Visualization of Research Collaboration Patterns in SDSS and SARS Research

Jian Zhang and Ting Zhang

Abstract— Various sciences have shown different research collaboration patterns. This study extended previous studies in terms of research domains and methods. We choose examples of large-scale scientific projects and natural disasters for comarision of different sciences. Besides traditional methods, information visualization method, the co-authroship network, was used in this study. Significant difference among the two domains were revealed. Research collaboration in SARS community is dispersive, with many small gropus inter-connected by a few transitional authors. Scientists in SDSS worked closely, showing heavily linked co-authorship networks. This study demonstrates the great potentials of information visualization in terms of revealing the different collaboration patterns among various sciences.

Index Terms— Research Collaboration, Informatino visualization, SARS, SDSS, Co-authorship Network.

1 Introduction

It has been documented that the level of research collaboration in different science domains varies a lot. For instance, Yoshikane and Kageura [8] compared the development of personal collaboration networks in four different domains, including electronic engineering, information processing, polymer science, and biochemistry, and found that in biochemistry domain researchers were collaborating with a relatively large number of partners, whereas in information processing the number was low. The difference, they thought, is caused by the difference in subject matter or research styles. In response to the tide of proving de Solla Price's prediction [3] that single authorship would extinct in the 1980s, O'neill [6] compared the authorship patterns in theory based journals versus research based journals in the education domain. His findings contradicted to Price's prediction since single authorship was still dominate the theory based journals. Abt [1] observed the multinational authorship in 16 sciences such as Astronomy/Astrophysics, Biology, Engineering, Mathematics, Medicine, and etc. The range of average percentage of multinational authored papers in the 16 disciplines' leading journals is from 13% in Surgery to 55% in Astronomy. No factors tested in his study clearly contributed to the difference. But Abt hypothesized that the objects studied in different disciplines might cause the difference of authorship pattern. He supported his hypothesis from a manually check of contents in Surgery and Astronomy literature, but no rigorous test had been conducted afterward.

These studies compared authorship patterns of various sciences in discipline levels or their sub-disciplines. In the current society, however, two events are strongly boosting the development of science research and impact research collaboration. They are large-scale scientific projects such as Human Gene Project and natural disasters like Mad Cow disease. Very few bibliometric researches have devoted to this two kinds of events in terms of their impact to the research collaboration except for [9]. It seemed appropriate to conduct a study in the fine granular level to compare impacts of these two kinds of events on the research collaborations.

- *Roy G. Biv is with Starbucks Research, E-Mail: roy.g.biv@aol.com.*
- *Ed Grimley is with Grimley Widgets, Inc., E-Mail: ed.grimley@aol.com.*
- *Martha Stewart is with Martha Stewart Enterprises at Microsoft Research, E-Mail: Martha.stewart@marthastewart.com.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

The previous studies mainly utilized the number of co-authorships as indicators of research collaboration and collect the statistics, presenting in abstract formats like numbers, tables, or diagrams. For example, Yoshikane presented four correlation diagrams of relation between the number of samples and transition variables, which are hard to be conceived even after carefully reading their descriptions. These abstract formats are good at telling the facts, but hardly showing readers the research collaboration in intuitive ways. In this study we advance one step further. We not only depict the authorship patterns in abstract formats like numbers, tables, and diagrams, but also delineate patterns the from information visualization perspective.

This study chooses Sloan Digital Sky Survey (SDSS) project as the example of large-scale scientific project and Severe Acute Respiratory Syndrome (SARS) as the example of natural disasters. The SDSS is an ongoing project in astronomy and astrophysics domain which aims to map the large-scale structure of the universe [7]. Previous studies [9] had found that the SDSSS project impact the research collaboration in astronomy domain. It believed that large-scale projects like the SDSS will change the way by which future astronomers conduct their research.

SARS is a respiratory disease in humans which is caused by the SARS coronavirus. It first appeared in December 2002 and soon spread to more than two dozen countries North America, South America, Europe, and Asia before the SARS global outbreak of 2003 was contained. [1] According to World Health Organization, 8096 people were infected and 774 died by April 2004. In addition, the outbreak of SARS in 2003 had cause national panic in China and great economic damage. But the outbreak of SARS also triggers a large number of researches on this topic. Given its impact and the number of literature, SARS could be considered as a good example of natural disasters occurred in the recent years.

Particular research questions this study tries to answers include:

1) Are the research collaboration patterns different in SDSS and SARS researches?
2) How did the collaboration patterns change along time?
3) Are the changes along time consistent with overall patterns?

