MODELING SCHOLARLY COMMUNICATIONS ACROSS HETEROGENEOUS CORPORA

Xin Shuai

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics & Computing,
Indiana University
August 2014
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Prof. Johan Bollen

Prof. Yong-Yeol Ahn

Prof. Ying Ding

Prof. Xiaozhong Liu

August 8th, 2014
ACKNOWLEDGMENTS

The completion of my PhD career and dissertation is a milestone in my career. I would never possible to achieve it without the generous help from my research committee, my colleagues, my family and my friends. I want to express my sincere gratitude to all people who ever helped my PhD research and dissertation writing.

First of all, I would like to thank my PhD advisor, Prof. Johan Bollen. Johan is very smart, knowledgeable and easy-going. He generously supported my overall PhD work both financially and spiritually. He’s always passionate about his research and striving for perfection. I have learned a lot from him, about how to formalize a scientific idea, how to write a scientific paper, how to testify a research hypothesis, and how to deliver scientific findings to others. Discussion with him always inspired me with new and exiting ideas.

Second, I would like express my thankfulness to all other other professors who gave me a lot of help in my PhD research. Especially, I want to thank Prof. Ying Ding and Prof. Xiaozhong Liu, with whom I collaborated in many research projects constituting the backbone of my dissertation. Both of them provided me with quite useful guidance on my study, and lots of research opportunities. In addition, I’m also very grateful to Prof. Yong-Yeol Ahn for his scientific advice and many insightful discussions with me.

Finally, I would like to thank my family, especially my mother Li Liu. She is a great mom who gives me life and lots of love. She is always supportive of my work and keeps encouraging me when I came across difficulties in life and study.
Scholarly communication plays a pivotal role in science. Recently, the process of scholarly communication has been fundamentally changed by the emergence of digital scholarship and social media. These changes have great potential for the democratization of science and scholarship yet challenge existing scientific norms and processes that rely on careful and deliberate review of scholarly value.

In this dissertation we introduce novel methods to study scholarly communication from large-scale, heterogeneous data sets that are generated as a product of digital scholarship, making fundamental contributions in modeling scholarly communities, the flow and exchange of scientific knowledge in scholarly networks, and the role of social media in shaping scholarly impact and communication. We outline our research in four sections of this dissertation.

- First, we investigate the emergence of scholarly communities by probabilistically modeling research topics from large-scale bibliometric data to provide a dynamic perspective on the evolution of topic- and author-based communities. This work addresses the difficult question of ranking authors and publication venues according to their impact within the context of a community, while accounting for the dynamics of topic changes.
• Second, we investigate how scientific knowledge propagates through scholarly
networks. By drawing an analogy between international trade and scientific
communication, we measure the flow of ideas between scientific domains, i.e.
intellectual trade, on the basis of the predominant direction of journal citations.

• Third, we analyze the online response of the scientific community to the publication
of scholarly articles using pre-print downloads, Twitter mentions, and early citations
data, using a single cohort of arXiv preprints. We find that Twitter mentions for the
selected cohort ramp up within days of article submission and most intriguingly that
they are correlated with later article downloads and early citations, indicating that
social media attention may shape scientific impact.

• Fourth, we compare the impact of papers, scholars, and topics as measured by
different measures, each derived from a different medium for the exchange of
scientific knowledge, in this case scholarly citations vs. Wikipedia mentions. The
results show that citation impact and Wikipedia impact are positively correlated.
CONTENTS

1 Introduction ........................................... 1
  1.1 Why Study Scholarly Communication? .................. 1
  1.2 Definition and Evolution of Scholarly Communication .... 3
  1.3 Bibliometrics and Scholarly Communication ................. 9
    1.3.1 Citation Analysis .................................. 9
    1.3.2 Altmetrics .......................................... 13
  1.4 Research Questions and Topics .......................... 14
    1.4.1 Community Based Scholarly Ranking .................. 15
    1.4.2 Intellectual Trade between Scientific Disciplines ....... 17
    1.4.3 The response of Twitter to Scholarly Articles ........... 18
    1.4.4 The Comparison between Wikipedia and Scholarly Ranking ... 20

2 Literature Review ...................................... 22
  2.0.5 Classic Bibliometrics ................................ 22
  2.0.6 Webometrics .......................................... 23
  2.0.7 Web Usage Data ....................................... 24
  2.0.8 Social Media and Science .............................. 25

3 Scholarly Community Detection based on Dynamic Research Topics 29
  3.1 Introduction .......................................... 30
3.2 Problem Definition ........................................... 31
3.3 Related Work ..................................................... 35
  3.3.1 Community Detection ........................................ 35
  3.3.2 Topic Modeling ................................................ 35
3.4 Methods ............................................................ 37
  3.4.1 Datasets ......................................................... 37
  3.4.2 Algorithms ...................................................... 37
3.5 Experimental Results ............................................ 42
  3.5.1 Analysis from Static Perspective ............................. 42
  3.5.2 Analysis from Dynamic Perspective .......................... 48
3.6 Discussion .......................................................... 51
  3.6.1 The Dynamic Functionality of DCTM ......................... 51
3.7 Conclusion and Future Work ..................................... 57

4 Modeling Intellectual Trade among Scientific Disciplines 58
4.1 Introduction ........................................................ 58
4.2 Methodology ....................................................... 61
  4.2.1 Citation and Classification Data .............................. 62
  4.2.2 Domain Network Generation ................................. 63
4.3 Results ............................................................. 64
  4.3.1 Domain size relative to trade volume ....................... 65
  4.3.2 Trade Balance .................................................. 68
  4.3.3 Dependence on External Trade ............................... 74
  4.3.4 Direction of Foreign Trade ................................... 75
  4.3.5 Trading Center ................................................ 80
  4.3.6 Bilateral Trade and Transit Trade ......................... 85
CHAPTER 1

INTRODUCTION

1.1 Why Study Scholarly Communication?

Science has propelled the single largest cultural, social, and economic expansion since the beginning of human history (Russell, 1952). It plays a pivotal, but perhaps insufficiently acknowledged, role in our civilization, improving and enabling the lives of billions of individuals, constantly accumulating, organizing, and developing new knowledge that expands and unlocks human potential.

To provide only a few notable examples, Newton’s theory (Newton, 1999) laid the foundation for modern engineering and space technology. Information Theory (Shannon, 2001) revolutionized communication technology and laid the foundation for the digital revolution of the past 40 years. Biological, medicinal, and pharmaceutical advances have lead to longer life spans, better nutrition, and better health for billions. The discovery of DNA (Watson & Crick, 1953) unlocked one of life’s most profound secrets and enabled entire new branches of medicine and biology, possibly leading to the creation of artificial life. Science has profoundly altered and improved the human condition, and will most likely continue to do so, potentially at accelerating rates.

Although science plays such a pivotal role, we understand very little of it. Pioneering studies attempted to characterize science form a qualitative and philosophical level (Popper,
2002; Kuhn, 1970), but in order to investigate science at a quantitative level, researchers have sought to develop systems to measure the key properties of science as a system. In Lord Kelvin’s words (Thomson, 1883): “I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind...”

Early pioneering efforts to study science from a scientific and quantitative perspective focused on the establishment of data representative of scholarly communication and methods to extract indicators of scientific impact or activity. Garfield (1955a) proposed to evaluate scientific outputs based on citation records and established the SCI citation index for retrieval and ranking. Solla Price (1974) measured the number of authors and number of publications by studying the distribution of scholar’s outputs demonstrating power-law distribution properties in author-publication data, for example 25% of authors published 75% of papers (Solla Price, 1965).

These pioneering works laid the foundation for bibliometrics and scientometrics, i.e. the scientific study of science itself, aimed at understanding how science works both at the micro- and macro-scale from objective data to measure research outputs, research activities, production, and common practices. Like the green revolution improved crop yields by leveraging plant science (Eicher, 1995), bibliometrics and scientometrics can help improve the effectiveness of science and the yield of scientific work by allowing scientific committees to assess research quality and better allocate research fundings\(^1\).

How is science measured? Science is a system of knowledge production and accumulation that builds upon scholarly communication. Therefore, the key to understanding and measuring science is to quantitatively study the process of scholarly communication. The

\(^1\)http://www.ref.ac.uk/
latter can be quantified by a number of features relating to its measurable outputs; papers, authors, citations, etc. It is therefore the focus of this thesis, in particular its changing role in the era of digital scholarship and online social media.

1.2 Definition and Evolution of Scholarly Communication

The process through which we acquire scientific knowledge has fundamentally changed over time, in particular through the formalization of the scientific method starting in the 17th century (Whewell, 1837). This dissertation is not concerned with a philosophy of science, nor a better understanding of the scientific method itself, but the central role that scholarly communication plays in shaping and changing science. Our work is situated in the emerging domain of Science of Science Policy (SOSP), a science concerned with studying the scientific enterprise in its own right. NSF has awarded grants from the Science of Science and Innovation Policy (SciSIP) program\(^2\) and US federal government published a report to emphasize the importance and applications of SOSP in 2008 (Science & Policy, 2008).

Modern scientific practice, i.e. the set of procedures and methods followed by most practicing scientists, now commonly proceeds along a number of well-defined steps, which Creswell (2004) can be outlined as follows:

1. Identification of research problem
2. Literature review
3. Specifying the purpose of research
4. Determine specific research questions or hypothesis
5. Data collection
6. Analyzing and interpreting the data
7. Reporting and evaluating research
8. Communicating the research findings and, possibly recommendations (publication)

\(^2\)http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501084
Most, if not all, of Creswell (2004)’s steps involve or require scholarly communication of some form or other. This is not surprising. The scientific method consists of a process in which existing theories and models are continuously subjected to empirical and theoretical scrutiny. Scientists can therefore not conduct their research without a strong awareness of existing work. Hence the scientific enterprise necessitates extensive communication and knowledge sharing (Lievrouw, 1992). This is the role of the scholarly communication process. In fact, we argue that it is impossible to study science without understanding the role of scholarly communication in shaping science.

According to Wikipedia, “scholarly communication is the process of academics, scholars and researchers sharing and publishing their research findings so that they are available to the wider academic community (such as university academics) and beyond”. Scholarly communication is not a monolithic, centralized system. It can take many forms. For example, scientists establish formal and informal social networks through which they communicate their findings. They may participate in workshops, seminars, and conferences, and disseminate the resulting exchanges in proceedings. Visiting scientists can make presentations to colleagues at the invitation of their respective institutions. Researcher can write white reports, or publish peer-reviewed journal articles. They can send emails, release software, and share data with other scientists.

Among all scholarly communication channels, the peer-reviewed journal publication has emerged as a canonical standard, most likely because of its formalized, systematic publication process that includes a form of stringent quality control, namely peer review, and formalized methods to ensure open, standardized delivery of scientific results (Spier, 2002). In most scientific domains a peer-reviewed journal publication is considered the most valuable, useful, and prestigious form of scholarly communication.

The publication of peer-reviewed journal publications rest on a number of key processes
which we will briefly elucidate below:

- Academic publishing: a system for the dissemination of scientific publications that relies on publishers, commercial or otherwise, to manage the submission, peer-review, and publication process. The traditional system has recently undergone significant changes with the advent of not-for-profit and Open Access publishing (Suber, 2010), but many of the latter still follow a similar process of submission, reviewing, and publication that is handled by a specific publishers that acts as a third party between the publishing scientist and its audience of other scientists. It must be noted that academic publishing is not merely focused on peer-reviewed journal publications, but may include book, report, dissertation, and news publishing, and remains largely a for-profit enterprise for better (Woll, 1998) or worse (Van Noorden, 2013).

- Peer-review: Peer review consists of a process in which a work is evaluated before being accepted for publication by the peers of its author (Spier, 2002). The “peers” are selected by the publisher, or the editor acting on behalf of the publisher, to be individuals who through their experience, background, and interests are most qualified to assess the scientific validity and value of a particular publication. Peer review thus functions as a threshold to publication, a pre-publication requirement, that is implemented by soliciting multiple reviews from an author’s peers. Although increasingly criticized for its role in the academic publication process (Smith, 2010) and its shortcomings as a means of performing a pre-publication evaluation of the quality, relevance, and significance of a scientific work (McCook, 2006), it remains the gold standard in the academic publication process.

- Citations: Given that science is an incremental, evolutionary process that builds on previous advances, challenging existing models and theories, confirming or invalidating them through continued empirical scrutiny, the citation remains a crucial element
of any scholarly publication. In its basic form a citation is a reference to previous
publications that the author wishes to acknowledge as a scholarly influence on his or
her own work (Neville, 2012), so that the reader of the paper in question can clearly
appreciate the originality and origins of the ideas that underpinned the author’s own
work. In more practical terms, a citation consists of metadata that is intended to
be sufficient for a reader to locate and retrieve the original work, often formatted as
an abbreviated key that may include the authors’ last names, year of publication,
journal, volume, issues, and page numbers\(^3\). Citations thus connect one article to the
other, by acknowledging the author’s intellectual debt to the precursors of his or her
own publication. In subsequent sections of this dissertation we will discuss in greater
detail the construction and use of large-scale citation network in bibliometric research.

The above three elements have remained vital to scholarly communication for decades
and perhaps centuries, but are now undergoing rapid changes. Many publications, and even
scientific research itself, are “born digital”, i.e. they are created, developed, and dissemi-
nated in digital form (Rumsey, 2011). Some have speculated this may lead to a profound
democratization of science, in terms of how scientific results are disseminated among scien-
tists and to the public (Hornby & Bussert, 2013).

Digital media are increasingly replacing brick-and-mortar libraries and printed mate-
rials (Hollander, Krugman, Reichert, & Adam Avant, 2011). This evolution is leading to
the emergence of so-called “digital scholarship”, i.e. the pervasive use of digital technol-
ogy for searching, publishing, and storing scientific works (Rumsey, 2011). In recent years
this evolution has extended beyond the use of digital resources into new online ecology
for scholarly communication. Digital scholarship is moving to online environments, such
as the WWW (Cullen & Chawner, 2011) and social media (Ullrich et al., 2008), further
\(^3\)http://web.archive.org/web/20071228053242/http://www.lesley.edu/library/guides/citation/
apa.html
accelerating a trend towards cheap, efficient, and open dissemination of scientific results, with serious implications to the formal publication and vetting cycle that has characterized scholarly communication for the past century, as well as for-profit academic publishing.

Particularly, the emergence of social media brings digital scholarship to a fundamentally new realm, where communication is even faster and open, and scientific materials are now discussed and reviewed by thousands of people instead of two or three scholarly peers. Social media creates an online virtual community where people are connected to each other and can create, share or exchange information within their community and larger network. Kaplan and Haenlein (2010) define social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.”

In the context of digital scholarship, many new elements are added into the traditional scholarly communications:

- Online viewing and downloading: The emergence of open-access journal (Suber, 2010) makes possible the unrestricted online access to peer-reviewed scholarly research. Authors can provide open access by self-archiving their journal articles in an OA repository (e.g., arXiv ⁴), or by publishing in an open access journal. (e.g., PLoS ⁵). Either way, readers, notably not limited to academics and scholars, can freely view and download articles online. Especially in OA repositories, scholars can get access to articles even before they are formally published.

- Hyperlink: A hyperlink is a reference to a resource in WWW that the reader can click to retrieve that resource. In digital scholarship, the same hyperlink approach can link news articles or blogs to scientific articles published in scientific journals. This is considered a new means of “citation” in the context of digital scholarship, enabling

---

⁴http://arxiv.org/
⁵http://www.plosone.org/
access to scholarly resources by the general public as well as those in the scientific community.

- Post: Users in social media can write short messages in their personal timelines to update their status or express interest. If a user is interested in science, he/she can post messages about current science progress or his/her own research. Particularly, users can attach a short URL to science news or articles they like. This is a big change in scholarly communication, since the public can freely participate in the dissemination of scientific thoughts or recommending scientific work without going through the peer-reviewed journal publication channel.

- Connection: In social media sites, in addition to posting personal messages, users can subscribe to other users’ timeline and receive their messages. Users who are interested in science, can subscribe to other scholar’s account to follow their research updates, receive their recommended papers and even discuss about future collaboration. The informal scholarly communication within the context of social networks constitutes a online scholarly community.

- Forward/Repost/Share: Once users see interesting content from their friends in social networks, they can further spread the content to their other friends. In social media environments, scientific information rapidly propagates at large scale to broad audiences. In fact, hyperlinks to an online manuscript can be posted to millions of social media users within a few hours of its online publication.

In summary, publishing, peer-review, and citation have a pivotal position in scholarly communication; Digital scholarship and Web enriches traditional scholarly communication, moving it from an off-line communication to an on-line one; Social media deepens this trend providing even more opportunities for scholars to recommend, disseminate, and discuss scientific materials at a unprecedented scale, speed and scope (see Table 1.1). Scholarly
communication is rapidly diversifying and leading to the emergence of large-scale of hetero-
geneous corpora that record scholarly activities along an increasing number of dimensions, e.g. citation data, readership, online citations, blog mentions, etc. This creates significant opportunities and challenges for scientists who study Science of Science.

Table 1.1: Evolution of scholarly communication

<table>
<thead>
<tr>
<th>Type</th>
<th>Traditional</th>
<th>Digital</th>
<th>Online</th>
<th>Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Library</td>
<td>Computer</td>
<td>Web</td>
<td>Web</td>
</tr>
<tr>
<td>Form</td>
<td>Print</td>
<td>CD-ROM</td>
<td>Open Access</td>
<td>Message, short URL</td>
</tr>
<tr>
<td>Interaction</td>
<td>Read, Play</td>
<td>View and Download</td>
<td>Post, Forward, Reply</td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td>Citation</td>
<td>N/A</td>
<td>Downloads, Click Logs</td>
<td>Number of tweets, Number of retweets</td>
</tr>
</tbody>
</table>

1.3 Bibliometrics and Scholarly Communication

The term bibliometrics was coined by Alan Pritchard in a paper published in 1969 (Pritchard, 1969). He defined the term as “the application of mathematics and statistical methods to books and other media of communication”. Bibliometrics, sometimes called “scientometrics”, aims to evaluate scientific outputs and investigate the structure and development of science by analyzing scholarly communication from both a quantitative and qualitative perspective. The main goals of bibliometrics are: one, to measure the scholarly impact of articles, scholars, journals, and institutions (impact analysis); and two, to investigate the relationships among different scientific disciplines (relational analysis) (Thelwall, 2007). Both of them provide insight about the development and output of science, based on which science policy-makers and funders can inform their decision-making.

1.3.1 Citation Analysis

Citation analysis plays an important role in bibliometrics which is often focused on the examination of citation frequencies, patterns, and graphs for collections of articles and books (Garfield & Merton, 1979a). To facilitate citation analysis, bibliographic databases
were created, commonly referred to as “citation indices”, that index citation data between publications. They provide users easily accessible information about the citation relationships among publications which is a necessity for citation analysis as part of bibliometric research. In 1960, Eugene Garfield’s Institute for Scientific Information (ISI) established the first citation index, e.g. the Science Citation Index (SCI). The Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI) were added later. The first online and automated citation index was created by CiteSeer in 1997 by crawling the web for scholarly publications. Google Scholar was introduced in 2004, operating on similar principles.

A common application of citation analysis is to extract indicators that reflect the scholarly impact of scientific papers, scholars, venues, or institutions. For example, the Journal Impact Factor (JIF) express the impact of a journal in terms of the average number of citations to its articles over a 2 year period (Garfield, 1972). It is frequently used as a proxy for the relative importance of a journal within its field, where journals with higher impact factors are considered more important than those with lower ones. The impact factor was first proposed by Eugene Garfield, the founder of the Institute for Scientific Information, in the 1950s, and has been published on a yearly basis in the Journal Citation Reports (JCR) since. Another example of a citation-based impact indicator is h-index (Hirsch, 2005), which attempts to measure both the productivity and impact of a scientist or scholar. The index is based on the distribution of the rank-ordered citations to a scientist’s bibliography. The index can also be applied to assess the productivity and impact of a group of scientists, such as a department, university, or country, as well as a journal. Bollen, Rodriguez, and Sompel (2006) compared two citation-based metrics, i.e. IF and Weighted Pagerank, and proposed a new indicator by considering both metrics to assess journal impact.

Impact indicators provide only a partial view of scholarly impact. The relations between
scholarly works and authors are equally important. Another important application of citation analysis is thus to characterize and visualize the topology of the relations between authors, papers, journals, and domains, the latter leading to the study of interdisciplinarity. This type of analysis is based on the assumption that scientific ideas are exchanged through citation links. Cawkell (2001) first analyzed the information flow between articles through their citations. It is now possible to map the structure of science by clustering authors, papers, and journals based on their co-citation or inter-citation similarities (Boyack, Klavans, & Borner, 2005). In fact, Leydesdorff (2006) compute network betweenness centrality metrics to characterize the interdisciplinarity of academic journals. Yan, Ding, Cronin, and Leydesdorff (2012) use a trading metaphor to study knowledge transfer in the natural and social sciences by tracking citation flows across 221 Web of Science subject categories.

Citation analysis constitutes the backbone of bibliometrics and establishes a well-recognized methodology to measure scholarly impact (e.g. IF, h-index) and track the topology of scholarly relations (e.g. maps of science). In spite of the many contributions of bibliometrics and citation analysis to the understanding of the scientific enterprise and scholarly activities, this paradigm has been challenged recently. As digital technology dramatically changes the scholarly communication and gives rise to new forms of data, it is time to re-examine the assumptions and practices of current bibliometric analysis. We identify the following gaps and issues:

- First, citation is now considered a delayed indicator of scholarly impact (Luwel & Moed, 1998). For citation data to be included in citation indices, a publication has to have entered the published record. In other words, it must be written, submitted, reviewed, and published. Afterwards, newly published articles must cite the given article and these citations need to be recorded as part of a citation index or database as well. This process can take years, and therefore citation data reflects scholarly
activity with considerable delays, sometimes up to 2 or 3 years after the original work was written.

- Second, citation as an indicator of knowledge flow has not been fully leveraged to study the relations across different disciplines. Several existing works (Boyack et al., 2005; Yan et al., 2012) leverage journal citation relationships to construct “maps of science” that visualize the structure and relations between scientific disciplines. However, to fully understand the flow of information in the scholarly community, we need more detailed analysis beyond visualization and the general study of the topology of citation networks. For instance, which domain mainly “export” knowledge? Which domains mainly “import” knowledge? Which domain serves as a bridge for the flow of information between other domains? Such useful information can be used by science policy makers to find emerging research areas that are in need of investment of further funding.

