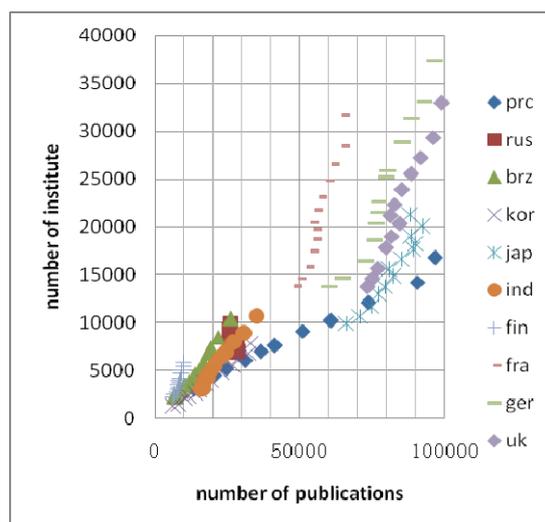

Figure 2: Distribution of the number of institute against the number of publications of 10 countries from 1995 to 2007.

The Gini Coefficient is defined as twice the area between the Lorenz curve and the diagonal line, or equivalently as the ratio of the aforementioned area to area of triangle below the diagonal line. Clearly this index is between zero and one, with larger values indicating greater concentration while a smaller on indicates greater uniformity.

Figure 3 presents the 11 countries' Gini Coefficient calculated with top 500 institutes from 1995 to 2007. It can be found from this figure that the values of USA、UK、France and Germany's Gini Coefficient are less than 0.6 in 2007, and such countries are traditional strong countries in science research, so we can say that in such countries, the degree of inequality amount different institutes are low. In other words, the entire S&T output level is stronger in such countries.

At the same time, the most countries' values keep decline trend in such 13 years. It means there is a generally trend from inequality to equality in the publications of institute, which reflex the degree of S&T output. However, USA and Japan's value is stable in this time span. As two most important strong countries in science research, their distribution of S&T institutes has established a balance state.
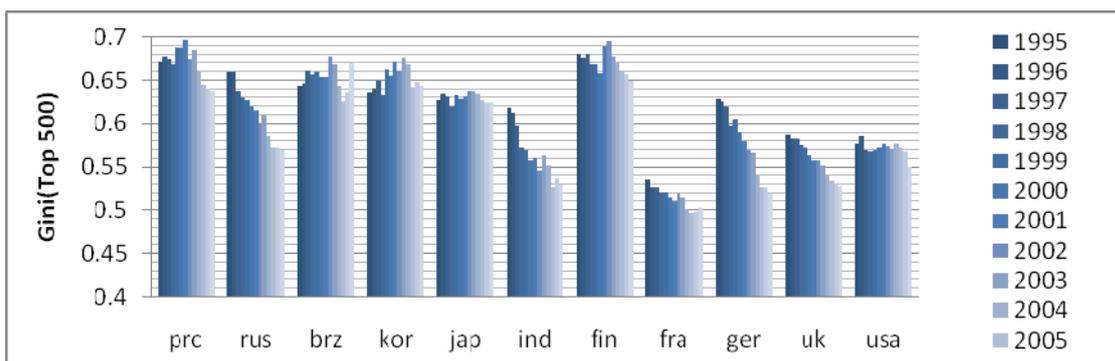
Figure 3: The Gini Coefficient calculated with top 500 institutes in 11 countries from 1995 to 2007

That is more clearly found from Figure 4, which presents the 11 countries' Average Gini Coefficient calculated with top 500 institutes from 1995 to 2007 per 5 years time span.

The Gini Coefficient of Japan is relatively stable. The 11 countries are divided into 3 groups based on the level of Japan.

Japan government still pays more attention on the development of unique, outstanding S&T. In 2006, Japan also sets the goal of "becoming an advanced science-and technology-oriented nation" as a national strategy in its science and technology basic plan. The 'Global Innovation Scoreboard' (GIS) Report compares the innovation performance of the EU25 to that of the other major R&D performing countries in the world. Japan is in the group of best performers in GIS 2006 report.

Finland is the global innovation leader in GIS 2006 report. Long term investment in science and technology is the key factor to Finland's success. It makes a 'science technology innovation' report to point out the development strategy. Republic of Korea performs better than the average performance of the EU25, and in the group of next-best performers in the GIS 2006 report. Brazil government announced 'innovation law' in the 2004 which encourages the research connection of universities, institutes and companies.

China's performance is quite different on each of the innovation dimensions in GIS2006. It is in the best performing countries for application. The SCI paper of China is growing rapidly compared to other countries. It can be seen from the graph that the curve of China is on the top over the period of 1995 to 2003.

