Science Without Borders: A Descriptive Mobility Study

Igor de Almeida Malheiros Barbosa[1], **João Martins de Oliveira Neto**[1], **Alexandre Nóbrega Duarte**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brazil

{igormalheiros92,martins.j.neto}@gmail.com, alexandre@ci.ufpb.br

Abstract. *The international student exchange market has grown significantly over the last decade. In an effort to improve the country's science and technology production, the Brazilian government created the Ciência sem Fronteiras program. The program provides funding for undergraduate and postgraduate students to study abroad. This paper analyzes data from this program, identifying patterns on distributions of students, their states in Brazil, destinations and areas of study. The results show that the long tail distribution appears ordinarily in the data. Furthermore, between pairs of data, the results showed that Quetelet's index was low in most cases, except in a few outliers, such as countries that have hosted a very small number of students.*

1. Introduction

One consequence of globalization is the rise of the internationalization of higher education. Universities and research institutions are sending and receiving larger numbers of students to and from different parts of the world. Traditional exchange of students in international higher education is usually financed by aid programs or inter-university partnerships for research [Bashir 2007]. In the worldwide domain, the number of mobile tertiary education students has reached more than 2.7 million in 2005. Almost 90% of exchange students have enrolled in institutions in countries belonging to the Organisation for Economic Co-operation and Development (OECD) with the main destinations (the US, the UK, Germany, France and Australia) recruiting over 70% of them [Verbik and Lasanowski 2007].

Student mobility constitutes a major of international activities in Europe [Murphy-Lejeune 2003]. Some countries tried to facilitate this activities through bilateral and unilateral agreements. The international student mobility within Europe had a massive growth after the Erasmus program, financed by European Union [González et al. 2011]. On the countries that receive students side, internationalization of education brings an extra money source and causes a positive impact on local economy [Tremblay 2002]. In Australia, the international higher education represents the fourth largest industry, widening opportunities for institutions and business to operate in worldwide scenario [Banks et al. 2007].

When considering a country to study in, many factors are taken into consideration by students such as the cost of education and the possibility of staying in the country after graduating. Identifying patterns and trends on what influences foreign students to choose one university over another is important. This information can be used to make universities more attractive[Choudaha and Chang 2012]. In order to become more competitive in

international student recruitment, many countries have also started making efforts to become more attractive to foreign students. Some of them created special visa policies for students. Australia, for instance, lets students stay in the country for 18 months after graduating. Institutions are also trying to improve the student living experience to meet the expectations of their current students and attract new ones [Verbik and Lasanowski 2007].

In 2011, the Brazilian government created a program to provide the opportunity of studying abroad to undergraduate and graduate students. The project, called *Ciência sem Fronteiras (Science without Borders)*, aims at promoting the consolidation, expansion and internalization of science and technology in Brazil through international student mobility. The program has financed over 70000 students on more than 2000 universities around the world.

This paper investigates trends and patterns of Brazilian students mobility financed by *Ciência sem Fronteiras* from 2011 to 2014. Data from all the students that joined the program in this period was collected. Relations between students that were financed by the program were also investigated, in terms of their states in Brazil, their chosen destination countries and universities for studying abroad and their area of study, for both undergraduate and graduate students. Another measure analyzed is the representativity of students from states in Brazil and their areas of study and chosen countries. All data used is available on *Ciência sem Fronteiras*'s official website. The purpose of this study is to analyze whether the place students choose to go is influenced by their area of study or the place they currently live, looking for patterns and trends on student mobility.

The remainder of this paper is organized as follows. Section 2 describes related works on international student mobility. Section 3 discusses the methodology used in this work. Section 4 shows the results found followed by a discussion. Section 5 provides final conclusions and directions for future work.

2. Related Work

The patterns and trends of exchange programs on higher education are focused by many researches that look for new ways to explain why some places are more attractive than others.

Verbik's work [Verbik and Lasanowski 2007] provides an extensive overview of patterns in international students exchange. It analyses policies taken by the most popular countries in terms of student preference and also by the destinations that have shown rapid growth in terms of international student numbers. The paper also examines strategies taken by some countries and institutions to become more competitive and attract a higher number of foreign students. It identifies that immigration procedures, overall student experience and cost of living are key elements that are likely to influence exchange students destinations and government policies over the next few years.

