

國立臺灣大學電機資訊學院資訊工程學系

博士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Doctoral Dissertation

資料引用之研究

A Study of Data Citation

黃曳弘

Yi-Hung Huang

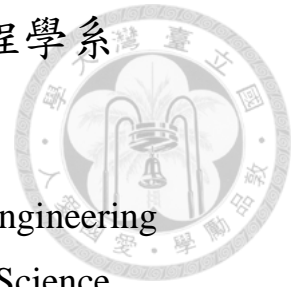
指導教授：許鈞南 博士，林軒田 博士

Advisor: Chun-Nan Hsu, Ph.D.,

Hsuan-Tien Lin, Ph.D.

中華民國 104 年 12 月

December 2015





國立臺灣大學博士學位論文
口試委員會審定書
資料引用之研究
A Study of Data Citation

本論文係黃曳弘君（學號D98922025）在國立臺灣大學資訊工程學系完成之博士學位論文，於民國 104 年 12 月 22 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

許鈞原 林軒田

(指導教授)

黃曳弘

陳信瑜 趙坤茂

劉家宏

趙坤茂

系主任





誌謝

回首過去這幾年的求學歷程，本篇論文能夠完成，實為仰賴諸位師長、同學、朋友與家人的指導、協助、鼓勵與祝福。首先，我要感謝指導教授許鈞南老師，感謝您領我進入AIIA實驗室，助我習得正確的做學術研究的技巧與態度；在我的博士學程開始後，也非常感謝您能繼續給予我關於研究方向的實質指導與意見，並提供非常寶貴的機會讓我能在美國一流的研究單位實習參學。我接下來要感謝的是共同指導教授林軒田老師，感謝您對於機器學習演算法的教導，以及感謝您治學嚴謹，為人清正的生活態度，為我們後學者立下良好的典範。在研究途中，我要感謝 Dr. Kristina Lerman 與 Dr. Peter W. Rose 對我的指導，協助發展本分析模型，引導至實際應用於蛋白質資料庫，並給予我適當的意見以分析探討實驗結果。我還要感謝許永真教授與Intel-臺大創新研究中心能提供我這一年的獎助金幫助，讓我能無顧忌地繼續做研究。

此外，我也非常感謝我的口試委員，趙坤茂教授、陳倩瑜教授、莊庭瑞研究員、曾宇鳳教授、林守德教授與劉家宏博士，感謝您們對於我的研究計劃的建議，並感謝您們能參予我研究過程的評斷，讓我能從全觀的角度重新審視自身的研究課題。

在此人生階段，我對於陪在我身旁的家人與朋友有著無限的感激，我要感謝我的父母，黃祥楊先生與古鳳妹女士，感謝您們支持著我的求學過程，以及對於我無償的付出與關懷。我還要感謝莊于穎小姐這些年的陪伴，在我最晦暗的人生階段，給予我相當的鼓勵。

最後在此，我想感念過去的恩師Prof. Alfonso Costanza(姜豐瑞教授)，感謝您對我們學生的付出，您留下來的的身影，總是在我最低潮困惑的時候，引導著我向正向方向思考。





摘要

在本文中，我們著重於分析數據資料庫之各種資料引用相關研究。我們認為，一致性的資料引用的實作將有助於推動的數據共享與增進數據重複使用性，因為它可被視為類比於期刊或其他出版物中的引用模式並受相關領域使用者的認可。

蛋白質資料庫 (Protein Data Bank, PDB) 為一個專門儲存蛋白質及核酸之三維結構資料的數據庫。他們大部份扮演了生物機制中關鍵的角色。這些資料數據主要經由世界各地的結構生物學家以X 射線晶體學或NMR 光譜學實驗所結構化而得。各個主要的科學雜誌要求科學家將自己的研究成果提交給PDB，並以獨立識別碼(PDB IDs) 存放到PDB 供公眾免費使用，是結構生物學研究中的重要資源。因此，PDB 是一個很好的實作對象用以進行資料引用之相關研究。我們的研究考慮PDB ID 在本文中提及的模式與其引用至參考文獻的模式之間的交互作用，並且藉由研究該資料引用模式來表達此兩種引用機制之間的相對重要性。

通過探索這些豐富的蛋白質結構資料和相關的引文中，我們可以從引文網絡的觀點來研究蛋白質結構之間的關係。此外，文獻和數據引網絡的分析可以顯示潛在的科學發展途徑，即知識和數據如何被用於推進結構生物學的發展之過程。基於這些分析的結果，我們可以提出適當的資料引用的實作方法，用以鏈接引用與資料兩者，以及衡量資料使用度量方式。這將有利於資料的重複使用，並有助於實驗過程的再現性，甚至提供機器可識別之資料使用追蹤能力。



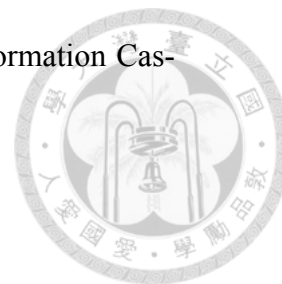


Abstract

In this thesis, we focus on analyzing the various of data citation to the data repository. We think consistent practice of data citation facilitates and incentivizes data sharing and reuse because it could be counted as professional recognition for data providers as citations of journal and other types publications. The Protein Data Bank (PDB) is the worldwide repository of 3D structures of proteins, nucleic acids and complex assemblies, most of which play essential biological roles. The major data of PDB are the experimentally determined structures of protein, and are provided by unique identifiers (PDB IDs) and corresponding primary citations that make them easier to be used as the referenced data. Therefore, it could be a good practice model for data citation research. Meanwhile, our studies focus on the interplay of PDB IDs mentions recognition and references cited of the literature, and the relative importance of these two mechanisms can be expressed by investigating the data citation patterns. By exploring rich structures and related citations of PDB, we can investigate the relationships between protein structures from the viewpoint of the citation network. Moreover, the analysis of the literature and data citation networks may demonstrate potential pathways of scientific discovery, that is, how knowledge and data were used to advance a particular field in structural biology. Based on the results of analyses, we could recommend data citation and provenance practices, approaches to discover data citations, ways of linking citations and data, and data access metrics. We hope our work will benefit the data reused, experiments reproduced, and even

provide machine readability for tracing the data usage.

Keywords: Data Citation, Citation Network Analysis, Information Cascade, Protein Data Bank.