The curves of France, the UK, Germany and USA are under the curve of Japan. All these countries are developed countries. They have relatively higher R&D and SCI papers every year. In the GIS 2006 report, France, the UK and Germany are in the next-best performers group. Their Gini indices are declining steadily and the rate of decrease is noticeable. USA is still the NO.1 in the rank of SCI papers. As we can see from the graph, the Gini Coefficient of USA is almost stayed the same.

In Indian science and technology policy 2003, it says that it is important for India to put all her acts together to become a continuous innovator and creator of science and technology intensive products. In 2006 India signed a 'Global Innovation & Technology Alliance' agreement. The curve of India is very similar to the UK and both of them almost overlap.
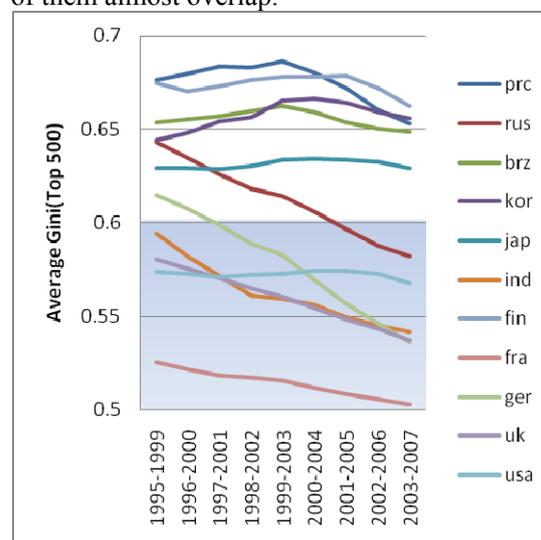

Figure 4: The Average Gini Coefficient calculated with top 500 institutes (1995-2007) for 5 years
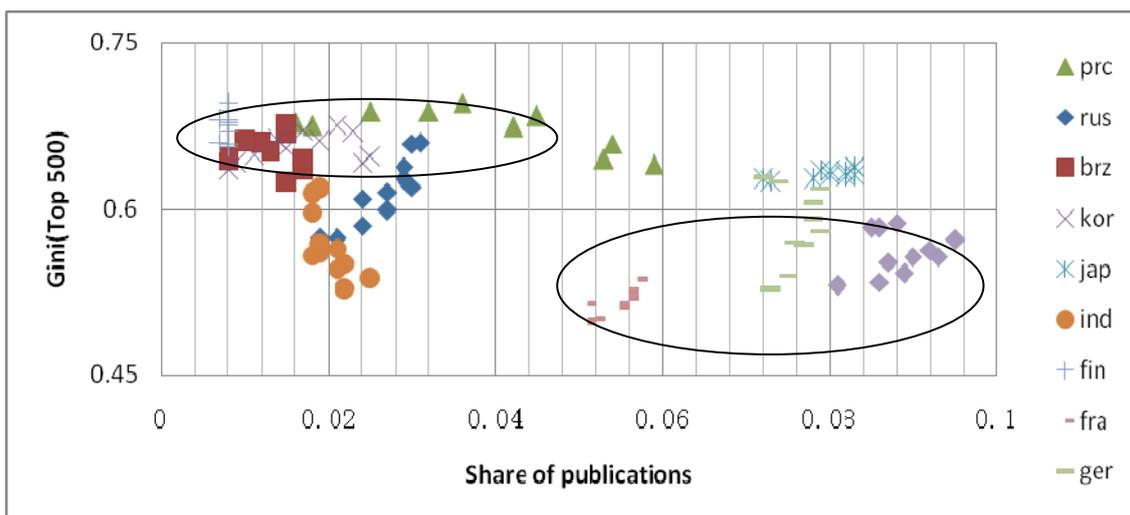
Figure 5: Distribution of 10 countries' Gini Coefficient calculated with top 500 institutes against to their share of all publications from 1995 to 2006

Figure 5 presents the distribution map of 10 countries' Gini Coefficient calculated with top 500 institutes against to their share of all publications from 1995 to 2006. As the same reason, USA share too many publications to show in one figure together with other countries, so there are only 10 countries' data points in this map. This paper try to define the value of "0.6" as a fuzzy key value to mark different group of countries with their inequality level of S&T output.

There is one observed cluster in the distribution map. It includes the data points from Finland, Brazil, Korea, and China, which are innovation countries or fast-growing developing countries. Most Gini Coefficient of such data points in this cluster is over 0.6 and which share of all publications is less. Japan as a innovative developed country, its Gini Coefficient is more than 0.6 too. There is another cluster include UK, Germany and France, the Gini Coefficient values of these traditional strong countries in S&T are less than 0.6, and even USA's data points never be shown in this figure, its Gini Coefficient values are also less than 0.6.

However, the data points of India and Russia seems not to comply by this rule, so the key value 0.6 is not sharp but fuzzy.

