Scientometrics (2011) 86:1–14
DOI 10.1007/s11192-010-0318-1

Scientometrics of big science: a case study of research in the Sloan Digital Sky Survey

Jian Zhang · Michael S. Vogeley · Chaomei Chen

Received: 30 September 2007 / Published online: 16 November 2010

Abstract Large-scale scientific projects have become a major impetus of scientific advances. But few studies have specifically analyzed how those projects bolster scientific research. We address this question from a scientometrics perspective. By analyzing the bibliographic records of papers relevant to the Sloan Digital Sky Survey (SDSS), we found that the SDSS helped scientists from many countries further develop their own research; investigators initially formed large research groups to tackle key problems, while later papers involved fewer authors; and the number of research topics increased but the diversity of topics remains stable. Furthermore, the entropy analysis method has proven valuable in terms of analyzing patterns of research topics at a macroscopic level.

Keywords Bibliometric · Entropy analysis · Publication analysis · Sloan Digital Sky Survey · Large-scale scientific project

Introduction

Large-scale scientific projects, such as the Human Genome Project (HGP) and the Sloan Digital Sky Survey (SDSS), have become major drivers of forefront scientific research. These projects benefit from advances in technology that allow collection and analysis of higher precision and more extensive data sets. In Astronomy, for example, the use of charge coupled devices (CCD) in detectors on space-based (e.g., the Hubble Space Telescope) and ground-based wide-field (e.g., SDSS) telescopes has allowed observations that have revolutionized the field. Thanks to Moore's Law, high-performance computers have become available to process and archive the resulting data sets, which may range to tera- or peta-byte size (the latter expected for the forthcoming Large Synoptic Survey

J. Zhang (✉) · C. Chen
College of Information Science and Technology, Drexel University,
3141 Chestnut Street, Philadelphia, PA 19104-2875, USA
e-mail: jz85@drexel.edu

M. S. Vogeley
Department of Physics, Drexel University, Philadelphia, PA, USA

Telescope and similar projects). As a result, "Astronomy faces a data avalanche" (Szalay 2001). Development of new instruments that are more precise and powerful means that the data sets are not only larger, but are of higher precision (Gray et al. 2005). Shifting from a data deficient discipline, astronomy becomes a data-intensive science. Astronomers no longer need to wait for several months to use a telescope, which might be located in a remote mountaintop, to obtain the data they want (Schilling 2001). But, to efficiently utilize the now readily-available archived data, they now need to learn how to analyze large volumes of data (Gray et al. 2005). An increasing number of disciplines are experiencing such a shift from not having enough data to having too much data to analyze.

Such large-scale scientific projects also share some characteristics in terms of cost, membership, time-span, and data impact. These projects typically require millions of dollars of funding (for example, HGP has cost three billion dollars[1]), have a large number of members from various countries and institutes (SDSS has at least 300 scientists at 23 institutes from more than eight countries, www.sdss.org), and last more than 5 years or even longer (HGP is a 13-year project). Data may be accessible only to project members at the beginning of a project, but soon are released into the public domain. More scientists then have opportunities to access those data and make their own contributions. Given the volume and high quality of these data, research based on these projects is expected to last a relatively long time. In Astronomy, data from the previous generation of photographic-plate based sky surveys (e.g., the Palomar Observatory Sky Survey) has been used for several decades.

Few studies have specifically investigated how these large-scale projects contribute to the advance of scientific discoveries. What is the specific pattern of scientific research based on large-scale projects? How do research efforts change in response to the availability of these data? How do research topics change as the project proceeds toward completion? Answers to these questions could guide organizers of future large-scale projects, and help to better understand the advance of scientific discoveries made possible by these projects.

In order to answer these questions, one of the most common methods is to ask domain experts' opinions. Interviewing domain experts normally helps to identify the advances of a certain scientific area. Small (2004) applied this strategy at the essential science indicators by asking highly cited authors why their papers obtained high citation counts. In addition, review papers like annual reviews normally fulfill some of the tasks. However, the subjectivity of those interviewees has limited the application of this method. Interviewees' limited acquaintance with the research outcomes made this method even more unreliable. One could easily be overwhelmed by the massive volume of publications.

In this paper, we investigate these problems from a scientometrics perspective. Scientific papers are seen as indicators of the advances of scientific research, thus reducing the bias of subjectivity in an interview. Current sophisticated bibliometric programs can handle a large amount of publications. Benefiting from this advantage, we can automatically analyze the research outcomes from large-scale projects. This study is the first step of a three-year research effort in an ongoing and ambitious project (NSF_0612129) to support scientific discoveries in astronomy. Our ultimate goal is to enable astronomers to explore and analyze across astronomical data and the corresponding literature.