Choudaha's report [Choudaha and Chang 2012] analyzes the growth of international student mobility numbers in the past years. It also investigates policies that have been adopted by the 4 major student destinations and the impact they might have on these numbers. Choudaha then examines the numbers inside the US, making an overview of major source countries and of the distribution of foreign students among the states, identifying rising destinations. It is also identified common recruitment practices that are becoming popular among institutions.

Bashir's work [Bashir 2007] also assesses trends in international higher education services and their value. It then investigates the factors that lead to the increasing number of international students, such as the possibility of entering the global market when having an internationally recognized qualification. It then analyzes some negative consequences of the trade in higher education. As the higher education services become a commodity, developing countries institutions may not be able to compete with the high quality institutions from richer countries. Domestic universities may also not be able to compete with foreign education providers, losing potential research students.

3. Methodology

3.1. Data Acquisition

All the data used on the analysis is available on the official *Ciência sem Fronteiras*'s web page in an unstructured format. An interactive map is provided where the user can select an university and see a list of their current and past *Ciência sem Fronteiras* students. A web scrapper was built to get the id codes given to the universities from this page. With those codes, web requests were made to get the students information from each of these universities. By the end of the process, data about 70984 students on 2060 universities around the world was acquired.

3.2. Data Scrubbing

In this work, data scrubbing techniques were used twice. The first time to correct the orthographic accents. Some universities and students names had problems with character encoding. To correct it, we used a text editor tool. The tool replaced all occurrences of problematic words with the correct word. The second time was to treat incomplete data. In some cases, pieces of information were missing. To treat these cases, data that had missing attributes was removed from the dataset.

3.3. Selection of Interesting Information

Pieces of information about students were selected. They were the state of student, current university in Brazil, university abroad, destination(country), study area, period of funding and whether the student was enrolled in an undergraduate or graduate programs. Knowing the state of each student it was also possible to get the student's region for further analysis.

3.4. Analysis

The analysis of this work was divided in four steps: the histograms of student frequency (1-D), their distribution analysis, the charts of students frequency (2-D) and the statistic summarizing between variable pairs.

First, we built histograms of student frequency for 1-D attributes. They were focused on three main attributes: origin state of the student, chosen country and area of study. For each set of attributes three different sets of data were used: only undergraduate students, only graduate students and the complete dataset. In total nine histograms were built. The average and standard deviation of these distributions were then computed and placed in a table.

In the second step, the long tail test was applied in each of the nine datasets from the previous step to check if they followed the long tail distribution [Anderson 2006].

For this, the Pareto principle was used [Pareto 1964]. The principle states that for many phenomena, 20% of the input is responsible for 80% of the results obtained. In this work it is verified that 20% of states, chosen country and area of study are responsible for 80% of *Ciência sem Fronteiras* students.

In the third step, pairs of attributes were crossed and bubble charts were generated to visualize and compare the distributions. The following pairs of attributes were crossed: origin states and study area; origin states and chosen country; study area and chosen country. For each pair we used three datasets: only undergraduate students, only graduate students and the complete dataset.

The bidimensional summarization was made between Area of study and destination for undergraduate and graduate students, between state and destination and also between the date when studies started and the state. For each of these pairs, the relative contingency [Pearson 1904], conditional probability, Quetelet's index [Quetelet 1835] and, to test the independence between variables, the difference between joint probabilities and product of probabilities.

Relative Contingency: count/total
Conditional Probability: $P\left(A\left|B\right.\right) = \left(A \bigcap B\right)$/P(B)
Quetelet's Index: $Q\left(A\left|B\right.\right) = \left[P\left(A\right|\ B\right) - p\left(A\right)\right]/P\left(A\right)$
Independence testing: $P(A\left|B\right.) - P\left(A\right) * P\left(B\right)$

3.5. Social Network

The information of Brazilians universities and foreign universities was used to model a social network. For this, we represent the nodes of a graph as universities and the edges as students, thus creating an undirected graph. The *Graph Modeling Language (GML)* was used to represent the graph and *Gephi* [Bastian et al. 2009] to generate the graph and calculate metrics. The data was divided in two sets, undergraduate and graduate students.

