How is an academic social site populated? A demographic study of Google Scholar Citations population

José Luis Ortega[1]

Cybermetrics Lab, CCHS-CSIC, Madrid, Spain,
jortega@orgc.csic.es

Abstract

This paper intends to describe the population evolution of a scientific information web service during 2011-2012. Quarterly samples from December 2011 to December 2012 were extracted from Google Scholar Citations to analyse the number of members, distribution of their bibliometric indicators, positions, institutional and country affiliations and the labels to describe their scientific activity. Results show that most of the users are young researchers, with a starting scientific career and mainly from disciplines related to information sciences and technologies. Another important result is that this service is settled by waves emanating from specific institutions and countries. This work concludes that this academic social network presents some biases in the population distribution that does not make it representative of the real scientific population.

Keywords: web bibliometrics; Google Scholar Citations, academic social networks, web demography

Introduction

[1] Cybermetrics Lab, CCHS-CSIC, Albasanz, 26-28 28037 Madrid, Spain, Tel. +34 916022603

The coming of the Web and Internet has created a transformation of the scientific communication, questioning traditional ways in which scientists interact among them and the appreciation of the research activity by the society. The term “Science 2.0” defines this new form of Science (Schneiderman, 2008) in which the collaborative activities and the free exchange of information are modelling new academic results (open access journals, academic repositories, etc.) and an alternative assessment system (altmetrics, webometrics, etc.). In this context, social networking sites such as Academia.edu, ResearchGate or Mendeley have recently raised as platforms to improve the social participation, the sharing of papers and the seeking of new collaborators. At the same time, academic search engines are broadening the publication outlets (repositories, digital libraries, etc.) at the expense of journals, while emphasize the role of authors and documents (Ortega, 2014). These new services are causing a challenge for research evaluation questioning the position of some agents (journals, publishers, journal level indicators, etc.), introducing open access products (repositories, web publishing, etc.) and suggesting new ways to measure science (altmetrics, webometrics, etc.). In this framework, studying population dynamics in those platforms would shed light on the representativeness of these sources and their reliability for research evaluation.

Google Scholar Citations (GSC) is a Google Scholar’s (GS) service that allows the building of a short personal page for free from the papers indexed in their databases, besides the addition of individual bibliometric indicators computed by the system. The novelty of GSC for research evaluation is that it makes possible the definition of specific research units, mainly researchers, which are able to be compared with others inside the same institution or research interest. In addition, the comprehensive coverage of research materials in GS favours that these pages offer a wide view of the research production and impact. And finally, the fact that these profiles are publicly available, it helps that an author can be appreciated for a broader range of academic activities.

However, GSC presents a singularity with respect to other academic search engines. Their profiles are directly created and made public by the researchers themselves. This causes that the population of GSC could be similar to academic social networking sites. This fact can have important consequences for research evaluation because it could produce unbalanced samples at disciplinary, country and institution level both in a static and longitudinal perspective. In this sense, this study pursues to observe dynamics on the use of social sites by researchers and how these services are settled along the time. Ultimately, to see whether the process of colonization of GSC –this is, the way in which GSC was taken up since their first moments– could shows important biases that influence the data collection and, in consequence, compel to adopt more precise sampling methods.

Related Research

Literature on demography in social network sites is rather scarce and in many cases, this makes up just descriptive reports about the geographical distribution of users. The most recent was the Duggan and Smith (2013) report which prompts important demographic differences between users from a social network site and another, signalling that each platform shapes its own population according to their services. In this sense, Boyd and Ellison (2007) already noticed the dissimilar successfulness of different services regarding to countries, gender or interests, which favours the changing nature of these sites. For example, Chang et al. (2010) described deep ethnicity changes in the American Facebook during three years, while Garcia et al. (2013) analysed the resilience of these sites facing the fast loss of users. Similar results were found by Mislove et al. (2011) on the United States population signed into Twitter.

However, literature on demographic aspect in academic social networks is even scanter. A few of papers have explored the presence of scientist in academic social sites. Haustein et al. (2014) followed the footprint of 57 bibliometricians on the Web, finding that 23% were in Google Scholar Citations and 16% had a Twitter account; whereas Mas-Bleda et al. (2014) tracked 1,517 researchers in several academic sites, detecting a low adoption rate and a limited overlapping between those sites. On the other hand, some reports, provided by the site itself, describe general statistics that illustrate the unbalanced distribution of researchers. Thus, a global report of Mendeley (2012) shows a strong presence of Biologist and Biomedicine users (31%) as well as a high weight of francophone countries and institutions. ResearchGate (2014) also presents a similar disciplinary distribution, with a hegemonic presence of Bio and Medicine users. Menendez et al. (2012) studied the positions and affiliations in Academia.edu finding that it is populated by young researchers and the presence of emergent countries is significant. As in generalist social networks, academic ones are also populated by different users from different countries, institutions and disciplines. Contrarily, most of the papers on academic social networks are focused on the use (Van Eperen and Marincola, 2011; Hogan and Sweeney, 2013). In this sense, Almousa (2011) observed disciplinary differences in the use of Academia.edu. Thelwall and Kousha (2014) described differences in the use of this site by gender and disciplines. Chakraborty (2012) compared Facebook and ResearchGate to detect the academic motivations to use both sites. And Ebner and Reinhardt (2009) studied the role of Twitter in scientific conferences.

But the most active interest on academic social networks is done from a research evaluation view, exploring the relationship between usage, followers, visits, etc., with citations and papers. In other words, examining the relationship of altmetric/webometric indicators with bibliometric ones. Li et al. (2012) found significant correlations between citations and numbers of bookmarked papers in Mendeley and CiteULike. Eysenbach (2011) observed that the tweet mentions can predict the future impact of highly cited papers. Contrarily, different results did not find a clear relationship between downloaded papers and their further scientific impact (Moed, 2005; Watson, 2009; Halevi and Moed, 2014; Glänzel and Heeffer, 2014).