- Third, due to different publication and citation practices within and between different research communities, citation data provides a biased and incomplete bases for the ranking of different publication venues (especially conferences) and scholars, in particular those involved in interdisciplinary research. For instance, Computer Science relies strongly on conference proceedings as a publication venue, which are however cited quite differently and are in fact not included in the Journal Citation reports, thus leading to significant issues in evaluating the impact of publications in the field of Computer Science from citation data. In addition, interdisciplinary scholars can have multiple research areas which are difficult to consolidate and disambiguate in citation data.
### 1.3.2 Altmetrics

Digital scholarship is in the process of revolutionizing scholarly communication, giving rise to many new opportunities in the study of bibliometrics, due to the availability of new scholarly data and new publication venues. The term Altmetrics was first introduced in Neylon and Wu (2009), as digital scholarship’s alternative to the widely used journal impact factor and h-index that are both derived from traditional citation databases. Priem, Taraborelli, Groth, and Neylon (2010) further developed the term and expanded its application to social media indicators. Although altmetrics are often thought of as metrics about articles, they can be applied to people, journals, books, data sets, presentations, videos, source code repositories, web pages, etc. Altmetrics can cover not just citation counts, but also other aspects of the impact of a work, such as how many data and knowledge bases refer to it, how many times the article has been read, viewed, or downloaded, and its mentions in social media and news media (Galligan & Dyas-Correia, 2013).

The promise of Altmetrics lies not in its objectives to leverage new data sources, but its aim to capture the impact of a variety of scholarly resources from perspectives that had not previously been considered in citation analysis.

A classification of altmetrics was proposed by ImpactStory in September 2012\(^6\), and a very similar classification is used by the Public Library of Science (Lin & Fenner, 2013):

- Viewed - HTML views and PDF downloads
- Discussed - journal comments, science blogs, Wikipedia, Twitter, Facebook and other social media
- Saved - Mendeley, CiteULike and other social bookmarks
- Cited - citations in the scholarly literature, tracked by Web of Science, Scopus, Cross-Ref and others

\(^6\)http://blog.impactstory.org/31524247207/
• Recommended - for example used by F1000Prime\(^7\)

Many projects are presently devoted to collect and compute altmetrics, cf. Altmetric.com\(^8\), and CitedIn\(^9\). In addition, many publishers have started to offer Altmetrics services, e.g. PLoS\(^10\), Frontiers\(^11\), Nature Publishing Group\(^12\), and Elsevier. Additionally, some science funders, like UK Medical Research Council (Viney, 2013) also show interests in alternative metrics.

Altmetrics indicators are generally based on web or social media usage, and are therefore considered leading indicators vs. citation data in the assessment of scientific impact (Brody, Harnad, & Carr, 2006b; Nielsen, 2007). However, the precise nature of the relation between the newly emerged Altmetrics and traditional citation is largely unknown. Given the large amount of Altmetrics (i.e., view and download, social media discussion, social bookmark, etc), we are eager to know which of these metrics can truly reflect scholarly value or impact. Since citation counts are generally considered a proxy of a paper/scholar’s scholarly value, it becomes crucial to investigate the correlation between Altmetrics and citation counts.

1.4 Research Questions and Topics

As argued above, the evolution of scholarly communication towards digital scholarship has introduced new challenges to bibliometric and scientometric studies, requiring new models, tools, and indicators to investigate scholarly communication from a deeper and more comprehensive level. Therefore, this dissertation raises the following three research questions as the beginning of our study:

• One, can we detect scholarly communities and rank scholars and venues within specific

---

\(^7\)http://f1000.com/prime  
\(^8\)http://www.altmetric.com/  
\(^9\)http://www.citedin.org/  
\(^10\)http://www.plosone.org/static/almInfo.action  
\(^11\)http://www.frontiersin.org/about/evaluationSystem  
\(^12\)http://www.nature.com/press_releases/article-metrics.html
community?

- Two, can we model the knowledge exchange among different scientific disciplines through citation relationships?
- Three, can we use other indicators from Web and social media other than citations to evaluate scientific outputs?

To address the above three research questions, this dissertation explores the modeling of scholarly communication across heterogeneous corpora in four ways: First, we investigate the emergence of scholarly communities by probabilistically modeling research topics from large-scale bibliometric data to provide a dynamic perspective on the evolution of topic- and author-based communities; Second, we investigate how scientific knowledge propagates through scholarly networks by drawing an analogy between international trade and scientific communication; Third, we analyze the online response of the scientific community to the publication of scholarly articles using pre-print downloads, Twitter mentions, and early citations data, using a single cohort of arXiv preprints; Fourth, we compare the impact of papers, scholars, and topics as measured by different measures, each derived from a different medium for the exchange of scientific knowledge, in this case scholarly citations vs. Wikipedia mentions.

1.4.1 Community Based Scholarly Ranking

In academia, scholarly communities are generally formed by a group of scholars who share research interests. Each community has its representative experts and publication venues that are recognized by scholars from the same community. Traditional bibliometric indicators such as the IF or h-index cannot evaluate a venue or scholar within the context of a specific community (i.e. topic). For instance, PLoS is an open-access journal with very high IF score, and its authority is well recognized in biology and medicine. However, many computer scientists do not consider PLoS a good venue for publication. As another example,
Jiawei Han is a computer scientist with a top-ranked h-index score. However, his reputation is largely built from data mining and network analysis areas, but not from other computer science areas, like image processing or hardware. To capture the interrelation between community and scholarly impact, we need a model to simultaneously detect scholarly impact and rank authors and venues within each community.

Much work in community detection has focused on the structural properties of communities neglecting other important aspects such as topic features (Erdős & Rényi, 1959; Girvan & Newman, 2002; Leskovec, Lang, & Mahoney, 2010). In addition, most structural-based community detection algorithms do not consider the fact that one node can belong to different communities at the same time, therefore they are not very suitable to detect communities from a set of scholars who are very likely to engage in multiple research areas at the same time. Moreover, the structural properties and topic aspects of communities may interact with each other. Common interests may drive the formation of communities, and in turn community structure may reinforce common interests. Few studies have systematically and quantitatively addressed the interaction between the structural and topic properties of communities (Ding, 2011).

Furthermore, social networks and their communities may change over time. Any effort to understand the formation of communities and their topical features needs to include the time dimension. Previous studies have used state space models on the natural parameters of multinomial distributions to analyze the time evolution of topics, or developed the continuous time dynamic model to mine the latent topics through a sequential collection of documents (Blei & Lafferty, 2006; Griffiths & Steyvers, 2004; Iwata, Yamada, Sakurai, & Ueda, 2010). Generally, these studies have applied a set of approaches to approximate posterior inference over the latent topics. However, none considered the community features of the network actors involved in their datasets that might reveal some hidden explanation.
To address those challenging problems for detecting communities by considering their topic features and ranking scholars and venue within each community, we propose the Community Topic Model (CTM), which leverages bayesian probabilistic model to group authors with similar research topics into the same community and rank them based on the probability of belonging to that community. To further capture the dynamic features of community evolution, we propose the Dynamic Community Topic Model (DCTM) by extending CTM with the time variables. Both CTM and DCTM were applied to large-scale scholarly dataset: Arnetminer (Scholarly publications in the area of Computer Science). The experiments show that both models can capture the topic features and dynamics changes of communities, and provide dynamic scholar and venue ranking in the context of communities.

1.4.2 Intellectual Trade between Scientific Disciplines

Scientific disciplines are not isolated but connected to each other through scholarly communications (e.g. citation). Several studies utilize journal-to-journal citations (Boyack et al., 2005) or user click stream (Bollen, Sompel, et al., 2009) to visualize the topological relation between science disciplines. Moreover, information is flowing dynamically from one discipline to another through the structural connections between disciplines, scientific ideas, methods or concepts. The "intellectual trade" occurring in science in the exchange of scientific knowledge, is very similar to international trade (Leontief, 1953) occurring between nations in the exchange of goods and capital. Some nations produce more goods than they import and vice versa. The flow of capital and goods in international trade is rarely symmetrical. Such asymmetries in international trade may exist in the scholarly community where some scholarly domains can be net exporters and/or consumers of scientific ideas and knowledge.
To study these kinds of asymmetries in scientific communication we propose to draw an analogy between international trade and the exchange of knowledge in the scholarly community, and utilize the network analysis techniques to investigate the direction, amount, and symmetry of knowledge flow across different scientific disciplines. We operationalize the production of knowledge as the publication of peer reviewed articles in scholarly journals. International trade aggregates trade between individual parties at the national level. We aggregate scientific trade at levels above that of individual journal articles, specifically at the level of scientific journals which are used as aggregations of individual articles. Scholarly domains are operationalized as aggregations of scientific journals demarcated by existing classification schemes.

Ideas can thus be “exported” to other domains where they are read and cited. Citations are as such considered an operationalization of the “acknowledgement” or in fact “importation” of these ideas. Asymmetries between the number of citations pointing from one domain to another can then be used to detect intellectual trade imbalances in the scientific community.

1.4.3 The response of Twitter to Scholarly Articles

Online social media, such as social networking and micro-blogging environments, have become a crucial component of public discourse. Scholars are becoming increasingly interested in leveraging user-generated data on social media platforms to study a multitude of social (Crandall et al., 2010), economic (Bollen, Mao, & Pepe, 2011), and political (Pepe & DiGennaro, 2009) phenomena. But the interest of scholars in social media is not limited to studying them; scholars are also increasingly using social media to acquire and disseminate scholarly information. Therefore the question becomes not so much how can social media be used in scientific study, but how is it affecting scholarly communication itself?
Some research has looked at how scientists use the microblogging platform Twitter during conferences by analyzing tweets containing conference hashtags (Letierce, Passant, Decker, & Breslin, 2010; Weller, Dröge, & Puschmann, 2011). Other research has explored the ways in which scholars use Twitter and related platforms to cite scientific articles (Priem & Costello, 2010a; Weller & Puschmann, 2011b). More recent work has shown that Twitter article mentions predict future citations (Eysenbach, 2011). Our work falls within, and extends, these lines of research by examining the temporal relations between quantitative measures of readership, Twitter mentions, and subsequent citations for a cohort of scientific preprints.

We employ the above mentioned indicators to perform two analyses. First, we investigate the temporal relationships that exist between preprint downloads and Twitter mentions. We study both the temporal delay and span of these reactions. The delay is the time difference between the date of an arXiv submission and a subsequent spike in downloads or Twitter mentions. The time span is measured as the time between the first and the last reaction (download or mention) for the article in question. In other words, we address the questions:

1. How long does it take for an article to receive its maximum volume of downloads and Twitter mentions? (Time delay), and
2. How long does that response activity last? (Time span).

Second, we investigate whether a correlation exists between how popular an article is on social media, as measured in the volume of Twitter mentions, and how much it gets downloaded and cited in the scholarly record.

We are concerned with the question whether the increasing role of Twitter and other social networking environments in the scholarly community can affect citation- or usage-based indicators of scholarly impact. In other words, we ask the question: Is there a correlation between the volume of Twitter mentions and the downloads and citations an article receives?
1.4.4 The Comparison between Wikipedia and Scholarly Ranking

Scholarship is increasingly becoming an online process, and social media are becoming an increasingly important part of the online scholarly ecology. Growing numbers of scholars discuss and share the research literature on Twitter and Facebook, organize it in social reference managers like Mendeley, and review it in Blogs (Priem & Costello, 2010a). The burst of scientific works in social media requires new measures to reassess their impact beyond traditional notions of scholarly citation and impact. We’ve already analyzed the statistical correlation between Twitter mentions and article citations, here we extend the investigation of the role of social media in scholarly evaluation, by quantifying the scholarly impact of articles, authors and topics through Wikipedia.

Wikipedia has now become an important source for creation, spread and learning of scientific knowledge. Kittur, Chi, and Suh (2009a) showed that over 25% articles in Wikipedia are related to natural or social sciences up to 2008. Large amount of scholarly entities, like papers, scholars, topics, are cited by Wikipedia editors, which can be defined as “wikipedia citation”, implying their values has been recognized by Wikipedia community. Unlike “academic citation” that demonstrates explicit recognition from expert scholarly peers, the authority of “wikipedia citation” needs further examination. Attempts have been made to study the relationship “academic citation” and “Wikipedia citation”. Nielsen (2007) shows that citations in Wikipedia correlate well with statistics from Journal Citation Report. Evans and Krauthammer (2011) investigated such relationships at the journal article level and found that PubMed journal articles that are mentioned in Wikipedia have significantly higher academic citation counts than an equivalent random article subset. Although these findings show that Wikipedia citation is also a good indicator of academic impact, their results are limited to journals or articles published in one journal.

As an extension along this line of work to larger-scale of data, broader research area and
more diverse scholarly entities, this dissertation makes an effort to quantitatively compare
the rankings of articles, authors, and topics selected from publications data of ACM Digital
Library (the world’s largest computing society), based on academic citation and Wikipedia
citation.
CHAPTER 2

LITERATURE REVIEW

2.0.5 Classic Bibliometrics

Bibliometrics provides powerful statistical tools to quantitatively analyze scholarly communications. One of the most important approaches in bibliometrics is “citation analysis” since the establishment of the Institute for Scientific Information (ISI) and Science Citation Index (SCI) by Garfield (1955b). In academia, citations have two important implications. First, scholars or articles being cited show intellectual merit. Second, scientific ideas spread from the cited articles/scholars to the citing articles/scholars. Consequently, the main goal of bibliometrics, which is mainly built on citation analysis, are twofolds: one, to measure the scholarly impact of articles, scholars, journals, and institutions (impact analysis); and two, to investigate the relationships among different scientific disciplines (relational analysis) (Thelwall, 2007). Both of them provide insight about the development and output of science, based on which science policy-makers and funders can inform their decision-making.

Many indicator for measuring the scientific impact of an article, journal, or scholar, are derived from citation data. Impact Factor is used to measure the impact of a journal and is calculated as the average number of citations received per paper published in that journal during the two preceding years (Garfield, 1972). A journal with a larger impact factor is considered more scholarly influential than those with lower values. H-index attempts to measure both the productivity and impact of a scholar’s published articles (Hirsch, 2005).
For example, a scholar with an index of $h$ has published $h$ papers, and each of them has been cited at least $h$ times. $H$-index is found to be a good indicator of predicting whether a scientist will win the Nobel Prize. Both Impact Factor and $H$-index are very important indicators of scholarly impact for journals or scholars.

Another important application of citation analysis is thus to characterize and visualize the topology of the relations between authors, papers, journals, and domains, the latter leading to the study of interdisciplinarity. This type of analysis is based on the assumption that scientific ideas are exchanged through citation links. Cawkell (2001) first analyzed the information flow between articles through their citations. It is now possible to map the structure of science by clustering authors, papers, and journals based on their co-citation or inter-citation similarities (Boyack et al., 2005). In fact, Leydesdorff (2006) compute network betweenness centrality metrics to characterize the interdisciplinarity of academic journals. Yan et al. (2012) use a trading metaphor to study knowledge transfer in the natural and social sciences by tracking citation flows across 221 Web of Science subject categories.

### 2.0.6 Webometrics

Previously, an article can only be referred by another published article in the reference part. Now any Web page can mention an article by adding URL to this article, which is called “web citation”. The web cyberspace provides a more open platform to disseminate scientific information, since the publish of web page does not need peer-review process. The availability of large-scale of web data enable us to derive new indicators to quantify the impact of scholarly article or work, as well as examine the scholarly communications across different scientific disciplines.

According to Björneborn and Ingwersen (2004), the definition of webometrics is “the study of the quantitative aspects of the construction and use of information resources, struc-
The emergence of webometrics provides a new angle to evaluate scholarly impact in bibliometrics, where the quantification and characterization of scientific and scholarly communications largely depend on web resources and digital traces. Thelwall (2007) concluded that webometrics mainly include three aspects: link analysis, search engine and web description.

Link analysis is the main task of webometrics, which is very similar to the citation analysis in traditional bibliometrics since linking to another website is just like citing another article. Similar to JIF, Ingwersen (1998) proposed Web Impact Factor (WIF), which measures the average number of external links per page to a web site. The hypothesis behind WIF is that the number of links targeting an academic web site is proportional to its research productivity and impact, at different level (i.e. university, department, individual, etc.). Besides link analysis, webometrics scientists also investigate the reliability and coverage of search engines and changes in the content and structure of web pages (Bar-Ilan, 2004).

2.0.7 Web Usage Data

The emergence of digital library and open-access journal, facilitates the spread and dissemination of scientific articles on Web. Scholars can log on to web portals, read scientific articles, and download them. Finally, all these user generated data are collected by web servers, constituting a different type of scholarly data than citation records.

A large amount of work has investigated how usage data can be leveraged to measure science. Kurtz and Bollen (2010a) formally defined the usage data model, and Bollen, Sompel, et al. (2009) generated the map of science using click-stream data. In addition, the relationship between usage data and citation data has been extensively studied, including readership and citation (Kurtz et al., 2005), as well as downloads and citation (Brody et
2.0.8 Social Media and Science

Social media has become an important tool for scholars to freely produce, discuss, and share scholarly publications. BioInformatics LLC conducted a survey in 2007 on scientists and social messaging\(^1\). 77% of life scientists participated in some type of social media. 50% viewed blogs, discussion groups, online communities, and social networking as beneficial to sharing ideas with colleagues. 85% saw social media affect their decision-making.

Several related works illustrate the influence of social media on scientific development. Shneiderman (2008) foresaw the age of ‘Science 2.0’ when traditional scientific methods will be revolutionized by social innovation and the rapid development of Web 2.0 technology. Ullrich et al. (2008) analyzed the pedagogical implications of Web 2.0 on research, teaching, and learning. They emphasized that the openness of Web 2.0 technology offers great facility for technology-based learning.

Some common social media tools used by scientists are introduced as follows:

**Microblogging** Twitter is the most popular microblogging website in the world, allowing users to post up to 140-character messages. There are over 500 million registered Twitter users and over 340 million “tweets” are posted every day\(^2\). Twitter recently has become an important tool for scientists to communicate\(^3\).

First, tweets can mention scientific articles by embedding URLs linking to those articles (Weller & Puschmann, 2011a; Priem & Costello, 2010b). Second, Twitter is frequently used in scientific conferences for scholarly communication (Reinhardt, Ebner, Beham, & Costa, 2009). Sopan, Rey, Butler, and Shneiderman (2010) developed a web app for the

---

\(^1\)http://scholarlykitchen.sspnet.org/2008/08/14/scientists-use-social-media/
\(^2\)http://www.jobstock.com/blog/social-media-statistics-2013/
\(^3\)http://www.sciencebase.com/scientists-on-twitter.html
real-time visualization of tweet conversation during conferences. Letierce, Passant, Breslin, and Decker (2010) analyzed Twitter activity during Semantic Web conference 2009. They further found that Twitter can help to identify trending topics, authoritative scholars, and conference attendees like to use hashtag for communication. Desai et al. (2012) analyzed the content, sentiment, and retweeting of tweets generated from Kedney Week 2011 and revealed a large number of educational tweets that were disseminated to the public.

**Social Bookmarking** Delicious is one of the the most popular social bookmarking websites. Ding et al. (2009) found that social tags related to bioinformatics, biology, and ecology are among the most frequently occurring tags in Delicious, implying the scholarly use of Delicious. In addition, two main types of social bookmarking tools for scholarly use are Mendeley and CiteULike. Mendeley, founded in 2007, is used for managing and sharing research papers, discovering research data and collaborating online. It combines Mendeley Desktop, a PDF and reference management application (available for Windows, Mac and Linux), with Mendeley Web, an online social network for researchers. CiteULike was launched in 2004, aiming to promote and develop the sharing of scientific references amongst researchers 4. CiteULike has been extensively used in medical science and around 1/6 new medical articles in MEDLINE are bookmarked in CiteULike (Good, Tennis, & Wilkinson, 2009). CiteULike can be utilized to recommend scientific articles to users based on their reference library (Bogers & Bosch, 2008) or to cluster academic resources like journals or authors (Jiang, He, & Ni, 2011).

**Wikipedia** Wikipedia, as a collaboratively edited, multilingual, and free Internet encyclopedia, has become an important source for the creation, distribution, and acquisition of scientific knowledge. Kittur, Chi, and Suh (2009b) shows that over 25% of pre-2008 arti-

---

4http://www.citeulike.org/
cles in Wikipedia are related to natural or social sciences. Black (2008) even discussed the feasibility of Wikipedia replacing the current peer-review journals as a recognized medium for scholarly publication, due to its rapid and reliable dissemination of scholarly knowledge. Wikipedia editors frequently reference scholarly entities, such as papers, scholars, and topics. We refer to such mentions as “Wikipedia citations”, implying that their value or influence has been explicitly recognized by the Wikipedia community. Unlike academic citations, which represent the explicit recognition of expert scholars, the authority of a Wikipedia citation is uncertain and needs to be examined further. Several studies have compared academic citations with Wikipedia citations. Nielsen (2007) showed that citations in Wikipedia correlate well with statistics from Journal Citation Reports. Evans and Krauthammer (2011) investigated this relationship at the journal article level and found that PubMed journal articles that are mentioned in Wikipedia have significantly higher academic citation counts than an equivalent subset of random article. Although these findings show that Wikipedia citations are an indicator of academic impact, their results are limited to journals or articles published in the same journal.