In this study, we choose the SDSS as an example of large-scale projects because the SDSS is one of the most important current projects in the astronomical community. The SDSS is the first large-area imaging and spectroscopic survey to use CCDs, so the images it

[1] http://www.ornl.gov/sci/techresources/Human_Genome/project/budget.shtml.

produces are substantially more sensitive and accurate than earlier surveys, which relied on photographic plates (cas.sdss.org/dr6/en/sdss/). In addition, SDSS has been considered a good example for serving diverse scientific communities. White (2007) says, "[I]n the case of SDSS, the release of the full survey data through a powerful, publicly accessible database has allowed astronomers across the world to carry out their own SDSS projects, thereby enhancing the whole community's opportunities for individual creativity." (p. 896).

The SDSS aims to map the large-scale structure of the universe seen in the distribution of galaxies, to characterize galaxy properties, and to study the properties of quasars and their evaluation. The project uses a 2.5-meter telescope equipped with a large-format mosaic CCD camera at Apache Point Observatory in New Mexico. The telescope uses a digital camera to image the sky in five optical bands, and employs two digital spectrographs to obtain spectra of about 1 million galaxies, 100,000 quasars, and 200,000 stars selected from the imaging data (York et al. 2000). First light of the telescope was achieved in 1998. Publications that used data from the project appeared soon afterward in the major astronomy journals. SDSS data are first released only to members of the project. After one year, data are made available to the public through an archive available on the Web. When the project is completed, half of the northern sky will be mapped. The total data collected by the SDSS project will exceed 40 terabytes (Szalay et al. 2000).

Previous researchers have also studied how astronomy research patterns change with time by examining statistics of publications (Abt 1981, 1990, 1994, 2000, 2007; Fernandez 1998). In his series of studies, Abt found a continuous increase of annual publication rate in the United States since WWII. Single authorship was replaced by double and triple authorship. In a 1994 study, he noticed that the availability of data obtained from spacecraft triggered a sharp increase of publications in *The Astrophysical Journal* after 1987.

The increase of multiple authorships comes along with the increased occurrence of international and national collaboration in astronomical publications. Abt (1990) concluded that in four major astronomy journals (*ApJ*, *AJ*, *MNRAS*, and *A&A*) international collaboration contributed one-quarter of all published papers. This finding was confirmed by Fernandez's (1998) study of two leading astronomy journals. Fernandez confirmed that collaboratively authored papers have formed the mainstream of astronomical publication in *The Astrophysical Journal* and *Monthly Notices of the Royal Astronomical Society*. Recently Abt (2007) found Astronomy has the highest multinational authorship among 16 disciplines including Surgery, Chemistry, and Geophysics. The multinational authorship papers constituted about 55% in astronomy, increasing 1.5% per year since the 1970s. On the other hand, multiple authorship papers distribute differently among journals. Basu and Lewison (2005) found that "internationally and nationally collaborative papers are in higher-impact journals, with a large number of single-author papers appearing in low impact journals." (p. 243).

However, to our knowledge, no studies to date have focused specifically on the impact of large-scale scientific projects on subsequent research. In this paper we examine the impact of the SDSS project on astronomical research through a scientometrics study of patterns in SDSS-based research papers. Based partially on authorship patterns noted by Abt in 1994 and the predictions of Szalay in 2001, we form the hypothesis about the impact of SDSS that:

(a) the availability of SDSS data to the scientific community would trigger a sharp increase of research publications related to the SDSS project; scientists from different countries would conduct their research based on SDSS data.

Studies cited above have shown that multiple authorships have become increasingly common in astronomy publications. However, it is the opinion of domain experts[2] that research efforts on large-scale projects such as SDSS may have special authorship patterns. Members of the project have worked for a long time before data becomes available. Therefore when the data first became available, the first publications that introduced the project and data included as co-author many of the one hundred project team members. Then later they formed smaller groups and focused on their own topics of interest. In light of this observation, our second hypothesis is that:

(b) the authorship of SDSS publications would start with a very large number, and decrease subsequently.

Discussions with domain experts reveal another possible pattern of research based on SDSS. As scientific groups using this survey increasingly focus on topics of their own expertise, we expect to see the increase of individual creativity as White (2007) expected. Therefore the third hypothesis is that:

(c) the number of research topics of SDSS increases and topics become more diverse with time. An operational definition of a topic is given in the "Methods" section.

Methods

In order to test our hypotheses and examine the advance of research based on SDSS, we collected bibliometric records of SDSS publications from ISI's *Web of Science* (WoS), and then analyzed the data with several well established software tools by using a new indicator of topics and a new measurement of topic diversity. Results were illustrated with standard desktop PC software.