2 Method

In order to test our questions, we first collected data from the Web of Science database. Then data were cleaned and divided by years.

[1] http://www.cdc.gov/ncidod/sars/factsheet.htm

Basic statistics were calculated. Later the co-authorship networks were created by CiteSpace [2].

2.1 Data Collection

The literature records of SDSS and SARS were retrieved from Thompson ISI's *Web of Science* (WoS) with search term 'SDSS' OR 'Sloan Digita*' for SDSS research and 'SARS' or 'Severe Acute Respiratory Syndrome' for SARS research. The time span of the retrieval was set up as from 2003 to 2006 because the SARS research mainly started from 2003. For the comparison purpose, this study only collect data before 2007 since records in 2007 may not completed. The records were stored in the format of "full records+ reference" into a local computer for further clean.

The abbreviation of 'SDSS' and 'SARS' could stand for many other phrases other than Sloan project or the diseases. So records were imported into HistCite [4] for cleaning up. According to the type of papers, in this study only Articles and Reviews were included. Papers from titles that clearly have nothing to do with the two events were excluded. The final dataset includes 1063 records for SDSS research and 2775 for SARS.

In order to compared with the previous studies, basic statistics of the authorship indicators were collected, such as the number of papers, journals, authors, unique authors, and the yearly output of those indices. Authors' names come from the ISI's Distinct Author identification system to 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- and SARS-based research.

2.2 Visualization of co-authorship network

Differ from the previous studies, this study not only presents the statistics of various co-authorships, but also depicts the research collaboration from information visualization perspective so that new insights can be introduced into the results.

This study used co-author networks for the visualization purpose. A co-author network is the collaboration graph where a node represents an author and links among nodes represent the co-author relation among authors [5]. Co-authorship networks bring new insights into the studies of research collaboration, such as the highly co-authored groups, the density of co-authorship, and etc.

A bibliometric software, CiteSpace, was used to create the co-authorship network. Given the volume of literatures, this study set up the threshold of co-authorship to 5, which means two scientists who have co-authored five times or more will be plotted on the network.

As the co-authorship network is a dynamic graph, changing along with time, the overall co-authorship graph of the four years may conceal the specific feature in each individual year. We plotted the co-authorship network in each year from 2003 to 2006 so that the comparison could occur in both overall level and fine granular level.

Co-authorship networks can tell the story intuitively. In order to supports the conception obtained from observing co-authorship network, we collected the number of nodes and links and calculate the link density (No. of Links divided by No. of Nodes) in each network.

3 Results

The results consist of three parts. The first part lists the similar indicators that had been employed in the previous studies, like the number of authors, journals, co-authorships and the distribution of journal-papers numbers. The second part lists the co-authorship networks of SDSS and SARS. Both networks covering all four years and networks of each year are shown. The third part shows the link density in each network in order to verify the results of visualization.

3.1 Indicators of authorship in SDSS and SARS Research

From 2003 to 2006 there are 1063 publication pertinent to SDSS projects, meanwhile 2775 publications pertinent to SARS. Table 1 and 2 summarize the number of publications, journals (including conference proceedings), unique authors, and average authorships.

Table 1. Indicators of authorship in SARS

Year	Papers	Journals	Unique authors	Average authorship
2003	420	206	2326	7.1
2004	818	336	4043	7.0
2005	795	395	4234	7.2
2006	742	361	3613	6.0

The SARS related publication shows a burst in 2003 and 2004. The total 420 published studies about this disease popped up in 2003 and almost doubled in 2004. But the number started to decrease in the rest two years. The number of journals and the size of SARS research community have the same trends, increasing from 2003, reaching the peak in 2005, and starting to drop down in 2006. The average authorship in SARS-related research did not change too much in the first three years, mainly around seven authors per paper. In 2006 the number dropped down to six.

Table 2. Indicators of authorship in SDSS

Year	Papers	Journals	Unique authors	Average authorship
2003	163	15	620	10.0
2004	213	19	838	8.9
2005	265	22	1034	8.0
2006	422	23	1703	7.4

The SDSS related publications show a different pattern compared with SARS authorship. All of the number of papers, journals, and unique authors keeps increasing. The average authorship, however, keeps decreasing from 10 to seven.

From 2003 to 2006, the number of SARS researches is almost triple of the SDSS researches. The number of journals that cover SARS is more than ten times of those cover SDSS research. The average authorship in SARS is a little bit lower than the number in SDSS.

