Chaomei Chen — Drexel University
Fidelia Ibekwe-SanJuan — Drexel University, University of Lyon 3
Roberto Pinho — University of São Paulo
James Zhang — Drexel University

The Impact of the Sloan Digital Sky Survey on Astronomical Research

The Role of Culture, Identity, and International Collaboration

Abstract

We investigate the influence of culture and identity (geographic location) on the constitution of a specific research field. Using as case study the Sloan Digital Sky Survey (SDSS) project in the Astronomy field, we analyzed texts from bibliographic records of publications along three cultural and geographic axes: US only publications, non-US publications and international collaboration. Using three text mining systems (CiteSpace, TermWatch and PEx), we were able to automatically identify the topics specific to each cultural and geographic region as well as isolate the core research topics common to all geographic zones. The results tended to show that US-only and non-US research in this field shared more commonalities with international collaboration than with one another, thus indicating that the former two (US-only and non-US) research focused on rather distinct topics.

1. Introduction

Culture and identity play a major role in the complex processes involved in knowledge creation, representation and acquisition. However, these two parameters have rarely been the focus of automated methods for knowledge representation. We aim to investigate if culture and identity (geographic location) influence the development of a specific research field. We take as case study the Sloan Digital Sky Survey (SDSS) project in the field of Astronomy. SDSS is funded by NASA and the National Science Foundation in the US and aims to collect high quality data for astronomical research. The availability of this data has led to an increasing number of discoveries in astronomical research. Given that this project is funded and operated in the U.S., a natural question would be whether the research themes undertaken by astronomers in the U.S. differ significantly from their counterparts in other countries and regions such as Europe and Asia.

2. Methodology

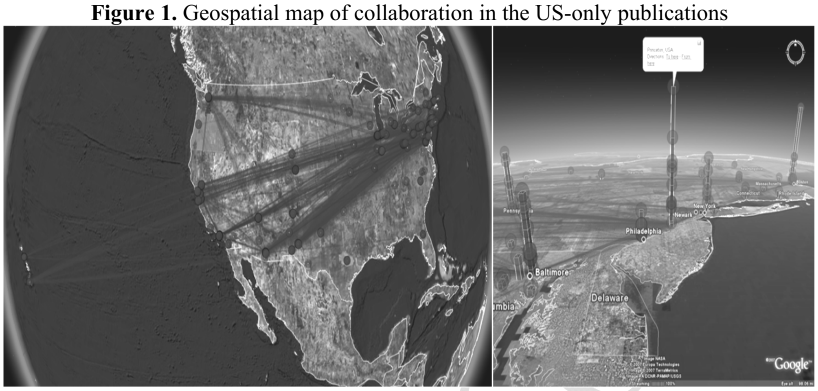
Our data consisted of a total of 1456 bibliographic records, retrieved from the Web of Science[1] database using a query containing the keywords "SDSS" or "SDSS Digit*". These records covered the period between 1998–2007. Among the 1456 publications, 379 were made by US institutions only, 459 by non-US institutions, and 618 are joint publications between US institutions and non-US institutions. We call this 3rd set International collaboration. We used three text mining systems to perform our analysis: CiteSpace (Chen 2006), Projection Explorer (Lopes et al., 2007) and TermWatch (SanJuan & Ibekwe-SanJuan 2006). These systems sought to highlight four aspects of the study: geo-spatial mapping; detection of salient topics; mining for association rules and finally a comparative analysis of the topics from each discourse community.

[1] http://scientific.thomson.com/products/wos/

3. Geo-spatial mapping of SDSS authors

A first level of macroscopic analysis is to visualize the geo-spatial distribution of publications across the world. This was done using Google Earth and CiteSpace (Chen 2006). Owing to reasons of space, we show maps for the US region only.