Data collection

The bibliometric records of SDSS-related publications were collected from WoS, which contains approximately 8,830 titles from 230 disciplines. WoS is one of the largest scientific publication index databases and covers the majority of SDSS-related research journals and conference proceedings.

In June 2007, the bibliometric records were retrieved with search terms: 'SDSS' OR 'Sloan Digita*' over a time span between 1990 and 2007. A total of 1,657 records were retrieved. Since the WoS has a multidisciplinary coverage, the dataset may include some records that are not relevant to the SDSS project. The abbreviation SDSS has been used for terms other than the Sloan Digital Sky Survey, for example, Strategy Decision Supporting System. In these cases, we removed these irrelevant records by using functions in the WoS, such as analyzing the "Document Types" and "Sources Titles".

The Document Types analysis shows that there are 1,548 articles, 95 reviews, 6 corrections, 5 letters, 2 editorial materials, and 1 meeting abstract, in the 1,657 records. Since we are interested in original research in astronomy, only records from the articles category were included in subsequent analysis. The 1,548 articles were from 194 titles. One of the authors, an astronomer, identified titles that clearly have nothing to do with the SDSS survey, like *Water Resource Management*, *Diabetologia*, and *Cancer*, reducing the number

[2] Including the second author, who is an astronomer, and also a member of the SDSS project.

to 53 journals and conference proceedings. In SDSS.org the first scientific publication of SDSS appeared in 1983, but the majority of publications started in 1997. Therefore records before 1997 were manually inspected. Three records were excluded because SDSS in these records meant “sodium dioctylsulphosuccinate” and “supplementary difference sets.”

The final data collection includes 1,453 bibliographic records of papers. This number is substantially larger than the 841 publications recorded by the SDSS consortium to have been submitted by consortium members. Searching NASA ADS, the major astronomy bibliographic database, with a query of SDSS or ‘Sloan Digital Sky Survey’ or ‘Sloan survey’, returned 1,668 records (http://tinyurl.com/42jxy). With reference to these two numbers, our dataset can be considered as a comprehensive collection about SDSS-related bibliographic records. All records were downloaded in the format of “Full Record + Cited Refs” for further analysis. The 1,453 records then were divided by years. Since the SDSS started to yield data in 1998, papers published before 1998 form one group; those after 1998 are divided year by year. In subsequent analysis we use one year as a unit of analysis.

The indicator of research topics, diversity, and entropy

Testing the third hypothesis requires measurement of two distinct properties of a research field: an indicator of research topic, and an objective measure of diversity.

The indicator of research topics could be key topic words in a paper, all single words in fulltext, or sentences in the abstract. Keywords or descriptors have been widely used as indicators of scientific topics, especially in medicine. The majority of medicine journals adopt the National Library of Medicine’s Medical Subject Heading (MeSH) thesaurus as their descriptor sources. Studies using the MeSH keywords as indicators of topics, like (Swanson 2006), have achieved successes. Astronomy, however, does not have such an extended subject heading system. A controlled vocabulary used by seven major astronomy journals is a category of subfields in astronomy. This vocabulary is not detailed enough to represent specific topics of individual papers. Previous studies (Leydesdorff and Hellsten 2005, 2006) used high-frequency single words in documents to detect research topics. Given the massive volume of SDSS research papers, this method requires full-text search of each paper, and will generate too many words to represent topics without proper aggregation. In addition, it is difficult to use single words to convey the context of their use for people to understand their specific meanings. Given their contexts, sentences are easy to understand, but are hard to automatically handle using computers.

In this study, we use noun phrases as a new indicator of research topics. A noun phrase is a phrase that is made of a number of nouns like “quasar luminosity function” and “star formation rate indicators.” We also allow a noun phrase to begin with an adjective such as “cosmic reionization epoch” and “dark matter halo.” Using noun phrases as the indicator of topics can avoid some of the problems with using single words to represent topics. For example, the noun phrase ‘dark matter halo’ corresponds to a concept that can be readily used to characterize a topic, whereas single-word terms such as ‘dark,’ ‘matter,’ and ‘halo’ on their own among many other single-word terms would be much harder to detect.

We use natural language processing techniques namely, part of speech (POS) tagging, to extract noun phrases from the titles and abstracts of bibliographic records. Statistics-based algorithms used in this study come from Stanford Natural Language Processing Group’s tool kits (Toutanova and Manning 2000; Toutanova et al. 2003). This algorithm had achieved high accuracy in tagging POS by using test data.

Noun phrases were tagged and extracted from the title and abstract sections of each bibliographical record. Authors of original papers may also provide a list of keywords. As discussed above those keywords come mainly from astronomy control vocabulary and cannot stand as indicators of topics. We claim that a set of noun phrases, such as all noun phrases from papers published in the same year, represents the research topics in that year. The software and detailed methods of collecting noun phrases are discussed in next section.