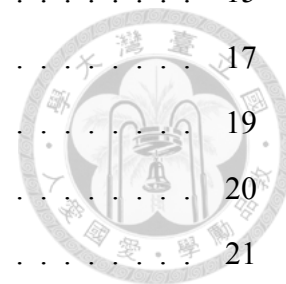




Contents

口試委員會審定書	iii
誌謝	v
摘要	vii
Abstract	ix
1 Introduction	1
1.1 Motivation and Overview of the Thesis	1
1.1.1 Data Citation	1
1.1.2 RCSB Protein Data Bank and Related Repository	2
1.1.3 Transformative Research	3
1.2 Organization of the Thesis	3
2 Identifying Transformative Scientific Research	5
2.1 Introduction	5
2.2 Related Work	8
2.3 Materials and Methods	10
2.3.1 Data	10
2.3.2 Cascade	11
2.3.3 Cascade Disruption	12
2.3.4 Computing Cascade Disruption	14
2.4 Evaluation	14

2.4.1	Validity	15
2.4.2	Reliability	17
2.4.3	Scalability	19
2.5	Results and Discussion	20
2.5.1	Physics	21
2.5.2	Computer Science	26
2.6	Summary	27
3	Citing the Protein Data Bank and Related Repository	29
3.1	Introduction	29
3.2	Related Work	30
3.3	Materials and Methods	31
3.3.1	Paper Citation Data	31
3.3.2	Mining URL Mentions	32
3.3.3	PDB Usage Statistics	33
3.3.4	Calibrated Disruption Score	33
3.4	Results and Discussion	34
3.4.1	Paper Citations	34
3.4.2	URL Mentions	36
3.4.3	Data Usage Statistics	38
3.5	Summary	40
4	Data Citation to the Protein Data Bank	43
4.1	Introduction	43
4.2	Materials and Methods	45
4.2.1	Citation data	45
4.2.2	Mention data	45
4.2.3	Mentions of issued PDB IDs	46
4.2.4	G-test of Independence	47
4.2.5	Pearson Correlation Coefficient	47



4.2.6	Co-citations/mentions between PDB Entries.....	48
4.2.7	Jaccard Index	48
4.3	Results and Discussion	49
4.3.1	User Tendency to the PDB Data Citation	49
4.3.2	Trends of Protein Structure Researches.....	49
4.3.3	Statistic Test to the Data Citation.....	51
4.3.4	Analysis of the Co-citation/mention Patterns	52
4.3.5	Identification of the Influential PDB Entries.....	54
4.3.6	If the Authors Clearly Cite Data Sources Will Also Help Improve Impact of Their Own Papers?	56
4.4	Summary	58
5	Summaries and Future Work	61
5.1	Summary of the results	61
5.2	Limitations.....	62
5.3	Future directions	63
	Appendices	65
	A Estimation of the Subsampling Error of Disruption Score	67
	B Model-based Approximation for Cascade Generating Function	73
	Bibliography	77

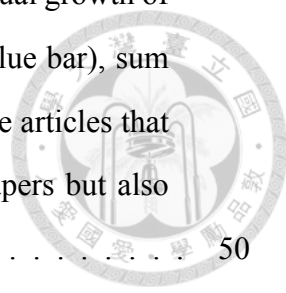




List of Figures

1.1	Consistent practice of data citation facilitates data sharing.	1
2.1	(A) Information disruption by a challenger in an information cascade. The seed of an established paradigm, marked in red, creates a cascade as the seed is cited by other papers, while a challenger, marked in blue, disrupts the cascade of the seed. (B) Disruption of the cascade of the seed paradigm (red) by the challenger paradigm (blue) can be visualized as the decline of Φ of the complement cascade (green).	7
2.2	An example of Cascade and their ϕ values.	13
2.3	Cascade disruption as an indicator of paradigm shift. The growth of the logarithm of average cascade function values per year for (A) superconductivity (BCS) [6, 5] and high-temperature superconductivity (HTS) [76, 77, 26] and (B) 30 control cases published in 1987 show no sign of cascade disruption against BCS. (C) The growth of the size of cascades shows no sign of disruption. (D) Another example of paradigm shift is conventional carbon nanotube [39] versus graphene [54, 55]. Unlike superconductivity, the cascades are not as large because they were published in recent years. Therefore, no logarithm of Φ is taken here.	16
2.4	Heatmaps of correlation between trials of 5-fold cross-validation for (A) “Fast Algorithms for Mining Association Rules in Large Databases (1994)” and (B) “Induction of Decision Trees (1986)”. “Full” is the result for the complete data, “Ex.CV fold i ” is the result of the i -th cross validation trial.	19

2.5	Distribution of PACS numbers of challenger papers identified by (A) our method and (B) baseline method.	23
3.1	Citation growth of the (A) PDB and (B) UniProt debut article and their follow-up articles.	34
3.2	Compare the growth of the (A) PDB and (B) UniProt debut paper's cascade with all the residue cascades created by its follow-up articles in 5 years ($\tau = 5$). The y-axis of both panels shows the logarithm of the annual average cascade function values Φ , defined in Eq. 2.3.	35
3.3	The residue cascades created by three 2003 follow-up articles.	36
3.4	(A) Annual growth of the citations to the PDB debut paper and the counts of the different PDB URL mentions. (B) Annual growth of the citations to the PDB debut paper (blue bar), sum of all PDB URL mentions (green bar) and the count of the articles that not only directly cite the PDB debut paper but also mention PDB URLs (red bar).	37
3.5	Growth of the cascade of the PDB debut article (black curve), the collection of PDB NAR update articles from 2002 to 2008, the PDB URL mentions articles, and their corresponding residue cascades. Notice the split between the black curve and green curve, indicating the cascade disruption.	39
3.6	The growth of citations of the PDB debut paper, PDB URL mentions, website downloads and views, and FTP archive access from 2008 to 2013. This analysis only considers citations and mentions available from the PubMedCentral archive.	40
3.7	The plots of the fitting of linear models between the PDB URL mentions c and the website downloads and views u , referred to by their case No.'s in Table 3.4: (A) Case No. 3, $y = u(t)$, and $x = c(t)$, (B) Case No. 7, $y = u(t)$, and $x = c(t - 1)$, (C) Case No. 11, $y = u(t) + u(t + 1)$, and $x = c(t) + c(t + 1)$, (D) Case No. 19, $y = c(t) + c(t + 1)$, and $x = u(t - 1) + u(t)$	42



4.1 (A) Growth of the depositions of new PDB entries. (B) Annual growth of the citations to the PDB entries' primary citation papers (blue bar), sum of all the PDB IDs' mention (green bar) and the count of the articles that not only directly cite the PDB entries' primary citation papers but also mention the PDB IDs (red bar). 50

4.2 (A) P-value of G-test of independence. (B) P-value of Pearson correlation coefficient. (C) The growth of Pearson correlation coefficient. (D) Q-Q plot between the distributions of citation and mention. 52

4.3 (A) Distribution of *Jaccard index*₃ of PDB IDs. (B) The average *Jaccard index*₃ ordered by the deposited time. 53

4.4 Heatmap of (A) co-citation degree between top cited categories of PDB IDs.(B) co-mention degree between top cited categories of PDB IDs. . . . 54

4.5 Venn Diagram of the selected papers. 57

4.6 The difference of two pattern articles in (A) Case 1: Both mention & citing (pattern 1) vs. Only mention (pattern 2), (B) Case 2: Both mention & citing (pattern 1) vs. Only citing (pattern 3), and (C) Case 3: Only mention (pattern 2) vs. Only citing (pattern 3). 59

4.7 The difference of two pattern articles in (A) Case 4: Both mention & citing (pattern 1) vs. Nor mention & citing (baseline), (B) Case 5: Only mention (pattern 2) vs. Nor mention & citing (baseline), and (C) Case 6: Only citing (pattern 3) vs. Nor mention & citing (baseline). 60

B.1 Average citations of papers in the APS citation network dataset. Each curve is for the papers published in one year from 1970 (darkest) to 2009 (brightest). The curve plots the change of the average citations to the past years. The horizontal-axis indicates how many year from the publication time *t*. 74

B.2 Estimating average cascade function values Φ by modeling citation counts Γ 75