For this work three types of centrality measures were used, degree [Hakimi 1962], closeness [Freeman 1979] and betweenness [Freeman 1977] [Brandes 2001]. The degree centrality measures the number of edges incident in a vertex. The normalized degree centrality is showed in equation 1.

$$C_D(i) = \frac{\sum_{j=1}^{n} a_{ij}}{n-1} \tag{1}$$

Where a_{ij} is 1 if *i* is connect to *j* and 0 if both are not connected. The variable *n* is the number of nodes on network.

The closeness centrality measure is the inverse of the sum of distances from a node to all other nodes, representing the importance of the node to its close neighbours and its importance in the complete network. The normalized betweenness centrality measure is illustrated in equation 2.

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} e_{ij}} \tag{2}$$

Where *n* is the number of nodes, e_{ij} is the number of vertex in shortest path from *i* to *j*.

Finally, the betweenness centrality quantifies how many times a node is used as a bridge in a shortest path between other two nodes. The normalized betweenness centrality is showed in equation 3.

$$C_B(i) = \frac{\sum_{j,k \wedge i \neq j \neq k} \frac{g_{jik}}{g_{jk}}}{\frac{(n-1)(n-2)}{2}} \tag{3}$$

Where n is the number of nodes, g_{jk} is the number of shortest path from j to k and g_{jik} is the number of shortest path from j to k through i.

4. Results and Discussion

The histograms made in the first step of the analysis of study areas are shown in Figure 1. Figure 1(a) presents the study areas of all students that joined the program *Ciência sem Fronteiras* (undergraduate and graduate students). Figure 1(b) shows the histogram only for undergraduate students. Finally, Figure 1(c) illustrates the histogram only for graduate students.

For undergraduate students, it is possible to see that engineering students have the biggest bar in the histogram. These students have more than the double of the second biggest bar, biology and health students. However, for graduate students, engineering students are in third place in terms of representativity. The first is taken by biology and health students and in the second are the exact sciences students.

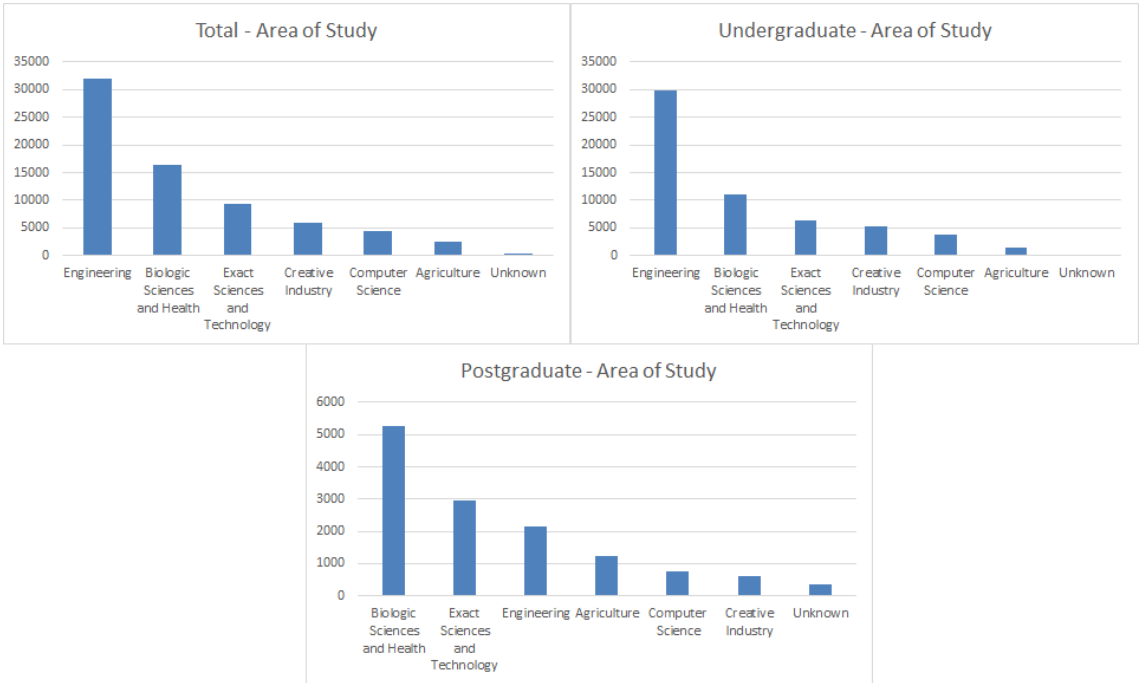

Figure 1. Histograms of Study Areas

Figure 2 shows the histograms of countries chosen by the students. Figure 2(a) shows the histogram for all students, both undergraduate and graduate students, that participated or participate by the Brazilian program to study abroad. Figure 2(b) presents the

frequency only for undergraduate students. Figure 2(c) shows the histogram for graduate students.