With regard to academic search engines, studies have been basically centred on Google Scholar and Microsoft Academic Search (MAS), the two most relevant engines that include author profiles. A comparative study showed that while MAS presented a balanced population, GSC was biased to computer-related disciplines (Ortega and Aguillo, 2014). Haley (2014) also compared both engines at journal level, finding correlations between bibliometric indicators (citations and h-index). More concretely on GS, some studies were focused on its coverage in relation to other citation databases (Bakkalbasi et al., 2006; Meho and Yang, 2007), its connection with web citations (Kousha and Thelwall, 2007) and its suitability to the scientific assessment (Jacsó, 2008; Aguillo, 2012).

More specifically, GSC profiles were studied almost since its begining (Pitney and Gilson, 2012; Huang and Yuan, 2012). Ortega and Aguillo (2012) mapped the labels included in each profile to build a Map of Science. They themselves analysed country and institutional collaboration networks using co-authors lists of these profiles (Ortega and Aguillo, 2013). On the other hand, Delgado López-Cózar et al. (2014) evidenced the possibility of manipulating bibliometric scores of profiles. However, no previous studies have addressed how this service was populated since their origins from a longitudinal view, discussing their implications for research evaluation. This papers attempt to represent the evolution of users by several demographic attributes (country, organization, subject matter, positions, etc.) as way to illustrate the representativeness of this population for research evaluation studies.

Objectives

The principal objective of this work is to describe the growth of GSC in its initial moments (2011-2012) through a set of personal attributes such as bibliometric indicators, positions, disciplines, organizations and countries. This objective aims to make clear the biases that could appear in this population and discuss how they would affect the research evaluation. Several research questions can be formulated from this primary objective:

- How is the growth of profiles in GSC and how can the number of profiles be estimated?
- How have the characteristics that define this population (bibliometric indicators, position, discipline, affiliation and country) evolved during this initial moment?
- What consequences could have this distribution of profiles for research evaluation?

Methods

Data obtaining and processing

The way in which this data was taken and processed was already detailed in previous works (Ortega and Aguillo, 2012; 2013). Data processing was developed in two stages: in the first one, a SQL script was written to crawl the entire service asking for the 25 letters of the Latin alphabet in groups of three for the first sample (December 2011) and in groups of two for the remaining ones. The objective was to identify as many profiles as possible and extract their author identification. Once the crawler finished, a second script harvested the fundamental data from each profile such as name, affiliation, labels, number of papers and citations. Five quarterly samples were taken from December 2011 to December 2012 in a unique attempt, which sum 191,858 unique profiles. The first sample in December 2011 did not extract the number of papers because the script was not developed at all.

However, one of the most important problems of GSC, from a bibliometric view, is that the information about each profile is filled out by the users themselves in a natural language. For this reason, this raw data has to be cleaned hard and normalized before any statistical analysis because it is possible, for example, that a same organization is written in multiple different forms. For instance, Universidade de São Paulo could be written more than 20 diverse ways such as University of Sao Paulo, Sao Paulo University, USP, U Sao Paulo, etc. This problem gets worse when positions, departments, faculties, etc., are included in affiliations. Another problem related with affiliations is that sometimes a user is appointed to several organizations because he/she is a visiting professor or works for various institutions. In this case the first organization was always adopted as a main affiliation. In instances where no affiliations were detected, the web domain of the e-mail was considered as an affiliation, although they didn't always coincide.

Similar inconsistencies occur in other fields. Labels can present a same keyword in different languages, abbreviated or in plural/singular form. Sometimes labels with imprecise meaning such as *control*, *reliability* or *assessment* were not classified. On the other hand, the existence of duplicated profiles –different profiles that correspond to the same author– is rather scarce because these are created and maintained by their own users. A search of similar names returned only 2.1% of duplicated profiles; notice that it includes many common names such as Wey Wang, John Smith or José López. Due to this, the real percentage of duplicated profiles could be under 1%.

To solve these problems Google Refine (Google Refine, 2015) was mainly used for organisations and labels to group similar variants of the same name or word.

Indicators

To test the reliability of the sample and to estimate the total population of GSC the Lincoln-Petersen formula was applied (Seber, 2002). This equation is widely used in Wildlife management and it is based on the mark and recapture method. This counting method assumes that a high proportion of repeated items would be an indicator of the completeness of the sample. As more samples are tested more consistency gains the population estimation.

$$N = \frac{\sum (M_i C_i)}{\sum R_i}$$

Where N is the total population to estimate, M is the total number of profiles retrieved by the crawler, C is the number of unique profiles and R is the number of repeated profiles that appear several times during the crawling process.

Compound annual growth rate (CAGR) was used to measure the increase rate of the profiles and their attributes. This formula was considered because it is suitable for models with exponential trends. Thus, V_1 is the initial observation, V_n the final one and n is the number of moments between the first and the last observation. Next, it was converted to percentage:

$$CAGR = \left[\left(\frac{V_n}{V_1} \right)^{\frac{1}{n}} - 1 \right] * 100$$

In addition, GSC calculates some bibliometric indicators that describe the performance of each profile and are analysed in this paper:

- Papers: number of items indexes in GS and included in each profile.
- Citations: total number of citations that receive those items from the indexed papers in GS.
- H-index: it is the largest amount of papers (*h*) which have received at least the same number of citations each (*h*). For example, an h-index=5 means that the one author has published at least five papers that have been cited five or more times.

Results

Samples

This part traces the growth of the successive samples obtained along 2011-2012 and the consequent estimations of the size of GSC in profiles.