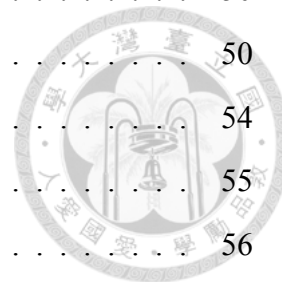




List of Tables

2.1	Statistics of the Test Data	10
2.2	Top ten challengers to the 1957 “Theory of Superconductivity” identified by (a) proposed method and (b) baseline method.	21
2.3	Top ten challengers to the 1967 “A Model of Leptons” by Steven Weinberg identified by (a) proposed method and (b) baseline method.	24
2.4	Top ten challengers (published after 1994) to the 1982 “Two-Dimensional Magnetotransport in the Extreme Quantum Limit” identified by (a) proposed method and (b) baseline method.	25
2.5	Top challengers to the 1986 “Induction of Decision Trees” paper identified by (a) proposed method and (b) baseline method.	27
2.6	Top challengers to the 1994 “Fast Algorithms for Mining Association Rules in Large Databases” paper identified by (a) proposed method and (b) baseline method.	28
3.1	Text patterns considered as PDB URLs.	32
3.2	5-year calibrated disruption scores of the PDB follow-up articles. The last column shows the average scores of randomly selected papers published in the same issue.	36
3.3	5-year calibrated disruption scores of the most highly cited articles in the database special issue of NAR.	37
3.4	The correlations between PDB data citations and PDB data usage statistics by linear modeling.	41
4.1	Mentions of Issued PDB IDs.	46

4.2	Top 10 PDB Entries. (Sorted by citation frequency)	50
4.3	Top 10 PDB Entries. (Sorted by mention frequency)	50
4.4	Highly cited category of PDB data.	54
4.5	Co-cited neighbor entries of the PDB entry-1AIK	55
4.6	Co-mentioned neighbor entries of the PDB entry-1AIK	56
4.7	Co-cited neighbor entries of the PDB entry-1F88	56
4.8	Co-mentioned neighbor entries of the PDB entry-1F88	56





Chapter 1

Introduction

1.1 Motivation and Overview of the Thesis

1.1.1 Data Citation

We think data itself should be considered as the citable products of research and cited in the same way as the academic paper, which have benefited from well-established bibliographic infrastructure that makes them easy to cite. Consistent practice of data citation facilitates and incentivizes data sharing and reuse because it could be counted as professional recognition for data providers as citations of journals and other types publications. However, currently no data citation practice has been commonly agreed. It is not clear which practice standard or policy gains the most adoption, nor is how they reflect the impact of the data being cited.

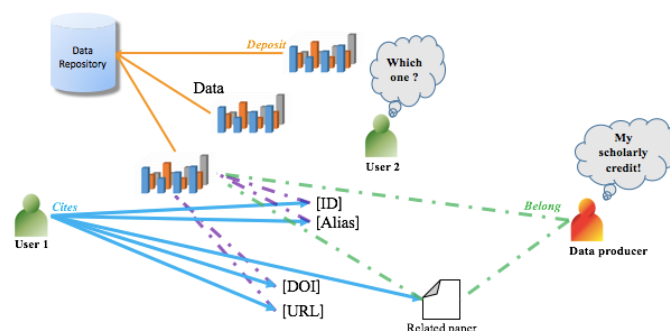


Figure 1.1: Consistent practice of data citation facilitates data sharing.

In the past few years, much of the studies on data citation have been received more attention in all disciplines of science as data become essential and ubiquitous in research. CODATA/ITSCI Task Force on Data Citation published a report on the current state of data citation in 2013 [69]. FORCE 11 (<http://www.force11.org>) has its final release of *Joint Declaration of Data Citation Principles* in 2014 [28], which identifies six principles as the guideline for the design of data citation standards and can be used as a good practice of data citation that contribute to data reuse, experiments reproduce, and even provide machine readability for tracing the impact of data.

1.1.2 RCSB Protein Data Bank and Related Repository

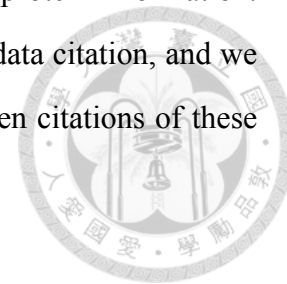
In this thesis, we focus on analyzing the various of data citation to the RCSB Protein Data Bank (PDB) and related repository. The Protein Data Bank (PDB) [11] is the worldwide repository of experimentally determined structures of proteins, nucleic acids, and complex assemblies, including drug-target complexes, most of which play essential biological roles and are the prime drug-targets in various diseases.

The major data of The PDB are the experimentally determined structures of protein. The PDB annotates structures according to standards set by the wwPDB [9] provides unique identifiers (PDB IDs) and digital object identifiers (DOIs) that make the data are accessible and persistence for researchers to use it as the referenced data. All journals require a prior submission of structures to the PDB as part of the publication process.

For a PDB entry, the primary citation papers is the study of crystallography process for a specific protein, and the primary citation should be declared when it was deposited to repository that have it be seen as legitimate, citable products of research. Hence, the data are easily to be given scholarly credit to all contributors to the data. All the characteristic make the PDB be a good practice model to help us study the behaviors that how the protein structure data being used by the researchers.

As a comparison of PDB, Uniprot is another comprehensive, high-quality and freely accessible repository that contains the protein sequences and functional annotation information [21]. It integrates, interprets and standardizes data from literature and numer-

ous resources to achieve the most comprehensive catalog possible of protein information. Therefore, it could be another good practice model for the study of data citation, and we will provide some comparison of similarities and differences between citations of these two resources.



1.1.3 Transformative Research

Transformative research refers to research that shifts or disrupts established scientific paradigms. Notable examples include the discovery of high-temperature superconductivity that disrupted the theory established 30 years ago. In Chapter 2, we will present a data-driven approach where citation patterns of scientific papers are analyzed to quantify how much a potential challenger idea shifts an established paradigm. The key idea is that transformative research creates an observable disruption in the structure of citation cascades. Citation cascades are chains of citations between two articles in a citation network that can be traced back to the papers establishing some scientific paradigm. Such a disruption is visible soon after the challenger's introduction. We define a *disruption score* to quantify the disruption and develop an algorithm to compute it from a large citation network that considers both the length of the chain and the number of paths. Identifying potential transformative research early and accurately is important for studying the data citation patterns. It also helps scientists identify and focus their attention on promising emerging works.

1.2 Organization of the Thesis

For our study, the main analysis tool is the citation cascade analysis. An important aspect is the interplay of literature and data citations, and the relative importance of these two mechanisms to make data discoverable. The analysis of the literature and data citation cascades demonstrates potential discovery pathways, that is, how knowledge and data were used to advance a particular field of science.

This idea is carried out as a pilot project in bioCADDIE, an NIH BD2K (Big Data to

Knowledge initiative) Data Discovery Index Coordination Consortium. The major aim is to correlate various metrics of citation networks with tangible impact indicators to determine empirically which metrics are more informative. Analysis of citation and data cascades of these networks will highlight putative pathways of how data and concepts led to high impact scientific discovery. Based on the results of these analyses, we will recommend data citation and provenance practices, approaches to data citation discovery, ways of linking citations and data, and data access metrics, for the NIH Data Discovery Index.

The rest of this thesis is organized as follows. In Chapter 2, we propose an approach called Disruption Score for identifying the transformative research and it will benefit us to study the data citation patterns. In Chapter 3, we focus on analyzing citations to the PDB data repository. In Chapter 4, we present a systematic investigation of how authors cite to individual structures and apply various network metrics to analyze different data citation practices to PDB. Finally, Chapter 5 summarizes this thesis and presents the future directions.