In all histograms it is possible to see a pattern of country choices. Undergraduate and graduate students are choosing the main technological powers in the world. The three histograms have similar distribution. The countries with most students are US, UK and France, traditional countries in terms of good quality in high education.

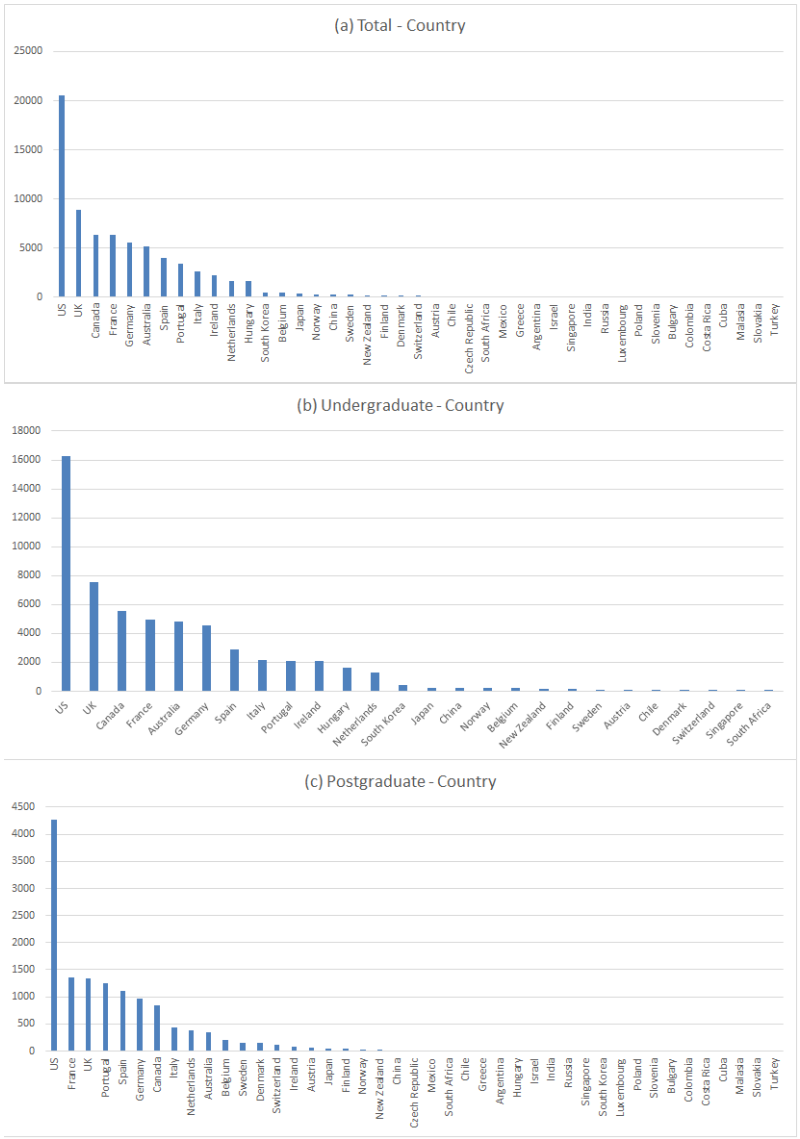

Figure 2. Histograms of Chosen Countries

Figure 3 presents histograms of the Brazilian states where *Ciência sem Fronteiras* students study. Figure 3(a) shows the histogram for all students (undergraduate and graduate). Figure 3(b) illustrates the histogram only for undergraduate students. In Figure 3(c) the histogram for the graduate students is shown.

For undergraduate courses, the number of students from São Paulo and Minas Gerais is almost the same (more than 11 thousand), but Minas Gerais has more students. Rio de Janeiro, Rio Grande do Sul and Paraná also have almost the same number of

students (4500 students). On the other hand, states such as Acre, Amapá, Rodônia and Roraima have less than 100 students. In terms of graduate students, São Paulo has the greater proportion of students, almost the double of the second, Rio de Janeiro. The same states that are underrepresented on the undergraduate histogram are also poorly represented on graduate histograms, with less than ten students for each state. Except from Roraima, where no postgraduate student joined the program.