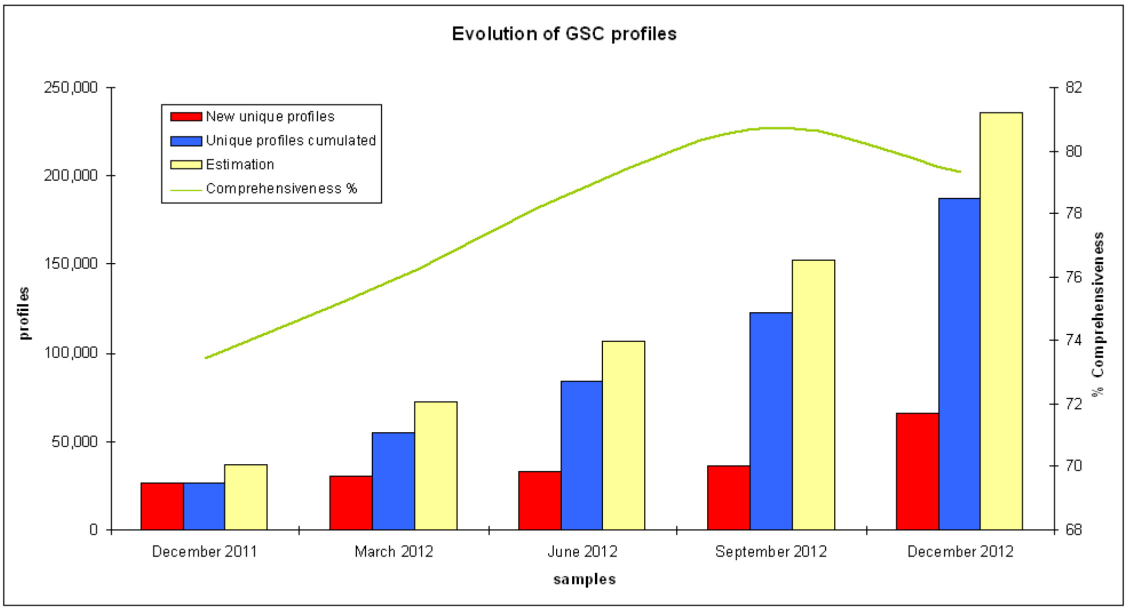

Figure 1. Growth and evolution of GSC by number of profiles

	December 2011	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*CAGR %*
Total retrieved	100,508	228,845	396,072	637,956	905,538	200.16
New unique profiles	26,682	30,588	33,233	35,655	65,699	56.92
Unique profiles cumulated	26,682	55,103	83,774	122,881	191,858	168.15
Repeated profiles	73,826	173,742	312,298	515,075	713,680	210.92
Estimation	36,325	72,579	106,246	152,196	243,435	158.87
Standard error	177.7	255.5	314	378.8	475.2	
Confidence intervals (95%)	35,977-36,673	72,078-73,079	105,630-106,861	151,454-152,939	242,503-244,366	
Comprehensiveness %	73.45	75.92	78.85	80.74	78.81	3.58

Table 1. Evolution of GSC profiles.

Figure 1 and Table 1 describe the evolution of GSC's profiles along each trimester, since December 2011 to December 2012. During this period, the number of unique profiles grew 164.9%, going from the 26,682 profiles in December 2011 to the 187,301 profiles in December 2012. At the same time, the number of estimated profiles increased 158.8%, from the 36,325 in December 2011 to the 243,435 in December 2012. It is interesting to notice that the new incorporations have remained stable (30,000 profiles approx.) until December 2012, when the number of new profiles was doubled. According to the comprehensiveness, which measures the percentage of unique profiles into the full estimation, it has been enhanced from 73.4% to 79.3 %. This high rate of completeness shows that these samples are enough representative of the total population.

Bibliometric indicators

Bibliometrics indicators (#papers, #citations and h-index) from each sample are graphed in a log-log plot to describe the evolution of the scaling exponent (α) and median of each distribution.

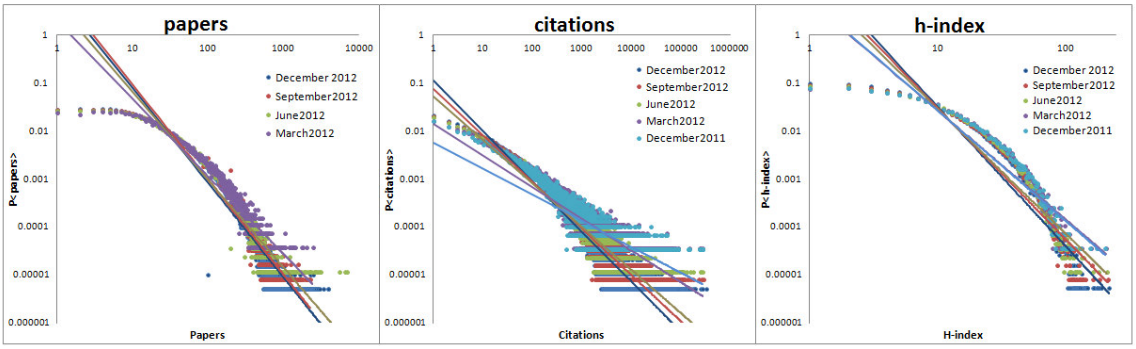

Figure 2. Papers, citations and h-indexes distributions by sample.