Chapter 2

Identifying Transformative Scientific Research

2.1 Introduction

Transformative research refers to research driven by ideas that lead to emerging concepts, approaches, and/or new subfields of research that shifts or disrupts an established scientific paradigm [70]. Thomas Kuhn’s influential book titled *The Structure of Scientific Revolutions* [44] describes the progress of science as non-linear, propelled by “paradigm shifts” in which scientists’ world-views, or paradigms, are altered dramatically by a new discovery, theory, or methodology. Recently, governments and industry R&D departments across the world are striving to maximize their return-of-investment in research budgets. Funding transformative research is generally agreed as an effective strategy and has been officially placed at the top priority of funding decision by the National Science Foundation (NSF) in the U.S. [70]. Systematically identifying transformative research accurately and early is therefore more critical than ever. This also applies to scientists, who need to constantly monitor the most recent transformative research in related fields to stay competitive in the forefront of their respective fields. Ability to systematically identify transformative research has numerous benefits, including helping funding agencies establish funding priorities, allowing individual scientists to better keep up with important new

research, and translating transformative research faster into practice.

However, identifying transformative ideas is not easy. The process by which such ideas are recognized and accepted by the scientific community is affected by a variety of factors, including cultural and cognitive biases, such as the well-documented “Matthew effect” [49, 50]. According to this effect, the scientific community pays disproportionate attention to the ideas of already-established scientists [1, 2], making it difficult for competing alternatives to gain attention [42]. These biases slow down the recognition and adoption of important new ideas, resulting in significant time delay in translating new research into new technologies and medical therapies [45, 51]. Yet, examples in which one theory, methodology, or line of inquiry overtakes an established one abound. One such case is the Nobel prize-winning discovery of high-temperature superconductivity in 1986 [8]. This breakthrough challenged the well-established theory of superconductivity [6], which explained how materials enter a superconducting state at low temperatures. Scientists who have been studying superconductors shifted their attention to new materials, which were shown to lose their electrical resistance at much higher temperatures than traditional superconductors, and, therefore, prove to be much more technologically useful. While such shifts are easily recognized in retrospect, many years or decades later, we claim that they are evident in citations patterns almost immediately after the paper describing the breakthrough is published.

Given the importance of timely identification of groundbreaking research, several studies have examined how scientific ideas are adopted by other scientists. Most of these studies analyze citations made by scientific papers, since scientists communicate, position their work, and allocate credit through citations. Using the number of citations, or its distribution, is an accepted way to calculate impact of a paper or a scientist [36]. However, this method is problematic, since it takes years for the citation count to reflect the status of a paper. Mazloumian et al. [48] argued that paradigm shifts occur because an author’s groundbreaking paper boosts attention given to his or her other publications. The boost establishes author’s “authority,” allowing his breakthrough to successfully compete for attention with an established paradigm. They proposed an automatic detection of such

boosts as a method for identifying an author’s seminal papers.

We view the process by which transformative research is recognized by the scientific community as a competition between paradigms for the attention of the scientific community. A paradigm is a theory of a phenomenon or a research method, *e.g.*, preparation of materials or a new experimental technique. A paradigm is established in one or more papers and supported in subsequent papers. The attention it receives can be measured by the structure of the information cascade the original papers create. The cascade consists of chains of citations that can be traced back to the original papers. We claim that transformative research shifts attention of the scientific community away from the established paradigm and that this is observable as a disruption of the growth of its citations cascade. Disruption occurs when the challenger paradigm can explain new citations received by the established paradigm. Our approach is general and can be applied to other domain, *e.g.*, social media, where ideas compete for attention of information consumers.

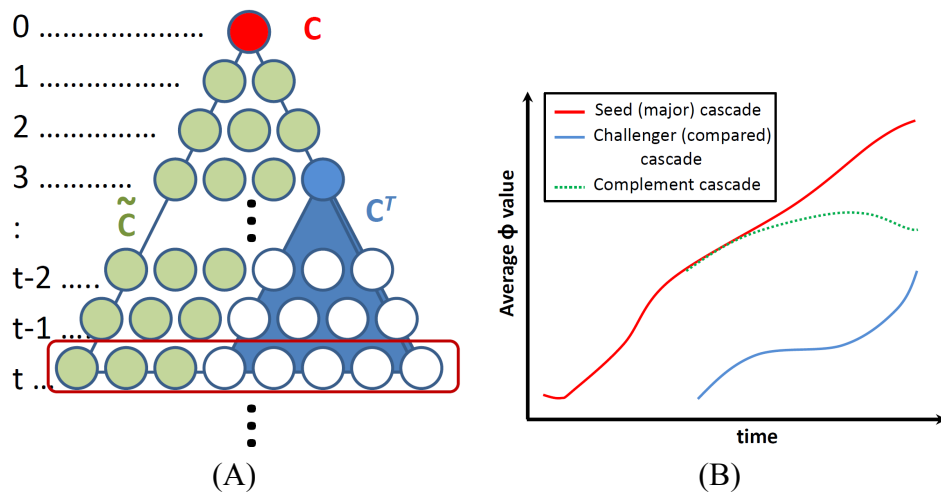


Figure 2.1: (A) Information disruption by a challenger in an information cascade. The seed of an established paradigm, marked in red, creates a cascade as the seed is cited by other papers, while a challenger, marked in blue, disrupts the cascade of the seed. (B) Disruption of the cascade of the seed paradigm (red) by the challenger paradigm (blue) can be visualized as the decline of Φ of the complement cascade (green).

Fig. 2.1(A) illustrates our idea. A *seed* (red node) represents a paper establishing some paradigm in a field of research. The paradigm’s influence grows over time as new papers cite it and are later cited by other papers, creating a *cascade* of citations that can be traced back to the seed. A *challenger* (blue node) is a paper that advocates a new paradigm. It

attracts new citations from papers shown as white nodes with blue background, leaving the *complement cascade* (green nodes) containing papers in the cascade of the seed that are not connected to the challenger. When the challenger represents a non-competing idea, though there will be papers that cite both seed and challenger, they will not interfere with the growth of the seed’s cascade. In contrast, a transformative challenger will disrupt the growth of the established paradigm. Without considering the challenger, it may appear that the established paradigm continues to prosper, as its cascade continues to grow, but subtracting part of the cascade taken over by the challenger will reveal that the growth of the complement cascade (green nodes) slows. In this case, the community’s attention shifts to the challenger paradigm. We propose a method to automatically identify such shifts.

In this chapter, we derive an error bound and empirically demonstrate the reliability of our method against sampling fluctuation of the citation network. This is important because complete citations information may not always be available. This property also allows us to scale the method up to large datasets by subsampling.

We illustrate the efficacy of the proposed approach with case studies. Specifically, we selected several highly influential papers from physics and computer science and showed that the proposed method is better able to identify successful challengers than alternative baselines that consider the number of citations received by the paper. Further, we demonstrate that our method identifies challengers that are more relevant to the topic of the seed paper than baselines. Moreover, challenger’s success is evident early on, allowing for early detection of transformative research. While the focus of this chapter is on scientific publications, the approach can be generalized to other areas where ideas compete to gain attention of information consumers.

2.2 Related Work

Much of bibliometric analysis uses citations count to measure a paper’s quality or scientist’s productivity [36]. Beyond simple citations count, researchers have explored methods that analyze the structure of citation networks to identify important papers [19, 29] or pre-

dict which papers will be important in the future [62].