**Blogging** Blogging has become a popular medium of informal scholarly communication and discussion. Blogs reduce the demand of the peer-review process for online publishing and thus encourage the spread of scientific information to the public. Two main research blog aggregators are Research Blog and Postgenomic. Martindale and Wiley (2005) described the use of blogs for scholarship. Groth and Gurney (2010) drew the keyword and citation similarity map of publications referred to by chemistry blogs in Researchblogging.org. They found that scientific discourse on the Weblog is more immediate, contextually relevant, and has a larger non-technical focus than the academic literature. Shema, Bar-Ilan,
and Thelwall (2012) found that bloggers in ResearchBlogging.org show a preference for papers from high-impact journals and blog mostly about research in the life and behavioral sciences. In addition, they found that most bloggers have Twitter accounts.
CHAPTER 3

SCHOLARLY COMMUNITY DETECTION BASED ON DYNAMIC RESEARCH TOPICS

Abstract

To provide scholarly rankings within the context of scholarly community is a useful but challenging task in bibliometrics. In this chapter, we investigate the emergence of scholarly communities by probabilistically modeling scholars and research topics simultaneously, and provide a dynamic perspective on the evolution of those communities. First, the Community Topic Model (CTM) is proposed to identify communities composed of scholars sharing similar topics. Second, the Dynamic CTM (DCTM) is proposed to capture the dynamic features of communities and topics based on the Bernoulli distribution that leverages the temporal continuity between consecutive timestamps. The experimental results on a large-scale of bibliographic data show that scholars with similar research topics are clustered together and ranked probabilistically. In addition, the co-evolution of communities, topics and rankings can be tracked as well. This study allows us to better understand the dynamic features of social networks and provide communitinized scholarly ranking.
3.1 Introduction

Ranking scholars and publication venues are crucial tasks in bibliometrics, and several ranking indicators like IF or h-index are proposed. However, most existing ranking indicators ignore the fact that academia consists of diverse scholarly communities, and each scholarly community has its own criteria to evaluate the scholarly impact. Particularly, a journal with extremely high IF does not mean that the authority of the journal is recognized in all communities; a scholar who is considered an expert in one community does not necessarily recognized in another community, even through the h-index score of the scholar is very high. Therefore, how to detect the scholarly communities and provide rankings for scholars and publication venues within the context of community is a useful but challenging task.

In addition, the study of social networks has enabled scientists to better understand social communication patterns and interpret social principles. Researchers have found that most real world networks, in contrast to random networks, exhibit three common properties: the small world property, power-law distributions, and community structure with relatively high clustering coefficient (Erdős & Rényi, 1959; Girvan & Newman, 2002). In fact, evidence of communities has been detected in a range of different domains and applications (Leskovec et al., 2010). Communities detected within a social network might correspond to a variety of social groupings affected by the heterogeneity of users as well as their interactions. The analysis of communities is therefore crucial to establish a better understanding and utilization of social networks. However, most of the existing work has focused on the structural properties of communities and neglects other important aspects such as their topic features. Moreover, the structural properties and topic aspects of communities may interact with each other. Common interests may drive the formation of communities, and in turn community structure may reinforce common interests. Few studies have systematically and quantitatively addressed the interaction between the structural
and topic properties of communities (Ding, 2011).

Furthermore, social networks and their communities may change over time. Any effort to understand the formation of communities and their topical features needs to include the time dimension. Previous studies have used state space models on the natural parameters of multinomial distributions to analyze the time evolution of topics, or developed the continuous time dynamic model to mine the latent topics through a sequential collection of documents (Blei & Lafferty, 2006; Griffiths & Steyvers, 2004; Iwata et al., 2010). Generally, these studies have applied a set of approaches to approximate posterior inference over the latent topics. However, none considered the community features of the network actors involved in their datasets that might reveal some hidden explanation for topical evolution.

To address those challenging problems for detecting communities by considering their topic features, we propose the Community Topic Model (CTM). To further capture the dynamic features of community evolution, we propose the Dynamic Community Topic Model (DCTM) by extending CTM with the time variables. Both CTM and DCTM were applied to large-scale scholarly dataset: Arnetminer (Scholarly publications in the area of Computer Science). The experiments show that both models can capture the topic features and dynamics changes of communities.

### 3.2 Problem Definition

In a co-author network, scholars may have different topic interests and therefore can be divided into different communities according to their topic distributions. For example, in Figure 3.1, the author has different research focuses: the Semantic Web and Text Mining, which means that he can belong to two communities with different topics. Most existing community detection methods focus on the topological structures of networks and ignore actors’ topic interests.
In order to detect communities from the topic level, the proposed algorithm should assign appropriate actors for each community based on matching topic interests. Unlike Author-Topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) that assigns authors to different topics based on the authors’ topic distributions, the proposed CTM assigns actors to different communities based on the similarity between authors’ topic distributions and community’s topic distributions. For example, researchers, who work in both areas of biology and the Semantic Web can be viewed as having similar topic distributions. Therefore, these researchers can be grouped into one community. It is hard to find the similar topic distributions of a group of authors, because the relationship between a group of authors and their topic distributions are latent variables. There are studies emphasizing how these two variables jointly affect the formation of links in the document graph (Zhu, Yu, Chi, & Gong, 2007). However, some questions are left unanswered, such as whether the structure of communities has an influence on the distribution of topics and how topic distribution determines the features of a community.

Additionally, an examination of changes over time is needed in order to discover the
dynamic relationship between communities and topics. Traditional methods treat different
timestamps independently and ignore the temporal continuity between consecutive times-
tamps (Li et al., 2010). These studies have two problems. The first problem is how to deter-
mine the corresponding relationship between latent variables from different timestamps (He
et al., 2009; Griffiths & Steyvers, 2004). For instance, for a certain community at time t,
it is hard to know which community it was derived from time t-1. The previous method
requires calculating the similarity between the current community in time t and all com-
munities in time t-1 in order to figure out the temporal inheritance. The community-topic
distribution and the topic-word distribution were needed for each calculation, which can be
computationally expensive. The second problem is that the temporal correlation between
consecutive timestamps was not considered. For example, an author’s previous research in-
terests may influence his current interests. The proposed DCTM can simulate the changes
of actors’ interests at different time periods, and observe the evolution of communities and
topics along the time.

Taking the example below, an scholar in a co-author network can be defined as actor=(a,
\{z_1, t_1\}, \{z_2, t_2\}, \{z_3, t_3\}, ...), where a means the author and \{z_i, t_i\} is the tuple which rep-
resents author a focuses on research topic z_i at time point t_i. We propose the following
function to define communities: \( c(a_i, a_j, c_k, t) = f(S_{a_i}, S_{a_j}, S_{c_k}, t) \). Here \( c(a_i, a_j, c_k, t) \), rep-
presents the decision to put two authors \( a_i \) and \( a_j \) into a community \( c_k \) at time \( t \), which
depends on the formula \( f(S_{a_i}, S_{a_j}, S_{c_k}, t) \), where \( S_{a_i}, S_{a_j}, S_{c_k} \) means the topic distribution
of \( a_i, a_j, c_k \) at time \( t \). Table 3.1 summarizes the mathematical notation used in this paper.
<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$</td>
<td>Paper</td>
</tr>
<tr>
<td>$w$</td>
<td>Word</td>
</tr>
<tr>
<td>$x$</td>
<td>Author (i.e. Scholar)</td>
</tr>
<tr>
<td>$z$</td>
<td>Topic</td>
</tr>
<tr>
<td>$r$</td>
<td>Publication venue (e.g., conference)</td>
</tr>
<tr>
<td>$c$</td>
<td>Community</td>
</tr>
<tr>
<td>$N_d$</td>
<td>The number of words in the current paper $d$</td>
</tr>
<tr>
<td>$N_D$</td>
<td>The number of words in the entire collection of documents</td>
</tr>
<tr>
<td>$a_d$</td>
<td>The set of co-authors in paper $d$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Hyperparameter for generating $\theta$ from Dirichlet Distribution</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Hyperparameter for generating $\chi$ from Dirichlet Distribution</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Hyperparameter for generating $\varphi$ from Dirichlet Distribution</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Hyperparameter for generating $\psi$ from Dirichlet Distribution</td>
</tr>
<tr>
<td>$\chi$</td>
<td>A multinomial distribution of authors over communities</td>
</tr>
<tr>
<td>$\theta$</td>
<td>A multinomial distribution of communities over topics</td>
</tr>
<tr>
<td>$\varphi$</td>
<td>A multinomial distribution of topics over words</td>
</tr>
<tr>
<td>$\psi$</td>
<td>A multinomial distribution of topics over publication venues</td>
</tr>
<tr>
<td>$D$</td>
<td>Collection of documents</td>
</tr>
<tr>
<td>$A$</td>
<td>Collection of authors</td>
</tr>
<tr>
<td>$T$</td>
<td>Collection of topics</td>
</tr>
<tr>
<td>$R$</td>
<td>Collection of conferences</td>
</tr>
</tbody>
</table>
3.3 Related Work

3.3.1 Community Detection

Researchers have used a number of methods to detect communities within networks. Two widely used approaches are those based on centrality and graph partitioning. Girvan and Newman (Girvan & Newman, 2002) used betweenness centrality to examine the community structure in large networks. The original algorithm was improved upon by Clauset, Newman, and Moore (Clauset, Moore, & Newman, 2008), who reduced the complexity from $O(m^2n)$ to $O(md\log n)$ (where $d$ is the depth of the dendrogram of the community structure). This algorithm has been tested empirically and validated as an appropriate model for community detection (Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004). Two standard examples of the graph partitioning approach are the local spectral partitioning algorithm (Andersen, Chung, & Lang, 2006) and the flow-based Metis MQI algorithm (Flake, Tarjan, & Tsioutsouliklis, 2004). These approaches were compared to the Girvan-Newman algorithm by Leskovec, et al. (Leskovec et al., 2010). In applying all of these algorithms against the same large-scale dataset, Leskovec, et al. found that the algorithms produced similar results and identified equally compact clusters at all scale sizes. However, none of these algorithms have taken into consideration the topic feature of communities.

3.3.2 Topic Modeling

Since the introduction of the LDA model (Blei, Ng, & Jordan, 2003), various extended LDA models have been used for automatic topic extraction from large-scale corpora. Rosen-Zvi et al. (Rosen-Zvi et al., 2004) introduced the Author-Topic model, which extended LDA to include authorship as a latent variable. Based on the Author-Topic model, Tang, et al. (Tang, Jin, & Zhang, 2008) further extended the LDA and Author-Topic model and proposed the Author-Conference-Topic (ACT) model, which is a unified topic model for
simultaneously modeling different types of information in academic networks. Nallapati and Cohen (Nallapati & Cohen, 2008) proposed a Link-PLSA-LDA model as a scalable LDA-type model for topic modeling and link prediction. Later, Si and Sun (Si & Sun, 2009) proposed a tag-LDA model, which extended the LDA model by adding a tag variable, and applied it to social tagging systems. The link structure of networks has served as an additional area for network research. Chang and Blei (Chang, n.d.) introduced the relational topic model (RTM) to model the link between documents as a binary random variable conditioned on their contents. Although research has been done in both areas of community detection and topic analysis, very few researchers have sought to combine the two. One notable exception is the work of Zhou, Manavoglu, et al (Zhou, Manavoglu, Li, Giles, & Zha, 2006), who used topic model for semantic community discovery in social network analysis. The other is the work of Liu, Niculescu-Mizil and Gryc (Liu, Niculescu-Mizil, & Gryc, 2009) who examined topic and author communities for a set of blog posts and citation data through jointly modeling underlying topics, author community, and link formation in one unified model. However, it was done synchronically, rather than diachronically. Therefore, it did not provide an evaluation of how the model functions in examining changes in topics over time.

As discussed above, studies on community detection haven’t taken other aspects of community profile into consideration, while research on topic modeling largely neglects potential relationships between topics and community structure. In this paper, we propose a different approach to address this question, by integrating dynamics and communities into the topic modeling algorithms.
3.4 Methods

3.4.1 Datasets

ArnetMiner (http://www.arnetminer.org) is an academic search system developed by the Tshinghua University (Tang, Zhang, et al., 2008). The Arnetminer dataset covers the major publications in the area of computer science. It was collected by using a unified automatic extraction approach on researcher’s profile pages from the Web and other online digital libraries. Currently, this dataset contains 629,814 publications, 12,609 conferences, and 595,740 authors covering the period of 2000-2010. Each publication has the information about abstract, authors, year, venue, and title. The abstracts and titles were pre-processed using a stemming algorithm and a stop word list.

3.4.2 Algorithms

In this paper, CTM (Community Topic Model) and DCTM (Dynamic Community Topic model) were proposed to capture the semantic relationships among communities and topics as well as their changes over time.

Background Knowledge

To better understand the algorithm of CTM, we will introduce the concepts of Dirichlet distribution, sKL, F1-measure and Gibbs sampling in this section.

1. Dirichlet Distribution: Dirichlet Distribution is a family of continuous multivariate probability distribution, which is used to denote the probability of a probability event. There are two main reasons for us to apply it to LDA(Latent Dirichlet Allocation): first, it is the conjugate prior of the categorical distribution and multinomial distribution, which can help us to solve the model by applying Gibbs sampling algorithm; second, it can provide initial parameters estimation, which can train LDA model to learn training data and analyze
new data automatically.

2. Gibbs sampling: Gibbs sampling is an efficient algorithm for solving MCMC (Monte Carlo-Markov Chain) problem. The process for LDA applying Gibbs sampling to learn topic distribution of training dataset can be seen as a MCMC process; for example, assuming a dataset has 10,000 documents, LDA can learn all the documents one by one and accumulate the domain knowledge of the dataset during that process. Gibbs sampling can help LDA to realize automatically studying, it randomly assigns the words to describe each topic by using sampling method at the beginning, when the iteration goes on, it will adjust the probability for a word appears in a certain topic and assign a word to a high related topic according to the updated probability. When the iterations are over, Gibbs sampling will provide bags of words for each topic to represent themselves.

3. sKL: sKL is a method to compute the similarity between two variables with vector as the description of their attributes. Assuming we have two authors $x_1$ and $x_2$ with interesting distribution on ten topics, then the sKL value between the two authors can be seen as below:

$$sKL(x_1, x_2) = n \sum_{i=1}^{n} x_{1i} \times \log \frac{x_{1i}}{x_{2i}} + x_{2i} \times \log \frac{x_{2i}}{x_{1i}}$$  \hspace{1cm} (3.1)

we can find from the formula that the lower the sKL is for two variables, the more similar the two variables are.

4. F1-measure: in many researches, we often use precision and recall to evaluate the performance of an algorithm’s prediction power. Precision summaries the total number of correct and wrong for the performance of target model, recall calculates the proportion of the right ones detected by target model in the whole results sets. F1-measure is an index, which can integrate precision and recall together to give a more accurate evaluation for the performance of target model.
Community Topic Model

The essential idea of CTM is to detect scholarly communities based on topic distributions over all authors and cluster authors with similar topic distribution together in one community. The community detection is achieved through a statistical learning process, during which the assignment of an author to a certain community (also including other types of assignments, like assigning topic to author, conference to topic, etc.) is implemented by sampling from several continuously updated and mutually related probability distributions.

The graphical representation of the learning process is shown in Figure 3.2, which can be explained by the following example. A group of authors, \( a_d \), collaborate on a paper \( d \). For each author \( x \) in \( a_d \), \( x \) first selects a community \( c \) from the author-community distribution \( \chi \); then select a topic \( z \) under community \( c \) from the topic-community distribution \( \theta \); and finally select a word \( w \) under topic \( z \) from the topic-word distribution \( \phi \) and a conference \( r \) related to \( z \) from the topic-conference distribution \( \psi \). The three shaded nodes, \( a_d \), \( w \) and \( r \), are all observable.

Gibbs sampling is used to estimate the four parameters \( \theta, \chi, \varphi \) and \( \psi \), whose initial values (i.e. prior probabilities) are determined by four hyperparameters: \( \lambda, \alpha, \beta, \) and \( \mu \), whose empirical values are given by \( \lambda = 50/C, \alpha = 50/T, \beta = 0.01 \) and \( \mu = 0.1 \) (Tang, Zhang, et al., 2008; Tang, Jin, & Zhang, 2008; Rosen-Zvi et al., 2004). The final values of the four parameters are obtained after 1000 iterations of sampling and estimation.

In each iteration, CTM assigns a community \( c \in C \) and a topic \( z \in T \), to each author \( x \in A \) and each word \( w \in V \), appeared in every paper \( d \in D \). For every possible assignment \( (c, z) \in C \times T \), the following probability is calculated:

\[
P_{c,z}(z^i = z, c^i = c, x^i = x | w^i = w, r^i = r, z^{i-1}, x^{i-1}, w^{i-1}, a_d) \propto \chi^{i-1} \cdot \theta^{i-1} \cdot \varphi^{i-1} \cdot \psi^{i-1} \tag{3.2}
\]

where \( i \) and \( i - 1 \) denotes the corresponding values of the current and previous step of iteration; \( P_{c,z} \) denotes the probability that \( (c, z) \) is assigned to \( x \) give the previous estimated
results and current observation. Based on the multinomial distribution \{P_{c,z}, (c, z) \in C \times T\}, a community \(c\) and a topic \(z\) will be randomly sampled and assigned to author \(x\) and word \(w\).

In addition, topic \(z\) is assigned to conference \(r\) without sampling but from direct observation.

After all assignments are done in the \(i\)th iteration, all probability matrices are updated as:

\[
\chi_{x,c} = \frac{n_{x,c} + \lambda}{\sum_{c' \in C} (n_{x,c'} + \lambda)}, \quad \theta_{c,z} = \frac{n_{c,z} + \alpha}{\sum_{z' \in T} (n_{c,z'} + \alpha)}, \quad (3.3)
\]

\[
\psi_{z,w} = \frac{n_{z,w} + \beta}{\sum_{w' \in V} (n_{z,w'} + \beta)}, \quad \varphi_{z,r} = \frac{n_{z,r} + \mu}{\sum_{r' \in R} (n_{z,r'} + \mu)}, \quad (3.4)
\]

\(\chi_{x,c}\) denotes the entry of author \(x\) and community \(c\) in matrix \(\chi\), and \(n_{x,c}\) denotes the number of times author \(x\) is assigned to community \(c\) at the current iteration. The similar denotation applies to \(\theta_{c,z}, \psi_{z,w}\) and \(\varphi_{z,r}\).

For each iteration, parameters estimated from previous iteration are utilized to make re-assignment for all the authors. The final probability distribution matrices, i.e. \(\theta, \chi, \varphi\) and \(\psi\), will be very close to the actual value when the iteration is done. In other words, the
assignment of authors to communities will be sufficiently accurate in the end. In the process of statistical learning, the assignments of topics to communities are determined by all the authors’ topic assignments, and authors with similar topic distribution are most likely to be assigned to the same community as the CTM estimation becomes more and more accurate. Besides, the rank of an author in a community (the probability of assigning the author to the community) is determined by his/her interest in the most popular topics (topics with high probability of being assigned to the community) in that community. For instance, if an author is interested in several topics and frequently writes papers to those topics, he/she is very likely to be assigned to the communities in which those topics rank high, and become a highly ranked author in that community. The algorithm in mathematical language is shown in Figure 3.3

1. For $i = 1:1000$ iterations:
2. For each document $d$:
3. For each author $x$ in document $d$:
4. For each word $w$ in document $d$:
5. Compute Equation (2) and sample a community $c$ and topic $z$ to the current author $x$ and word $w$; also assign topic $z$ to conference $r$.
6. Update $\lambda$, $\phi$, $\psi$ and $\theta$ according to Equation (3);
7. End for word $w$.
8. End for author $x$.
9. End for document $d$.
10. End for iterations.
11. Output final $\lambda$, $\phi$, $\psi$ and $\theta$.

Figure 3.3: Algorithm of Community Topic Model (CTM).

Dynamic Community Topic Model

To model the evolution of a community, we assume that the distributions of communities are based on a Bernoulli trial. When time goes from one time slice to another, we flip a coin for each author. If the coin lands as a head, the previous community distribution will be
kept. Otherwise, a new distribution will be sampled for that author. To determine authors’ current interest, a switch variable $s$ is introduced. The value of $s \in \{0, 1\}$ is sampled based on a Bernoulli distribution. When the sampled value of $s$ equals 1, author’s current interest is determined by his status in the last time period; when the sampled value of $s$ equals 0, author’s current interest is not influenced by his previous status but his current status. In Figure 3.4, the results from previous time point $t-1$ are used as prior knowledge to train the current training dataset at the time point $t$, and the Bernoulli trial is applied to simulate the changes of authors’ interests. The dynamic model assigns a unique id for each author, community, topic, word and conference at the first time period, and passes these to the next time period after the iteration of the first time period is finished. Therefore, all communities and topics from different time slices can be consistently tracked. The pseudo-code of DCTM can be seen in Figure 3.5.

### 3.5 Experimental Results

#### 3.5.1 Analysis from Static Perspective

For the Arnetminer dataset, the whole time span was divided into three periods: 2000-2003, 2004-2007, and 2008-2010. In each time period, CTM was used to calculate the topic distribution of author, community and conference. The probability distribution of author for a given community was used to assign authors to different communities. 20 communities and 30 topics were extracted using the CTM. Authors in each community detected by CTM have similar topic distributions.

CTM can calculate the author community distribution and community topic distribution, while other existing models could not. These distributions provide enriched information to analyze the relationships among author, community and topic. Figure 3.6(a) displays the author community distributions for the selected 10,000 authors during the pe-
period of 2008-2010. This can be explained as the authors’ community preference. The value in the y-axis indicates the probability of an author choosing a community. Some authors have very high preference for certain communities such as Community 1 (computer system, network), Community 2 (intelligence system, parallel and distributed systems, semantic web, neural, wireless network, fuzzy), and Community 15 (image recognition, knowledge management, mathematics, machine learning, user interface and collaboration system). The Figure 3.6(b) shows the community topic distribution. For each community, the probability of a given topic (i.e., the value in y-axis) indicates the significance of this topic in that community. Some communities have strong preferences on Topic 4 (manufacturing optimization), Topic 15 (parallel and distributed systems), Topic 21 (embedded systems), and
Figure 3.5: Algorithm of Dynamic Community Topic Model (CTM).

**Topic 26 (knowledge management).**

The value of y-axis in Figure 3.7 represents the mean probability of all authors choosing a community, which can be interpreted as the popularity of a community. Figure 3.7 shows the popularity of communities during the period of 2008-2010. The range of popularity for all communities is between 0.049 and 0.053, indicating that the popularity differences among all communities are small. Community 15 is the most popular community in 2008-2010.

Figure 3.8 shows the topic distribution for Community 15. Topic 16 and 26 are the most popular topics in Community 15, followed by topic 17, 25 and 28. Table 3.2 illustrates the top words, conferences for above popular topics and the top authors for the Community 15. Topics in Community 15 are diverse including image recognition, software development,
Figure 3.6: (a) Author-community distribution for 10,000 selected authors in Arnetminer and (b) Community-topic distribution of Arnetminer in 2008-2010 wireless network, information management and mathematic algorithms. The listed top journals and conferences are consistent with the content of the topics, for example, PAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence) in Topic 16, SSEN (ACM SIGSOFT Software Engineering Notes) in Topic 17, and WN (wireless network) in Topic 25.