	December 2011	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*Total*
Papers						
α		1.617	1.826	1.98	1.965	1.89
Median		27	26	25	23	26
<10 (%)		24.1	25.6	26.7	28.1	25.23
<100 (%)		83.6	84.5	85.2	86.7	83.9
Citations						
α	0.539	0.657	0.902	0.965	1.045	.974
Median	212	224	180	151	132	154
<10 (%)	13.7	13.6	16.3	16.6	17.7	15.42
<100 (%)	36.2	37.1	40.8	43.5	45.6	39.58
h-index						
α	2.297	2.273	2.593	2.756	2.894	2.722
Median	7	7	7	6	6	6
<10 (%)	63.6	65.8	69.1	71.0	70.1	64.93
<100 (%)	95.2	100.0	100.0	100.0	100.0	99.94

Table 2. Principal parameters of papers, citations and h-indexes distributions by samples

Figure 2 plots the frequency distribution of papers, citations and h-indexes of each sample. Table 2 contains the main parameters that describe these distributions as well. These parameters were only obtained for descriptive purposes and not for estimation attempts, which is the reason why these distributions were not logarithmically binned (Milojević, 2010). In general, it is perceived that the scaling exponents (α) grow as time goes by, mainly since June 2012 when an important leap is perceived. This means that the differences between profiles increase in each sample, causing that the distributions of papers, citations and h-indexes are more and more unbalanced. In addition, median values gradually descend which indicates that the new added profiles in each sample correspond with small users in bibliometric terms. This is confirmed by the increasing values of percentages less than 10 papers, citations and h-indexes.

Academic positions

From the total 191,858 unique profiles, 88,335 (46%) profiles showed an academic status. The aim is to present the scholar position as a way to describe the youthfulness or maturity of the population in academic terms. Six professional categories, as close as possible to the academic hierarchy, are defined to group these academic statuses (Table 3). Thus Professor is the position most frequent (38%), being followed by Assistant Professor (18.4%) and Doctoral Student (16.3%). These two categories could correspond to young professional statuses which suggest that GSC is being settled more by young researchers than recognised professionals such as Professors. This is confirmed if Research Fellow is added to this group of young scholars (46.1%). This explains the low proportion of Associate Professor (15.2%), an intermediate scale, or Emeritus Professor (.7%). In line with this, the academic positions that most rise are Doctoral Student (Δ18.84%) and Assistant Professor (Δ12.48%) as well. This confirms that young researchers and professors are getting a considerable presence in this service.

Academic position	*December 2011*	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*Profiles*	*CAGR %*
Professor	5,478 (37.45%)	5,653 (35.76%)	6,536 (41.15%)	6,194 (40.32%)	9,721 (36.47%)	33,582 (38.02%)	12.15
Assistant Professor	2,744 (18.76%)	3,162 (20%)	2,822 (17.77%)	2,617 (17.04%)	4,940 (18.53%)	16,285 (18.44%)	12.48
Doctoral Student	2,059 (14.08%)	2,492 (15.76%)	2,222 (13.99%)	2,707 (17.62%)	4,880 (18.31%)	14,360 (16.26%)	18.84
Associate Professor	2,365 (16.17%)	2,530 (16%)	2,493 (15.7%)	2,068 (13.46%)	3,994 (14.99%)	13,450 (15.23%)	11.05
Research Fellow	1,854 (12.68%)	1,865 (11.8%)	1,689 (10.63%)	1,701 (11.07%)	2,924 (10.97%)	10,033 (11.36%)	9.54
Emeritus Professor	126 (.86%)	108 (.68%)	122 (.77%)	75 (.49%)	194 (.73%)	625 (.71%)	9.01
Total	14,626	15,810	15,884	15,362	26,653	88,335	12.75

Table 3. GSC profiles grouped by academic statuses.

Labels

Labels that describe the research activity of each profile were counted and classified to study the evolution of GSC according a subject matter view. Scopus Subject Area scheme was used to group each label and show hence an easier disciplinary evolution.

Subject class	*December 2012*	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*Total*	*CAGR %*
Computer Sciences	9376 (19.73%)	8270 (17.7%)	6880 (14.63%)	6633 (14.18%)	10838 (13.24%)	41997 (15.56%)	2.94
Engineering	3395 (7.14%)	3738 (8%)	3691 (7.85%)	3454 (7.38%)	6264 (7.65%)	20542 (7.61%)	13.03
Physics and Astronomy	2625 (5.52%)	2776 (5.94%)	3082 (6.55%)	3030 (6.48%)	5991 (7.32%)	17504 (6.48%)	17.94
Mathematics	2916 (6.14%)	3125 (6.69%)	2900 (6.17%)	2687 (5.74%)	4780 (5.84%)	16408 (6.08%)	10.39
Medicine	2474 (5.21%)	2390 (5.12%)	2825 (6.01%)	2978 (6.36%)	5048 (6.17%)	15715 (5.82%)	15.33
Agricultural and Biological Sciences	2460 (5.18%)	2527 (5.41%)	2695 (5.73%)	2816 (6.02%)	4849 (5.92%)	15347 (5.69%)	14.54
Biochemistry, Genetics and Molecular Biology	2814 (5.92%)	2355 (5.04%)	2475 (5.26%)	2677 (5.72%)	4375 (5.34%)	14696 (5.44%)	9.23
Social Sciences	2833 (5.96%)	2595 (5.55%)	2405 (5.11%)	2414 (5.16%)	3986 (4.87%)	14233 (5.27%)	7.07
Environmental Science	1633 (3.44%)	1954 (4.18%)	2090 (4.44%)	2125 (4.54%)	3824 (4.67%)	11626 (4.31%)	18.55
Multidisciplinary	2363 (4.97%)	2005 (4.29%)	1691 (3.6%)	1707 (3.65%)	2765 (3.38%)	10531 (3.9%)	3.19
Total	47523	46721	47026	46790	81869	269929	3.19

Table 4. Evolution of the new labels added in each moment by research classes in GSC

Descending on the subject class level (Table 4), it can be valued that the disciplines with highest number of labels are Computer Sciences (15.56%), followed far by Engineering (7.61%) and Physics and Astronomy (6.48%). However, the disciplines that get the most joining up to GSC are Environmental Science (Δ18.55%) and Physics and Astronomy (Δ17.95%), while Computer Science (Δ2.94%) is the field that increases most slowly, missing the beat with the rest of disciplines. This suggests that a disciplinary change could be happening, where information technologies disciplines are given away to the biological and physical subject matters.