Figure 1. Geospatial map of collaboration in the US-only publications



4. Structure of SDSS research by cultural and geographic regions

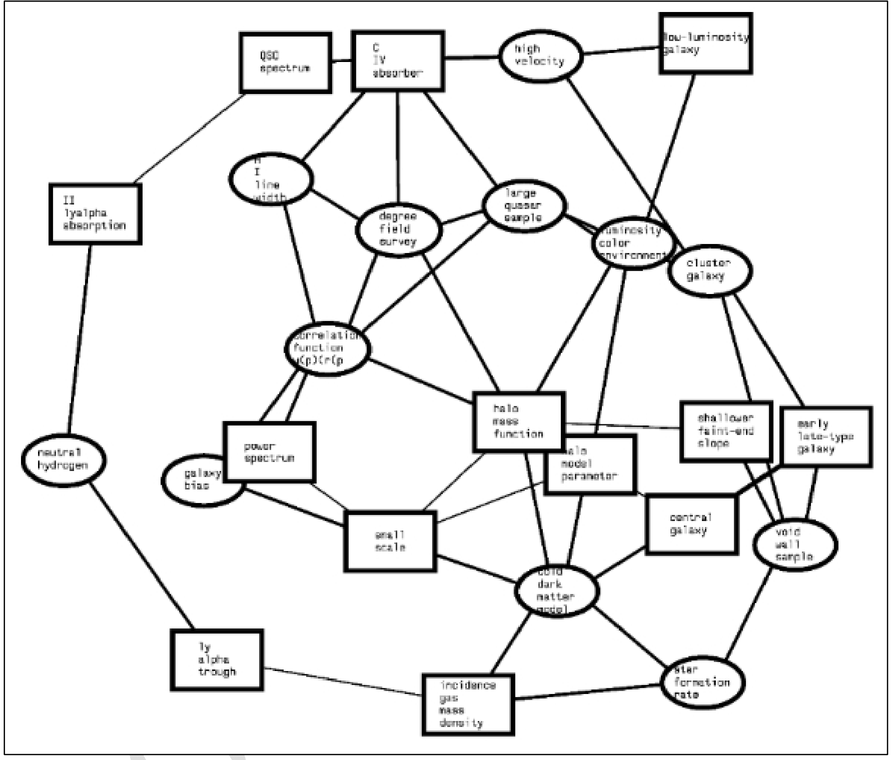
Salient topics were identified by applying natural language processing and information extraction techniques to SDSS-related publications. Salient topics are represented in terms of author-defined keywords, noun phrases extracted from the title and abstract fields of each record by CiteSpace. TermWatch was then used to obtain a global view of research topics in the three data sets based on terminological variations.

4.1 Research topics structure in the US-only institutions

375 publications were made by US-only authors in our data set. At the term variation level, a map of research topics was obtained with TermWatch (figure 2). Green nodes denote more recent topics (2005–2007). Pink color are topics whose terms appeared between 2002–2004. Shades of blue indicate topics found in earlier publications made between 1996–1998 (light blue) and 1999–2001 (deep blue). TermWatch identified central and peripheral atoms using the graph decomposition algorithm described in Biha et al., (2007). The most central cluster labeled "*halo mass function*" seems to be focused on galaxy clustering and formation models. On the whole, the majority of the topics appeared in the most recent period of the corpus (2005–2007).

CiteSpace was used to obtain association networks from titles and abstracts fields of the publications in this dataset. For reasons of space, we cannot show the network obtained from the US-only publications. However, the first cluster includes topics related to "*black holes*" (bh) such as "*velocity dispersion*, *local bh mass density*, *bh mass*, *bh merger*, *cluster galaxy evolution*". A second middle cluster focused on *star formation*, including terms like "*poststarburst galaxies*, *emission line*, *strong balmer absorption line*". A third cluster dealt with "*galaxy formation model*", including terms like "*quasar luminosity function*, *halo mass*, *satellite galaxies*, *dark matter halo, host galaxies*". The two systems, CiteSpace and TermWatch highlighted some common or semantically close terms even though they used different techniques to extract the terms. The following terms in CiteSpace's association networks "*cdm model, velocity dispersion, velocity distribution, cluster galaxy evolution, star formation*" appeared either in the exact form or as semantic variants in TermWatch's central atom (*cold dark matter model, high velocity, cluster galaxy, star formation rate*).

Figure 2. The major group of highly connected topics in the US only publications



4.2. Structure of SDSS research outside the US

Four hundred and fifty-nine publications were made by non-US authors. The titles and abstract fields were analyzed by TermWatch. The map obtained showed that there is no one central atom in contrast to the US-only research. SDSS research outside the US seem to be articulated around five major topics labeled "*supernova type ia*, *star formation rate, black hole, syfert galaxy* and *nearby cluster*". To gain further insight into the particularities of research on SDSS outside the US, a country-by-country analysis was performed by CiteSpace. We show results for two other prominent countries: UK and China. SDSS research in the UK seem to be characterized by the terms "*CDM cosmology* (CDM = *cold dark matter*), *sdss spectra* and *wavelength range*. SDSS research in China contains terms such as *sdss spectrum*, *sdss dr1* (data release 1) and *stellar velocity dispersion*. The cluster on the right deals with topics such as *double-peaked broad emission line*, *dimensionless accretion rate*, *black hole mass*, *eddington ratio* and *oxygen abundance*.