Diversity is a complex concept, used across a range of scientific fields such as information theory, ecology, and economics. Purvis and Hector (2000) considered that biological diversity includes three different attributes of a system: Variety, Balance, and Disparity. Variety measures the number of distinctive categories. Balance measures the evenness of distribution of species. Disparity measures the degree to which the categories are different from each other. Normally variety can be easily identified. Counting the number of categories can fulfill this function. Balance and disparity are often measured in terms of entropy (Grupp 1990; Schmidt et al. 2006).

For the purpose of this study, we follow the concept of biological diversity. We use the number of noun phrases as the indicator of variety and use entropy to measure the balance and disparity. Following Grupp's (1990) formula, we define the entropy, E, of each set of noun phrases as follows:

$$E = \sum_{i}^{N} p_i \ln p_i$$

where p_i is the proportion of noun phrase i in a set of N noun phrases. In general, with a fixed N, when the entropy increases, the diversity increases. In this study, however, the comparison of entropy requires a different measure. As the number of noun phrases, N, increases, the entropy always increases. In this sense, comparing two entropies based on different Ns cannot quantify differences in diversity.

In this study, we use relative entropy, also known as the Kullback–Leibler distance (Cover and Thomas 1991), KL(E), to measure the changes of diversity. The KL(E) compares the distance[3] between two distributions. We compare the actual entropy and the maximum possible entropy of the same number of N noun phrases. The KL(E) was defined as follows:

$$\mathrm{KL}(E) = \sum_{i}^{N} p_i \ln \frac{p_i}{\mathrm{Max}(p_i)}$$

where Max(p) represents the maximum probability by which the number of N noun phrases could achieve their maximum entropy. Theoretically, a distribution has its maximum entropy when all of its units have the same proportion, which means $p_1 = p_2 = \ldots = p_{n.}$ This is the most balanced state. KL(E) avoids the possible bias caused by different Ns. The larger the relative entropy KL(E) is, the smaller the diversity.

Automatic analysis process and software

Given the volume of data, two bibliometric software packages that can handle very large datasets are used in this study: CiteSpace (Chen 2006) and HistCite (Garfield 2003).

To address hypothesis (a) and (b), the data files were imported into HistCite, which can automatically collect the statistics of publications, authors, unique authors, institutes, and countries. Authors' names come from the ISI's Distinct Author identification system to

[3] Even here it was called "distance," "it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often useful to think of relative entropy as a "distance" between distributions." (Cover and Thomas 1991, p. 18)

minimize errors due to variations of authors' name. Among these indicators, unique author means no matter how many times one author appeared in one year's dataset, he or she is counted as only one person. Based on this indicator we measure the size of the research community that was involved in SDSS-based research. Those statistical results are then analyzed with Excel and SPSS.

Noun phrases are generated by CiteSpace' Data Exchange function. In CiteSpace users can control the range of the number of words in phrases by specifying the minimum and maximum number of words allowed in a phrase. By default, the minimum number of nouns in a phrase is two and the maximum is four. Here, the minimum of two is used to filter out many single words, which are usually part of a longer phrase. In our dataset, CiteSpace collected 23,108 single words which form 60 percent of the total noun phrases (39,141). Some of them are "noise" like numbers. In this study we chose not to use word stemming, which is often used to merge singular and plural forms of a word, such as galaxy and galaxies as galax-. However, the information retrieval literature has suggested that stemmed words are harder to read and the advantage of using stemming is not significant in our case. After the noun phrases were extracted, they were checked manually by the authors to remove obvious anomalies and combine identical terms in different variants. The final noun phrase list is available online for further reference.[4] CiteSpace also collects the statistics of those noun phrases. Based on the statistics, the entropy of each set of noun phrases is calculated in SPSS and Excel.

Results

Corresponding to our hypotheses, the results are presented in three sections, namely, SDSS research range and the researcher community, SDSS research authorship, and SDSS research topics and their diversity.

SDSS research range and the researcher community

The research based on SDSS includes 1,453 bibliographic records, published in 53 journals and conference proceedings by a total of 3,245 authors from 693 institutions in 48 countries. Figure 1 summarizes the statistics of SDSS publications. Because the data were collected up to June 2007, the numbers for 2007 are partial.

All of these numbers support the first hypothesis. From 1999 to 2006, the number of publications steadily increased. The percentages of papers, unique authors, and institutions sharply increased from 2005 to 2006. The number of journals and conference proceedings that published SDSS research papers also increased. The number of unique authors reflects the size of the research community that is devoted to research based on SDSS, which has continually expanded. Scientists doing research related to the SDSS are spread among an increasing number of institutes and countries. The number of countries increased from three countries before 2000 to more than 15 countries in 2006.