Few works have explicitly studied transformative research or develop methods to automatically identify such research. Mazlounian et al.[48] characterized how a publication of a landmark paper increases attention paid to author’s other papers, leading to a paradigm shift, which may eventually be recognized with a Nobel prize. Chen [17] described the use of a dynamic co-citation network to reveal “intellectual turning point” papers. Our approach differs from related work in that, first, we explicitly target papers that disrupt established works, and second, we consider cascades, which take chains of citations into account. Next, it is well-known that citation counts decay over time even for a highly influential work [3]. Therefore, it is important to consider its continuing influence of cascades, which provide indirect exposure to the work. Ghosh and Lerman [30] developed a function to quantify the structure of a growing cascade of information spreading in social media, which we use to measure the size of evolving cascades. Here, we propose an efficient method that use this function to identify transformative scientific research.

How information spreads in a network of information ecosystems like social media and scientific publications has been heavily studied. Various models are available to explain and predict information diffusion [30, 46, 31, 47]. Widely spread information may be disrupted by the presence of another piece of information that competes to gain attention from information consumers [52]. In scientific publications, information diffusion is usually measured by counting citations, and citations-based measures, such as the h-index [36], are widely used to evaluate the productivity of scientists. Disruptions of citation cascade growth of well-established, field-defining papers usually represent an event of paradigm shift, breakthrough, emergence of a disruptive idea, and a successful transformative research. Similarly, in social media, the flow of a dominant topic may be disrupted by a challenger, which will gain attention from crosstalk information consumers, who have been following the dominant topic but now switch their attention. A challenger successfully disrupts the dominant topic when the challenger substitutes the attention of a sufficient number of crosstalk information consumers.

2.3 Materials and Methods

2.3.1 Data

We use two large citation network datasets in the empirical evaluation. One of them is the dataset of the journals published by the American Physical Society (APS) [61], which consists of articles published from 1893 to 2009. The APS dataset contains important physics papers that announced a new discovery or a new technique, many of which were recognized by the Royal Swedish Academy of Sciences with a Nobel prize, the highest honor in physics. APS is perfect for our study because it contains many examples of successful transformative research and recognized paradigm shifts and makes them available for analysis.

The other dataset is the DBLP-Citation-network V5 (DBLP) available at Arnetminer.org [68, 65, 67, 66], which consists of two major computer science bibliographic datasets, DBLP and ACM, covering publications from 1936 to 2011. The DBLP dataset contains some of the important papers in computer science that describe widely used techniques and algorithms. Table 2.1 summarizes the statistics of these datasets. The difference of the network structure and the scale of these two datasets reflects the difference in citation culture between these two disciplines of science.

Table 2.1: Statistics of the Test Data

dataset	# paper	# citations	avg. degree
APS	449,667	4,710,548	20.91
pruned	115,753	1,153,967	19.02
DBLP	1,572,278	2,083,947	2.65
pruned	82,762	414,776	10.46

To reduce noise, we pruned low-citation publications as a pre-processing step. Papers must be cited more than 10 times in APS and 5 in DBLP to be included in our evaluation. We considered a citation to a more recent paper as an error and removed 284 from APS and 20,418 from DBLP, respectively. In addition, we excluded 2,555 review articles from the APS dataset that were published in *Reviews of Modern Physics*, since their citation patterns are different from regular research papers [18]. Review papers never start a new

paradigm or become a challenger by definition and thus are not in the scope of our search.



2.3.2 Cascade

We start by defining cascades in citation networks. A citation network is essentially a directed graph $G = (V, E)$ where V is the set of papers and E is the set of edges indicating citations made by papers. A link $(i \leftarrow j) \in E$ denotes that paper j cites paper i , $cite(j)$ denotes the set of all papers that j cites and $cited(i)$ the set of all papers that cite i . V_t is the set of papers published at time t . We assume that if $(i \leftarrow j) \in E$ and $i \in V_t$ and $j \in V_{t'}$ then $t < t'$. That is, no new paper should be cited by an older paper.

Given one or more papers $\mathcal{S} \in G$, a cascade C is a subgraph that contains all citation chains that end at \mathcal{S} . The set \mathcal{S} is called the *seed* or *root* of the cascade. The seed indirectly exerts influence on all papers in the cascade, but influence decays with the distance to the seed [13]. For a node j in the cascade, the cascade generating function $\phi(j)$ summarizes the structure of the cascade [30], i.e., all existing citation chains. The cascade generating function quantifies the influence of \mathcal{S} on node j , and is defined recursively by

$$\phi(j) := \begin{cases} 1 & \text{if } j \in \mathcal{S} \\ \sum_{i \in cite(j)} \alpha \phi(i) & \text{otherwise,} \end{cases} \quad (2.1)$$

where α is a constant damping factor. Fig. 2.2 shows an example cascade and the ϕ values for its nodes. For a paper j published after T time steps (e.g., years) from the publication of the seed, $\phi(j)$ can be written as follows:

$$\phi(j) = \sum_{p=0}^T a_p \cdot \alpha^p, \quad (2.2)$$

where the coefficient a_p is the number of distinct paths of length p from one of the seeds to j . The impact of α is that the smaller the value of α , the higher the penalty against long paths. It is also possible to assign a unique α_{ij} for each link but we found that it is simpler to assign a constant 0.5 for all links to control its impact.

2.3.3 Cascade Disruption

Consider Fig. 2.1(A). C is the full cascade originated by the seed paper. Let $C^{(\mathcal{T})}$ denote the cascade originating from the challenger \mathcal{T} . We define the *complement cascade* \tilde{C} as the subgraph of C obtained by subtracting $C^{(\mathcal{T})}$ from C , *i.e.*,

$$\tilde{C} := C - (C \cap C^{(\mathcal{T})}) = C \setminus C^{(\mathcal{T})}.$$

By definition, references of papers in \tilde{C} can only be traced back to the seed papers but not the challenger. Thus, they represent the influence of \mathcal{S} that cannot be attributed to. We note that it is not necessary for the challenger \mathcal{T} to be in C . The blue nodes in Fig. 2.1(A) are the root node(s) of the intersection of C and $C^{(\mathcal{T})}$. These nodes can be considered as “cross-talk” between the seed and challenger paradigms.

We say that challenger \mathcal{T} disrupts the growth of \mathcal{S} when new papers in the cascade of $\mathcal{S}(C)$ can be explained by the cascade of $\mathcal{T}(C^{(\mathcal{T})})$. This will result in a shrinking complement cascade \tilde{C} . Next, we present a procedure to measure disruption.

Let C_t be the set of papers in cascade C published at time t , *i.e.*, nodes in the bottom red box in Fig. 2.1(A). The average of the cascade function ϕ of papers in C_t is defined by

$$\Phi_t(C) := \frac{1}{|C_t|} \sum_{j \in C_t} \phi(j) = \sum_{p=0}^t \bar{a}_p \cdot \alpha^p, \quad (2.3)$$

where \bar{a}_p is the average of the coefficient a_p in Eq. (2.2) for j in C_t , and \bar{a}_p indicates on average number of distinct citation chains of length p from papers published at time t to the seeds. The variable Φ_t can be interpreted as an indicator of the seed papers’ influence at time t .

Fig. 2.1(B) shows the growth of $\Phi_t(C)$ (red curve), $\Phi_t(\tilde{C})$ (green curve), and the challenger cascade $\Phi_t(C^{(\mathcal{T})})$ (blue curve). The value of Φ_t for both the seed (red) and challenger (blue) papers may both grow rapidly, but the growth of the complement cascade flattens and drops once the challenger successfully shifts the attention of the community. Otherwise, the green curve will continue to grow. In other words, successful information disruption is associated with a declining values $\Phi_t(\tilde{C})$ of the complement cascade.