In Table 3.3, the top ranked authors tend to have different research areas. For example, Metin Demiralp’s research focus is mathematic algorithms. His work has been published in different conferences and journals specialized in microelectronics, applied mathematics,
Table 3.2: The description of Topic 16, 17, 25, 26 and 28.

<table>
<thead>
<tr>
<th>Topic</th>
<th>16</th>
<th>17</th>
<th>25</th>
<th>26</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Imaging</td>
<td>Software</td>
<td>Networks</td>
<td>Information</td>
<td>Computer</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(0.046)</td>
<td>(0.083)</td>
<td>(0.024)</td>
<td>(0.019)</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>Development</td>
<td>Wireless</td>
<td>Management</td>
<td>Algorithm</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.015)</td>
<td>(0.035)</td>
<td>(0.017)</td>
<td>(0.014)</td>
</tr>
<tr>
<td></td>
<td>Detection</td>
<td>Engine</td>
<td>Mobile</td>
<td>Systems</td>
<td>Polynomial</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.015)</td>
<td>(0.022)</td>
<td>(0.015)</td>
<td>(0.014)</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>Oriented</td>
<td>Sensor</td>
<td>Business</td>
<td>Algebra</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.021)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td></td>
<td>Feature</td>
<td>Architecture</td>
<td>Routing</td>
<td>Knowledge</td>
<td>linearization</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.008)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Conference</td>
<td>IEEE PAMI</td>
<td>IEEE Software</td>
<td>IEEE/ACM Networking</td>
<td>SS</td>
<td>JCAM</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.017)</td>
<td>(0.031)</td>
<td>(0.028)</td>
<td>(0.041)</td>
</tr>
<tr>
<td></td>
<td>PRL</td>
<td>ACM SSEN</td>
<td>CN: CTN</td>
<td>Computer</td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.016)</td>
<td>(0.030)</td>
<td>(0.017)</td>
<td>(0.025)</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>JSS</td>
<td>WN</td>
<td>JASIST</td>
<td>JSC</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.014)</td>
<td>(0.030)</td>
<td>(0.014)</td>
<td>(0.025)</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td>IEEE SE</td>
<td>CC</td>
<td>IM</td>
<td>TCS</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.012)</td>
<td>(0.016)</td>
<td>(0.013)</td>
<td>(0.024)</td>
</tr>
</tbody>
</table>

Table 3.3: Top 10 authors in community 15.

<table>
<thead>
<tr>
<th>Scholar</th>
<th>Probability * 10^{-4}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metin Demiralp</td>
<td>6.75</td>
</tr>
<tr>
<td>Nikos Fakotakis</td>
<td>4.11</td>
</tr>
<tr>
<td>Yang Liu</td>
<td>3.82</td>
</tr>
<tr>
<td>Virginie Govaere</td>
<td>3.23</td>
</tr>
<tr>
<td>Michael McAleer</td>
<td>3.23</td>
</tr>
<tr>
<td>Marc Moonen</td>
<td>3.23</td>
</tr>
<tr>
<td>Ibrahim Busu</td>
<td>2.94</td>
</tr>
<tr>
<td>Nico Mastorakis</td>
<td>2.93</td>
</tr>
<tr>
<td>Thierry Martin</td>
<td>2.93</td>
</tr>
<tr>
<td>Zhong Liu</td>
<td>2.93</td>
</tr>
</tbody>
</table>
engineering, informatics and communications, and signal processing. The function of the latent variable of community is to group the authors with similar topic distributions into one community. This function can better discover authors with similar research interests
and therefore can be used to make personalized recommendations. For example, according to the topic distribution of Community 2, its main topic is Topic 1 (network system), Topic 15 (parallel and distributed systems), Topic 25 (wireless network), Topic 16 (image recognition), Topic 8 (intelligence system and semantic), and Topic 11 (fuzzy, neutral, system control). Authors who have high ranks in Community 2 may be interested in publishing articles with several of those topics. In CTM, each author has a community distribution and each community has a topic distribution. The assignment of an author to a community is based on the similarity between the topic distribution of an author and the topic distribution of a community. This is different compared to other existing LDA models. In other LDA models, each author also has a topic distribution. But authors are only grouped based on their probabilities on a single topic rather than being grouped by their probability distribution over all topics. In other words, other LDA models can automatically define a topic by using a set of words and their probabilities in that topic, while CTM can automatically define a community by using a set of topics and their probabilities in that community.

3.5.2 Analysis from Dynamic Perspective

DCTM has the built-in functionality to simultaneously track the temporal changes of topics and community structures, which can identify the hidden dynamic relationships between topics and communities. Here, DCTM was tested on the Arnetminer datasets to unveil their community evolution patterns.

Figure 3.9 displays the evolution of all communities and topics along three time periods in Arnetminer. Figure 3.9(a) displays the changes of all authors’ preferences for each community. In period 1 and 3, authors have significant preferences for some communities, while in period 2, this phenomena is not obvious. In Figure 3.9(b), most of the topics exhibit a smooth increase or decrease along the time. Among them, Topic 15 (parallel
Figure 3.9: The evolution of (a) communities and (b) topics during three time periods in Arnetminer computation/distribution systems) is significantly higher than other topics during all three time periods.

CTM ranks authors and topics in each community. The ranking of a topic in a community depends on the probability that the topic is assigned to the community. The ranking of an author in a community depends on two factors: first, the active level of the author during a certain time period; second, the ranking of the representative topics of the author in the community. Intuitively, if an author writes a lot of papers in the highly ranked topics in a
In our experiment, we selected top 100 ranked authors in each community, and found that few top ranked authors remain the same in a community across all three different time periods. The main reason is that the ranking is based on active level of each author in a certain time period but not the influence of that author in related research area. In other words, even the ranking of a well-known author in some community will drop as long as the author does not write as many papers as before.

Another finding is that highly ranked topics for a community do not change significantly along the time. Specifically, in Community 1, the top 3 ranked topics (i.e. Topic 9: Database, 14: Machine Learning, 25: Network and Wireless) in the first time period still ranked relatively high during the next two time periods. However, the overall topic distribution is still changing and the ranking of some topic does not remain the same. For instance, in the third time period, Topic 1 (i.e., information management) is ranked much higher than the previous two time periods).

Figure 3.10 displays the changes of its top ranked authors and topics in Community 1. The topic distribution of Community 1 does not change significantly for the entire time periods (For example, the representative topics in blue box is mainly about database, machine learning, and clustering algorithm in first time period; distribute system, system performance in second time period; and user interface, agent, optimization, and intelligence in the third time period), while the composition of top ranked authors is experiencing relatively big changes. Very few authors remain in the same community over the three time periods. The probability of an author for a given community is associated with the yearly productivity of this author. For example, for two authors with similar topic distributions, if an author published more papers than the other in a certain time period, he will be ranked higher.
3.6 Discussion

3.6.1 The Dynamic Functionality of DCTM

The dynamic function of DCTM can estimate the current topic and community distributions based on the priori knowledge from the previous time period. Here the experiment was
Table 3.4: F1-measure of CTM and DCTM.

<table>
<thead>
<tr>
<th>Arnetminer</th>
<th>CTM</th>
<th>DCTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time period 2 (2004-2007)</td>
<td>0.1051</td>
<td>0.1102</td>
</tr>
<tr>
<td>Time period 3 (2008-2010)</td>
<td>0.0944</td>
<td>0.1073</td>
</tr>
</tbody>
</table>

designed to demonstrate its dynamic function. A subset of the Arnetminer that includes 10,000 publications, 5307 conferences, 166774 authors, and 26617 words from 2000 to 2010, were selected. The experiment contains the following steps:

- Step 1: 10% of the papers from the dataset of each time period were randomly selected as a testing data; the rest was used as a training data.
- Step 2: DCTM was applied to the entire training dataset and generated the author-community, community-topic, topic-word, and topic-conference distribution matrixes for each time period.
- Step 3: The training data was divided into three time periods and the CTM was applied on the training data in each time period.
- Step 4: For each time period, the results of CTM and DCTM were used to recommend a conference or journal for each paper in the testing data separately. F1-measure was used to evaluate the recommendation results of the CTM and DCTM correspondingly.

Table 3.4 shows that the dynamic function can significantly improve the performance of the DCTM for conference and journal recommendations. For each time period, since it can use priori knowledge to estimate the new dataset and adjust the results.

We applied "Statistical Significance Test" to further prove the outperforms of Dynamic Mechanism. In order to get enough samples, we assigned the number of time periods as 10 (it means that each year can be seen as a time period) and re-did the experiment for calculating F1-measure of CTM and DCTM in each time period. t-test is applied to compare the results of CTM and DCTM. The p-value of t test is smaller than 0.05, which means that there
exists differences of performance between DCTM and CTM. The average deviation is bigger than zero, which means that DCTM outperforms CTM on F1-measure.

**Dynamic analysis of topic and community**

Griffiths and Steyvers (2004) pointed out that an important part of realizing dynamic topic model is to build up the consistency for latent variables between adjacent timestamps. The dynamic function of DCTM can automatically generate communities and topics at the first time point and guarantee the consistency of latent variables for other time points. Based on the Arnetminer dataset, the similarity of all communities and topics between each two adjacent time periods was calculated and displayed in below heat-maps (Figure 3.11 and 3.12):

In Figure 3.11 and 3.12, the dark color means that two variables have a high similarity. All the heat-maps exhibit a high similarity on the diagonal, which means that the same latent variables can be assigned to a unique id through the whole time period. To compare with results generated by CTM based on the same dataset (see Figure 3.13), it is clear that there is no consistency for communities for different time periods. DCTM demonstrates the clear advantage on identifying the consistency for the latest variables.

**Community Content and Structure Analysis**

Most community detection algorithms are based on the graph topology of nodes and edges. The members in a community identified by CTM and DCTM demonstrate the strong topic similarities. Therefore, authors in such community may not coauthor with each other but do share common topic interests. Conductance was used to measure the quality of different communities, which is defined by Leskovec, et.al (2008):

\[
f(C) = \frac{s_c}{2m_c + s_c}
\]  

(3.5)
Figure 3.11: Community Similarity between each adjacent time periods based on DCTM

where $C$ denotes the set of nodes in a community, $m_c$ as the number of edges in $C$, and $s_c = |(u,v)|u \in C \& v \notin C|$ is the number of all $(u,v)$ that satisfies the condition. According to the definition of conductance, a community of high quality should have a low conductance value. Girvan-Newman community detection algorithm was applied to the coauthor network from the small dataset of Arnetminer (including 10,000 papers) (Girvan and Newman, 2002). The detected communities were compared with the communities derived from CTM. Table 3.5 lists the conductance of top five communities derived from the Girvan-Newman approach and CTM correspondingly.

The conductance of communities identified by CTM is higher than communities derived
Figure 3.12: Topic Similarity between each adjacent time periods based on DCTM by the Girvan-Newman approach, which means that the number of coauthor connections in a CTM community is fewer than that in a Girvan-Newman community. Average sKL divergence was calculated for the 500 authors in CTM communities and 500 authors in Girvan-Newman communities. The results are summarized in Table 3.5 as well.

The low sKL divergence means the nodes in a community have high topic similarity. The average sKL derived from CTM is higher than that from Girvan-Newman approach, indicating that authors in Girvan-Newman community tent to share more similar topics than those in the CTM community, because the co-author relationship reflects a strong
Figure 3.13: Communities similarity between two adjacent time periods based on CTM. semantic connection among different nodes. However, the results in Table 3.5 still indicate that DCTM can discover authors with relative high similarity of topic distribution in a detected community, while those authors may have few co-author relationships compared with the communities derived from Newman-Girvan (the authors in the DCTM community tent not to collaborate with each other). So the nodes in the community identified by the DCTM model embedded the feature of sharing similar topics but collaborating in a limited manner. That can provide meaningful recommendations for authors who would like to find potential cooperators that they do not know before.

We also calculated conductance of all communities detected by CTM and Girvan-Newman algorithm separately, and then got average conductance as 0.03343 from 132 communities detected by Girvan-Newman algorithm, got average conductance as 0.7742.
Table 3.5: Comparison between Girvan-Newman approach and CTM

<table>
<thead>
<tr>
<th></th>
<th>Community ID</th>
<th>14</th>
<th>4</th>
<th>20</th>
<th>9</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM</td>
<td>Conductance</td>
<td>0.754</td>
<td>0.765</td>
<td>0.772</td>
<td>0.795</td>
<td>0.795</td>
</tr>
<tr>
<td></td>
<td>Average sKL</td>
<td>0.043</td>
<td>0.038</td>
<td>0.045</td>
<td>0.038</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>Community ID</td>
<td>7</td>
<td>17</td>
<td>8</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>N-G</td>
<td>Conductance</td>
<td>0.0223</td>
<td>0.0326</td>
<td>0.0354</td>
<td>0.0368</td>
<td>0.0377</td>
</tr>
<tr>
<td></td>
<td>Average sKL</td>
<td>0.0252</td>
<td>0.0326</td>
<td>0.0192</td>
<td>0.0272</td>
<td>0.0308</td>
</tr>
</tbody>
</table>

from 20 communities detected by CTM. We found that the conductance of communities identified by DCTM is higher than communities derived by the Girvan-Newman approach, which means that the number of coauthor connections in a DCTM community is fewer than that in a Girvan-Newman community.

3.7 Conclusion and Future Work

In this paper, we present the CTM and DCTM to detect communities and topics. The CTM contains four observed variables and two latent variables. It can discover topic features from the four observed variables and uses the relationships to define communities. The dynamic function of DCTM takes into account the temporal continuity between consecutive timestamps that ensures the consistency for each community and topic during the whole time period. Experiments show that the CTM can find communities sharing similar topics, while the DCTM can identify the dynamic features of communities and topics. In the future, we will integrate a supervised model into the DCTM in order to capture the highly cited authors instead of highly productive authors. In order to improve computational efficiency, we will consider adopting parallel computing technology such as the MPI (Message Passing Interface) to accelerate the process.
CHAPTER 4

MODELING INTELLECTUAL TRADE AMONG SCIENTIFIC DISCIPLINES

Abstract

When nations trade they exchange goods and capital. These exchanges can be highly asymmetric, e.g. one nation could import more goods and thus export more capital than its trading partners. In this chapter we draw an analogy between international trade and scientific communication. By applying concepts from the study of international trade we measure the flow of ideas between scientific domains, i.e. intellectual trade, on the basis of the predominant direction of citations. Domains that import ideas mostly cite articles published in other domains. Vice versa, domains that export ideas have their publications more frequently cited or read by other domains than they cite or read those of other domains. We find that intellectual trade exhibits its own characteristics as a special type of knowledge exchange. To the best of our knowledge, this work is the first large-scale analysis of the citation interactions in science from the perspective of international trade.

4.1 Introduction

International trade (Leontief, 1953) occurs when nations engage in the exchange of goods and capital. International trade is not fundamentally different from trade between individual
parties in which goods are exchanged for capital. It is defined at the national level and thus pertains to the aggregate sum-total of the trade between various partners in their respective nations within a specific period of time.

Some nations produce more goods than they import and vice versa. The flow of capital and goods in international trade is rarely symmetrical. The trade relation between the United States of America (US) and the People’s Republic of China (PRC) is an often discussed (and lamented) case in point to illustrate such asymmetries (Groenewold & He, 2007). The US is a net importer of PRC goods, i.e. over the past decades it has imported more goods from PRC than it exports to the PRC. As a result, the flow of capital is also highly asymmetrical with the PRC accumulating capital in exchange for the goods that it exports to the US.

A situation similar to the asymmetries in international trade may exist in the scholarly community where some scholarly domains can be net exporters and/or consumers of scientific ideas and knowledge. For example, scientists in physics could rely disproportionally on ideas developed in mathematics whereas the opposite may not not the case. Much like a trade imbalance the flow of ideas would thus predominantly point from mathematics to physics. Citations on the other hand would mostly point from physics to mathematics as they serve to acknowledge idea originating in the latter.

To study these kinds of asymmetries in scientific communication we propose to draw an analogy between international trade and the exchange of knowledge in the scholarly community. The analogies we draw are summarized in Table 4.1. Intellectual trade is viewed from the perspective of the exchange of ideas vs. acknowledgements between intellectual trading partners who are both to some, but not necessarily equal, degree consumers and producers of information (Egghe, 2009).

We operationalize the production of knowledge as the publication of peer reviewed
articles in scholarly journals. International trade aggregates trade between individual parties at the national level. In a similar manner as shown in Fig. 4.1, we aggregate scientific trade at levels above that of individual journal articles, specifically at the level of scientific journals which are used as aggregations of individual articles. Scholarly domains are operationalized as aggregations of scientific journals demarcated by existing classification schemes. Ideas can thus be “exported” to other domains where they are read and cited. Citations are as such considered an operationalization of the “acknowledgement” or in fact “importation” of these ideas. Asymmetries between the number of citations pointing from one domain to another can then be used to detect intellectual trade imbalances in the scientific community.

Figure 4.1: Measuring imbalances of intellectual trade between journals, and domains as aggregations of journals, by means of citation asymmetries.

Although several studies have utilized journal-to-journal citations (Boyack et al., 2005) or user click stream (Bollen, Sompel et al., 2009) to visualize the topological structure of science disciplines, our study extends this line of research by analyzing the disciplinary relations from “trade” perspective. Particularly, we study intellectual trade on the basis of the 2006 Journal Citation Reports (Science and Social Science indices) which provide citation numbers at the journal level, i.e. it lists the number of citations that point from one journal to another at given point in time. We define a set of scholarly domains into which we aggregate journal citation levels to study domain-level citation imbalances from the perspective of intellectual trade. We define a range of indicators of intellectual trade
Table 4.1: Intellectual trade versus international trade

<table>
<thead>
<tr>
<th>Intellectual trade</th>
<th>International trade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idea</td>
<td>Good</td>
</tr>
<tr>
<td>Citation</td>
<td>Trade</td>
</tr>
<tr>
<td>Journal</td>
<td>City</td>
</tr>
<tr>
<td>Domain</td>
<td>Nation</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of journals in one domain</td>
<td>Size of a country</td>
</tr>
<tr>
<td>Citation frequency</td>
<td>Trade volume</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations within one domain</td>
<td>Domestic trade</td>
</tr>
<tr>
<td>Citations between one domain to other domains</td>
<td>Foreign trade</td>
</tr>
<tr>
<td>Directly cited by other domains</td>
<td>Export trade</td>
</tr>
<tr>
<td>Directly citing other domains</td>
<td>Import trade</td>
</tr>
<tr>
<td>Citation between two domains themselves</td>
<td>Bilateral trade</td>
</tr>
<tr>
<td>Citations between two domains via a third domain</td>
<td>Transit trade</td>
</tr>
</tbody>
</table>

4.2 Methodology

We extract journal to journal citation data from the Journal Citation Reports (JCR) that were issued by Thomson-Reuters (TR) in 2006. This results in a journal to journal citation network. Each journal in this network is mapped to a JCR classification code, allowing us to map the journal citation network to a domain citation network. We examine the latter for asymmetries using a variety of indicators of that reveal how a domain is involved with intellectual exports or imports to neighboring domains.
4.2.1 Citation and Classification Data

We start from the Journal Citation Reports (JCR) that were issued by Thomson-Reuters (TR) in 2006. The JCR consists of a set of adjacency lists that consist of journal pairs, i.e. journal 1 and journal 2, and the number of citations that passed from articles published in journal 1 and 2 respectively. Each adjacency list is separated according to the date in which the outgoing citations occurred and which publication year they point to. In this case we limit the JCR to 2006 citations pointing to journal publications in 2004 and 2005.

Our study focuses on intellectual trade among different scientific domains. We therefore adopt a journal classification that maps each journal to a specific domain, using the classification function $A(v) \rightarrow d$. To account for various levels of domain abstraction, e.g. physics vs. theoretical physics, we adopt the Getty Research Institute’s Art and Architecture Thesaurus (AAT) which was earlier leveraged as a journal classification system by (Bollen, Sompel, et al., 2009).

The AAT is structured as a taxonomical tree that starts with general and broad categories which are refined into increasingly finer subject areas. The AAT Journal Classification taxonomy as developed by Bollen, Sompel, et al. (2009) contains for 4 taxonomic levels. The 1st level is composed of 3 categories, i.e. natural sciences, social sciences and humanities. The 2nd level comprises 10 categories, including science, health sciences, materials science, etc. The 3rd level comprises 29 categories, including mathematics, engineering, education, chemistry, etc. The 4th level comprises 262 categories, including —it electric engineering, clinical neurology, biochemical, sport sciences, etc. Here we focus on the 3rd level of the AAT taxonomy since it provides classifications at a level that is overly broad (such as those defined at level 1 and 2) nor being too narrow to be tractable (level 4).
4.2.2 Domain Network Generation

Combining the JCR citation data and the AAT taxonomy, we can now define a Domain citation network which aggregates the citations between journals within particular domains, much like international trade is defined on the level of individual nations which group trade relations between parties that each pertain to their own nation.

The JCR citation network can be mathematically modeled as a directed, weighted graph $G$, i.e.

$$G = (V, E, W)$$

where $V$ is the set of $N$ journals, $E \subseteq V^2$ and $W$ a weight function that maps each edge $(v_i, v_j)$ to an positive, whole number $w(v_i, v_j) \in \mathbb{N}^+$ that corresponds to the number of citations passing from $v_i$ to $v_j$. The edge $(v_i, v_j)$ indicates that one or more citation edges point from journal $v_i$ to $v_j$ and thus that the articles published in journal $v_i$ imported knowledge from journal $v_j$ and conversely that $v_j$ exported knowledge to $v_i$.

We define the AAT classification function $A_k : v_i \rightarrow a_k$ which maps each journal $v_i$ to an AAT classification at level $k \in \{0, 1, 2, 3, 4\}$, the latter an integer value that represents the distance from the root of the AAT taxonomy from which classifications have been retrieved.

We can now define the domain citation network

$$G_k = (V_k, E, W_k)$$

where $V_k$ represents the set of AAT domains at classification level $k$. Each domain $d_i \in V_k$ is a subset of $V$, i.e. $d_i \subset V$, the set of all journals in the JCR. The edges $(d_i, d_j \in E = V_k^2$ thus represent the citations that pass from the journals that pertain to
Table 4.2: Network parameter for the JCR citation graph \((G)\) and the domain citation graph \((G_d)\). For diameter and average clustering coefficient we ignore the edge directionality of the JCR citation graph \((G)\).