Affiliations

Processing and analysing affiliations makes it possible to know the origin of each profile and above all to know how the working place influences the settlement of an academic service. Figure 3 and Table 5 describe the number of new added profiles in each sample by country. Recognised countries in the scientific world such as the United States (25.78%) and the United Kingdom (7.85%) occupy the first positions, as well as emerging countries such as Brazil (6.6%) and India (2.8%) which are taking important places. The rest of the countries, such as Italy (5.24%), Australia (4.08%) or Canada (3.57%), are important scientific countries that have relevant positions in most of the research rankings.

But perhaps the most important fact is that the proportion of profiles from each country has changed as samples were taken. Thus, the first sample in December 2011 shows a high proportion of Anglo countries such as the United Kingdom, Australia and Canada, besides other important scientific countries such as Spain and Germany. Next, the sample of March observes the emergence of other European countries such as Italy and France, while in the sample of June and September 2012 it occurs the explosion of Brazil. This shows that the addition of new profiles is not done in a constant way but by following waves. According to the growth rate, Italy (Δ52.09%) and Brazil (Δ43.57%)

are countries with the most new profiles added to GSC in this period.

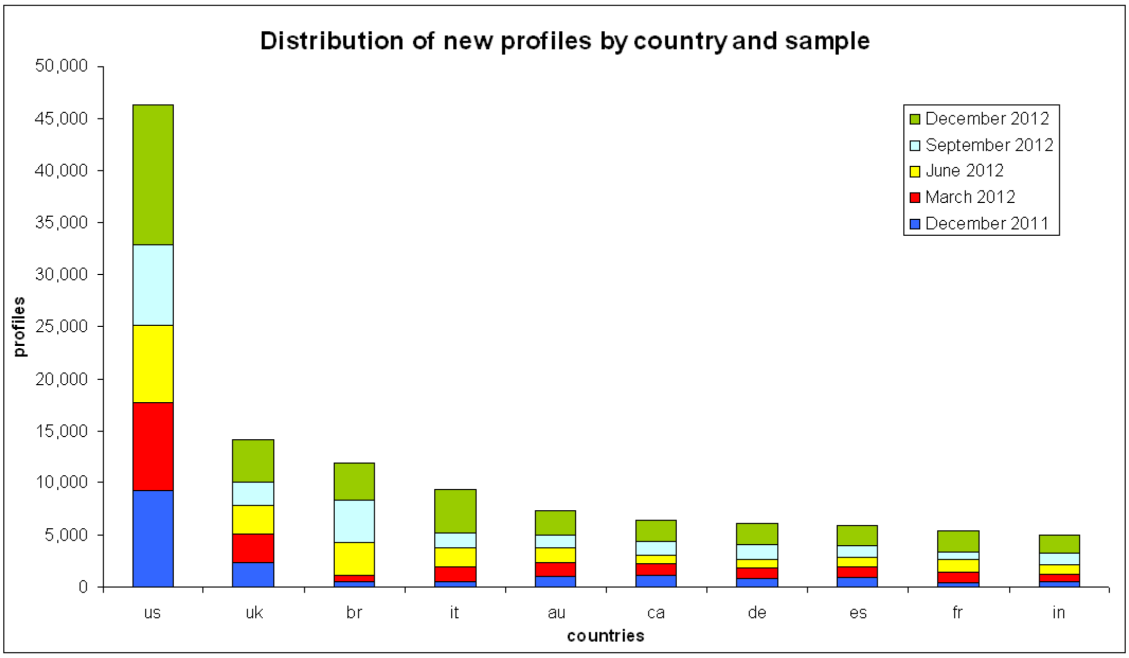

Figure 3. Distribution of new profiles by country and sample

Country	*December 2011*	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*Total*	*CAGR %*
United States (us)	9,340 (35.37%)	8,328 (28.11%)	7,430 (23.55%)	7,743 (23.33%)	13,550 (22.9%)	46,391 (25.78%)	7.73
United Kigndom (uk)	2,442 (9.25%)	2,679 (9.04%)	2,747 (8.71%)	2,203 (6.64%)	4,061 (6.86%)	14,132 (7.85%)	10.71
Brazil (br)	573 (2.17%)	626 (2.11%)	3,070 (9.73%)	4,117 (12.4%)	3,495 (5.91%)	11,881 (6.6%)	43.57
Italy (it)	512 (1.94%)	1,504 (5.08%)	1,700 (5.39%)	1,550 (4.67%)	4,167 (7.04%)	9,433 (5.24%)	52.09
Australia (au)	1,068 (4.04%)	1,296 (4.38%)	1,358 (4.3%)	1,333 (4.02%)	2,284 (3.86%)	7,339 (4.08%)	16.42
Canada (ca)	1,174 (4.45%)	1,081 (3.65%)	786 (2.49%)	1,372 (4.13%)	2,007 (3.39%)	6,420 (3.57%)	11.32
Germany (de)	854 (3.23%)	1,064 (3.59%)	728 (2.31%)	1,451 (4.37%)	2,082 (3.52%)	6,179 (3.43%)	19.51
Spain (es)	975 (3.69%)	1,025 (3.46%)	775 (2.46%)	1,158 (3.49%)	2,044 (3.45%)	5,977 (3.32%)	15.96
France (fr)	437 (1.65%)	1,039 (3.51%)	1,089 (3.45%)	809 (2.44%)	2,028 (3.43%)	5,402 (3%)	35.93
India (in)	491 (1.86%)	731 (2.47%)	987 (3.13%)	1,054 (3.18%)	1,776 (3%)	5,039 (2.8%)	29.32
Total	26,407	29,622	31,548	33,192	59,182	179,951	17.52