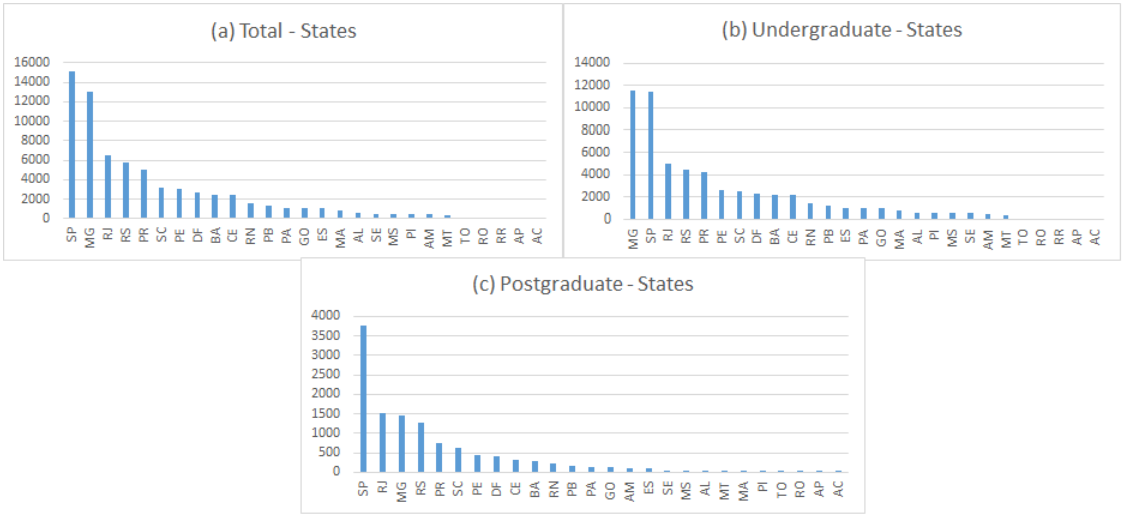

Figure 3. Histograms of Brazilian States

Table 1 presents the results of the Pareto principle test to verify if the data follows a long tail distribution. In the complete set, 20% of the states and areas have almost 80% of the total students, following a long tail distribution. For countries, 20% received 88.4% of the students, more than necessary to follow a long tail distribution. In the undergraduate set, the states and countries do not seem to follow the long tail distribution. However, 20% of study areas hold 81.7% of students, indicating the long tail distribution. Finally, for the graduate set, 20% of the states hold 73.5% of students, which seems like a long tail distribution. 20% of countries hold 89.7% of students, almost all students, following a long tail distribution. 20% of study areas have 63.2% not indicating a long tail distribution.

Table 1. Result of Pareto Principle

Set	**20% of States have**	**20% of Countries have**	**20% of Areas have**
Total	78.5% of students	88.4% of students	78.5% of students
Undergraduate	63.6% of students	67.7% of students	81.7% of students
Postgraduate	73.5% of students	89.7% of students	63.2% of students

Figure 4 provides a comprehensive visualization of undergraduate students distribution over different countries and areas of knowledge. Figure 5 provides the same visualization but for graduate students. It is observed that graduate students are more evenly distributed between different areas of study and countries than undergraduate students. To support this observation, the average frequency and standard deviations of students are shown on Table 2 and Table 3.