Figure 1 shows the journal-paper distribution. The *x*-axis is the journal ID sorted according to the number of papers published in each journal (from high to low). The *y*-axis is the number of papers in each journal.

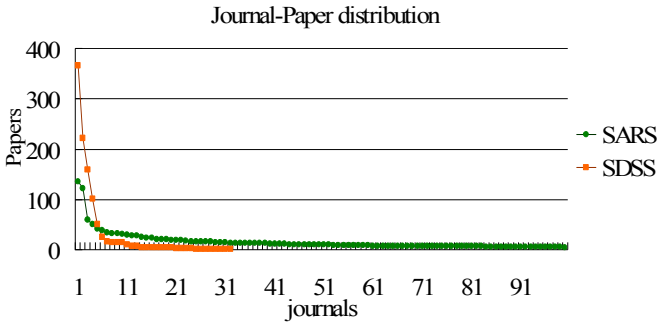

Fig. 1. Journal-Paper distribution from 2003 to 2006.

The journal-paper distribution of SARS research has different pattern compared with SDSS research. In SARS research domain, few journals have published more than 100 papers, but the majority of journals stay in the "long tail" with two-digit number of papers published in the four years. The SDSS research domain has several

"core" journals that published hundreds of papers in the four year period, meanwhile the majority of journals published less then ten papers. Most of them published only one paper about SDSS project.

3.2 The co-authorship networks

Based on the threshold (c, cc, ccv = 5, 5, 20%), the co-authorship networks were created in CiteSpace. Figure 2 shows the co-authorship of scientists working on the SARS in the overall four years; meanwhile figure 3 lists the co-authorship network in each year.[2]

In the co-authorship networks, each circle stands for one author. The size of a circle denotes the number of articles that a given author published in the dataset. The colour of a link shows the first time the two connected authors published an article together. The colours of an author show the number of articles the author published over time. The time-coloured rings progress inside out {Chen, 2005 #11}. Nodes with a thin bright pink circle on their outer layer are the transitional scientists who connect two or more groups, like Yuen_KY, Chen_PKS, and Peiris_JSM in figure 2.

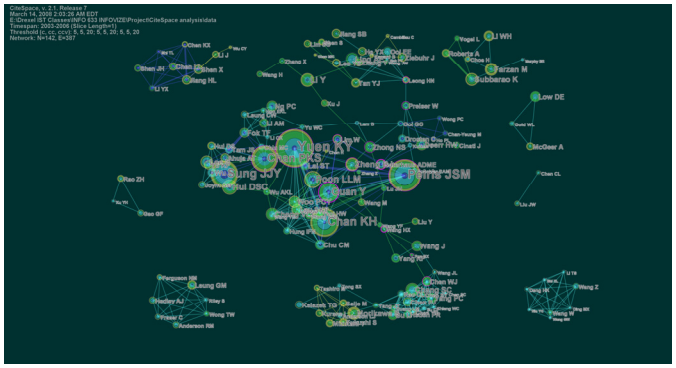

Fig. 2. The overall co-authorship network in SARS research.

The overall co-authorship network of SARS scientists shows a discrete pattern. There are tens of individual groups, which have many inner group co-author relations and have no or just a few external co-author links.

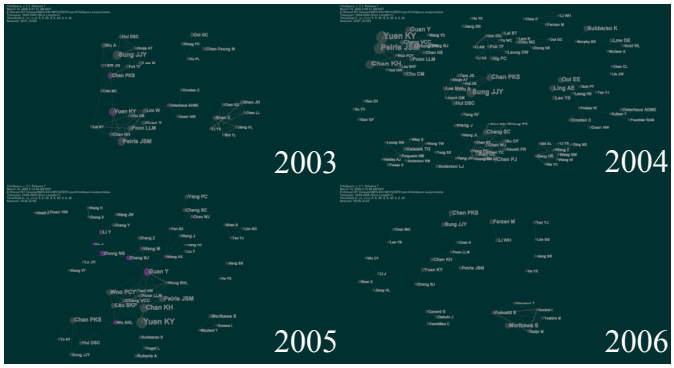

Fig. 3. Co-authorship network in SARS research from 2003 to 2006.

In figure 3, all of the co-authorship networks in each year share the similar patterns with others and the overall network. Scientists devoted to SARS research tended to work "locally," merely working within small groups. Only a few scientists have external co-author relations with other groups, forming the transitional nodes, such as Guan_Y in the 2005 network who was the only one that connect the upper groups and lower groups. The evolution of the SARS co-authorship experienced a boom in from 2003 to 2004 with more groups of authors showing on networks, but from 2005 to 2006, the number of groups decreased. There were no transitional authors in both 2004' and 2006's co-authorship network.