The distribution of papers among journals varies. Figure 2 shows the distribution of papers among ten major journals that carry on SDSS-related papers. The journals and conference proceedings other than those ten journals were grouped into the "Others" section. More than half of the papers in the dataset were published in two journals:

[4] The final list is available at http://nevac.ischool.drexel.edu/~james/scientometrics_paper/SDSS_Noun_Phrase_ALL.xls.

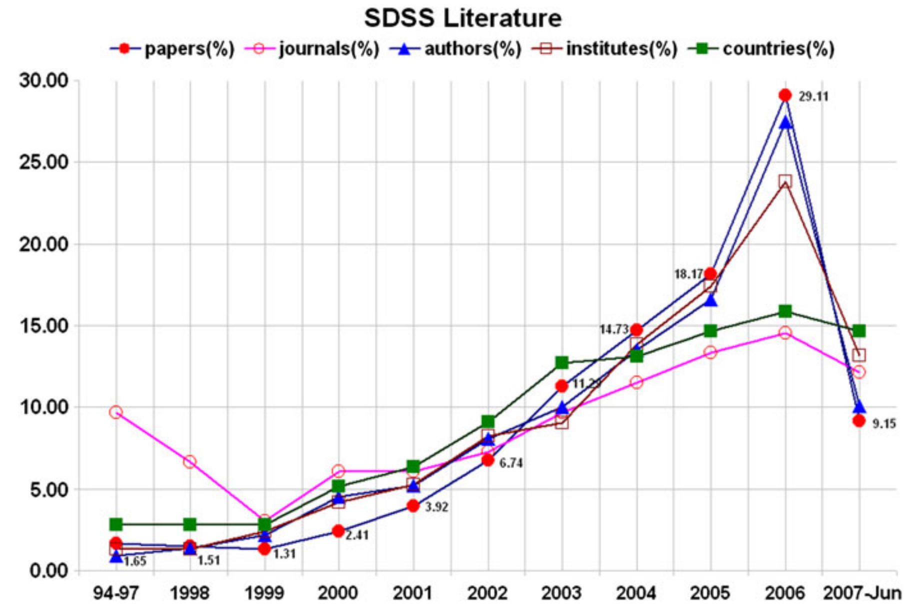

Fig. 1 The percentage growth of the SDSS literature. For example, 29.11% of the total number of papers in the dataset were published in 2006

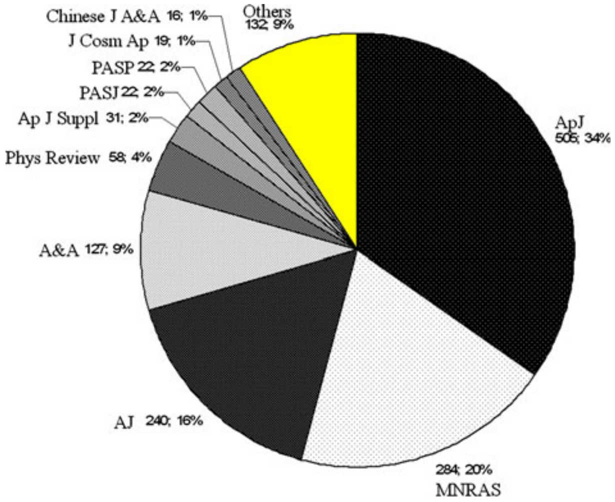

Fig. 2 Paper distributions among the ten major journals

Astrophysical Journal and *Monthly Notices of the Royal Astronomical Society*. More precisely, these two journals published 505 SDSS papers (34%) and 284 SDSS papers (20%), respectively. Close behind is the *Astronomical Journal* (16%). In this study's dataset, journals are obviously the major publication tools. All papers that appeared in conference proceedings were grouped into the Others section (a total of 132 papers, which is 9% of the entire dataset).

Table 1 Summary of authorship in SDSS research by years

Year	No. of papers	No. of authors	Average authorship
1994–1997	24	65	2.71
1998	22	68	3.09
1999	19	197	10.37
2000	35	622	17.77
2001	57	932	16.35
2002	98	1,419	14.48
2003	164	1,640	10
2004	214	1,892	8.84
2005	264	2,113	8.00
2006	423	3,139	7.42
2007 (June)	133	863	6.49

SDSS research authorship

Although we measure an increasing number of scientists involved with SDSS-related research, the multiple authorship experienced a shift after its peak in year 2000; thereafter authorship started to decline from a peak of nearly 18 authors per paper to six authors in 2007. Table 1 lists the details of the authorship, which support our second hypothesis. One interesting observation is that before 1999, the size of authorship was very small (two or three co-authors per paper), but the number increased abruptly to more than ten in 1999. Although the average size of authorship decreased to six in the first half of 2007, the number is still higher than normal authorship identified by previous studies (Abt 1981, 1990, 1994, 2000).