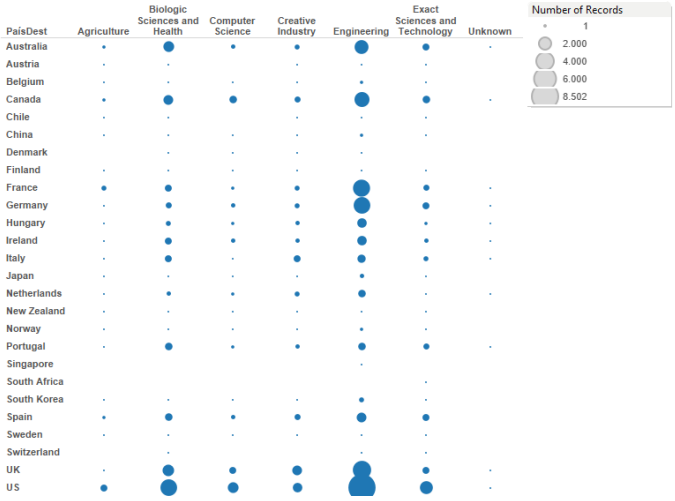

Figure 4. Undergraduate students bubble chart: Destination x Area of study

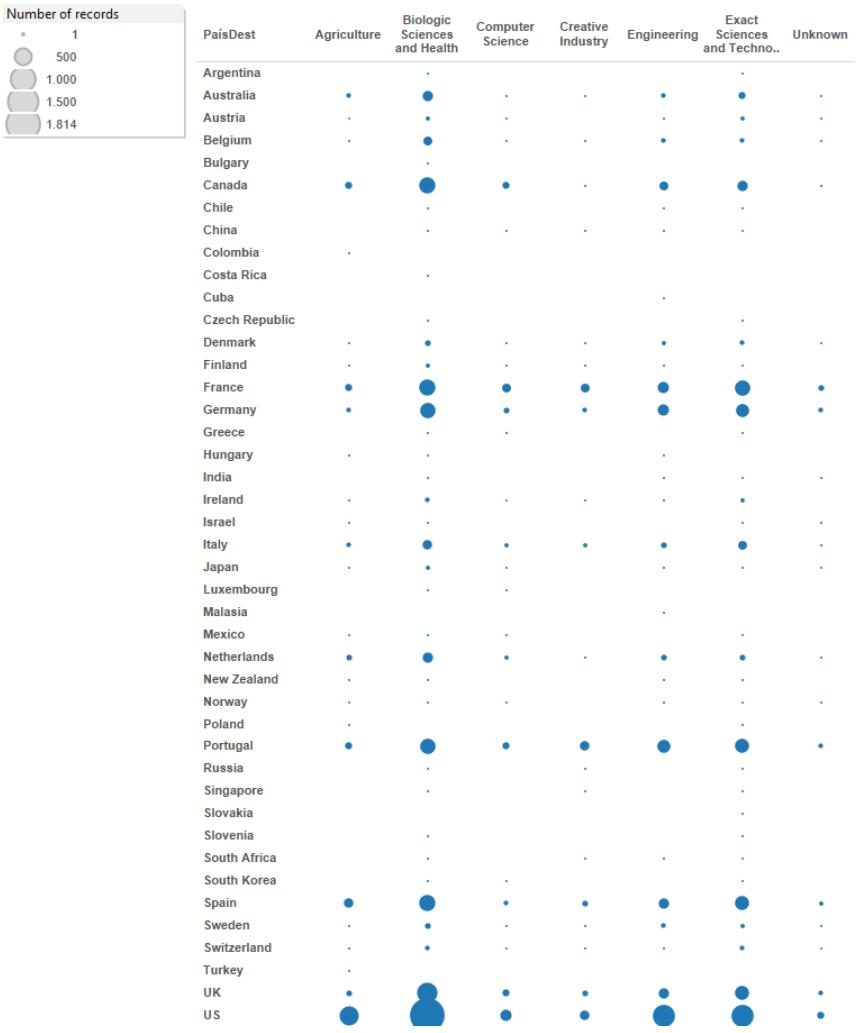

Figure 5. Postgraduate students bubble chart: Destination x Area of study

Table 2. Area of study frequency means and standard deviations

Area of Study	**Undergraduate**	**Graduate**	**Total**
average	8239.285714	1901.285714	10140,57143
standard deviation	10159.15603	1753.407037	10928.69059

Table 3. Destination country frequency means and standard deviations

Destination	**Undergraduate**	**Graduate**	**Total**
average	2218.269231	309.5116279	1650.790698
standard deviation	3573.385012	735.4631962	3670.657879

One possible way to explain this behaviour is that graduate students need to have a project and make contact with the foreign universities before applying, while undergraduates do not need to make this contact. This makes graduate students to choose universities that have partnerships with their current university, or even look for less popular universities. Another behaviour that is explained by this is the presence of countries with a very small number of students (graduates only), such as Colombia and Turkey (1 student each).