<table>
<thead>
<tr>
<th>Network parameter</th>
<th>(G)</th>
<th>(G_d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>7388</td>
<td>24</td>
</tr>
<tr>
<td>Edges</td>
<td>637691</td>
<td>478</td>
</tr>
<tr>
<td>Average Degree</td>
<td>86.314</td>
<td>19.917</td>
</tr>
<tr>
<td>Density</td>
<td>0.012</td>
<td>0.866</td>
</tr>
<tr>
<td>Diameter</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0.395</td>
<td>0.939</td>
</tr>
</tbody>
</table>

Domain \(d_i\) to those in domain \(d_j\) respectively. The weight function \(W_k\) maps each domain citation edge \((d_i, d_j)\) at the classification level to a whole number \(w_k(d_i, d_j) \in \mathbb{N}^+\) such that

\[
    w_k(d_i, d_j) = \sum_{\{(v_i, v_j) : v_i \in d_i \land v_j \in d_j\}} w(v_i, v_j)
\]

When \(k = 3\) this procedure produces a AAT domain citation graph \((G_3)\) whose characteristics listed in Table 4.2. The structure of the domain citation network \(G_3\) is visualized in Fig. 4.2.

4.3 Results

Our motivation is to analyze intellectual trade relations by using recognized indices and concepts from the study of international trade. The following section will discuss the results of our analysis along these various indices, in particular trade volume, balance, distribution, indicators of influence in trade network, and finally bilateral and transit trade.
Figure 4.2: $G_3$, i.e. the domain citation graph produced by aggregating journal citations at $k = 3$ distance from the root contains 24 domains and 478 edges. Arrows represent the direction of citations whereas the width of edges corresponds to the number of citations. For clarity, we only showed the 250 largest weighted edges in CDN. Nodes are colored according to their PageRank, i.e. darker means higher PageRank.

### 4.3.1 Domain size relative to trade volume

The size of a nation is closely related to its significance in the international trade network. Large nations can have access to abundant resources and a sizable domestic market. These factors thus affect their ability to engage in foreign and domestic trade. We hypothesize that a similar situation occurs in intellectual trade. We define domain size as the total
number of journals in that domain, i.e. $||d_i||$. We then define trade volume as the total number of citations involving the journals of that domain, which includes domestic trade volume (citation within the same domain) and foreign trade volume (citation to and from different domains). In international trade, we refer to countries with large trade volumes as a *trading powers*. Similarly we refer to domains with a large volume of citations, regardless of whether they pertain to intra- or extra-domain citations, an intellectual *trading power*. However a distinction needs to be made with respect to foreign trade volume, i.e. the sum-total of trade between a given nation and its trading partners, and domestic trade volume, i.e. the sum-total of trade within the given nation. In a similar fashion we can make the same distinction of within- and between citation volume, and define a domain $d_i$’s external and internal trade, $T_x$ and $T_n$ respectively, as follows:

\[
T_n(d_i) = w_k(d_i, d_i)
\]

\[
T_x(d_i) = \sum_{d_j} w_k(d_i, d_j) + w_k(d_j, d_i)
\]

The scatterplot of domain size versus domestic trade volume in shown in Fig. 4.3(a), and domain size versus foreign trade volume in shown in Fig. 4.3(b). Their values of Pearson’s $r$ are 0.9 and 0.814, respectively. We can see that the domain size are highly positive correlate with both domestic and foreign trade volume.

Table 4.3 lists the top 10 domains ranked in terms of size, foreign trade and domestic trade volume. Particularly, 8 of the 10 largest domains in terms of size (number of journals), are ranked in the top 10 largest foreign trade volume. This could be expected since larger domains contain more journals and thus have a higher probability of being cited as a group from other domain. Similarly, 8 of top 10 largest domains are also ranked in the top 10 largest domestic trade domains. The largest domain, health sciences, is also ranked top 2 largest foreign trade volume as well as domestic trade volume domain, implying
its superpower position in the intellectual trade. Although the size of domain is highly correlated with trade volume (i.e. trade volume) in general, there’re some exceptions. For example, physics is ranked only 10th in the domain size, but ranked as high as 3rd in domestic trade and 6th in foreign trade. It is very similar to Japan, a small-sized country yet among one of the most important trading powers. Besides, computer science is among
Table 4.3: Top 10 scientific domains ranked by domain size, trade volume, export volume and import volume

<table>
<thead>
<tr>
<th>rank</th>
<th>Domain size</th>
<th>Foreign trade volume ($T_x$)</th>
<th>Domestic trade volume ($T_n$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>health sciences</td>
<td>chemistry</td>
<td>health sciences</td>
</tr>
<tr>
<td>2</td>
<td>engineering</td>
<td>health sciences</td>
<td>chemistry</td>
</tr>
<tr>
<td>3</td>
<td>chemistry</td>
<td>science</td>
<td>physics</td>
</tr>
<tr>
<td>4</td>
<td>science</td>
<td>biology</td>
<td>science</td>
</tr>
<tr>
<td>5</td>
<td>social sciences</td>
<td>engineering</td>
<td>engineering</td>
</tr>
<tr>
<td>6</td>
<td>sociology</td>
<td>physics</td>
<td>biology</td>
</tr>
<tr>
<td>7</td>
<td>biology</td>
<td>materials science</td>
<td>materials science</td>
</tr>
<tr>
<td>8</td>
<td>computer science</td>
<td>social sciences</td>
<td>sociology</td>
</tr>
<tr>
<td>9</td>
<td>mathematics</td>
<td>sociology</td>
<td>earth sciences</td>
</tr>
<tr>
<td>10</td>
<td>physics</td>
<td>environmental sciences</td>
<td>mathematics</td>
</tr>
</tbody>
</table>

the 8th largest domains but ranked neither the top 10 foreign trade volume nor domestic trade volume domains. This is a little like Russia, with the largest size in the whole world but plays a minor role in international trade. Whether the domain contain some “active” journals or not also affects the place of the domain in the trade. Such “active” journal to the domain is equivalent to some commercial trading center city to its country.

4.3.2 Trade Balance

Trade balance is an important index in international trade and is usually operationalized in terms of the comparison of export vs. import volume for a particular nation. Intellectual trade balance occurs when a domain’s export and import of ideas are balanced. Here we operationalize intellectual trade balance on the basis of the domain citation network $G_3$. We distinguish between intellectual trade symmetry across all relationships in the network, i.e. as a network parameter, and within a particular domain, i.e. as a node parameter.
Trade balance in the domain citation network is defined as the degree to which its edges overall exhibit symmetry. For any pair of domains \((d_i, d_j)\) at classification level \(k = 3\), we denote the weight of the citation edges \((d_i, d_j)\) as \(w(d_i, d_j)\). Intellectual trade in the domain citation network is perfectly balanced when for all pairs of domains \((d_i, d_j)\) it is true that \(w(d_i, d_j) = w(d_j, d_i)\). To assess the degree to which this symmetry holds we perform a linear regression analysis across all pairs of \((d_i, d_j)\), and determine its coefficient of determination \(R^2\), the latter serving as an indicator of the degree to which the edges in the domain citation network do conform to \(w(d_i, d_j) = w(d_j, d_i)\).

Trade balance between domains, is defined as the ratio of a domain \(d_i\)'s import vs. its export with another domain \(d_j\), i.e.

\[
T_b(d_i, d_j) = \frac{w(d_i, d_j)}{w(d_j, d_i)}
\]  

When \(T_b(d_i, d_j) > 1\) the domain \(d_i\) maintains an intellectual trade surplus to domain \(d_j\); if \(T_b(d_i, d_j) < 1\), the domain maintains an intellectual trade deficit to domain \(d_j\); if \(T_b(d_i, d_j) = 1\), the pair of domains has reached trade balance.

Fig. 4.4 shows the scatterplots of the reciprocal edges (raw weights) in \(G_3\) and \(G_4\). Each point on the scatterplot represents an edge of the domain citation network which is positioned such that its x-coordinate is \(w(d_i, d_j)\) and its y-coordinate is \(w(d_j, d_i)\). The color bar on the right side of the figure provides a color legend for the density in log scale.

We find that in Fig. 4.4(a,b) edges of \(G_3\) and \(G_4\) are closely clustered along the diagonal indicating a high degree of Trade Balance in the domain citation network. Furthermore, the amount of scatter decreases as \(w_{i,j}\) values increase along the diagonal implying that as the edge weights grow the degree of symmetry grows correspondingly. This could be the
<table>
<thead>
<tr>
<th>Raw edge weight ($G_3$)</th>
<th>$r$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.973</td>
<td>0.937</td>
</tr>
<tr>
<td>Normalized edge weights ($G_3$)</td>
<td>0.210</td>
<td>-0.224</td>
</tr>
<tr>
<td>Raw edge weights ($G_4$)</td>
<td>0.858</td>
<td>0.727</td>
</tr>
<tr>
<td>Normalized edge weights ($G_4$)</td>
<td>0.194</td>
<td>-0.689</td>
</tr>
</tbody>
</table>

Table 4.4: Pearson’s $r$ and $R^2$ values for linear regression over all reciprocal edges in the domain citation network at classification level 3 and 4, i.e. $G_3$ and $G_4$.

result of large domains connecting to other large domains. Since large domains are more likely to have high magnitude citation edges, across the network the magnitude of citation edge weights would be strongly related to that of their reciprocal edges.

In fact, the scatterplots in Fig. 4.4(a,b) are based on raw citation numbers, i.e. $w_{i,j}$ values. Large domains will however contain larger numbers of journals than smaller domains, and therefore accumulate higher citation numbers. Comparing the citations weights of connected domains with differing size will thus confound their relative sizes and ability to reach trade balance. For example, if A cites B 10 times while being cited by B 10 times, their trade seems symmetric in terms of citation frequency. However, if A cites all other domains 100 times, while B cites all other domains 20 times, then the rate of A citing B is 0.1 while the rate of B citing A is 0.5. A is thus a very important knowledge source to B while the reverse is not true, but this is not obvious from the raw domain citation counts.

To mitigate this effect we normalize the weight of an outlink edge by the sum of all outlink weights for each domain. The resulting normalized weights indicate the relative importance of a domain to its exporting trade partners.
Figure 4.4: (a) Scatterplot of $w_{i,j}$ vs. $w_{j,i}$ in $G_3$; (b) Scatterplot of $w_{i,j}$ vs. $w_{j,i}$ in $G_4$; (see Eq. 4.4).

The original domain citation weight $w_{i,j}$ thus reflects the actual “citation frequency” between two scientific domains while the normalized citation weight $w'_{i,j}$ characterizes the “citation rate” or “citation probability” of one domain citing another.

We show the result of this citation weight normalization in Fig. 4.5(a,b) for $G_3$ and
$G_4$, i.e. at the 3rd and 4th level of AAT classification. The degree of symmetry is greatly reduced after normalization, implying that the observed “big domain effect” does not apply to the relative importance of intellectual trading relations which are found to be highly asymmetric. We list the corresponding $R^2$ of the subsequent linear regression in Table 4.4 which confirms these visual observations.

![Figure 4.5](image1.png)  
(a) Scatterplot of $w'_{i,j}$ vs. $w'_{j,i}$ in $G_3$; (b) Scatterplot of $w'_{i,j}$ vs. $w'_{j,i}$ in $G_4$; (see Eq. 4.4).

At this point we’ve only assessed trade balance as a network parameter of $G_3$. However, using Eq. 4.2 we can track the degree of intellectual trade balance for individual
domains which can yield information on which scientific domains are the largest exporters and importers of knowledge.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>law</td>
<td>3.039</td>
</tr>
<tr>
<td>2</td>
<td>information science</td>
<td>2.305</td>
</tr>
<tr>
<td>3</td>
<td>communications</td>
<td>1.543</td>
</tr>
<tr>
<td>4</td>
<td>geography</td>
<td>1.536</td>
</tr>
<tr>
<td>5</td>
<td>planning disciplines</td>
<td>1.364</td>
</tr>
<tr>
<td>6</td>
<td>education</td>
<td>1.241</td>
</tr>
<tr>
<td>7</td>
<td>materials science</td>
<td>1.159</td>
</tr>
<tr>
<td>8</td>
<td>engineering</td>
<td>1.157</td>
</tr>
<tr>
<td>9</td>
<td>biology</td>
<td>1.111</td>
</tr>
<tr>
<td>10</td>
<td>social sciences</td>
<td>1.105</td>
</tr>
<tr>
<td>11</td>
<td>environmental sciences</td>
<td>1.105</td>
</tr>
<tr>
<td>12</td>
<td>earth sciences</td>
<td>1.102</td>
</tr>
<tr>
<td>13</td>
<td>earth sciences</td>
<td>1.094</td>
</tr>
<tr>
<td>14</td>
<td>mathematics</td>
<td>1.076</td>
</tr>
<tr>
<td>15</td>
<td>chemistry</td>
<td>1.055</td>
</tr>
<tr>
<td>16</td>
<td>physics</td>
<td>1.024</td>
</tr>
<tr>
<td>17</td>
<td>philosophy</td>
<td>0.967</td>
</tr>
<tr>
<td>18</td>
<td>computer science</td>
<td>0.947</td>
</tr>
<tr>
<td>19</td>
<td>history</td>
<td>0.938</td>
</tr>
<tr>
<td>20</td>
<td>classics</td>
<td>0.926</td>
</tr>
<tr>
<td>21</td>
<td>sociology</td>
<td>0.811</td>
</tr>
<tr>
<td>22</td>
<td>political science</td>
<td>0.723</td>
</tr>
<tr>
<td>23</td>
<td>science</td>
<td>0.648</td>
</tr>
<tr>
<td>24</td>
<td>economics</td>
<td>0.497</td>
</tr>
</tbody>
</table>

Figure 4.6: Domains $T_b(d_i)$ ranked by intellectual trade balance as defined in Eq. 4.2.

In Table and Fig. 4.6 we ranked domains in $G_3$ according to their intellectual trade balance $T_b(d_i)$ as defined in Eq. 4.2. We observe that law and information science are highly export-oriented; they both exhibit uniquely high trade surpluses in knowledge exchange meaning that they export knowledge to other domains 3 to 2 times more than they import ($T_b(d_i) = \{3.039, 2.305\}$).

The same ranking shows that Economics and political science are strong importers and maintain high trade deficit indicating that they have a much higher propensity to borrow knowledge from other domains than they are to export knowledge.

Just like a country in the international trade, a scientific domain with either high trade surplus or import-substitution will impede its development. Those domains with the ratio of export to import slightly more than 1.0, like health science, social sciences, are right
those currently highly developed domains and trading powers.

### 4.3.3 Dependence on External Trade

Nations can be dependent on foreign trade to various degrees. Some nations hardly engage in foreign trade, others are strongly dependent on their ability to export and import goods. Trade dependence is frequently operationalized in terms of the ratio value of the foreign trade over the Gross Domestic Product (GDP) or Gross National Product (GNP). In other words, it is defined as the ratio of the sum of import and export volume to GDP.

Similarly, for intellectual trade we characterize the dependence on knowledge exchange with other domains as the ratio of a domain’s total “trade” volume to total trade volume. If a domain has a highly foreign dependence degree, we think of it as *exogamous*, otherwise we think of it as *endogamous*.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>planning disciplines</td>
<td>0.937</td>
</tr>
<tr>
<td>2</td>
<td>communications</td>
<td>0.891</td>
</tr>
<tr>
<td>3</td>
<td>classics</td>
<td>0.888</td>
</tr>
<tr>
<td>4</td>
<td>philosophy</td>
<td>0.876</td>
</tr>
<tr>
<td>5</td>
<td>biology</td>
<td>0.856</td>
</tr>
<tr>
<td>6</td>
<td>history</td>
<td>0.843</td>
</tr>
<tr>
<td>7</td>
<td>social sciences</td>
<td>0.841</td>
</tr>
<tr>
<td>8</td>
<td>geography</td>
<td>0.837</td>
</tr>
<tr>
<td>9</td>
<td>environmental sciences</td>
<td>0.836</td>
</tr>
<tr>
<td>10</td>
<td>information science</td>
<td>0.830</td>
</tr>
<tr>
<td>11</td>
<td>engineering</td>
<td>0.803</td>
</tr>
<tr>
<td>12</td>
<td>materials science</td>
<td>0.766</td>
</tr>
<tr>
<td>13</td>
<td>education</td>
<td>0.765</td>
</tr>
<tr>
<td>14</td>
<td>science</td>
<td>0.751</td>
</tr>
<tr>
<td>15</td>
<td>sociology</td>
<td>0.740</td>
</tr>
<tr>
<td>16</td>
<td>computer science</td>
<td>0.739</td>
</tr>
<tr>
<td>17</td>
<td>earth sciences</td>
<td>0.685</td>
</tr>
<tr>
<td>18</td>
<td>physics</td>
<td>0.581</td>
</tr>
<tr>
<td>19</td>
<td>political science</td>
<td>0.575</td>
</tr>
<tr>
<td>20</td>
<td>mathematics</td>
<td>0.569</td>
</tr>
<tr>
<td>21</td>
<td>chemistry</td>
<td>0.569</td>
</tr>
<tr>
<td>22</td>
<td>economics</td>
<td>0.489</td>
</tr>
<tr>
<td>23</td>
<td>health sciences</td>
<td>0.378</td>
</tr>
<tr>
<td>24</td>
<td>law</td>
<td>0.363</td>
</tr>
</tbody>
</table>

**Figure 4.7: Percentage of Foreign Trade in Total Trade**

From Figure 4.7, we can see that *planning disciplines* is the most exogamous domain,
while law is the most endogamous domain. We can infer that the development of planning disciplines mainly depends on external influence, while law mainly relies on self-development, i.e., academic communication within law journals. It is kind of surprising that the two largest trading power health science and chemistry have very low foreign dependence degree. On the contrast, some “small” domains depend heavily on foreign trade, like communication, classics. Similar situation occurs in international trade as well. Some trading power also has large size and population, whose domestic needs constitute its main source of total trade. USA is the best example. More small countries, like korea, singapo only can survive through trading with other nations because of their limited resources and population. Similarly, large scientific domains can develop by scholarly communication between journals within domains, while small domains have to exchange with other domains. Additional interesting observation is that: most of the natural sciences disciplines tends to have low foreign dependence degree, like mathematics, chemistry and physics; while most the social sciences and humanity disciplines tends to have high foreign dependence degree, like planning disciplines, classics, and philosophy. It implies that social sciences disciplines rely more on cross-domain communication than natural sciences to sustain disciplinary development.

4.3.4 Direction of Foreign Trade

The direction of a foreign trade of a country refers to the distribution of its importing and exporting target countries. It reflects the main trading partners as well as their relative importance to this country. When it comes to intellectual trade, we studied how a domain exchange knowledge with its relatively fixed neighbor domains. Gini coefficient as a economic index has already been used in citation pattern analysis. It is a measure of inequality of distribution, a value of 0 expressing total equality and a value of 1 expressing maximal
inequality, which can be calculated as (Buchan, 2002):

$$G_D = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{(n - 1)\sum_{i=1}^{n}x_i}$$

where $G_D$ is the Gini coefficient of domain $D$, $n$ is the number of domains that $D$ citing (if calculating export Gini) or being cited (if calculating import Gini), and $x_i$ is the number of citations with $i$ being the rank of the number of citations in ascending order.

![Figure 4.8: Export Gini Coefficient](image)

Figure 4.8 and 4.9 show that although different domains have different export/import gini coefficients, overall the average export and import gini coefficient is above 0.6. It implies that, for most of domains, the direction of foreign trade is far from evenly distributed among all scientific domains. Instead, each domain has its VIP partners, less important partners as well as rare interactive partners. Again, the similar phenomenon is spotted in internal trade, since each country will tends to trade with some fixed countries. Another interesting finding is that export gini coefficient is highly positive correlated with import gini coefficient (shown in Figure 4.10), implying that a domain’s export trading direction is large overlapped with
Figure 4.9: Import Gini Coefficient

Figure 4.10: Correlation between export gini and import gini. The correlation factor is 0.868

its import trading direction.

We selected *information science* domain as the case study for the direction of foreign
trade. From the Figure 4.11 we can see that, computer science plays the most important role in the knowledge exchange with information science, regardless of import or export. information science import amounts of knowledge from some domains, like physics, history, mathematics while never export to them any knowledge. It implies that a lot of journals in information science citing journals in those domains but seldom being cited. Besides, information science has very limited export or import activities with economics, biology,
materials science, etc, all of which play a minor role in the development of information science. All of above give us an impression that both the export and import structure of information science have its characteristic hierarchy.

Besides Gini coefficient, Information Entropy is also a useful index that characterizes the distribution. However, the main difference between Gini coefficient and Information Entropy is that high Gini coefficient reflects high inequality, while high Information Entropy reflects high equality. Therefore, we define an exchange entropy (EE) for a domain D:

$$EE_D = -\sum_{i=1}^{N-1} P(d_i) \ast \log_2(P(d_i))$$  \hspace{1cm} (4.6)

$$P(d_i) = \frac{E(d_i)}{\sum_{i=1}^{N-1} E(d_i)}$$  \hspace{1cm} (4.7)

in above equations, $d_i, i = 1, 2, ..., N - 1$ represents the all the domains except $D$, and $N$ represents the total number of domains. $E(d_i)$ represents the total number of exchange weights between $D$ and $d_i$. After we obtain $EE_D$, we normalize it by maximum entropy $\log_2(N - 1)$ and obtain Normalized Exchange Entropy (NEE). If $NEE_D$ is close to 1, it means $D$ has evenly distributed exchange with all other domains; if $NEE_D$ is close to 0, it means $D$ almost only exchanges with 1 or 2 domains.

The ranking of NEE is shown in Figure 4.12. What can NEE actually tell us about how a domain exchange knowledge with its neighboring domains? The example of planning disciplines and materials sciences is shown in Figure 4.13 to explain why the NEE of the former is larger than the latter. Domain materials sciences interacts with other 17 domains but almost concentrate its interactions with chemistry and physics. By contrast, the number of neighbors of planning disciplines is a litter smaller than materials sciences, but most of the interactions between planning disciplines and neighbors are relatively uniformly distributed among social science, economics, political science, geography, environmental sciences as well.
Figure 4.12: Normalized Exchange Entropy

as health sciences. In other words, the main direction of foreign trade of planning disciplines are actually broader than materials sciences.