SDSS research topics and their diversity

We hypothesized that the research topics would increase and diversify. To answer the question, first we collected noun phrases (NP) from the title and abstract of each bibliographic record. Table 2 summarizes the number of those NPs in each year's data set.

In Table 2, the average number of noun phrases reflects the length of the title and abstract in a paper at the noun-phrase level. The unique NPs means that no matter how many times the same noun phrase appeared in a given data set, it was only counted as once. In this sense, the list collects the number of unique concepts in each year's set, which indicates the size of topics in each year's dataset.

Along with the increase in the number of publications, the number of noun phrases also increased. In this sense, the variety of topics increased. In each paper, the average contribution of noun phrase increased very slightly from 1998 to 2001, afterward varied between 14 and 16. The average number of unique noun phrases stays stable in the range between 10 and 11, except for a slight drop in 1998 and an increase in 2007.

The number of NPs shows an increase of research topics. Next we examine the diversity of topics by means of entropy analysis. Table 3 shows the summary of entropy analysis. Figure 3 shows the curves of relative entropy by year. Due to the incompleteness of data in the year of 2007, we omit data for 2007 from Fig. 3.

Comparing to the maximum entropy with the same number of noun phrases, the relative entropy KL(E) shows some initial change followed by a steady slight increase in recent years.

Table 2 Summary of noun phrases (CiteSpace) in SDSS research by years

Year	No. of papers	No. of NPs	Average no. of NPs	No. of unique NPs	Average no. of unique NPs
1994–1997	24	307	12.79	251	10.46
1998	22	262	11.91	203	9.23
1999	19	251	13.21	212	11.16
2000	35	483	13.80	350	10.00
2001	57	859	15.07	669	11.74
2002	98	1,494	15.24	1,120	11.43
2003	164	2,496	15.22	1,811	11.04
2004	214	3,149	14.71	2,274	10.63
2005	264	4,314	16.34	2,973	11.26
2006	423	6,564	15.52	4,422	10.45
2007 (June)	133	2,284	17.17	1,705	12.82

Table 3 Noun phrases entropy of SDSS research by years

Year	*E*	Max(*E*)	KL(*E*)
1994–1997	5.44	5.53	0.92
1998	5.07	5.31	1.05
1999	5.20	5.36	0.92
2000	5.50	5.86	1.20
2001	6.20	6.51	1.12
2002	6.67	7.03	1.16
2003	7.06	7.07	1.27
2004	7.30	7.73	1.28
2005	7.57	8.00	1.29
2006	7.94	8.39	1.36
2007 (June)	7.16	7.44	1.11

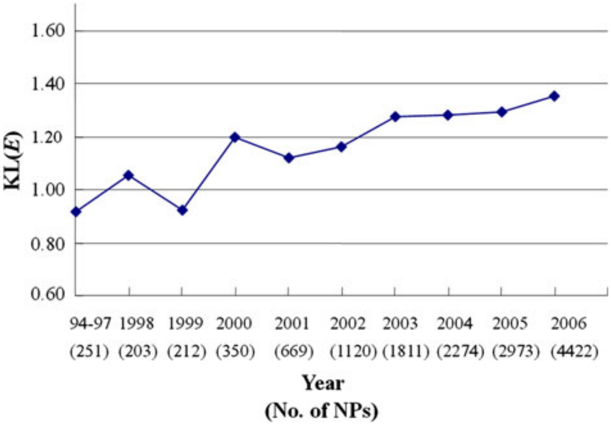

Fig. 3 The curve of Kullback–Leibler distance of noun phrase entropy by years

It is clear that the curve of relative entropy of SDSS noun phrases started with a fluctuation before 2001. The KL(E) value dropped twice in 1999 and 2001, indicating that topic diversity in the 2 years increased. After 2001 the KL(E) value kept increasing, but only slightly.

Both the average number of unique noun phrases and KL(E) curve tell the same story. The topic diversity went up and down in the first few years in SDSS-based research. Later the diversity grew only slightly, as shown in the slow increase of KL(E).

Discussion

Using automatic bibliometric methods, we explore the patterns of scientific research based on the Sloan Digital Sky Survey, a large-scale scientific project. Unlike general astronomy research, research based on this large-scale project displays some special patterns.

The SDSS has laid down the foundation for many scientific investigations. Scientists around the world are now involved in research relying on the data produced by this project. This result supports White's (2007) statement that SDSS "has allowed astronomer across the world to carry out their own SDSS projects." Our findings show that the number of authors, institutes, and countries using SDSS data increased in the past 10 years. Corresponding to the expanding number of people involved in the SDSS-related research, publications based on SDSS increased abruptly. The year of 2006 witnessed a burst of papers, almost double the number of papers in 2005. Future research is needed to compare this pattern with other astronomical surveys, like the 2df Redshift Survey, so that we can test if the above pattern is unique to the SDSS or if it is common across recent surveys.