To measure the relation between different variables, some 2D statistical summarizing was made. The contingency, relative contingency, conditional probability, Quetelet's index and the difference between the joint probability and the product of probabilities were analyzed. In this analysis, the following pairs of data were used: States and Date, States and Destination, Area of study and Destination for both undergraduate and graduate students.

Very little correlation between these pairs was found. Quetelet's index was low in most cases, except in some outliers, such as countries that have hosted a very small number of students. The difference between the joint probability and the product of probabilities was also very small in every case. By definition, this shows that these variables are independent.

Using the social network modeling, it was possible to measure centrality metrics such as degree, closeness and betweenness. Table 4 and Table 5 show these metrics for the 5 nodes with the highest degree for undergraduate and postgraduate students, respectively.

Table 4. Centrality Metrics for Undergraduate Students

University	**Degree**	**Closeness**	**Betweenness**
USP	652	2.037	143788.720
UFMG	607	2.078	99599.305
UNB	574	2.123	84844.624
UFRJ	533	2.176	88545.916
UFSC	530	2.174	78386.616

Table 5. Centrality Metrics for Graduate Students

University	Degree	Closeness	Betweenness
USP	586	2.188	472915.351
UFRJ	349	2.503	208045.776
UFRGS	344	2.512	198109.271
UNICAMP	331	2.501	191007.914
UNESP	330	2.541	182344.682

The tables 6 and 7 show some general information about the universities graph for undergraduate and graduate students. The average degree, graph radius and diameter and the number of shortest paths are described in these tables.

Table 6. Information of Complete Undergraduate Graph

Average Degree	43,547
Diameter	7
Radius	4
Average Path Length	2.85336828184578
Number of Shortest Path	3446594

Table 7. Information of Complete Graduate Graph

Average Degree	19,548
Diameter	8
Radius	0
Average Path Length	3.38573090801817
Number of Shortest Path	3769462

These tables show the importance of some universities in the *Ciência sem Fronteiras* program. USP (*Universidade de São Paulo*) is the most important university for graduate and also for undergraduate in the social network in terms of degree and betweeness. The UFRJ (*Universidade do Rio de Janeiro*) is also represented in both sets, graduate and undergraduate. The other universities, even though are from south and southeast of Brazil, are different for the two sets. The exception is UnB that is from the center-western region of Brazil and is the third with most undergraduate students.

5. Conclusion

This work presented a study of Brazilians student mobility financed by the program *Ciência sem Fronteiras*. All data was collected from the official website of the program. The information of more than 70 thousands students was analyzed, using distribution tests and correlation tests, building graphs and histograms, in order to find trends and patterns of students mobility.

The results shows that, in most part of tests, the long tail distribution is present in many different information of data. Furthermore, the test of correlation between variable shows that the variables are independent.

For future works, we intend to make a deep analyze on data using others information as chosen university and funding period. Furthermore, another idea is collect others data as human development index (IDH) of Brazilian states and countries, data from best Brazilians universities and foreign universities. These data could be crossed with *Ciência sem Fronteiras* data and analyzed.

References

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. Hyperion.

Banks, M., Olsen, A., and Pearce, D. (2007). Global student mobility: An australian perspective five years on. *IDP Education*.

Bashir, S. (2007). Trends in international trade in higher education: Implications and options for developing countries. education working paper series, number 6. *World Bank Publications*.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177.

Choudaha, R. and Chang, L. (2012). Trends in international student mobility. *World Education News & Reviews*, 25(2).

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

González, C. R., Mesanza, R. B., and Mariel, P. (2011). The determinants of international student mobility flows: an empirical study on the erasmus programme. *Higher Education*, 62(4):413–430.

Hakimi, S. L. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial & Applied Mathematics*, 10(3):496–506.

Murphy-Lejeune, E. (2003). *Student mobility and narrative in Europe: The new strangers*. Routledge.

Pareto, V. (1964). *Cours d'économie politique*. Librairie Droz.

Pearson, K. (1904). *Mathematical contributions to the theory of evolution*, volume 13. Dulau and co.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés ou essai de physique sociale*. Bachelier.

Tremblay, K. (2002). Student mobility between and towards oecd countries: a comparative analysis. *International mobility of the highly skilled*, pages 39–67.

Verbik, L. and Lasanowski, V. (2007). International student mobility: Patterns and trends. *World Education News and Reviews*, 20(10):1–16.