4.3.5 Trading Center

Trading centers are those countries or regions in the core role of facilitating and organizing the whole international trade. Trading centers are considered to have very broad commercial ties to most of the countries all over the world. In complex network, centrality is such a indicator to characterize “how central a node is” within network. There are three main measures of centrality that are widely used in network analysis: degree centrality, closeness centrality and betweenness centrality. Degree centrality is defined as the number of links a node has. If the links are directed, each node has two measures of degree centrality, namely indegree and outdegree. Closeness centrality is defined as the inverse of average shortest path length from one node to all other nodes. Betweenness centrality of a node is defined as
Figure 4.13: Comparison of exchange with neighbors between "planning disciplines" and "materials science". For both subfigure, we combine the output weights and input weights for the center domain thus the directed edges become indireced.

Let’s again make a comparison between intellectual trade and international trade to
interpret the implication of centrality measures. Domain with high indegree centrality is very similar to import center country like India; Domain with high outdegree centrality is very similar to export center country as Germany; Domain with high closeness centrality is very similar to the international center trade country like China and US; Domain with high betweenness is very similar to famous port like Hong Kong. In addition, as a measure of communication between different scientific domains, the centrality is closely associated with other two important concepts: multidisciplinary and interdisciplinary. More specifically, we think that degree centrality or closeness centrality is the indicator of multidisciplinary, because it measures how frequent a domain interact with various other domains; betweenness centrality is the indicator of interdisciplinary, because it evaluate whether the position of a domains lies in the intersect region of different domains.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>engineering</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>education</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>social sciences</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>biology</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>health sciences</td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>science</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>environmental sciences</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>sociology</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>computer science</td>
<td>0.936</td>
</tr>
<tr>
<td>7</td>
<td>political science</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>chemistry</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>economics</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>communications</td>
<td>0.956</td>
</tr>
<tr>
<td>14</td>
<td>mathematics</td>
<td>0.913</td>
</tr>
<tr>
<td>15</td>
<td>information science</td>
<td>0.869</td>
</tr>
<tr>
<td>16</td>
<td>earth sciences</td>
<td>0.826</td>
</tr>
<tr>
<td>17</td>
<td>geography</td>
<td>0.783</td>
</tr>
<tr>
<td>17</td>
<td>physics</td>
<td>0.783</td>
</tr>
<tr>
<td>17</td>
<td>history</td>
<td>0.783</td>
</tr>
<tr>
<td>20</td>
<td>philosophy</td>
<td>0.739</td>
</tr>
<tr>
<td>21</td>
<td>materials science</td>
<td>0.652</td>
</tr>
<tr>
<td>21</td>
<td>classics</td>
<td>0.652</td>
</tr>
<tr>
<td>23</td>
<td>planning disciplines</td>
<td>0.565</td>
</tr>
<tr>
<td>24</td>
<td>law</td>
<td>0.522</td>
</tr>
</tbody>
</table>

Figure 4.14: Indegree

Figure 4.14 and 4.15 show the indegree/outdegree centrality of all the domains. Figure 4.16 shows the correlation between the two types of degree centrality. Comparing to
the export and import gini correlation, the outdegree and indegree centrality correlation is relatively low. A typical case is information science, which ranked nearly in the middle of position in terms of indegree centrality but ranked the last in terms of outdegree centrality.
Figure 4.17 and 4.18 show some interesting findings, especially when compare with tabel 1. *Health science* ranked 1st in both closeness and betweenness centrality, which is quite natural because the domain has the largest size and the second largest foreign trade volume. In addition, *information science* ranked almost at the bottom in both closeness and betweenness centrality because of its small size and trade volume, reflecting its marginal position in the whole intellectual trade. However, the size and trade volume of a domain cannot always reflect its centrality. For example, as a medium-sized domain, *social sciences* has smaller trade volume than some other larger domains, such as *engineering, chemistry*, but is still located in the top centrality position. Similarly, *environmental sciences* lies in a more centralized position than those big domains and trading powers, even though it has much smaller size and trading volume. Besides, as the largest foreign trading power with the 3rd largest size,*chemistry* is ranked 4th in closeness centrality but only 11th in betweenness centrality. It shows that, although *chemistry* focus on multi-disciplinary communication with
its “disciplinary intimate” domains, but it itself is not a very important bridge connecting relatively dissimilar domains.

4.3.6 Bilateral Trade and Transit Trade

The trade exclusively occurring between two nations is called bilateral trade. In intellectual trade, bilateral trade is the direct citation between two domains. That is almost the most common and explicit way of knowledge exchange. The edge weight is just the representation of the amount of bilateral trade in the intellectual trade. As we described before, there’re two kinds of weight we used: one is the raw citation frequency count, the other is the output citation rate or probability, i.e. normalized weight. In the following table, we listed the top 10 bilateral trade partners ranked by either raw count or normalized weight. From the above table, as we expected, the largest bilateral trade amount in terms of raw citation count are dominated by those communication between trading powers, like health sciences,
Table 4.5: Top 10 bilateral trade

<table>
<thead>
<tr>
<th>Raw Weight</th>
<th>Normalized Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>domain pair</td>
<td>value</td>
</tr>
<tr>
<td>health sciences $\rightarrow$ chemistry</td>
<td>114667</td>
</tr>
<tr>
<td>chemistry $\rightarrow$ health sciences</td>
<td>110048</td>
</tr>
<tr>
<td>chemistry $\rightarrow$ science</td>
<td>79806</td>
</tr>
<tr>
<td>health sciences $\rightarrow$ science</td>
<td>73936</td>
</tr>
<tr>
<td>science $\rightarrow$ chemistry</td>
<td>51870</td>
</tr>
<tr>
<td>biology $\rightarrow$ chemistry</td>
<td>44598</td>
</tr>
<tr>
<td>biology $\rightarrow$ health sciences</td>
<td>44441</td>
</tr>
<tr>
<td>health sciences $\rightarrow$ biology</td>
<td>43535</td>
</tr>
<tr>
<td>science $\rightarrow$ health sciences</td>
<td>43111</td>
</tr>
<tr>
<td>chemistry $\rightarrow$ physics</td>
<td>41959</td>
</tr>
</tbody>
</table>

chemistry, biology, etc. However, the bilateral trade in terms of normalized count shows some interesting result, because you can found some trade between trading powers and small domains. For example, the trade from philosophy to health sciences ranked 2nd in the list. It seems a little surprising that philosophy and health sciences have close bilateral relation. We checked it at the journal level and found out that the two most cited journals in philosophy are Journal of Medical Ethics and Journal of Law Medicine and Ethics. The two are mainly cited by medical related journals from health sciences domain, which can explain the bilateral relation between the two domains.

Sometimes, the trade between two countries are not directed but through one or more third-party countries. Such types of trade are called transit trade. This is also a very common phenomenon in any flow network and a theoretical assumption for well-known max flow problem. For example, in CDN, there exists such a transit trade path: engineering $\rightarrow$ education $\rightarrow$ sociology. It implies that it is possible that the idea originated in engineering may finally reach to sociology via education. It is true that such remote influence between two domains doesn’t always hold but at least it provides a possibility for indirect trade.
between the two.

To quantify the “transit trade” volume in scholarly communication between two domains, we multiple all citing probabilities from all possible paths between the two domains, excluding direct path between the two. In addition, if the number of intermediate domains between the two domains is more than one, the indirect trade between the two becomes so weak that can be ignored. So we only consider the transit trade between two domains connecting by one domain only. In the following table, the rank of transit trade volume is shown. Now let’s compare transit trade in Table 4.6 with bilateral trade in terms of normalized weight in Table 4.5. Both of the two trades share a overall tendency that the largest trades always flows from small domains, like *classics, philosophy, sociology*, to trading giants, like *health sciences* and *chemistry*. However, different from bilateral trade where the trade is almost between two domains sharing disciplinary similarity, the transit trade often occurs between two domains with different disciplinary property. In other words, the transit trade always flows from a domain of social sciences to the other domain of natural sciences.

<table>
<thead>
<tr>
<th>domain pair</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>classics − − − &gt; health sciences</td>
<td>0.404</td>
</tr>
<tr>
<td>political science − − − &gt; health sciences</td>
<td>0.289</td>
</tr>
<tr>
<td>philosophy − − − &gt; chemistry</td>
<td>0.288</td>
</tr>
<tr>
<td>education − − − &gt; health sciences</td>
<td>0.285</td>
</tr>
<tr>
<td>sociology − − − &gt; chemistry</td>
<td>0.279</td>
</tr>
<tr>
<td>environmental sciences − − − &gt; chemistry</td>
<td>0.257</td>
</tr>
<tr>
<td>economics − − − &gt; health sciences</td>
<td>0.256</td>
</tr>
<tr>
<td>social sciences − − − &gt; chemistry</td>
<td>0.254</td>
</tr>
<tr>
<td>law − − − &gt; health sciences</td>
<td>0.241</td>
</tr>
<tr>
<td>earth sciences − − − &gt; health sciences</td>
<td>0.232</td>
</tr>
</tbody>
</table>
sciences.

4.4 Conclusions and Implications

In this paper, a large-scale study on the citation interactions in the whole science filed through comparative analysis with international trade is shown. Our main findings include:

- Intellectual trade and international trade shares a lot of similarities. Most recognized statistic indices and concepts from in international trade, such as export, import, trade volume, foreign dependence degree, bilateral trade, etc, have counterpart in intellectual trade.
- Trade balance are more easily to be established between large trading amount domains. If we eliminate the trading volume effect, the whole scientific communication is highly asymmetric.
- Domains with small trading volume tend to have bigger foreign dependence degree than large domains.
- The exchange between different domains is far from evenly distributed but shows apparent hierarchical structure. For each domain, its connected domain can be ranked from most intimate trading partner, frequent trading partner, infrequent trading partner, to never trading partner. The factors that will affect the frequency of two different domains lies on its similarity.
- Generally speaking, centrality is positively correlated with the size and trading volume of domains. However, some exceptions exist, like chemistry, social sciences and environmental sciences.
- Bilateral trade and transit trade shows an overall idea flow direction after eliminating the trading volume effect. That is, small domains spread their intellectual influence to the whole science via extensively exporting their ideas to large domains.
The analysis taken in this paper can help some domain leaders establish a global view of the domain’s communication with other domains and make up future development policy. All of the indices listed in our paper shows that health sciences is the most powerful and advanced scientific domain. It has very frequent and broad interactive citation with other domains. At the same time, it does not purely depend on cross domain citation but emphasize on citation between journals within its domain. In addition, it focus on multi-disciplinary and interdisciplinary academic communication. All of these contribute to the confidence of domain leaders about its bright future development. On the contrast, another seemingly well-developed domain, information science actually exposes many weakness. It has relatively low intellectual trade volume, in both domestic and foreign trade; Its high trade surplus makes it lack external knowledge support; Its high foreign dependence degree exhibit the shortness of self-development capability; The measure of gini coefficient, entropy as well centrality all show, that the range of its communication with other domains is narrow and limited. All of these are warnings that needs attention from the information science leaders and policy makers.
CHAPTER 5

THE RELATION BETWEEN SOCIAL ATTENTION AND SCHOLARLY IMPACT

Abstract

We analyze the online response to the preprint publication of a cohort of 4,606 scientific articles submitted to the preprint database arXiv.org between October 2010 and May 2011. We study three forms of responses to these preprints: downloads on the arXiv.org site, mentions on the social media site Twitter, and early citations in the scholarly record. We perform two analyses. First, we analyze the delay and time span of article downloads and Twitter mentions following submission, to understand the temporal configuration of these reactions and whether one precedes or follows the other. Second, we run regression and correlation tests to investigate the relationship between Twitter mentions, arXiv downloads and article citations. We find that Twitter mentions and arXiv downloads of scholarly articles follow two distinct temporal patterns of activity, with Twitter mentions having shorter delays and narrower time spans than arXiv downloads. We also find that the volume of Twitter mentions is statistically correlated with arXiv downloads and early citations just months after the publication of a preprint, with a possible bias that favors highly mentioned articles.
5.1 Introduction

The view from the “ivory tower” is that scholars make rational, expert decisions on what to publish, what to read and what to cite. In fact, the use of citation statistics to assess scholarly impact is to a large degree premised on the very notion that citation data represent an explicit, objective expression of impact by expert authors (Rubin, 2010). Yet, scholarship is increasingly becoming an online process, and social media are becoming an increasingly important part of the online scholarly ecology. As a result, the citation behavior of scholars may be affected by their increasing use of social media. Practices and considerations that go beyond traditional notions of scholarly impact may thus influence what scholars cite.

Recent efforts have investigated the effect of the use of social media environments on scholarly practice. For example, some research has looked at how scientists use the microblogging platform Twitter during conferences by analyzing tweets containing conference hashtags (Letierce, Passant, Decker, & Breslin, 2010; Weller et al., 2011). Other research has explored the ways by which scholars use Twitter and related platforms to cite scientific articles (Priem & Costello, 2010a; Weller & Puschmann, 2011b). More recent work has shown that Twitter article mentions predict future citations (Eysenbach, 2011). This article falls within, and extends, these lines of research by examining the temporal relations between quantitative measures of readership, Twitter mentions, and subsequent citations for a cohort of scientific preprints.

We study how the scientific community and the public at large respond to a cohort of preprints that were submitted to the arXiv database (http://arxiv.org), a service managed by Cornell University Library, which has become the premier pre-print publishing platform in physics, computer science, astronomy, and related domains. We examine the relations between three types of responses to the submissions of this cohort of pre-prints, namely the number of Twitter posts (tweets) that specifically mention these pre-prints,
downloads of these pre-prints from the arXiv.org web site, and the number of early citations that the 70 most Twitter-mentioned preprints in our cohort received after their submission. In each case, we measure total volume of responses, as well as the delay and span of their temporal distribution. We perform a comparative analysis of how these indicators are related to each other, both in magnitude and time.

Our results indicate that download and social media responses follow distinct temporal patterns. Moreover, we observe a statistically significant correlation between social media mentions and download and citation count. These results are highly relevant to recent investigations of scholarly impact based on social media data (Priem & Hemminger, 2010a; Priem, Piwowar, & Hemminger, 2011) as well as to more traditional efforts to enhance the assessment of scholarly impact from usage data (Bollen, Van de Sompel, et al., 2009; Bollen, Van de Sompel, & Rodriguez, 2008; Brody, Harnad, & Carr, 2006a; Kurtz & Bollen, 2010b).

5.2 Data and Study Overview

5.2.1 Data collection

Our analysis is based on a corpus of 4,606 scientific articles submitted to the preprint database arXiv between October 4, 2010 and May 2, 2011. For each article in this cohort, we gathered information about their downloads from the arXiv server weekly download logs, their daily number of mentions on Twitter using a large-scale collection of Twitter data collected over that period, and their early citations in the scholarly record from Google Scholar. Table 5.1 summarizes the discussed data collection and Figure 5.1 provides an overview of the data collection timelines.

The datasets employed in this study are:

- **ArXiv downloads**: For each article in the aforementioned cohort we retrieved their weekly download numbers from the arXiv logs for the period from October 4, 2010 to
Table 5.1: Overview of data collected for a cohort of 4,606 articles submitted to the preprint database arXiv between October 4, 2010 and May 2, 2011.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>N</th>
<th>Articles</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>arXiv downloads</td>
<td>2,904,816</td>
<td>4,606</td>
<td>October 4, 2010 to May 9, 2011</td>
</tr>
<tr>
<td>Twitter mentions</td>
<td>5,752</td>
<td>4,415</td>
<td>October 4, 2010 to May 9, 2011</td>
</tr>
<tr>
<td>Early citations</td>
<td>431</td>
<td>70</td>
<td>October 4, 2010 to September 30, 2011</td>
</tr>
</tbody>
</table>

Figure 5.1: Timeline of data collection. Our cohort consists of all papers submitted to arXiv between October 4, 2010 and May 2, 2011. Weekly article downloads and daily Twitter mentions were recorded after the article’s submission date, up to May 9, 2011. Early citation counts for each article were manually recorded from Google Scholar on September 30th, 2011.

May 9, 2011. A total of 2,904,816 downloads were recorded for 4,606 articles.

- **Twitter mentions**: Our collection of tweets is based on the Gardenhose, a data feed that returns a randomly sampled 10% of all daily tweets. A Twitter mention of arXiv article was deemed to have occurred when a tweet contained an explicit or shortened link to an arXiv paper (see “Materials” appendix for more details). Between October 4, 2010 and May 9, 2011 we scanned 1,959,654,862 tweets in which 4,415 articles out of 4,606 in our cohort were mentioned at least once, i.e. approximately 95% of the cohort. Such a wide coverage of arXiv articles is mostly due to specialized bot accounts which post arXiv submissions daily. The volume of Twitter mentions of arXiv papers
was very small compared to the total volume of tweets in period, with only 5,752
tweets containing mentions of papers in the arXiv corpus. We found that 2,800 out of
5,752 tweets are from non-bot accounts. After filtering out all tweets posted by bot
accounts, we retain 1,710 arXiv articles out of 4,415 that are mentioned on Twitter by
non-bot accounts. Including or excluding bot mentions, the distribution of number of
tweets over all papers was very skewed; most papers were mentioned only once, but
one paper in the corpus was mentioned as much as 113 times.

- **Early citations**: We manually retrieved citation counts from Google Scholar for the
  70 most Twitter-mentioned articles in our cohort. Citation counts were retrieved on
  September 30, 2011 and date back to the initial submission date in arXiv. All 70
  articles combined were cited a total of 431 times at that point. The most cited article
  in the corpus was cited 62 times whereas most articles received hardly any citations.

  By the nature of our research topic, we are particularly focused on *early* responses to
  preprint submissions, i.e., immediate, swift reactions in the form of downloads, Twitter
  mentions, and citations. Therefore, we record download statistics and Twitter mention
data only one week over the submission period itself (up to May 9, 2011). As for citation
data, we are aware that citations take years to accrue. We do not explore here long-term
citation effects, but only the early, immediate response to pre-print submission in the form
of citations in the scholarly record. Our citation data pertains to a time period that spans
from 5 months to 1 year: it is a fraction of the expected amount of “maturation time” for
citation analysis. Citation data must therefore be considered to reflect “early citations”
only, not total potential citations.
5.2.2 Definitions: Delay and Time Span.

Twitter mentions and arXiv downloads may follow particular temporal patterns. For example, for some articles downloads and mentions may take weeks to slowly increase after submission, whereas for other articles downloads may increase very swiftly after submission to wane very shortly thereafter. The total number of downloads and mentions is orthogonal to these temporal effects, and could be different in either case.

The two parameters that we use to describe the temporal distributions of arXiv downloads and Twitter mentions are delay and the time span, which we define as follows. Let \( t_0 \in \mathbb{N}^+ \) be the date of submission for article \( a_i \). We represent both arXiv downloads and Twitter mentions for article \( a_i \) as the time series \( T \), the value of which at time \( t \) is given by the function \( T(a_i, t) \in \mathbb{N}^+ \). We then define the time of the first, maximum, and last arXiv download of article \( a_i \) as \( T_{\text{first}}(a_i) \), \( T_{\text{max}}(a_i) \), and \( T_{\text{last}}(a_i) \) respectively:

\[
T_{\text{first}}(a_i) = \min\{t : T(a_i, t) > 0\}
\]

\[
T_{\text{max}}(a_i) = t : \max(T(a_i, t))
\]

\[
T_{\text{last}}(a_i) = \max\{t : T(a_i, t) > 0\}
\]

The delay, \( \Theta(a_i) \), and span, \( \Delta(a_i) \), of the temporal distribution of arXiv downloads for article \( a_i \) will then be defined as:

\[
\Theta(a_i) = T_{\text{last}}(a_i) - t_0
\]

\[
\Delta(a_i) = T_{\text{last}}(a_i) - T_{\text{first}}(a_i)
\]

To distinguish between the delay and span of arXiv downloads and twitter mentions, we simply denote \( \Theta_{\text{ax}}(a_i) \), \( \Delta_{\text{ax}}(a_i) \), \( \Theta_{\text{tw}}(a_i) \), \( \Delta_{\text{tw}}(a_i) \) respectively which are defined according to the above provided definitions.

As shown in Figure 5.2, the delay is thus measured as the time difference between the date of a preprint submission and a subsequent spike in Twitter mentions (the day in
which an article receives the highest volume of related tweets) or arXiv downloads (the
day in which it receives the highest volume of downloads). The time span is the temporal
“duration” of the response, measured as the time lag between the first and the last Twitter
mention or download of the article in question.

Figure 5.2: Span and delay of temporal distribution of arXiv downloads or Twitter mentions
over time expressed in terms of time passed between submission of article and peak and
time passed between first and last event, respectively.

To illustrate delay and span, we examine in detail the response dynamics for an article
in the corpus, in Figure 5.3. The article in question was submitted to arXiv on October
14, 2010. Time runs horizontally from left to right. Downloads and Twitter mentions are
charted over time (weekly for downloads, daily for mentions). As Figure 5.3 shows, the
Twitter response to submission occurs within a day, reaching a peak of nearly 40 daily
mentions within several days, and then slowly dies out over the course of the following
week. The peak of arXiv downloads, with over 16,000 weekly downloads, occurs a couple of
weeks after submission, and continues to be marked by downloads for months. From a post
hoc, ergo propter hoc point of view, in this case the Twitter response occurs immediately
and nearly exactly before the peak in arXiv reads, suggesting that social media attention
may have led to subsequently higher levels of arXiv downloads.
Figure 5.3: Response dynamics (Twitter mentions and arXiv downloads) for a selected arXiv preprint. As shown, for this particular example, Twitter mentions spikes shortly after submission and publication, and wane quickly with very mentions after the initial burst. ArXiv downloads peak shortly afterwards but continue to exhibit significant activity many weeks later.

5.3 Results

In this section, we present three results: descriptive statistics of arXiv downloads and Twitter mentions, a temporal analysis of time span and delay in arXiv downloads and Twitter mention, and a regression analysis between arXiv downloads, Twitter mentions, and early citations. For the descriptive statistics, we keep all 5,752 tweets and 4,415 articles mentioned on Twitter, since we want to show a full picture of our data. For the subsequent temporal and regression analysis we only focus on the 2,800 tweets and 1,710 arXiv articles mentioned by non-bot accounts to avoid spurious effects introduced by automated bot
accounts.