The SDSS data archive will stay online for a long time (Szalay 2000), and much of these data remains to be analyzed by astronomers. We can anticipate that the number of publications, institutes and countries will keep increasing. An interesting question is when the growth of the SDSS literature will slow down and eventually reach its peak. Future study to answer this question will help to measure the impact of the SDSS project both in terms of overall volume of publication and timescale of impact.

Our findings on authorship reveal collaboration patterns of research efforts in the SDSS research community. Coauthorship in the SDSS literature started with a small number, two to three authors per paper, and increased to more than ten authors per paper. This finding supports domain experts' conscious understanding of the changes of research efforts. Large-scale projects like SDSS normally require collaboration among members from different institutes. At the very beginning of the project, a few key members wrote papers about the principles, purposes, and plans of the project. Later, when data first were successfully taken, members of the project were co-authors on the first several publications that introduced the project and data.

There are other factors that may also influence the pattern of authorship. One is that before funding can be obtained for a large project, the proposers must specify in considerable detail various scientific projects that they hope to solve with the new facilities. When these facilities are complete, scientists are compelled to work on those "key" projects which, because they often require large amounts of data and extensive analysis, are not necessarily the first ones finished. Thus the pattern of most of the proposers working as a large team during the first 5–10 years is expected.

One of the findings in this study differs greatly from Abt's and other scholars' findings. Even though double and triple authorship became common in astronomy publications, the authorship in SDSS literature is still larger than the average even in the most recent years examined, when the authorship decreased significantly. One explanation of the high authorship might come from the changes of the ways by which astronomers conduct their research in the "data avalanche" era. Gray (Gray et al. 2005) thought that current astronomers would need to learn how to analyze those volumes of data and develop automated methods to "find needles in the haystack." Because current astronomers may not all be good at handling such volumes of digital data, extra research members become necessary in order to find useful information from huge databases, partially contributing to the increasing number of authors. An interesting research topic for future study is to examine whether this pattern of increased authorship is shared by other large-scale projects. Further study is needed to compare the SDSS with other surveys, and with general authorship in astronomy.

In terms of diversity of research topics of SDSS research, our findings reveal a complex scenario. In terms of the numbers of topics, diversity increased as the number of noun phrase kept increasing. New concepts emerge every year. In this sense, the variety of research topics increased. The statistics of noun phrases and entropy analysis, however, reveal that balance and disparity did not change very much in recent years.

Table 2 shows that the average number of noun phrases experienced a slow increase from 1998 to 2002, and then fluctuated only slightly in the remaining years. Generally this means CiteSpace extracted nearly the same average number of noun phrases from datasets taken from recent years. In terms of new concepts, the average of unique noun phrases remained almost constant during the SDSS project. Each paper's title and abstract supplied about 11 unique noun phrases. This trend only changed in the year of 1998, which we would expect to be anomalous because the survey was just starting. The KL(*E*) curve indicates that the diversity of research topics initially fluctuates in the years 1998–2000, and is then quite stable, with only a slight, if any, increase with time of KL(*E*), which shows that the diversity of topics stays stable.

The third hypothesis we proposed is answered partially here. The number of research topics did increase, but our findings do not indicate a strong increase of diversity. We leave this question open for future studies that might include tracking the changes of topic terms to examine if topics are fixed or keep growing. It is important to note that the entropy analysis views the diversity problem at a macroscopic level; it tells whether there is dramatic change among different years, but it is hard to identify what causes the change. In order to explain what exactly has changed in terms of scientific discoveries, future studies are needed to explain the evolution of scientific knowledge though a combination of qualitative and quantitative approaches, such as clustering related terms into groups based on statistical metrics and visualizing the progress for intuitive understanding of the evolution of topics.

Conclusion

When large-scale scientific projects become the main impetus of scientific discovery, we need to know how they change the research and scientific community. As perhaps the first scientometrics study of the research patterns based on a large-scale scientific project in this era, we focus on the impact of the Sloan Digital Sky Survey on the patterns of research.

Our study reveals a number of patterns. First, large-scale scientific projects such as SDSS help a wide range of scientists from different countries in the world carry out their own scientific research. Second, scientists initially formed large research groups, but then worked in smaller-sized groups. Nevertheless, the average authorship per paper in SDSS research is still higher than the number of co-authors reported in previous studies of astronomical literature. Third, the number of research topics of SDSS-based research grew rapidly, but the diversity of research topics on SDSS varied little in recent years, according to the measures of diversity that we used.