4.2 China's Gini Coefficient calculated with CSTPCD

Calculate the annual Gini Coefficient (1991-2006) of Chinese institutes which publish papers on Chinese journals. Sample interval, 1994, 1998, 2002, 2006, and get figure 6. Gini Coefficient keep increase year by year, but the different of internal imbalance between different years' data is small.
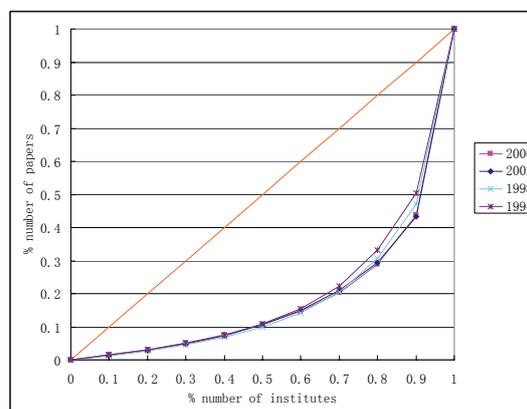

Figure 6. The calculate of Gini Coefficient about Chinese institutes which publish papers on Chinese journals and the changes of the index.

The Gini Coefficient of every year are showed in table 1.

Table 1: Gini Coefficient of Chinese institute which publish papers on Chinese journals. (G2 are the Gini Coefficient of all institutes. On the contrary, G1 are them of top 5% institutes.).

Year	G1	G2
1991	0.585	0.750
1992	0.581	0.763
1993	0.596	0.769
1994	0.620	0.770
1995	0.605	0.799
1996	0.598	0.810
1997	0.604	0.821
1998	0.643	0.825
1999	0.696	0.813
2000	0.701	0.820
2001	0.735	0.824
2002	0.652	0.888
2003	0.614	0.899
2004	0.685	0.898
2005	0.660	0.902
2006	0.652	0.902

Number stream G1 is very fluctuant and can be roughly divided into four parts. Part A contains data from 1991 to 1997. The data almost remains constant around 0.600.Part B contains data from 1997 to 2001. Data of this part keeps increase annually, from 0.604 of 1997 to 0.735 of 2001. Part C contains data from 2001 to 2004, Data of this par seems like a "V" character. It reduces from 0.735 of 2001 to 0.614 (approximating data in part A) of 2003. After that, the number come s back to 0.685 of 2004. Part D contains data from 2004 to 2006. Data of this part keeps reduce slightly, from 0.685 of 2004 to 0.652 of 2006.

Number stream G2 is large than curve G1 in every year. The data of G2 keeps slightly increase with few fluctuation from 0.750 to 0.902.

To present the trend clearly, 5 years average numbers of G1 and G2 are calculated and showed in figure 7.
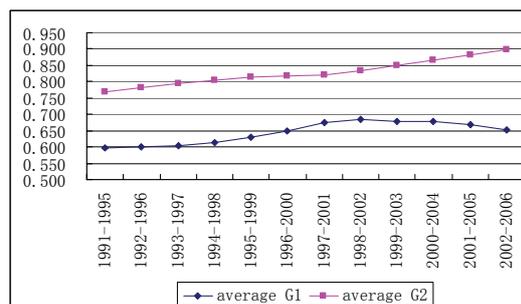

Figure 7. 5 years average numbers of G1 and G2.

Data of Gini Coefficient is smoothed by average method. And there is a dramatic inflexion in curve average G1. Data keeps increase before the inflexion of "1998-2002". On the contrary, data keeps reduce after the point.

The inflexion of curve G1 in figure 7 is corresponding of the period from 1998 to 2002. During this time, there were many remarkable policy changes in China. These changes maybe interrelated of the inflexion in this research.

Higher education reform in China was begun from 1998.Many mergers between colleges happened. The reform brings not only the changes of colleges number, but also the increase of concentrations of research capability. That is partly because of the most mergers are happened between first-class colleges, such as Peking University and Beijing Medical University.

Number1, the numbers of colleges which going to merger, can be find in the website of the Ministry of Education of the People's Republic of China. And 5 years average numbers of Number1 are calculated .After the curve of average Number1 is combined with figure 7, figure 8 is showed below.
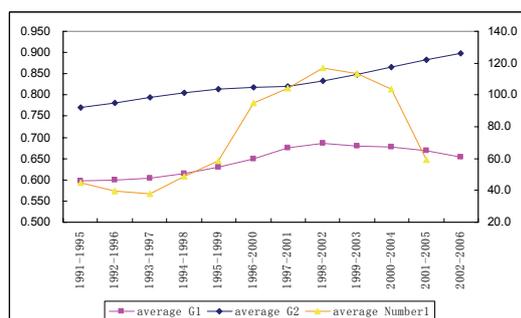

Figure 8. Contrast between average Number1 and Figure 7.