5.3.1 Domain-level Descriptive Statistics

Some descriptive statistics about the datasets analyzed in this article are presented in Figure 5.4. The first row of plots in Figure 5.4 displays the arXiv subject domains of (a) downloaded, and (b) Twitter mentioned papers (by percentage). A full list of the subject domain abbreviations used in these plots is available in the Materials section, Table 5.3. We observe a broad and evenly spread distribution of subject domains for downloads and mentions: most papers downloaded and mentioned on Twitter relate to Physics, in particular Astrophysics, High Energy Physics, and Mathematics. The second row of plots in Figure 5.4 displays the temporal distributions of (c) downloads, and (d) Twitter mentions (the dotted line in both figures is obtained by fitting a 3rd order polynomial function for smoothing). As shown in Figure 5.4(c), download counts of articles increase over time. This may be partly caused by a cumulative effect: papers that were published earlier have had more time to accumulate reads than papers that were published later. Figure 5.4(d), however, shows that the total number of tweets that mention arXiv papers decreases over time.

In order to better understand how Twitter mentions vary across domain, we show the Complementary Cumulative Distribution Functions (CCDF) of Twitter mentions for all articles in the five most frequently observed subjects domains of Figure 5.5. We find that within each domain few papers receive relatively many mentions whereas the majority receive very few. The frequency-rank distribution is thus strongly skewed towards low values indicating that most articles receive very few Twitter mentions. Note that we rely on the so-called Twitter Gardenhose, a random sample of about 10% of all daily tweets, and may thus underestimate the absolute number of Twitter mentions by a factor of 10.
In Figure 5.6, we plot the distributions of $\Theta_{ax}(a_i)$, $\Delta_{ax}(a_i)$, $\Theta_{tw}(a_i)$, and $\Delta_{tw}(a_i)$ following article submission. We can see that the distributions of $\Theta_{ax}(a_i)$, $\Delta_{ax}(a_i)$, $\Theta_{tw}(a_i)$, and $\Delta_{tw}(a_i)$ are highly skewed towards very low values, with very few cases characterized by extensive delays or time spans. In Figure 5.6(a), the distribution of $\Theta_{ax}(a_i)$ curve shows that
Figure 5.5: Complementary Cumulative Distribution Functions (CCDF) of Twitter mentions for all articles in the 5 most frequently observed subjects domains.

nearly all articles take at least 5 days to reach the peak of arXiv downloads \((x \leq 4 : y = 1)\), i.e., all articles take more than 4 days to reach peak downloads. In addition, the distribution of \(\Delta_{ax}(a_i)\) curve shows that most of the articles are downloaded persistently for over 100 days \((x \leq 100 : y > 0.6)\).

From Figure 5.6(b), it emerges that nearly 80\% of the articles in the corpus reach the peak of Twitter mention just one day after they are submitted, as is shown on the distribution of \(\Theta_{tw}(a_i)\) curve \((x = 2 : y \simeq 0.8)\). Over 70\% articles reach the peak of Twitter mention within 5 days of submission \((x = 5 : y < 0.3)\). However, the distribution of \(\Delta_{tw}(a_i)\) curve shows that over 80\% of arXiv.org articles are mentioned one and one day only \((x = 2 : y < 0.2)\), i.e., one or multiple tweets about an article are posted within the time range of 24 hours and then are never mentioned again. Overall, compared with arXiv downloads, the Twitter response to scientific articles is typically swift, yet highly ephemeral,
Figure 5.6: Distributions of $\log(P(\text{days} \geq x))$ for delay and span values in $\log(\text{days})$ for (a) arXiv downloads and (b) Twitter mentions, recorded for all arXiv submissions in our cohort.
a pattern indicative of a process in which the news of a publication is quickly passed around and very little in-depth discussion taking place afterward.

5.3.3 Regression between article downloads, Twitter mentions, and citations

We investigate the degree by which article citations, denoted $C$, can be explained in terms of article-based Twitter mentions, denoted $T$, and arXiv downloads, denoted $A$, by means of a multi-variate linear regression analysis. This analysis is limited to a cohort of the 70 most mentioned articles on Twitter that were submitted to arXiv.org from October 4, 2010 to March 1, 2011 (5 months). This limitation is due to the extent of work involved in manually collecting early citation data as well as to the fact that a cohort of articles submitted earlier in the timeline can provide a fuller coverage of Twitter mentions and arXiv downloads. For each article, we retrieve the total number of Twitter mentions and arXiv downloads 60 days after submission, and their total number of early citation counts on September 30, 2011 (7 months later after submission of the latest paper).

Given that each article could have been submitted at any time in a 5 month period, i.e. October 4, 2010 to March 1, 2011, on September 30, 2011 some articles could have had 5 more months than others to accumulate early citations. Therefore the citation counts observed on September 30, 2011 may be biased by the submission date of the article in question. We must therefore include the amount of time that an article has had to accumulate citations since their submission date as an independent variable in our regression models.

Let $P$ represent the number of days between the submission time of the article and September 30, 2011. We thus define the following multivariate linear regression models:
Table 5.2: Multi variant linear regression analysis of article citations $C$ vs. twitter mentions $T$, article arXiv downloads $A$, and time in days elapsed between beginning of our test period and submission of article, $P$.

<table>
<thead>
<tr>
<th>model</th>
<th>$\beta_1$ (st. error)</th>
<th>$\beta_2$ (st. error)</th>
<th>$\beta_3$ (st. error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C = \beta_1 T + \beta_2 P + \varepsilon_1$</td>
<td>0.150*** (0.035)</td>
<td>0.044***(0.019)</td>
<td>-</td>
</tr>
<tr>
<td>$C = \beta_1 A + \beta_2 P + \varepsilon_2$</td>
<td>2e-04*** (7e-05)</td>
<td>0.038*(0.020)</td>
<td>-</td>
</tr>
<tr>
<td>$C = \beta_1 T + \beta_2 A + \beta_3 P + \varepsilon_3$</td>
<td>0.120*** (0.040)</td>
<td>1e-04(8e-05)</td>
<td>0.041***(0.019)</td>
</tr>
</tbody>
</table>

*: p<0.1, **: p<0.05, ***: p<0.01, ****: p<0.001

\[
C = \beta_1 T + \beta_1 P + \varepsilon \quad (5.1)
\]

\[
C = \beta_1 A + \beta_2 P + \varepsilon \quad (5.2)
\]

\[
C = \beta_1 T + \beta_2 A + \beta_3 P + \varepsilon \quad (5.3)
\]

where $\beta_i$ denotes the corresponding regression coefficient.

From Table 5.2, we observe that publication period $P$ is certainly a non-neglectable factor to predict the citation counts $C$ but also that Twitter mentions $T$ shows equally significant correlations. Moreover, Twitter mentions seem to be the most significant predictor of citations, compared to arXiv downloads and time since publication. This is not the case for arXiv downloads which, when accounting for Twitter mentions and arXiv downloads, do not exhibit a statistically significant relationship to early citations.

In Figure 5.7 we show the bivariate scatterplots between Twitter mentions, arXiv downloads and citations. The corresponding Pearson’s correlation coefficients are shown as well. Figure 5.7(b) and 5.7(c) again show that Twitter mentions are correlated with citations better than arXiv downloads, which matches our results obtained from multivariate linear regression analysis. In addition, Twitter mentions are also positively correlated with arXiv
downloads as is shown in Figure 5.7(a), suggesting that the Twitter attention received by
an article can be used to estimate its usage data, but usage, in turn, does not seem to
correlated to early citations. Given the rather small sample size and the unequally dis-
tributed scatter, we performed a delete-1 observation jackknife on the Pearson’s correlation
coefficient between Twitter mentions and early citations (N=70). This yields a modified
correlation value of 0.430 vs. the original value of 0.4516 indicating that the observed cor-
relation is rather robust. However, dropping the top two frequently tweeted articles does
reduce the correlation to 0.258 (p=0.016) implying that the observed correlation is strongest
when frequently mentioned articles on Twitter are included, matching the results reported
by (Eysenbach, 2011).

Figure 5.7: Log-log scatter plots of (a) Twitter mentions vs. arXiv downloads, (b) Twitter
mentions vs. citations and (c) arXiv downloads vs. citations for 70 most mentioned articles
on Twitter indicate statistically significant correlations. Marginal densities of distributions
are shown as well, indicating strongly skewed distributions of arXiv article downloads,
Twitter mentions and citations.
5.4 Discussion

The ongoing move to online scholarly communication has introduced new possibilities for measuring scholarly impact. At the same time, it has become more difficult to determine which communities drive a particular form of online impact. For example, usage data, measured as volume of downloads, is generally assumed to reflect the interests and preferences of the general public, but what if the particular online service for which usage data was recorded is dedicated to serving scientists only? What if an online service for scientists increasingly becomes a tool for the general public to learn about scientific findings? The online user communities associated with particular services may in fact overlap to various degrees as the scholarly community progressively moves online and the online public moves toward scholarly information services. Naturally, scholarly impact metrics should acknowledge this new reality.

The research presented in this paper is based on data from two services which are arguably associated with and intended for two different audiences. ArXiv.org is focused on offering scientists an online platform to publish pre-prints. Twitter is designed to serve as a micro-blogging services for the public. In this study we did, however, not attempt to conceptualize arXiv downloads solely as scientific impact, and Twitter mentions solely as public chatter. Rather, we measured the correlation and temporal differences between these forms of responses, working under the assumption that these services naturally have overlapping and interacting user communities.

Our results, though preliminary, are highly suggestive of a strong tie between social media interest, article downloads, and even early citations. We find that Twitter mentions and arXiv downloads of scholarly articles follow two distinct temporal patterns of activity, with Twitter mentions having shorter delays and narrower time spans than arXiv downloads. We also find that volume of Twitter mention is statistically correlated with that of both...
downloads and “early” citations, i.e., citations in the scholarly record occurring less than 7 months after the publication of a preprint.

We can think of two possible explanations for these results. First, the manner in which Twitter mentions, arXiv downloads and article citations are correlated could indicate a causal relation. Scholars are increasingly exposed to social media such as Twitter, and therefore their scholarly download and citation behavior is unavoidably affected. A paper submitted to arXiv that happens to receive high levels of mentions in social media will, as a result, receive greater exposure among both the general public and scholars. As a consequence, it will receive greater levels of scholarly interest, and higher volumes of downloads and subsequent citations. Our results indeed indicate that early Twitter mentions of a paper seem to lead to more rapid and more intense download levels and subsequently higher citation levels. Second, an equally plausible, alternative explanation for our results lies in the intrinsic quality or popular appeal of different manuscripts. A manuscript of greater quality or appeal, either among the public or the scholarly community, will by virtue of this characteristic enjoy higher levels of mentions on Twitter, higher levels of downloads on arXiv, and higher levels of later citations. As a result these indicators will seem to be correlated, and even causative of each other.

We therefore acknowledge that these observations can be the result of a number of distinct or overlapping factors which our methodology confounds and fails to distinguish. Consequently, we caution against drawing the unwarranted conclusion that these results indicate that the scholarly impact of an article can be fully determined by its social media coverage, nor that one could increase the citation rate of an article by merely tweeting about it. The fact that some correlation — no matter how small — was observed between social media coverage, usage, and early citations may nevertheless indicate that the scientific communication process is increasingly affected by the growing societal importance
of social media. In future research we will therefore continue to focus on unraveling the potential mechanisms that tie these various factors together. These efforts might shed light on whether and how social media is becoming a component of academic and scholarly life.

5.5 Materials

5.5.1 Abbreviations

Table 5.3 presents a list of the subject domain abbreviations used in this article.

<table>
<thead>
<tr>
<th>Subject Abbr.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>astro-ph</td>
<td>Astrophysics</td>
</tr>
<tr>
<td>hep</td>
<td>High Energy Physics</td>
</tr>
<tr>
<td>physics</td>
<td>Physics</td>
</tr>
<tr>
<td>math</td>
<td>Mathematics</td>
</tr>
<tr>
<td>cond-mat</td>
<td>Material Science</td>
</tr>
<tr>
<td>cs</td>
<td>Computer Science</td>
</tr>
<tr>
<td>quant-ph</td>
<td>Quantum Physics</td>
</tr>
<tr>
<td>gr-qc</td>
<td>General Relativistic and Quantum Cosmology</td>
</tr>
<tr>
<td>nucl</td>
<td>Nuclear</td>
</tr>
<tr>
<td>q-bio</td>
<td>Quantitative Biology</td>
</tr>
<tr>
<td>math-ph</td>
<td>Mathematical Physics</td>
</tr>
<tr>
<td>nlin</td>
<td>Nonlinear Science</td>
</tr>
<tr>
<td>stat</td>
<td>Statistics</td>
</tr>
<tr>
<td>q-fin</td>
<td>Quantitative Finance</td>
</tr>
</tbody>
</table>

5.5.2 Data Collection

Our process of determining whether a particular arXiv article was mentioned on Twitter consists of three phases: crawling, filtering, and organization. Tweets are acquired via the Streaming API from Twitter Gardenhose, which represents roughly 10% of the total tweets from public time line through random sampling. We collected tweets whose date and time
The goal of the data filtering process is to find all tweets that contain a URL that directly or indirectly links to any arXiv.org paper. However, determining whether a paper has or has not been mentioned on Twitter is fraught with a variety of issues, the most important of which is the prevalence of partial or shortened URLs. Twitter imposes a 140 character limit on the length of Tweets, and users therefore employ a variety of methods to replace the original article URLs with alternative or shortened ones. Since many different shortened URLs can point to the same original URLs, we resolve all shortened URLs in our Twitter data set to determine whether any of them point to the articles in our arXiv cohort.

We distinguish between four general types of scholarly mentions in Twitter, based on whether they contain:

1. a URL that directly refers to a paper published in arXiv.org.
2. a shortened URL that upon expansion refers to an arXiv.org paper
3. a URL that links to a web page, e.g. a blog posting, which itself contains a URL that points to an arXiv.org paper.
4. a shortened URL that links to a type (3) mention after expansion.

In order to detect these four types of Twitter mentions, we first expand all shortened URLs in our crawled public tweets. We select the top 16 popular URL shortening services, including bit.ly, tinyurl.com, and ow.ly, and expand the shortened URLs in our collection of tweets using their respective APIs. As such, we resolved 98,377,880 short URLs, which were mostly generated by the following URL shorteners: bit.ly (61.3%), t.co (15.2%), fb.me (6.5%), tinyurl.com (6.1%) and ow.ly (4.4%). (We acknowledge that this procedure will not identify all Twitter mentions of a given arXiv.org paper, but it will however capture most.)
From the resulting set, we retain all tweets that contain the term ‘arXiv’ and at least one URL. Next, we associate tweets to arXiv papers by extracting the arXiv ID (substrings matching ‘dddd.dddd’) from any papers mentioned in those tweets. (Note that in the case of the third and fourth type of Twitter mention the arXiv paper ID is not explicitly shown in the tweet itself, but needs to be extracted from the web pages that the tweet in question links to.)
CHAPTER 6

THE RELATION BETWEEN WIKIPEDIA RANKING AND SCHOLARLY RANKING

Abstract

In addition to its broad popularity, Wikipedia is also widely used for scholarly purposes. Many Wikipedia pages pertain to academic papers, scholars, and topics, providing a rich ecology for scholarly uses. Scholarly references and mentions on Wikipedia may thus shape the “societal impact” of a certain scholarly communication item, but it is not clear whether they shape actual “academic impact”. In this paper, we compare the impact of papers, scholars, and topics according to two different measures, namely scholarly citations and Wikipedia mentions. Our results show that academic and Wikipedia impact are positively correlated. Papers, authors, and topics that are mentioned on Wikipedia have higher academic impact than those that are not mentioned. Our findings validate the hypothesis that Wikipedia can help assess the impact of scholarly publications and underpin relevance indicators for scholarly retrieval or recommendation systems.

6.1 Introduction

Science Citation Index established by Garfield and Merton (1979b) makes citation statistics a gold standard for the assessment of scholarly impact. Citation data is held to be a valid
and reliable indicator of scholarly impact because it represents an explicit and objective acknowledgment of influence by expert authors. Yet, the Web 2.0 is revolutionizing scholarly practices. A growing number of scholars discuss and share the research literature on Twitter and Facebook (Shuai, Pepe, & Bollen, 2012), organize it in social reference managers like Mendeley, and review it in blogs (Priem & Costello, 2010a). The increasing role of social media in scholarship requires new ways to assess impact beyond traditional approaches on the basis of citation data. Wikipedia, as a collaboratively edited, multilingual, and free Internet encyclopedia, has become an important source for the creation, distribution, and acquisition of scientific knowledge. Kittur et al. (2009a) shows that over 25% of pre-2008 articles in Wikipedia are related to natural or social sciences.

Wikipedia editors frequently reference scholarly entities, such as papers, scholars, and topics. We refer to such mentions as “Wikipedia citation”, implying that their value or influence has been explicitly recognized by the Wikipedia community. Unlike “academic citations” that represent the explicit recognition of expert scholars, the authority of a “Wikipedia citation” is uncertain and will need to be examined further.

Several studies have compared “academic citations” with “Wikipedia citations”. Nielsen (2007) showed that citations in Wikipedia correlate well with statistics from the Journal Citation Reports. Evans and Krauthammer (2011) investigated this relationship at the journal article level and found that PubMed journal articles that are mentioned in Wikipedia have significantly higher academic citation counts than an equivalent random article subset. Although these findings show that Wikipedia citations are an indicator of academic impact, their results are limited to journals or articles published in the same journal.

Here we extend this line of work to larger-scale data and across a broader research area for a more diverse set of scholarly entities. This paper makes an effort to quantitatively compare the rankings of articles, authors, and topics selected from ACM Digital Library
publication data on the basis of their academic citations and Wikipedia citations. Our major findings include:

- Academic and Wikipedia rankings exhibit a positive correlation across all types of scholarly entities.
- Papers, authors, and topics that are mentioned on Wikipedia have a higher average academic impact than those that are not mentioned.
- Wikipedia mentions are biased towards high-impact scholars who publish many papers.
- Wikipedia mentions are biased towards trending topics that occur in many articles.

6.2 Related Work

Studying the influence of social web on the scholarly community opens a new direction for scientometrics. Priem and Hemminger (2010b) propose scientometrics 2.0 and Altmetrics to measure scholarly impact from social media data. Shuai et al. (2012) compare different types of online responses to newly submitted preprint publications, namely article downloads, Twitter mentions and citations. Nielsen (2007), Evans and Krauthammer (2011) correlate journal citations to Wikipedia and academia. Our study extends this work by comparing the rankings of three different scholarly entities, i.e. papers, authors, and keywords, from a large-scale dataset in the field of Computing between scholarly sources and Wikipedia.

6.3 Problem Definition

The framework of our study is shown in Figure 6.1. $P$ is a set of scientific papers, $A$ is a set of authors, $K$ is a set of keywords (topics) and $W$ is a set of Wikipedia articles. $X = P \cup A \cup K$ is a set of heterogeneous scholarly entities.
**Academic Citation** AC includes paper citation \( AC_p = \{(p_i, p_j)|p_i \text{ cites } p_j, p_i, p_j \in P\} \), author citation \( AC_a = \{(a_i, a_j)|a_i \text{ cites } a_j, a_i, a_j \in A\} \) and keyword citation \( AC_k = \{(k_i, k_j)|k_i \text{ cites } k_j, k_i, k_j \in K\} \).

**Wikipedia Citation** WC = \{\( (w_i, x_i)|w_i \text{ mentions } x_i, w_i \in W, x_i \in X\}\} represents the acknowledgement of the scholarly entity \( x_i \) from Wikipedia article \( w_i \).

**Paper Citation Network** \( G_p = \{P, A, K, AC_p\} \) is a directed and unweighted heterogeneous network.

**Author Citation Network** \( G_a = \{A, AC_p, F_a\} \) is a directed and weighted network derived from \( G_p \), where \( F_a \) is a weight function that maps each edge \( (a_i, a_j) \) to a positive integer \( f(a_i, a_j) \in \mathbb{N}^+ \) that corresponds to the number of citations passing from \( a_i \) to \( a_j \).

**Keyword Citation Network** \( G_k = \{K, AC_k, F_k\} \) is a directed and weighted network derived from \( G_p \), where \( F_k \) is a weight function that maps each edge \( (k_i, k_j) \) to a positive integer \( f(k_i, k_j) \in \mathbb{N}^+ \) that corresponds to the number of citations passing from \( k_i \) to \( k_j \).

**Wikipedia Interlinking Network** \( G_w = \{W, E\} \) is a directed and unweighted network composed of Wikipedia articles and their internal links, where \( (w_i, w_j) \in E \) indicates that \( w_i \) contains a hypertext linking to \( w_j \).

**Ranking Function** \( R(X, \Theta) \) maps \( X = \{x_i\} \) into a sorted permutation \( X^* \) where \( \Theta(x_i) \geq \Theta(x_j) \).

**Academic Ranking** \( AR(X, \Theta) \) ranks \( X \) based on statistics calculated from ACM database, where \( \Theta(x_i) \) can be \( \Theta_{af} \) (the frequency of \( x_i \)), \( \Theta_{ac} \) (the citation of \( x_i \) in \( G_x \)), or \( \Theta_{ap} \) (the Pagerank of \( x_i \) in \( G_x \)).
Wikipedia Ranking \( WR(X, \Theta) \) ranks \( X \) based on statistics calculated from Wikipedia, where \( \Theta(x_i) = \sum_{w_j} S(w_j) \) satisfying that \( (w_j, x_i) \in WC \). Especially, \( S(w_j) \) is a score attached to each Wikipedia article, which can be constantly equal to one, or the total count of edits of \( w_j \), or the Pagerank score of \( w_j \) in \( G_w \). Therefore, we have three types of \( \Theta \) for WR: \( \Theta_w \), \( \Theta_{we} \) and \( \Theta_{wp} \).

The main goal of this study is to examine the relationship between \( AR(X, \Theta) \) and \( WR(X, \Theta) \). Specifically, we investigate three questions:

- What’s the correlation between \( AR(X) \) and \( WR(X) \)?

Figure 6.1: Academic ranking and Wikipedia ranking
• How do different types of $\Theta$ affect above correlation?

• Are those scholarly entities mentioned in Wikipedia ranked higher in $AR(X)$ than those not mentioned?

6.4 Methods

Two types of data sources, i.e. the ACM database and Wikipedia data, are involved in our study. Our analysis consists of three steps. First, we select a set of papers, authors, and keywords from the ACM database. Second, we build a Wikipedia search index and query it for these set papers, authors, and keywords. Third, we rank all papers, authors, and keywords based on $AR(X, \Theta)$ and $WR(X, \Theta)$, and compare the resulting rankings.