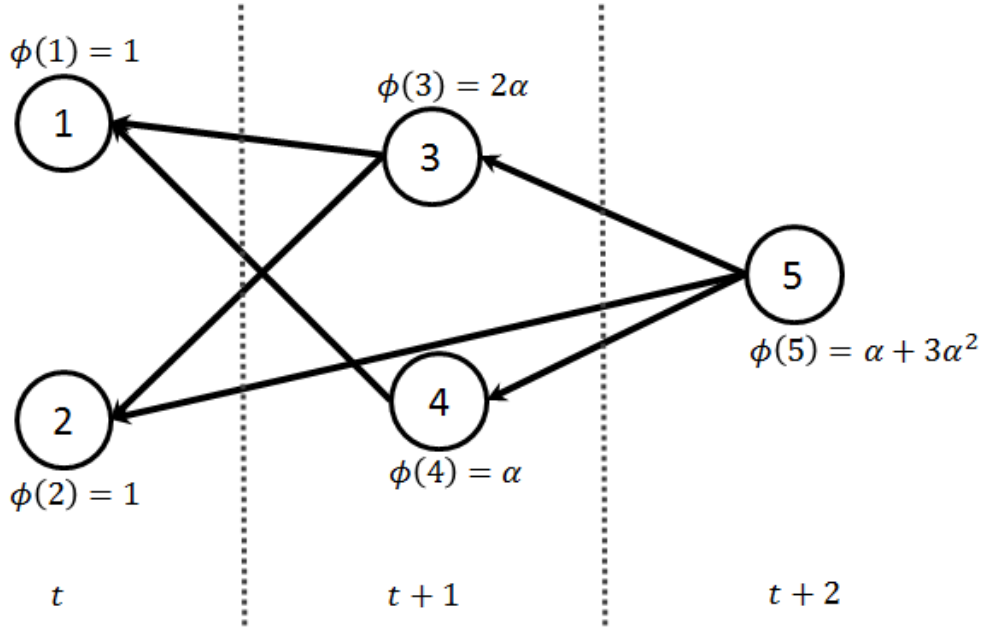


Figure 2.2: An example of Cascade and their ϕ values.

We quantify this decline by the *disruption score* $\delta(\tau)$, which is a function of the time interval of τ given the seed and challenger cascades. Let t_0 be the publication time of the challenger paper,

$$\begin{aligned} \delta(\tau) &:= \sum_{t=t_0}^{t_0+\tau} \log \frac{\Phi_t(C)}{\Phi_t(\tilde{C})} \\ &= \sum_{t=t_0}^{t_0+\tau} \left(\log \Phi_t(C) - \log \Phi_t(\tilde{C}) \right). \end{aligned}$$

The disruption score can be visualized as the area between the red and green curves in Fig. 2.1(B) from t_0 to $t_0 + \tau$. The disruption score allows us to identify and measure the impact of the challenger paper.

When comparing candidate challengers published too long apart over time, the cascade of the seed paper may be so different that might give unfair advantages to old challengers. For example, the cascade of a seed paper published in 1950's may grow many-fold from 60's to 90's. For a new paper to disrupt the same proportion of the cascade as an old paper may require a much larger number of citations. The disruption score is immune from this problem because ϕ will be smaller after 30 years as citation paths to the seed stretch. More importantly, we consider the average, not sum. Also, the number of publications and thus

citations to new papers grow faster in recent years and may compensate for the difference in cascade size.



2.3.4 Computing Cascade Disruption

To obtain the disruption score, we need to compute ϕ of the nodes in the cascade. A citation network is a directed acyclic graph if cycles are considered as errors. From Eq. (2.1), traversing the citation network in a topological order [41] and updating ϕ values along the way will guarantee that no backtracking is necessary to compute all ϕ values for all nodes. Therefore, we can apply topological sorting to compute ϕ and obtain the disruption scores. The time complexity of topological sorting is $O(|V_C|+|E_C|)$, which is linear to the sum of the number of nodes and edges in cascade C .

Cascade generating function ϕ can measure information cascades not only in citations networks, but also in other domains, such as information diffusion in social media or influence in social networks. The method for measuring disruptions of cascade growth should, therefore, carry over to these domains as well. This could lead to numerous other applications, such as comparing competing memes that are spreading in social media to determine which one is attracting more attention, or which person is becoming more influential.

2.4 Evaluation

According to the classical test theory, a quantitative measure must be both valid and reliable. The notion is closely related to bias and variance in statistical data mining and pattern recognition [34]. As a quantification measure of transformative research, the disruption score must be *valid*, in the sense that truly transformative research will be scored higher than others, and *reliable*, in the sense that the score is robust against incomplete subsampling of citation network data. In addition, computation of the score must scale up to large citation network data. In this section, we evaluate the validity, reliability, and scalability of the disruption score as a detector of transformative research.

2.4.1 Validity

The disruption score is a valid indicator of paradigm shift if the score distinguishes truly transformative research papers from the rest with high sensitivity and specificity. However, unlike well-defined data mining problems, it is difficult to create a large gold standard of truly transformative research to quantitatively assess the validity of the proposed method. Therefore, we focus on a few well-known cases of transformative research to evaluate our method's validity. Section 2.5 reports detailed results of applying our method to APS and DBLP datasets.

Consider superconductivity. The 1957 theory of superconductivity by Bardeen, Cooper, and Schriffer (BCS) [6, 5] was a dominant paradigm in this field until the discovery of high-temperature superconductivity [8] (HTS) in 1986, an indisputable transformative research accomplishment for which the authors were awarded the Nobel Prize in Physics the next year.

Fig. 2.3(C) shows evolution of the cascade size, *i.e.*, the number of papers in the cascade, of the BCS cascade, and the HTS cascade rooted at three pioneering APS papers in this field [76, 77, 26]. One may expect that discovery of HTS would slow down the growth of the BCS cascade, but Fig. 2.3(C) shows otherwise. The cascade size, in terms of the cumulative number of papers in the cascade each year, continues to grow, though at a slower pace than HTS. HTS might surpass BCS soon, but the impact of paradigm shift is hardly observable 20 years later if we use the cascade size as an indicator.

Fig. 2.3(A) compares the growth of the logarithms of $\Phi_t(C^{(bcs)})$ (red), $\Phi_t(C^{(hts)})$ (blue) and $\Phi_t(\tilde{C})$ (green) as computed from the APS dataset. We see a pattern identical to the one shown in Fig. 2.1(B), a vivid demonstration of cascade disruption and paradigm shift. Moreover, the disruption starts immediately after the publication of HTS.