Figure 4 shows the co-authorship of scientists working on the SDSS project in the overall four years; meanwhile figure 5 lists the co-authorship network in each year.

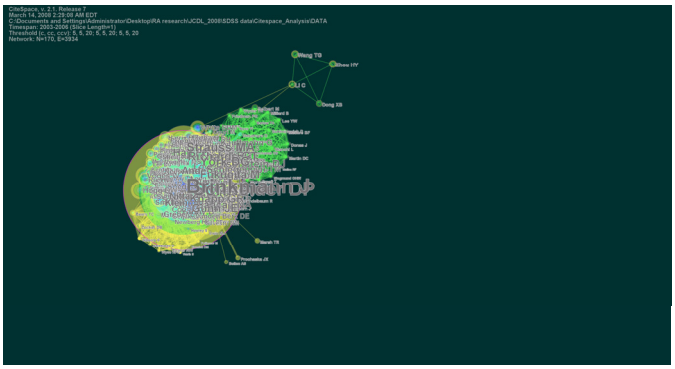

Fig. 4. The overall co-authorship network in SDSS research.

The co-authorship networks of SDSS scientists depict a totally different collaboration patterns compared with SARS domain. The overall network in figure 4 has two major author groups with heavily inter-connected links. Within each of the two groups the inner links are dense, which makes identifying individual author hard. Four authors form a small co-authorship group on the top-left corner in figure 3.

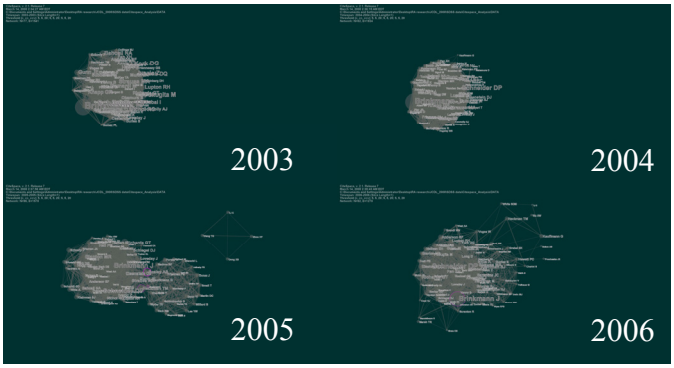

Fig. 5. Co-authorship network in SDSS research from 2003 to 2006.

Figure 5.shows the evolution of the SDSS co-authorship networks. In 2003 and 2004 the SDSS co-authorship networks are both one heavily inter-connected "ball." SDSS Scientists frequently co-authored, and no transitional scientists existed in the two years. From 2005, the "ball" started to divide into two large groups, and the small group showing on top-right corner of figure 4 also formed in this year. In the middle of the two large groups are three transitional authors, Szalay_AS, Heckman_TM, and Budavari_T. In the 2006, the three groups merged a lot, overlapping more than they did in 2005.

3.3 Comparison of co-authorship network density

Based on the same threshold, we collected the number of authors (nodes) and the co-authorship relations (links) in each networks. The densities of network were calculated too. Table 4 lists the numbers.

Table 4. Number of nodes and links in co-authorship networks

	SARS			SDSS		
Year	Nodes	Links	Density	Nodes	Links	Density
2003	31	85	2.7	77	1541	20
2004	97	263	2.7	82	1924	23

[2] The high-resolution version of all figures is available at http://nevac.ischool.drexel.edu/~james/images/COA-JPG.rar

2005	48	85	1.7	96	1570	16
2006	29	33	1.1	92	1270	13
Overall (03-06)	142	387	2.7	170	3934	23

It is clear that SDSS co-authorship networks are denser (nearly times) than the SARS networks. Networks in both domains experienced the same trend, increasing in 2003 and 2004, and then dropping down.

4 Discussions

Responding to our research questions, the results supported that the research collaboration patterns are different in SDSS and SARS researches. The co-authorship networks reveal significant difference between research collaboration in SARS and in SDSS. In each domain the collaboration patterns did not change a lot and the changes along time are consistent with overall patterns.