Figure 3. Clusters derived from papers by institutions in the U.K (*p* = .05).

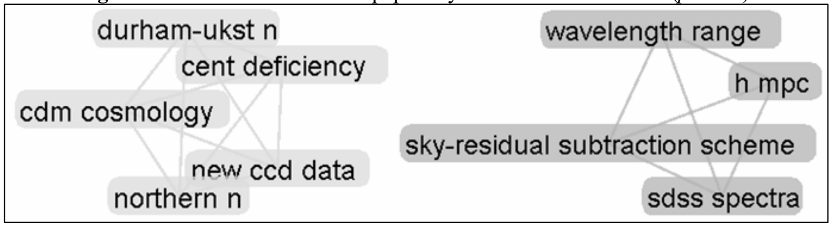


Figure 4. Clusters derived from papers by institutions in China (*p* = .05).

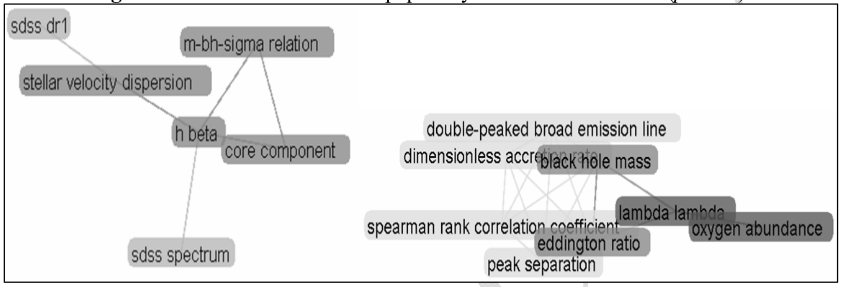
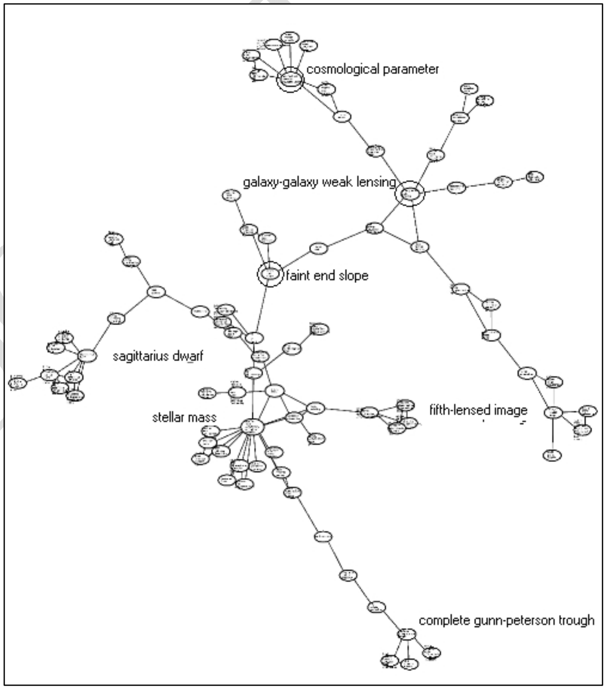


Figure 5. Topic map for international collaboration



4.3. Structure of international collaboration

This third set are publications made simultaneously by US and non-US authors. 618 records were concerned. The following map obtained by TermWatch shows the layout of research topics (figure 5). Like the non-US research, international collaboration on SDSS is not articulated around a unique center. Several subgroups of research topics are connected through chains of intermediary topics. We have circled and labeled the cluster at the center of the different subgroups: *cosmological parameter, galaxy-galaxy weak lensing, faint end slope, sagittarius dwarf, stellar mass, fifth lensed image, complete gunn-peterson trough*.

5. Comparative analysis of topics across geographical and cultural regions

Here, we seek to determine if there is a core set of research concerns shared by authors regardless of geographical origins, across the three data sets. To this end, we utilized the two systems PEx and TermWatch to perform a more detailed analysis.