To test the specificity of cascade interruption, we randomly selected 30 papers published in 1987, the same year as HTS seeds, from the APS dataset as negative controls and plotted the growth of their cascades as shown in Fig. 2.3(B), where the blue curve shows the means and standard deviations of the average cascades of these 30 challengers and the green curve shows those for their complement cascades. The curves show that though the

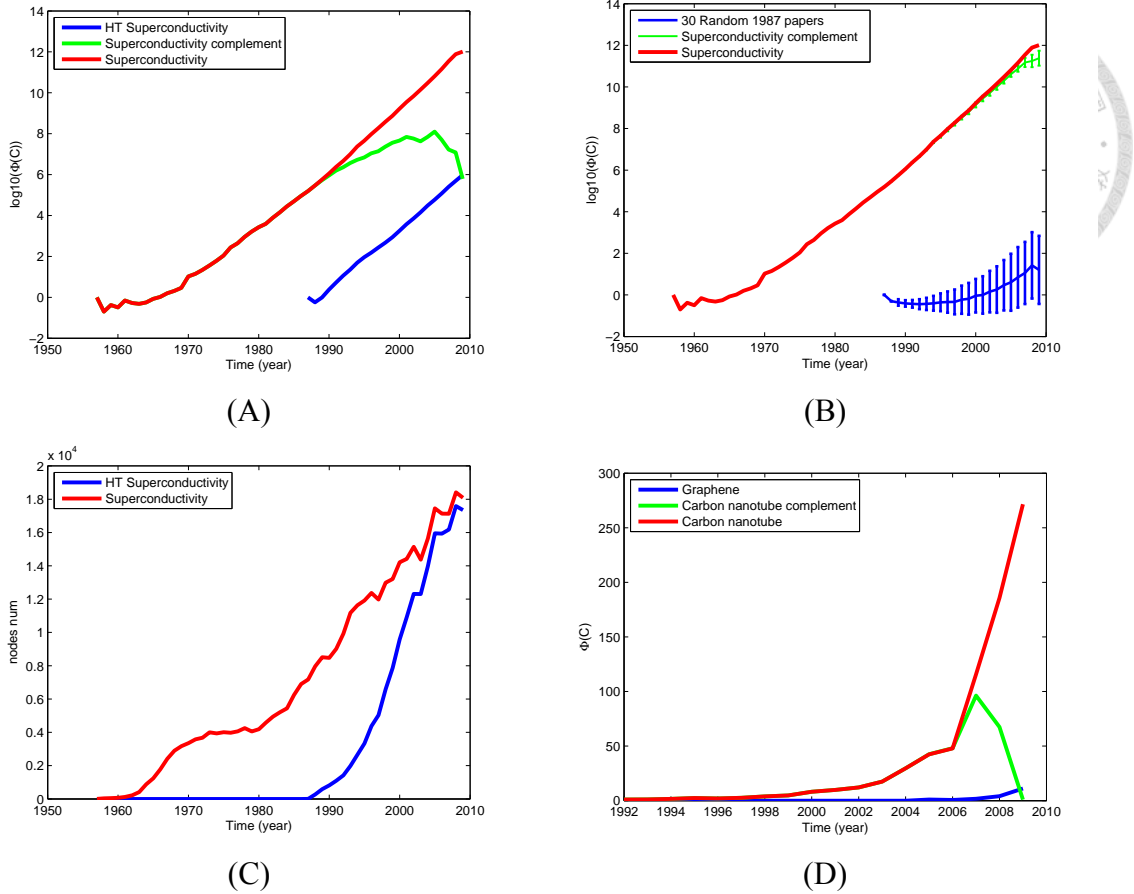
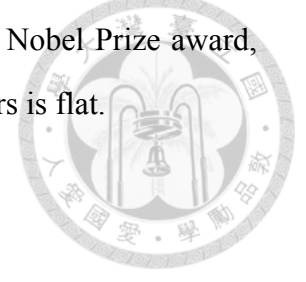


Figure 2.3: Cascade disruption as an indicator of paradigm shift. The growth of the logarithm of average cascade function values per year for (A) superconductivity (BCS) [6, 5] and high-temperature superconductivity (HTS) [76, 77, 26] and (B) 30 control cases published in 1987 show no sign of cascade disruption against BCS. (C) The growth of the size of cascades shows no sign of disruption. (D) Another example of paradigm shift is conventional carbon nanotube [39] versus graphene [54, 55]. Unlike superconductivity, the cascades are not as large because they were published in recent years. Therefore, no logarithm of Φ is taken here.

growth of their cascades varies widely, the complements of the BCS cascade are hardly disrupted, unlike the HTS papers.

Another example of transformative research is the development of graphene in 2004 [54, 55], which was considered a breakthrough both for the materials fabrication technology, focused on carbon nanotubes [39] and as a system for studying properties of 2-dimensional electron systems. The developers of graphene were awarded Nobel Prize in Physics in 2010. Fig. 2.3(D) shows the cascade growth and disruption in this case. Again, the disruption is observable starting in 2006, right after their publication. This is as fast as possibly

detectable because we removed all citations between papers published in the same year. The disruption then drops sharply in 2007, three years before their Nobel Prize award, even though the growth of the average cascade of the graphene papers is flat.



2.4.2 Reliability

Existing datasets of citation networks are inevitably incomplete and only contain a subset of all related papers and citations. It is important that the proposed disruption score produces consistent results given different subsamples of citation network data.

Here we show that it is possible to derive a theoretic error bound of the disruption score given a subsample of citation network data, compared to the score obtained from the complete citation network. We observed empirically in our preliminary study that if the average cascade function values $\Phi_t(C) > \Phi_t(\tilde{C})$, the relation will maintain when they are estimated from a subsampled cascade, *i.e.*, $\Phi_t(C') > \Phi_t(\tilde{C}')$. In other words, if we observe cascade disruption in a subsampled cascade, then it is almost certain that cascade disruption will also present in a complete cascade.

To see why this is the case, let C' be the subsampled cascade from the complete cascade C with a constant node sampling ratio ρ . The citation links in C adjacent to nodes not in C' are removed from C' . Since Φ essentially is the true mean of the cascade function values ϕ given a complete cascade C , if ϕ of the nodes in the subsampled cascade C' are identical to their ϕ values in the complete cascade C , then according to Hoeffding's inequality, which states that the probability that the difference of sample mean and true mean is large is less than a formula that is roughly proportional to the exponential of the inverse of the sample size, we can show that $\Phi_t(C') \approx \Phi_t(C)$ and this is similarly the case for the complement cascade \tilde{C} and its subsample \tilde{C}' and hence the inequality relation will maintain.

However, the new ϕ of the nodes in the subsampled cascade C' will be different, because citations to the removed nodes are absent. Also, ϕ of those being cited will be

smaller due to the removal of the unselected nodes. Therefore,

$$\begin{aligned}\Delta\phi_C(j) &= \phi_C(j) - \phi_{C'}(j) \\ &= \alpha \left(\sum_{i \in C} \phi_C(i) I(i \in \text{cite}(j) \& i \notin C') \right. \\ &\quad \left. + \sum_{i \in C} \Delta\phi_C(i) I(i \in \text{cite}(j) \& i \in C') \right),\end{aligned}$$



and its expectation will be

$$\begin{aligned}\mathbb{E}[\Delta\phi_C(j)] &\approx \\ &\alpha \mathbb{E}(|\text{cite}(j)|) |C| ((1 - \rho) \mathbb{E}[\phi_C(i)] + \rho \mathbb{E}[\Delta\phi_C(i)]),\end{aligned}\quad (2.4)$$

where $|C|$ is the number of nodes in cascade C . This applies to the complement cascade \tilde{C} and its subsample \tilde{C}' as well. Since Φ is the expectation of ϕ for papers published at the same time, from the Hoeffding's inequality and Eq. (2.4), we can conclude that with a high probability proportional (roughly speaking) to the sampling size of the subsampled cascade, the difference

$$|\Delta\Phi_t(C) - \Delta\Phi_t(\tilde{C})| = |\Phi_t(C') - \Phi_t(\tilde{C}') - (\Phi_t(C) - \Phi_t(\tilde{C}))|$$

will be very small. The following theorem establishes a bound for the sampling error of the average cascades.