6.4.1 ACM Data Selection

The ACM database stores over 200K papers published in computer science journals and conferences, from 1980 to 2010. Most papers contain three types of metadata: a title, author, and keywords field. Each type of metadata needs careful filtering before we search for them in Wikipedia, since short queries (title, author names, keywords) may lead to false positives. We therefore decide to remove titles, authors, and keywords from our ACM database that match the properties in Table 6.1.

<table>
<thead>
<tr>
<th>metadata</th>
<th>removed type</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>less than three non-stop words</td>
</tr>
<tr>
<td></td>
<td>e.g. &quot;On the Level&quot;, “Behavioral synthesis”</td>
</tr>
<tr>
<td>author</td>
<td>less than six tokens or one-letter first name</td>
</tr>
<tr>
<td></td>
<td>e.g. “C. Richard” or “Li Li”</td>
</tr>
<tr>
<td>keyword</td>
<td>single word or occurs only once in database</td>
</tr>
<tr>
<td></td>
<td>e.g. “program”, i.e. “system initiated dialogue”</td>
</tr>
</tbody>
</table>

Table 6.1: Criteria to remove titles, authors, and keywords from our ACM database.
Once $P$, $A$, and $K$ are selected, we construct $G_p$, $G_a$ and $G_k$, based on which $AR(X, \Theta_{af})$, $AR(X, \Theta_{ac})$ and $AR(X, \Theta_{ap})$ are obtained.

### 6.4.2 Wikipedia Search

The export data of the English version of Wikipedia is publicly available in XML format\(^1\). We use a Perl script WikiPrep\(^2\) to parse this data and remove all discussion and talk pages, leaving only article pages. We further remove category, file, disambiguation, and redirect pages to make sure that each article page is about an explicit concept.

Since the to-be-ranked sets of papers, authors, and keywords come from the ACM database, the Wikipedia articles that are used to build the search index should match the topics of the ACM repository. Our preliminary tests showed that including Wikipedia articles that do not match ACM fields in the process of searching will enormously reduce the retrieval precision. Therefore, we adopt a two-step Wikipedia index building process on the assumption that the set of keywords $K$ selected from ACM database are representative topics in computer science. First, all Wikipedia article pages are indexed using Lucene\(^3\). Second, we run a full-text search in the complete Wikipedia index for every keyword from $K$. In the retrieval results, only those Wikipedia articles containing at least two keywords are selected to construct $G_w$ and build a computing-exclusive index named Wiki\(_{cs}\). For each node in $G_w$, we compute the PageRank score and query its historical edit frequency using WikiMedia API\(^4\). Finally, we search $P$, $A$ and $K$ in Wiki\(_{cs}\) and generate $WR(X, \Theta_{w1})$, $WR(X, \Theta_{we})$ and $WR(X, \Theta_{wp})$.

---

6.4.3 Evaluation Measures

When \( X \) is searched in Wiki_{cs}, only a subset of \( X \) is found in Wikipedia and denoted as \( X_{\text{wiki}} \). The complement of this subset \( X \) is denoted \( X_{\text{nowiki}} \). We propose two measures to perform two types of ranking comparisons: (1) the ranking of \( X_{\text{wiki}} \) in ACM and Wikipedia, and (2) the ranking of \( X_{\text{wiki}} \) and \( X_{\text{nowiki}} \) in ACM.

First, we use Spearman’s rank correlation coefficient (Spearman, 1987) to measure the quantitative relation between \( AR(X, \Theta) \) and \( WR(X, \Theta) \):

\[
\rho = \frac{\sum (I_{x_i} - \bar{I}_{x_j})(I_{y_i} - \bar{I}_{y_j})}{\sqrt{\sum (I_{x_i} - \bar{I}_{x_j})^2(I_{y_i} - \bar{I}_{y_j})^2}}, x_i, y_i \in X_{\text{wiki}}
\]  

(6.1)

where \( I_{x_i} \) and \( I_{y_i} \) are the index of \( x_i \) and \( y_i \) in \( X^* \) ranked by \( AR(X, \Theta) \) and \( WR(X, \Theta) \), respectively, and

\[
\bar{I}_{x_i} = \frac{\sum x_i \in X_{\text{wiki}} I_{x_i}}{|X_{\text{wiki}}|}, \bar{I}_{y_i} = \frac{\sum y_i \in X_{\text{wiki}} I_{y_i}}{|X_{\text{wiki}}|}
\]

(6.2)

Next, we define a Normalized Average Ranking (NAR) to measure the average ranking positions of a subset of entities in the whole ranking list.

\[
\text{NAR}(X_{\text{sub}}) = \frac{\bar{I}_{x_i}}{\max(I_{x_j})}, x_i \in X_{\text{sub}}, x_j \in X, X_{\text{sub}} \subseteq X
\]

(6.3)

where, \( X_{\text{sub}} \) can be either \( X_{\text{wiki}} \) or \( X_{\text{nowiki}} \), and \( \max(I_{x_j}) \) is the maximum ranking index in \( X \).

6.5 Results and Discussion

We obtain 122,350 papers, 163,172 authors\(^5\) and 35,518 keywords from the ACM database. In addition, we select 357,345 Wikipedia articles from the Sep 2, 2012 Wikipedia dump, and build the search index Wiki_{cs}. After querying all selected papers, authors, and keywords in Wiki_{cs}\(^6\), we find 3836 papers, 17564 authors and 20891 keywords that are mentioned on

\(^5\)We do not consider name ambiguity here and leave it for future work

\(^6\)Titles and keywords use exact match, and authors use fuzzy match
Wikipedia. We can see that only a very small portion of papers (0.03) and authors (0.1) but more than half of keywords occur in Wikipedia.

The statistics of $G_p$, $G_a$, $G_k$ and $G_w$ are shown in Table 6.2.

<table>
<thead>
<tr>
<th>measure</th>
<th>$G_p$</th>
<th>$G_a$</th>
<th>$G_k$</th>
<th>$G_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>num of nodes</td>
<td>122,350</td>
<td>163,172</td>
<td>35,518</td>
<td>357,345</td>
</tr>
<tr>
<td>num of edges</td>
<td>395,152</td>
<td>2,877,723</td>
<td>1,746,734</td>
<td>17,947,480</td>
</tr>
<tr>
<td>density</td>
<td>$2.64 \times 10^{-5}$</td>
<td>0.0001</td>
<td>0.0014</td>
<td>0.0001</td>
</tr>
<tr>
<td>clustering</td>
<td>0.1326</td>
<td>0.2754</td>
<td>0.4784</td>
<td>0.2903</td>
</tr>
</tbody>
</table>

Table 6.2: Descriptive statistics of paper, author, keywords, and Wikipedia citation networks.

The Spearman rank correlations between $AR(X)$ and $WR(X)$ using different $\Theta$, i.e. $\Theta_{af}$, $\Theta_{ap}$ and $\Theta_{ac}$, for $AR(X, \Theta)$ and $\Theta_{w1}$, $\Theta_{wp}$ and $\Theta_{we}$ for $WR(X, \Theta))$ are shown in Table 6.3.

The overall scholarly ranking and Wikipedia ranking are positively correlated. The p-values for all correlations are less than 0.001, hence we can reject the null-hypothesis that the two variables are not correlated. However, none of correlation coefficients are greater than 0.5, indicating significant differences between the scholarly community and the community of Wikipedia editors in recognizing the impact or importance of scholarly entities.

We note several interesting results. First, the inclusion of Wikipedia article weight does not improve the ranking correlation. Wikipedia ranking using $\Theta_{w1}$ consistently shows better correlation with the academic ranking than $\Theta_{wp}$ and $\Theta_{we}$. It seems that “Authority” of Wikipedia page cannot provide information to better weight the importance of scholarly outcomes. Second, Wikipedia pages tend to cite scholars of high scholarly reputation more than those who are “merely” productive. $\Theta_{af}$ tends to rank productive authors higher while $\Theta_{ap}$ and $\Theta_{ac}$ tend to rank the most cited authors higher. Ranking with $\Theta_{ap}$ shows higher

---

7It is also possible that we did not find the correct way to measure Wikipedia article importance.
Table 6.3: The Spearman rank-order correlation between scholarly ranking and Wikipedia ranking using different ranking criteria. The academic ranking is based on frequency of occurrence ($\Theta_{af}$), citation PageRank ($\Theta_{ap}$), and citation frequency (cited) ($\Theta_{ac}$). The Wikipedia ranking takes into account Wikipedia page weight, including one ($\Theta_{w1}$), editing frequency ($\Theta_{we}$) and inter-linking PageRank ($\Theta_{wp}$).

correlation with Wikipedia ranking than $\Theta_{af}$ and $\Theta_{ac}$, implying that reputable authors who published influential works are more favored by Wikipedia community than those that are new and productive. Third, Wikipedia tends to mention trending research topics more than classical or traditional topics. The situation of keyword ranking is different than author ranking, i.e. frequently occurring keywords are more likely to be mentioned in Wikipedia than most cited keywords (“classical” topics). It implies that Wikipedia editors prefer to discuss trending scientific topics when they create new, or edit existing Wikipedia pages.

Table 6.4 shows the differences in academic ranking between scholarly entities that are mentioned and those that are not mentioned in Wikipedia. It is apparent that papers, authors, and keywords that are mentioned on Wikipedia are ranked higher in the scholarly community than those are not mentioned. We use t-test to examine the differences and all
Table 6.4: The average academic ranking comparison between Wikipedia mentioned ($X_{wiki}$) and non-mentioned ($X_{nowiki}$) scholarly entities

corresponding p-values (consistently smaller than 0.001) confirm the statistical significance. It indicates that Wikipedia is a good social filtering system that recommends high impact papers, authors, and topics to the public.

6.6 Conclusions

This paper compares the scholarly ranking and Wikipedia ranking of a set of selected papers, authors, and keywords (topics), from the field of Computer science. We run a full-text search of those papers, authors, and keywords in Wikipedia, and separate the sets into two subsets: $X_{wiki}$ denotes the set of papers, authors, and keywords that are mentioned in Wikipedia, while $X_{nowiki}$ represents the set that are not mentioned. First, we compute the Spearman rank-order correlation coefficient between scholarly and Wikipedia rankings of $X_{wiki}$ using different ranking methods. We find that the two rankings are statistically significantly correlated and that the Wikipedia community favors reputable authors and trending topics. Second, we compare the average ranking of $X_{wiki}$ and $X_{nowiki}$ and found
the former had much higher values than the latter. This implies that Wikipedia does serve as a collaborative social filtering system which is able to favor “classical” papers, authors, and topics, and recommend them to the general public.
CHAPTER 7

DISCUSSION AND CONCLUSION

Scholarly communication plays a pivotal role in understanding science. This dissertation proposes novel indicators as well as computational methods to quantitatively measure and model scholarly communication. The development of technology and the web is revolutionizing scholarly communication, bringing about a shift from brick-and-mortar libraries to online web services to social media platforms, giving rise to what is now commonly referred to as “digital scholarship”. This diversification of scholarly communications leads to heterogeneous publication practices, heterogeneous corpora, and heterogeneous bibliographic data, challenging traditional bibliometrics which has strongly relied on publication and citation data. New data has become available that is extending the ability of bibliometricians to study scholarly communication as it pertains to digital scholarship, such as citation indices, bibliographic data, log data, clickstream data, and social media messages. These data sets enable us to model and analyze scholarly communications from a more comprehensive and systematic perspective, thereby providing deeper insights into what shapes scientific discovery. The latter may provide tools for scientific committees and funders to better evaluate scientific outputs and thereby make more effective policy and funding allocation decisions.

Bibliometrics provides powerful statistical tools and methods to quantitatively analyze scholarly communication. Citation data served as a “ground truth” which respect to scholarly communication because it is extracted from bibliographic records, and it has high face
validity with respect to providing an accurate record of the flow of ideas and recognition in the scholarly literature. From a more pragmatic perspective, large databases such as Thomson-Reuters Web of Science and Elsevier Scopus have provided extensive coverage of citation patterns throughout the bibliographic record. Naturally, bibliometrics has historically converged on a reliance on citation data that persists to this moment. Although citation analysis has been central to bibliometric science for decades, that role is now being challenged on scientific and pragmatic grounds. First, in the age of digital scholarship, citation data, due to its origin in publication records, provides a delayed and biased perspective of scientific developments. First, only after review, publication, and the appearance of the publication in citation databases, it is reflected in citation data, well after the ideas, writing, and other forms of scholarly activities have already taken place. In fact, the emergence of Institutional and pre-print repositories has sped up scholarly communication to the degree that the final, citable publication can emerge in the scholarly records years after the idea has already been disseminated and published. Second, scholarly communication has become more heterogeneous. Scientific publication is no longer the purview of book and journal publishers, but takes place on blogs, social media, web sites, and involves a much wider community and audience than the scientific community. In addition, the material that is published has changed as well; data sets, software, tools, databases, etc are now an important part of scholarly communication that is not acknowledged or reflected in citation data. Impact indicators derived from citation data therefore provide only a partial view of scholarly activity. For example, the h-index does not take into account the research area or domain of scholars in its impact assessment. The Journal Impact Factor is designed to only rank journals but does not acknowledge conferences and many other formal and informal venues. Maps of science derived from bibliographic or citation data provide a delayed snapshot of what science looked like months or years ago, due to the delay with which citation
data becomes available.

To explore the possibility of enriching citation-based bibliometrics by leveraging large-scale and heterogeneous data, in order to more completely and exhaustively model scholarly communications in the era of digital scholarship, my dissertation focused on the following research questions:

- One, can we detect scholarly communities and rank scholars and venues within specific community?
- Two, can we model the knowledge exchange among different scientific disciplines through citation relationships?
- Three, can we use other indicators from Web and social media other than citations to evaluate scientific outputs?

To address those research questions, this dissertation investigates scholarly communication from the perspective of using large-scale, heterogenous data sets. My work revolved around four major scientific fronts.

First, we investigate the emergence of scholarly communities by probabilistically modeling research topics from large-scale bibliometric data to provide a dynamic perspective on the evolution of topic- and author-based communities; Second, we investigate how scientific knowledge propagates through through scholarly networks; Third, we analyze the online response of the scientific community to the publication of scholarly articles using pre-print downloads, Twitter mentions, and early citations data, using a single cohort of arXiv preprints; Fourth, we compare the impact of papers, scholars, and topics as measured by different measures, each derived from a different medium for the exchange of scientific knowledge, in this case scholarly citations vs. Wikipedia mentions.

Our findings can be summarized as follows:

- The impact of journals (conferences) and scholars becomes more meaningful within the
context of specific research communities. As scholar’s research focus change, impact rankings change as well.

- Science is a dynamic and connected system, where idea flows between different domains in an asymmetric manner. Particularly, different domains play different roles, and exhibit different properties in organizing cross-domain scholarly communications.
- Social media has become an important factor in scientists’ social as well as their professional lives. We found that new indicators derived from social media may be used to evaluate the scholarly impact of paper in the early stage.

There already exist many to investigate and modeling scholarly communications beyond the scope of traditional citation analysis (Bollen, Sompel, et al., 2009; Shneiderman, 2008; Groth & Gurney, 2010). As an extension of the line of research, this dissertation paves a path towards the investigation of scholarly communication in the age of “digital scholarship” and provides models and metrics to evaluate and characterize scientific outputs and value from a new angle.

- First, when assessing the scholarly impact of articles, scholars or venues, topics and domain must be taken into consideration. It is unfair and in accurate to assess scholarly impact by only considering raw citation counts, since the distribution of citation counts is quite unbalanced across different research domains.
- Second, the direction of citation flow yields important information to help us better understand the role that domains plays in scientific development. Some domains mainly import knowledge, some domains mainly export knowledge while some domains serve as a bridge to connect different domains. Understanding the topology of science can thus help science policy makers to make more effective funding decision about funding and resource allocation.
• Third, “altmetrics” can serve as the leading indicator to assess the scholarly impact of scientific works. Social media can reflect the collective attention about a given scientific article or scholar. As such this signal leads traditional citations and citation-based impact indicators. Given this information, policy makers may decide whether to allocate money in a early stage.

This dissertation addressed a number of pertinent scientific issues in bibliometrics, in particular in relation to the advent of digital scholarship. We sought to model scholarly communication using new model and metrics and were able to make advanced on a number of front. However, I conceive this dissertation as only a first step in a larger program to extend and refine the science of bibliometrics as it evolves to address the newly emerging heterogeneity of scholarly communication. Areas that will merit additional scientific work in the near future are:

• First, are Altmetrics valid and legitimate indicators to measure scholarly impact? This question is still outstanding and in needs of significant further investigation. Although some preliminary studies show that Altmetrics exhibit positive correlations with citation counts (Eysenbach, 2011), recent studies have challenged these findings (Haustein, Peters, Sugimoto, Thelwall, & Larivière, 2013). In our study, we too observed significant and telling correlations between social media-based indicators and more traditional citation-based indicators, but the correlation coefficients were weak and characterized by high dispersion. I expect that more extensive studies, along a variety of different scholarly communication items, and social media platforms, will be necessary to test the validity of Altmetrics as a valid and accurate addition to the bibliometric toolbox.

• Second, further study is needed to understand how scientific articles propagate through social networks and how such propagation will affect their scholarly value. Regardless
of the data source, citation data or social media activity, the metrics or indicators themselves need further refinement. Many Altmetrics indicators are presently defined as raw counts which are susceptible to gaming, bias, and a variety of other distortions already observed in citation-based indicators. Further work is needed to study the dynamic and topological features of articles propagation in social media to arrive at more sophisticated, accurate, and valid Altmetrics indicators.

• Third, bibliometrics indicators have increasingly been used by policy-makers to inform promotion, tenure, funding, and resource allocation decisions. I believe this evolution will continue, regardless of the bibliometric validity and accuracy of the indicators involved. Our community has a strong two-fold responsibility with respect to this ongoing trend. First, we need to make sure our scientific work is focused not merely on the conceptual development of indicators, whether they be derived from citation data or other sources, but on their development as accurate and valid tools, with proven performance. Second, our community needs to ensure that metrics and indicators that emerge from our scientific work are properly understood and correctly applied by decision-makers so that their decisions are well-informed and unbiased, and rely on the best scientifically valid practices. Many lessons can be learned from evidence-based medicine, where medical and pharmaceutical research is frequently translated into best practices for practitioners to ensure the best possible outcomes.
REFERENCES


(http://arxiv.org/abs/0911.1583)


Institute for BioHealth Informatics.


Garfield, E. (1955a, July 15). Citation indexes for science; a new dimension in document-


Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2012). A bird’s-eye view


Shuai Xin

Department of Informatics, Indiana University Bloomington

Email: xshuai@indiana.edu

Homepage: http://www.cs.indiana.edu/~xshuai

Education

• Ph.D. Informatics, Indiana University Bloomington, 2008-2014. GPA: 3.862/4.0
• M.S. & B.S. Automation & Control Science, Zhejiang University, 2006-2008. GPA: 3.9/4.0

Research

Interests

• Social Media Mining, Predictive Modeling, Information Retrieval and Network Analysis

Working Papers


Publications


• Shuai, X., Chen, S. S., Ding, Y., Sun, Y. Y., Busemeyer, J. R and Tang, J. (2012) There is more than complex contagion: an indirect influence analysis on Twitter, Workshop on \textit{Mining Data Semantics} in conjunction with \textit{KDD2012}


• Shuai, X., Liu, X. Z., and Bollen, J. (2012) Improving News Ranking by Community Tweets, Workshop on \textit{Mining Social Network Dynamics} in conjunction with \textit{WWW2012}

• Shuai, X., Ding, Y., Busemeyer, J. R.(2012) Multiple spreaders affect the indirect influence on Twitter (poster, accept ratio: 28.1\%) [pdf], \textit{WWW2012}


Projects

• Graduate Research Assistant. Indiana University Bloomington, September 2011 – August 2014
  – Investigated social media impact on scholarly communications using regression and correlation analysis.
  – Designed a binary text classifier to distinguish sensitive tweets leaking privacy from general tweets.
  – Improve Google & Yahoo news ranking by modeling geographical community interests on Twitter.
  – Predicted users’ opinions by jointly considering topic and social influence in Chinese microblogging site.
  – Analyzed and visualized data using Python and R.

• Insight Data Science Fellows Program at Mountain View, August 2013 – October 2013
  – Developed a web app TwiNews (http://xintwinews.com) in Python to re-rank news articles from Google & Yahoo according to Twitter geographical popularity.
  – Crawled news articles and collected public tweets using Twitter API, stored the data in MySQL, ranked news articles based on cosine similarity scores with local tweets, and evaluated the ranking on MTurk.
  – Implemented the web interface using Flask and deployed it using Amazon Web Service.

• D4D Data Challenge, Indiana University Bloomington, October 2012 – May 2013
  – Inferred socio-economic index from mobile phone usage patterns using correlation analysis.
  – Detected the digital divide phenomenon and communication gap between rich and poor areas in an African country through network analysis.
  – Visualized the geographical mobile communication flow using R and JEQL.
• Internship @ Yahoo! Research at Barcelona, June-September 2012
  – Applied Hadoop Map-Reduce framework for large-scale data processing of news photo metadata to predict social event attendees using language model from Terrier IR platform.
  – Incorporated textual, network and temporal information together into language model to improve prediction.
  – Analyzed the effects of topical, temporal and scale properties of events on prediction performance using R.

• Internship @ Dialogue Group at Avaya Lab (a heritage from Bell Lab), June-August 2011
  – Developed an IR system based on web service technology to support proactive search to enrich enterprise communication and collaboration, including:
    * Disambiguated employee names using co-author social networks.
    * Recommend experts given a query based on relevance, reputation and authority.
    * Annotated text stream relying on Wikipedia using key phrases extraction and concept disambiguation.

• Google Summer of Code, May-August 2009
  – Collaborated with a team to develop an open source software package using GitHub.
  – Programmed Python/Perl language interfaces to libsequence (a bioinformatics C++ class library) on multi-operational systems to facilitate the biological data analysis using scripting language.

Skills
  • Languages: Python, R, C/C++, Java, Perl
  • Tools: Scikit-learn, NetworkX, NumPy, SciPy, MATLAB, MySQL, Hadoop, Pig, Flask, Git, SVN