From the statistics of publications in SARS and SDSS, there are different publication environments in the two domains. The community of scientists working on SARS abruptly increased in the first two years, which might results from the high mortality rate of SARS and its sudden outbreak. When the disease was under controlled, and passion on this topic cools down, the number of research decreased. In contrast, the SDSS research experiences stead expansion [9]. This study, however, have observed only four year data, future studies are needed to view the longitude trends of SARS publications.

An interesting finding regard to the publication environment is that the SARS community have a long list of journals and proceedings to choose for submission. It seems astronomers have to compete in a relatively small number of journals. Future studies are needed to find out if this factor impacts the research collaboration patterns.

The average authorship can hardly tell the differences of research collaborations in the two domains. Zhang [9] had found that the SDSS average authorship peaked in 2000 with 17 authors per paper, and then keeps decreasing to seven, which is very close to the average authorship in SARS literature. Merely based this number, there is no clear clue how the research collaboration in SARS community differs from SDSS community.

The co-authorship networks, however, reveal the difference between the two domains. The overall co-authorship network in SARS represents the dispersive pattern of research collaborations. Scientists in this domain tend to work "locally" within their groups, only few people play the transitional role among different groups. The SDSS co-authorship networks reveal a highly collaborated community. Link density of co-authorship networks bolsters this observation. The SDSS community have co-author links nearly ten times as the SARS community, which results in the heavily inter-connected "balls" showing in figure 4 and 5.

This difference among research collaboration is constant in both domains, changing very little along the four years. The individual co-authorship networks in figure 3 depict similar patterns, dispersive groups with few transitional authors. Even though the SDSS co-authorship network in 2005 split into two large groups, the inter-connection among the two groups is still strong. In 2006, the two groups merged and reduced the transitional players between them from three to one, Brinkmann_J.

The results in this study reveal the potential of information visualization when this method is employed in comparative studies of different domains. The results shown in numbers, tables, and diagrams do tell the facts. Large amount of information, however, have been lost. The co-authorship networks shown above not only clearly reveal the differences between SARS and SDSS research collaboration patterns, but bring new insights into this study. For example, in SARS community, the transitional authors could be considered as gatekeepers who control the information flow in and out between different groups, especially their own groups. If they block the information flows, no more information share become available to different groups. As SDSS members share the data information through a powerful database, the SDSS Archive, their collaboration faces fewer obstacles than does the SARS community. Therefore, few transitional players exist in the SDSS co-authorship networks.

5 Conclusion

This study analyzes the difference of research collaboration patterns in SARS and SDSS studies. Through combination of the traditional methods used in the previous studies of research collaboration in various sciences and the method of information visualization, we found that:

1) The research collaboration pattern in SARS is significantly different from the patterns in SDSS community.
2) Scientists working on SARS research tend to work locally within their own research groups, and a few people play the transitional role among the community.
3) SDSS scientists are like to work closely with others, having higher co-author relations than the SARS community.
4) Information visualization shows great potential in terms of revealing the collaboration patterns in this study.

This study also needs further work for improvement. For example, the sample size of publications is different in SARS (2775 papers) and SDSS (1065 papers) literature. Will this affect the results? To verify the co-authorship networks, this study uses link density as the prove method. This method is simple in terms of comparison of different graph structures. Further studies on the validation methods are needed.

Acknowledgements

This work was supported in part by a grant from the NSF under grant # IIS-0612129.

References

[1] H.A. Abt, The Frequencies of Multinational Papers in various Sciences, Scientometrics 72 (1) (2007) 105-115.

[2] C. Chen, 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) (2006) 359-377.

[3] D.J. de Solla Price, Little Science, Big Science ... and Beyond, Columbia University Press, New York, 1986.

[4] E. Garfield, A.I. Pudovkin, V.S. Istomin, Why Do We Need Algorithmic Historiography?, Journal of the American Society for Information Science and Technology 54 (5) (2003) 400-412.

[5] E.L. Logan, W.M.J. Shaw, An Investigation of the Coauthor Graph, Journal of the American Society for Information Science and Technology 38 (4) (1987) 262-268.

[6] P.G. O'neill, Authorship Patterns in Theory based versus Reseach based Journals, Scientometrics 41 (3) (1998) 291-298.

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

[8] F. Yoshikane, K. Kageura, Comparative analysis of coauthorship networks of different domains: The growth and change of networks, Scientometrics 60 (3) (2004) 433-444.

[9] J. Zhang, M. Vogeley, C. Chen, Scientometrics of big science: A case study of research in the Sloan Digital Sky Survey, Scientometrics (Accepted) (2008).