5.1 Association rules derived from the SDSS literature by geographic regions

Association rules (ARs) are implications extracted from transaction databases. In the framework of automatic discourse analysis, ARs can be used to detect implications between domain terms supported by their co-occurrences. For instance, an association rule might find that the word "*matter*" always implies "*halo*" because the two co-occur more often than not. This is indeed supported by eight documents in the corpus. Locally Weighted Association Rules (LWR) (Lopes et al., 2007) were preliminarily used to identify salient topics from the three data sets. To selectively extract rules, LWR gives more weight to association rules in which words are specific (local) to a subset of documents from the corpus. Table 1 shows the intersections in terms of ARs found in the three data sets. The first row contains terms found uniquely on one of the sets of rules. Terms on the other rows were found on two or more of the datasets. Some remarks can be made from the list produced by the ARs: (i) as the algorithm privileges words that are specific to each subset of the corpus, it is no surprise that there are few common terms; (ii) Cosmic Microwave Background (CMB) is found to be common to both US and non-US, however WMAP (a probe) which detects CMD is found only in the collaboration subset, thus indicating a more widespread subject; (iii) studies regarding quasars, halos and satellites seem to be particular to US-only research; (iv) studies using or related to photometry (photometric, photometry, imaging, luminosity) are the prominently found in international collaboration.

5.2. Similarities in research topics

Here we used the terminology extraction and terminology variation identification components in TermWatch. This analysis is carried at two levels: topics (cluster labels) and topics contents (cluster contents).

By comparing cluster labels, we found topics that were common to authors from different geographic areas.

Table 1. Words specific in different data sets

US_only	**non_US**	**International Collaboration**
catalog; class; classes; correlate; degrees; demonstrate; disk; due; dwarfs; gamma; halo; halos; kpc; matter; previously; primary; properties; quasar; quasars; relative; satellite; satellites; sources; suggest; surveys; variables; wide;	circle; dot; early; equation; field; groups; method; models; obtained; order; parameter; parameters; power; ratio; ray; redshifts; scale; state; stellar; universe;	absolute; based; discovery; discuss; emission; estimate; galactic; imaging; inflation; law; low; luminosity; magnitude; observations; photometric; photometry; report; selected; simple; system; telescope; variation; wmap;
US & non_US	**US & International**	**non_US & International**
background; cosmic; cosmological; dark; microwave; observed; release; results; spectrum;	color; consistent; dwarf; function; high; mass; objects; stars; type	density; distribution; model; optical; range; spectra
Inter, USA, non_US	large; line; observed; present; redshift; show; star	

Table 2. Overlap in cluster labels by geographic and cultural zones

	Non_US	**US_only**	**Inter**
Nb_clusters	163	119	240
	Total clusters	**Overlap (%)**	
US, NonUS, Inter	552	6 (1%)	
US vs Non_US	282	10 (4%)	
US vs Inter	359	22 (6%)	
Non_US vs Inter	403	29 (7%)	

Table 3. Content overlap across geographic and cultural zones

	Non_US	**US_only**	**Inter**
Nb_Terms	442	342	683
	Total terms	**Overlap (%)**	
US, NonUS, Inter	1467	72 (5%)	
US vs Non_US	784	86 (11%)	
US vs Inter	1025	137 (13%)	
Non_US vs Inter	1125	153 (14%)	

The overlap in cluster labels among the three data sets and between pairs of data sets is very low, thus pointing to significant differences in SDSS research across different geographic and cultural regions. It appears from the above figures that both US_only and non_US share more common points with international collaboration research than with each other as indicated by the very close overlaps (6% and 7%) with topics in international collaboration. The six labels common to all three geographic zones are: *star formation rate, emission line, surface brightness, black hole, rest frame, large scale structure*.

Comparison of the clusters contents obtained for each data set gives a measure of their overlap across the three data sets. Table 3 gives details of this comparison.

The proportion of overlap in cluster contents echoes the ones found among cluster labels. Thus, similarities are consistent whether we look at the topic labels alone or into their contents.

6. Conclusion

The results obtained here are encouraging for identifying the impact of the SDSS survey on the global research community of astronomers and the uniqueness of each cultural or geographic region in contributing to knowledge discovery and dissemination in this field. These results have shown that three geographical zones have distinct research pre-occupations characterizing them but that each region (US and non-US) are brought together by international collaboration on some common research topics. This is remarkable considering that the terms were extracted automatically from the text fields of the bibliographic records. The systems have been able to automatically isolate the core set of shared knowledge among SDSS researchers worldwide without resorting to a human perusal of the publications which would be too time consuming.

Acknowledgement

The work reported here is in part supported by the National Science Foundation under Grant No. 0612129.

References

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.

Lopes A. A., Pinho R., Paulovich F. V., Minghim R. (2007) Visual text mining using association rules. *Computers & Graphics*, 31, 316–326.

SanJuan E., Ibekwe-SanJuan F., (2006). Textmining without document context. *Information Processing & Management*, Special issue on Informetrics II, Elsevier, 42(6), 1532–1552.