There are clear privities between curve average G1 and average Number1. And these two curve have a same inflexion in 1998-2002.

On the other hand, reforms of scientific research institutes in China, CAS (China Academy of Science) for example, also happened after 1998. The representative event is Knowledge Innovation Project in CAS. A result of this project is some institutes merger into academy, so the research capability is concentrated.

So there is a conjecture, after 1998, The major contribution of the constant growth data in curve average G2 is did by the "higher" institutes as papers' writers, but not by the "highest".

This "higher" institutes locate in "rich" positions when 100% Gini Coefficient are calculated (G2) and in "poor" positions when top 5% Gini Coefficient are calculated (G1). So the increase of these institutes can lead the both results, the growth of average G2 and the reduce of average G1.

5   Discussion

To compare the 11 countries' Gini Coefficient calculated with top 500 institutes from 1995 to 2007，Finland, Brazil, Korea, and China, which are innovation countries or fast-growing developing countries，have similar character in series Gini Coefficient. At the same time, USA, UK, Germany and France, which are developed countries, keep a low degree of inequality in S&T output.

This paper tries to define the value of "0.6" as a fuzzy key value to mark different group of countries with their inequality level of S&T output, but it need to be confirm with more data from more countries in following works.

The annual Gini Coefficient (1991-2006) of Chinese institutes which publish papers on Chinese journals keeps increase year by year. On the contrary, this index of top 5% institutes has a inflexion around 1998-2002. Data keeps increase before this point and reduce after it

The inflexion of the 5 years' average numbers of colleges which going to merger is the same with the inflexion of the 5 years' average annual Gini Coefficient of top 5% institutes. It can be conjectured that the decrease of the imbalance of top 5% institutes' research capability is relevant with the high education reform and the scientific research institutes reform.

Acknowledgement

References

Lee W.C.(1996). Analysisi of Seasonal Data Using the Lorenz Curve and the Associated Gini Coefficient. *International Journal of Epidemiology* 25(2):426-434

Barry C. Arnold. (2005).The Lorenz Curve: Evergreen after 100 years. http://www.unisti.is /eventi /ginilorenz05 25%20may%20paper /paper_arnold.pdf

Bogaert, J., Zhou, L., Tucker, C.J. , Myneni, R.B., & Ceulemans, R. (2002). Evidence for a persistent and extensive greening trend in Eurasia inferred from satellite vegetation index data. *Journal of Geophysical Research* 107 (ACL 4-1):4-14.

Egghe, L. (2005). Power Laws in the Information Production Process. *Lotkaian Informetrics*. Amsterdam: Elsevier.

Lorenz, M.O. (1905). Methods of measuring concentration of wealth. *Publications of the American Statistical Association* 9: 209-219.

Lambert. P.J. (2001). The distribution and redistribution of income (3rd edition).Manchester (UK): Manchester University Press.

Kleiber, C. & Kotz, S. (2003). Statistical size distributions in economics and actuarial sciences. Hoboken (NJ): Wiley.

Dalton, H. (1920). The measurement of the inequality of incomes. *Economic Journal* 30:348-361.

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

Gini, Corrado (1921). Measurement of Inequality of Incomes. *The Economic Journal* 31: 124–126.

Yao, Shujie (1999). On the Decomposition of Gini Coefficients by Population CLass and Income Source: A Spreadsheet Approach and Application. *Applied Economics* 31: 1249–1264.

Rao, V. M. (1969). Two Decompositions of Concentration Ratio. *Journal of the Royal Statistical Society* Series A, 132: 418–425.

Jenkins, S.P., & Lambert, P.J. (1997). Three 'I's of poverty curves, with an analysis of UK poverty trends. *Oxford economic Papers*, 49:317-327.

Weiner, J. (1985). Size hierarchies in experimental populations of annual plants. *Ecology* 66:743-752.

Zheng, B. (2000). Poverty orderings. *Journal of Economic Surveys* 14(4), 427-466

Nijssen D., Rousseau, R., & Van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses* 13(1): 33-38.

Rousseau, R. (1998). Evenness as a descriptive parameter for department or faculty evaluation studies. In E. de Smet (ed.), Informatiewetenschap 1998:135-145, Antwerp, Werkgemeenschap Informatiewetenschap.

Egghe, L. (2002). Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorenz curve. *Mathematical and Computer Modelling* 36: 587-602.

Egghe, L., R. Rousseau, et al. (2007). TOP-curves. *Journal of the American Society for Information Science and Technology* 58(6): 777-785.

Yao, QJ, (2004). Some review on high. education reform. *Journal of Beihua University (Social Sciences)* 5 (2), 2–5.

MOE. (2008). Education Development Columns, *http://www.moe.edu.cn/*