Table 5. Distribution of new profiles by country and sample

Going into further detail, the distribution by organisations fits more clearly with the statement that this service is settled by waves and that these could come from certain countries. In general terms, the principal institutions by number of profiles are the Brazilian Universidade de São Paulo (1.83%) and Universidade Estadual Paulista (.77%), followed by Harvard University (.53%) from the United States and the Universidade Estadual de Campinas (.53%), again a Brazilian university. This ranking confirms the huge increase of the Brazilian profiles. However, this process is not sequential but abrupt. Figure 4 and Table 6 illustrate how the first sample is occupied mainly by American universities (Harvard University, Massachusetts Institute of Technology and University of Michigan), but it is in the third and fourth sample when the Brazilian universities blast off taking the hegemony of Google's service. Thus, for example, the universities that most increase their profiles are Universidad Estadual Paulista (Δ116%), Universidade Estadual de Campinas (Δ68.6%) and the Universidade de São Paulo (Δ59.2%). On the contrary, it is surprising to notice that important international universities such as Harvard University (Δ-3.61%) and Massachusetts Institute of Technology (Δ-.19%) are slowed down the inclusion of profiles.

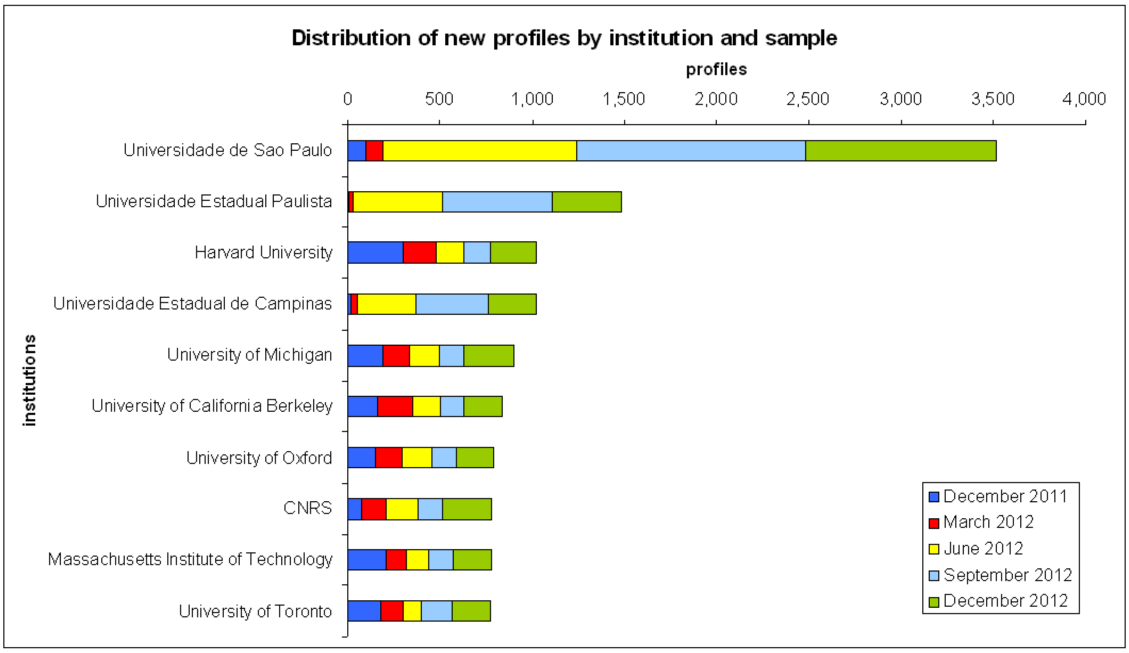

Figure 4. Distribution of new profiles by institution and sample

Affiliation	*December 2011*	*March 2012*	*June 2012*	*September 2012*	*December 2012*	*Total*	*CAGR %*
Universidade de São Paulo	101 (.38%)	91 (.3%)	1,049 (3.16%)	1,242 (3.48%)	1033 (1.57%)	3516 (1.83%)	59.20
Universidade Estadual Paulista	8 (.03%)	23 (.08%)	483 (1.45%)	594 (1.67%)	376 (.57%)	1484 (.77%)	115.98
Harvard University	304 (1.14%)	177 (.58%)	147 (.44%)	142 (.4%)	253 (.39%)	1023 (.53%)	-3.61
Universidade Estadual de Campinas	19 (.07%)	36 (.12%)	310 (.93%)	396 (1.11%)	259 (.39%)	1020 (.53%)	68.62
University of Michigan	188 (.7%)	145 (.47%)	168 (.51%)	128 (.36%)	273 (.42%)	902 (.47%)	7.75
University of California Berkeley	162 (.61%)	188 (.61%)	152 (.46%)	126 (.35%)	209 (.32%)	837 (.44%)	5.23
University of Oxford	146 (.55%)	152 (.5%)	161 (.48%)	128 (.36%)	205 (.31%)	792 (.41%)	7.02
CNRS	76 (.28%)	134 (.44%)	169 (.51%)	136 (.38%)	265 (.4%)	780 (.41%)	28.38
Massachusetts Institute of Technology	209 (.78%)	112 (.37%)	121 (.36%)	130 (.36%)	207 (.32%)	779 (.41%)	-0.19
University of Toronto	176 (.66%)	125 (.41%)	98 (.29%)	162 (.45%)	209 (.32%)	770 (.4%)	3.50
Total	26,682	30,588	33,233	35,655	65,699	191,857	19.75

Table 6. Distribution of new profiles by institution and sample

Discussion

Methodologically, this work presents the challenge of estimating the population of GSC using a capture-recapture method. The principal weakness of this study is that it only has a sample for each moment, because the data processing and obtaining require a great technical effort and time-consuming. This affects the Lincoln-Petersen formula because it produces overestimations when few samples are used (Tilling, 2001). This recommends taking these estimations with caution and considering lower values. A previous study (Radicchi and Castellano, 2013), crawling profiles from labels in common, obtained similar figures – 49,365 for March and 89,786 for July 2012. This lets us suppose that the real population could be slightly under our estimations and close to the retrieved profiles by the crawler.