Contributions of this study are twofold: the statistical results and the method of entropy analysis. First, the statistical results will help to provide an overview of the progress of research based on the SDSS project and to evaluate the impact of the SDSS project. When combined with other factors like data release policies, the SDSS consortium members' and non-members' achievements, etc., the SDSS project managers, PR staff, and funding agencies might find the results of this study useful for their evaluation and decision making. Second, the entropy analysis offers one way to view large volume research outcomes from a macroscopic perspective. By analyzing the changes of relative entropy, KL(E), we can identify a basic pattern of the changes of research topics. Given the volume of research papers related to large-scale scientific projects, this method may help scholars to obtain a brief overview of the changes before they jump into the details. Therefore the entropy analysis can serve as a starting point for future studies on large-scale scientific projects.

This study also has its limits. Because the bibliometric data were retrieved by searching certain key words, the data may not represent all research publications based on SDSS. Further research needs a better way to locate comprehensive publications based on large-scale projects. The results of statistics of noun phrases and entropy analysis were built up using state-of-the-art natural language processing techniques integrated in CiteSpace, but these are emerging technologies and so there are remains much room for these techniques to improve. We expect that our approach will become increasingly efficient and effective as the underlying enabling techniques improve.

Acknowledgments Special thanks go to the two reviewers, Helmut Abt and Virginia Trimble, who offered many great suggestions and insights to this study. This study is supported by the NSF under grant # IIS-0612129. Thanks also to Thomson ISI for providing the bibliometric data. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

References

Abt, H. A. (1981). Some trends in American astronomical publications. *Publications of the Astronomical Society of the Pacific, 553*, 269–272.

Abt, H. A. (1990). Trends toward internationalization in astronomical literature. *Publications of the Astronomical Society of the Pacific, 102*, 368–372.

Abt, H. A. (1994). The current burst in astronomical publications. *Publications of the Astronomical Society of the Pacific, 106*, 1015–1017.

Abt, H. A. (2000). Astronomical publication in the near future. *Publications of the Astronomical Society of the Pacific, 112*, 1417–1420.

Abt, H. A. (2007). The frequency of multinational papers in various sciences. *Scientometrics, 72*, 105–115.

Basu, A., & Lewison, G. (2005). Going beyond journal classification for evaluation of research outputs—A case study of global astronomy and astrophysics research. *ASLIB Proceedings, 57*(3), 232–246.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (99th ed.). New York: Wiley.

Fernandez, J. A. (1998). The transition from an individual science to a collective one: The case of astronomer. *Scientometrics, 42*, 61–74.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology, 54*, 400–412.

Gray, J., Liu, D. T., Santisteban, M. N., Szalay, A. S., DeWitt, D., & Heber, G. (2005). Scientific data management in the coming decade. *Microsoft Research Technical Report MSR-TR-2005-10*. Redmond, WA. http://research.microsoft.com/~Gray/. Accessed 27 May 2007.

Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics, 18*, 219–239.

Leydesdorff, L., & Hellsten, I. (2005). Metaphors and diaphors in science communication. *Science Communication, 27*(1), 64–99.

Leydesdorff, L., & Hellsten, I. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies', 'Frankenfoods', and 'stem cells'. *Scientometircs, 67*(2), 231–285.

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature, 405*, 212–219.

Schilling, G. (2001). The virtual observatory moves closer to reality. *Science, 289*(5477), 238, July 14.

Schmidt, M., Glaser, J., Havemann, F., & Hewz, M. (2006). A methodological study for measuring the diversity of science. In *Proceedings International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting* (pp. 129–137), Nancy, France.

Small, H. (2004). Why authors think their papers are highly cited. *Scientometrics, 60*(3), 305–316.

Swanson, D. R., Smalheiser, N. R., & Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of American Society of Information Science and Technology, 57*(11), 1427–1439.

Szalay, A. S. (2001). The national virtual observatory. In F. R. Harnden, F. A. Primini, Jr., & H. E. Payne (Eds.), *Astronomical Data Analysis Software and Systems X ASP Conference Series,* 238, pp. 3–12.

Szalay, A. S., Kunszt, P., Thakar, A., Gray, J., Slutz, D., & Bruner, R. J. (2000). Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey. In *Proceedings of the ACM SIGMOD*, Austin, TX, May 2000.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 252–259), Edmonton, Canada, May 2003.

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)* (pp. 63–70), Hong Kong, China, October 2000.

White, S. D. M. (2007). Fundamentalist physics: Why Dark Energy is bad for astronomy. *Reports on Progress in Physics, 70*, 883–897.

York, D. G., Adelman, J., Anderson, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., et al. (2000). The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal, 120*, 1579–1587.