Results on GSC point out a good evolution of this service during 2012, with a CAGR of 159% of estimated profiles which represents a seven-fold increase in a year. Although it is necessary to be reminded that these services suffer from a high volatility (Garcia et al., 2013), in fact, a recent crawler operated in December 2013 brought just an 11.7% of annual increase which supposes a growing stabilisation of profiles.

The longitudinal analysis of the population that was settling GSC along 2012 has made it possible to build a standard profile of the users of this service. The great majority is researchers with a short curriculum because the median is 26 articles, 154 citations and 6 h-index, low numbers that describe an incipient research activity. Even more, these figures decrease as time goes by which suggests that new added profiles in each sample are mainly researchers with a short career. This observation fits with academic positions where more than 34% of the profiles correspond to young academic categories (Doctoral Students and Assistant Professors) that have just started their academic careers as well as being the most increasing posts. This youthfulness is a characteristic of other academic sites where "graduated students" prevail (49%) (Menendez et al., 2012). This same occurs in generalist social network sites (Duggan and Smith, 2013) where most of the users are younger than 30 years old.

According to the thematic distribution, GSC is dominated by computer science researchers and other professionals related with information technologies and web environments, being the 15.56% of the total profiles. This fact was already observed in a previous study on GSC, where a Map of Science showed a core of computer science labels centring the picture (Ortega and Aguillo, 2012; Radicchi and Castellano, 2013). However, the disciplinary evolution of the service draws that other research fields such as Environmental Science (Δ18.55%) and Physics and Astronomy (Δ17.94%) are quickly growing, while Computer Science becomes stabilised with the lowest growing rate (Δ2.9%). This suggests that GSC advances toward a thematic equilibrium with a fairer proportion of researchers from all disciplines. Even so, subject matter distributions are also unbalanced in other academic services. Thus, Mendeley (2012) and ResearchGate (2013) bring very different figures with a strong presence of Bio and Medicine users.

One of the most interesting aspects of the population of this social platform is that this is done by waves of researchers from different countries and institutions. In the first stages, this service was settled by researchers from English-speaking countries such as the United States (35.4%) or the United Kingdom (9.25%) (December 2011). But in following rounds, European countries such as Italy (5.1%) and France (3.5%) (March 2012) strongly emerged (Ortega and Aguillo, 2013); and in the last samples, it shows emergent countries such as India (3.8%) and, above all, Brazil (12.4%) that is one of the countries with the highest growth (September 2012). These continuous series of users are better observable at institutional level. Thus, while the first period (December 2011-March 2012) is dominated by American universities such as Harvard University (1.14%) and Massachusetts Institute of Technology (.8%), in June 2012, abruptly Brazilian universities turn up such as Universidade de São Paulo (3.2%) and Universidade Estadual Paulista (1.45%), taking up the service (Ortega, 2014). These sudden changes and unexpected distributions of countries and institutions were already reported in early studies on social networks, where the successfulness of these services differs from one country to another (Boyd and Ellison, 2007) and where the fast emergence of different groups is usual (Chang et al., 2010). For example, Menendez et al. (2012), analysing Academia.edu, found similar figures for the United States and the United Kingdom but, however, detected important differences regarding Brazil and India. Mendeley's (2012) fact sheets described a singular presence of francophone countries and institutions. These population biases could be motivated by external reasons such as certain institutional policies or styles between scientists inside a country which cause a non-random occupation of these services.

This evidence a volatile reality, where country, institutional and thematic distributions frequently fluctuate along the time, provoking heterogeneous populations. This fact has important implications for bibliometric studies because these profiles are not representative of the total population of researchers in the world. On the contrary, they make clear the influence of specific institutional politics for the use and population of these services that cause intentional alteration of the population distribution. In this way, macro studies at institutional, country or subject matter level can not be extrapolated to the global scientific performance due to GSC represents only a specific group of researchers that jointed this platform for particular reasons. In this case, stratified approach would be recommended to select representative samples instead of random selections.

Conclusions

Several conclusions can be extracted from the results:

GSC was growing very fast during 2012, going from 26,600 profiles in December 2011 to 187,301 in December 2012. At least from the harvested data, because our estimations suggest 236,000 profiles, which is close to 10 times of the initial size.

According to bibliometric indicators, GSC is getting settled by young researchers with a starting career which boost a low bibliometric performance. The low median values and the increasing differences between the same parameters along the time, evidences the strong irruption of these new researchers. This is confirmed by the high presence of Assistant Professors and Doctoral Students.

From the subject matter point of view, GSC is dominated since its beginnings by researchers close to Computer Science and related disciplines. However, the last samples appreciate the emergence of researchers from Physics and Environmental Sciences and Medicine that balance the thematic distribution of the service.

Both country and institutional distributions exhibit evidence that this service is getting populated by waves of researchers, firstly from English-speaking countries where Harvard University and Massachusetts Institute of Technology were outstood; then from European countries and finally from emergent countries, highlighting Brazil and their Universidade of São Paulo and Universidade Estadual Paulista.

Finally, these results have important implications for research evaluation because they evidence that GSC's profiles, created by the scholars' will, generate a population biased towards any aspect (disciplinary, organization, country, etc.) and with rapid and strong fluctuations. This suggests that the use of this source for research evaluation should not be done randomly, but selecting precise strata of population.

Acknowledgements

I would like to thank Jennifer Carranza her helpful recommendations on the English version of this paper and the interesting suggestions of the reviewers.

References

Aguillo, I.F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. Scientometrics, 91(2), 343-351

Almousa, O. (2011). Users' classification and usage-pattern identification in academic social networks. IEEE Jordan conference on applied electrical engineering and computing technologies AEECT (p. 1-6). New York: IEEE.

Anderson, K. (2008). Scientists Use Social Media, The Scholarly Kitchen. Retrieved from http://scholarlykitchen.sspnet.org/2008/08/14/scientists-use-social-media/

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006).Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomedical Digital Libraries, 3(7), http://www.bio-diglib.com/content/3/1/7

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication, 13(1), 210–230.

Chakraborty, N. (2012). Activities and Reasons for Using Social Networking Sites by Research Scholars in NEHU: A Study on Facebook and ResearchGate. 8th Convention PLANNER-2012, Sikkim University, Gangtok. Ahmedabad, IN: IFLIBNET
Retrieved from http://ir.inflibnet.ac.in/bitstream/handle/1944/1666/3.pdf?sequence=1

Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). ePluribus: Ethnicity on Social Networks. Fourth International Conference on Weblogs and Social Media (ICWSM-10). Washington DC: AAAI Press.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology, 65(3), 446–454.

Duggan, M., & Smith, A. (2013). Social Media Update – 2013. Washington DC: Pew Research Center
Retrieved from http://pewinternet.org/Reports/2013/Social-Media-Update.aspx

Ebner, M., & Reinhardt, W. (2009). Social networking in scientific conferences–Twitter as tool for strengthen a scientific community. 4th European Conference on Technology Enhanced Learning, EC-TEL 2009. Nice: Springer

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet Research, 13(4), e123.

Garcia, D., Mavrodiev, P., & Schweitzer, F. (2013). Social Resilience in Online Communities: The Autopsy of Friendster. Retrieved from http://arxiv.org/pdf/1302.6109.pdf

Glänzel, W., & Heeffer, S. (2014). Cross-national preferences and similarities in downloads and citations of scientific articles: A pilot study. Proceedings of the Science and Technology Indicators Conference. Leiden: Universiteit Leiden

Google Refine (2015). Google Refine, a power tool for working with messy data (formerly Freebase Gridworks) - Google Project Hosting.
https://code.google.com/p/google-refine/

Halevi, G., & Moed, H. (2014). Usage patterns of scientific journals and their relationship with citations. Proceedings of the Science and Technology Indicators Conference. Leiden: Universiteit Leiden

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. Scientometrics, 1-19. http://dx.doi.org/10.1007/s11192-013-1221-3

Hogan, N. M., & Sweeney, K. J. (2013). Social networking and scientific communication: A paradoxical return to Mertonian roots? Journal of the American Society for Information Science and Technology, 64(3), 644–646.

Huang, Z., & Yuan, B. (2012). Mining Google Scholar Citations: An Exploratory Study. Lecture Notes in Computer Science, 7389, 182-189

Jacsó, P. (2008). Google Scholar revisited. Online Information Review, 32(1), 102 - 114

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web-URL citations: A multi-discipline exploratory analysis. Journal of the American Society for Information Science and Technology, 58(7), 1055-1065

Khabsa, M., & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. PLoS ONE, 9(5), e93949.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. 17th International Conference on Science and Technology Indicators (p. 451-551). Montréal: Science-Metrix and OST.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. Scientometrics, 91(2), 461-471.

Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. F. (2014). Do Highly Cited Researchers Successfully use the Social Web? Scientometrics http://dx.doi.org/10.1007/s11192-014-1345-0

Meho, L. I., & 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.

Mendeley (2012) Global Research Report. Retrieved from http://www.mendeley.com/global-research-report/#.UsbnMLQ5s4M

Menendez, M., de Angeli, A., & Menestrina, Z. (2012). Exploring the virtual space of academia. In: J. Dugdale et al. (eds.) From research to practice in the design of cooperative systems: Results and open challenges. London: Springer-Verlag.

Milojević, S. (2010). Power law distributions in information science: Making the case for logarithmic binning. Journal of the American Society for Information Science and Technology, 61(12), 2417-2425

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. 5th International AAAI Conference on Weblogs and Social Media (p. 554-557). Barcelona: AAAI Press

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. Journal of the American Society for Information Science and Technology, 56(10), 1088–1097.

Ortega, J. L. (2014), Academic Search Engines: A quantitative outlook. Cambridge, UK: Chandos Publishing, pp. 200 ISBN 1843347911

Ortega, J. L., & Aguillo, I. F. (2012). Science is all in the eye of the beholder: keyword maps in Google Scholar Citations. Journal of the American Society for Information Science and Technology, 63(12), 2370-2377

Ortega, J. L., & Aguillo, I. F. (2013). Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations. Journal of Informetrics, 7(2), 394-403

Pitney, W. A., & Gilson, T. A. (2012). Educational technology: Using Google Scholar Citations to support the impact of scholarly work. Athletic Training Education Journal, 7(1), 38-39

Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. Scientometrics, *97*(3), 627-637.

ResearchGate (2014). Main Page. Retrieved from http://www.researchgate.net/

Seber, G. A. F. (2002). The Estimation of Animal Abundance and Related Parameters. Caldwel, New Jersey: Blackburn Press.

Shneiderman, B. (2008). Science 2.0. Science, 319(5868), 1349-1350

Thelwall, M., & Kousha, K. (2014), Academia.edu: Social network or Academic Network? Journal of the Association for Information Science and Technology, 65(4), 721–731.

Tilling, K. (2001). Capture-recapture methods—useful or misleading? International Journal of Epidemiology, 30(1), 12-14.

Van Eperen, L., & Marincola, F. M. (2011). How scientists use social media to communicate their research. Journal of Translational Medicine, 9(1), 1-3.

Watson, A. B. (2009). Comparing citations and downloads for individual articles at the Journal of Vision. Journal of Vision, 9(4), article i