Theorem 1. *For any strictly positive constant ε , with probability greater than*

$$1 - 2e^{-2\varepsilon^2|C'_t|} \frac{1}{4\varepsilon|C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} - 2e^{-2\varepsilon^2|\tilde{C}'_t|} \frac{1}{4\varepsilon|\tilde{C}'_t|} \sqrt{\frac{\pi}{2|\tilde{C}'_t|}},$$

if $(\Phi_t(C') - \Phi_t(\tilde{C}')) > 0$, then

$$\exists S > 0, (\Phi_t(C) - \Phi_t(\tilde{C})) \in (S + \varepsilon, S - \varepsilon).$$

Proof. Appendix A provides a detailed proof by the Hoeffding’s inequality. □

We also empirically tested the reliability of the disruption score with subsampling. We chose the top highly cited papers in DBLP and ranked the papers in their cascades according to their disruption scores. Next, we assessed the reliability of our method by a 5-fold cross validation sampling test, where we divided all papers in the dataset into five subsets and used four of them to assign the ranks. Then we used the Spearman’s rank correlation coefficient to measure the similarity of the ordering of the top 1000 articles in the five trials. The similarity tests show that using 80% of the data yields similar disruption scores and similar rankings. Fig. 2.4 shows the heatmaps of correlations for two well-known papers in the data mining community as the seeds. We also observed that the differences between the disruption scores of the top 5 challengers computed from the cross-validation subsamples and from the complete dataset are small and with negligible variance (data not shown). We set $\tau = 4$ years when computing the disruption score in all trials. Using other highly-cited papers in DBLP gives similar results.

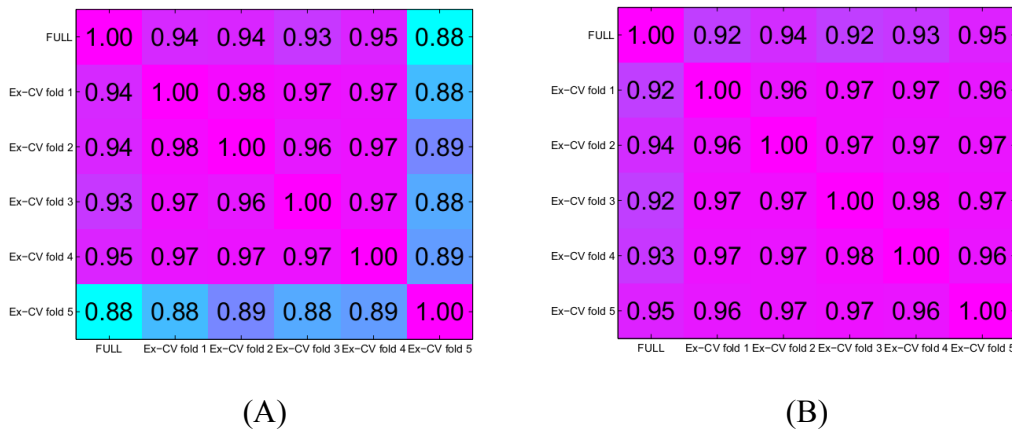


Figure 2.4: Heatmaps of correlation between trials of 5-fold cross-validation for (A) “Fast Algorithms for Mining Association Rules in Large Databases (1994)” and (B) “Induction of Decision Trees (1986)”. “Full” is the result for the complete data, “Ex.CV fold i ” is the result of the i -th cross validation trial.

2.4.3 Scalability

We already showed that the algorithm to compute the disruption score is linear in the size of the citation network. As the number of publications grew geometrically in recent years,

and to apply the algorithm to even larger networks of social media, the scalability of the algorithm has to improve further. One of the options is to explore subsampling. Theorem 1 and the empirical results show that the disruption score can be estimated reliably from a subsampled citation network. This useful property allows us to further accelerate computation. With a suitable sampling, computing the disruption scores can be more efficient in both computation time and memory space. According to our execution time statistics with different ratios of node sampling from APS, the time drops nearly exponentially because the number of citations decreases exponentially as the number of nodes decreases linearly: e.g., sampling 80% of papers can save 55% of the time.

When the task is to rank a large number of candidate challengers by their disruption scores, it is possible to avoid exhaustive pairwise comparison by reusing intermediate results. Suppose we would like to rank 100 candidate challengers by their disruption scores. A brute-force approach is to compute the complement cascades for each of the candidates. By sorting these candidates in their topological order in the citation network, the ϕ values computed for the upstream candidates can be reused for the downstream candidates and significantly reduce the computational costs.

2.5 Results and Discussion

In previous section, we report an evaluation of our method by testing if the disruption scores are high for known examples of transformative research when they are scored against the representative papers of the paradigms that were disrupted. In this section, we report a further test, where a highly cited paper is chosen and the goal is to use our method to rank all the papers in its cascade by their disruption score and see whether the highest scoring paper represents the best transformative research, under the condition that the system is blind about which papers are transformative. We note that in this case, it is possible that no challenger is sufficiently transformative against selected high cited papers but the highest scoring ones may still hint us about which papers are emerging. Again, since it is difficult to create a large set of the “ground truth” of transformative research, we will not to provide a quantitative evaluation, such as measuring error rates or area-under-

Table 2.2: Top ten challengers to the 1957 “Theory of Superconductivity” identified by (a) proposed method and (b) baseline method.

Year	Cites	Title
(a) our method: sorted by disruption score		
1958	14	Meissner Effect
1958	307	Random-Phase Approximation ... Superconductivity
1959	40	Evidence for Anisotropy of the Superconducting Energy...
1989	574	Phenomenology of ...Cu-O high-temperature supercon...
1987	368	Antiferromagnetism in $\text{La}_2\text{CuO}_{4-y}$
1987	281	Two-dimensional antiferromagnetic quantum ...
1988	149	$\text{Ba}_2\text{YCu}_3\text{O}_7$: Electrodynamics of Crystals ...
1990	156	High-resolution angle-resolved photoemission ...
1988	399	Low-temperature behavior of two-dimensional quantum ...
1995	95	Momentum Dependence of the Superconducting ...
(b) baseline: sorted by cover ratio		
1958	307	Random-Phase Approximation ... of Superconductivity
1958	14	Meissner Effect
1958	63	... States in the Theory of Superconductivity...
1958	93	Paramagnetic Susceptibility in Superconductors
1958	14	Meissner Effect and Gauge Invariance
1960	246	Quasi-Particles and Gauge Invariance ... Superconductivity
1959	36	Impurity Scattering in Superconductors
1959	37	Collective Excitations in the Theory of Superconductivity
1958	119	... Spectra of Nuclei ... the Superconducting Metallic State
1960	32	... Solution and ... Superconducting Transition Temperature
(c) baseline: sorted by citations		
1981	3191	Self-interaction correction to density-functional approx...
1996	3088	Generalized Gradient Approximation Made Simple
1980	2651	Ground State of the Electron Gas by a Stochastic Method
1976	2569	Special points for Brillouin-zone integrations
1996	2387	Efficient iterative schemes for ab initio total-energy...
1990	1951	Soft self-consistent pseudopotentials in a generalized...
1991	1950	Efficient pseudopotentials for plane-wave calculations
1975	1597	Linear methods in band theory
1992	1567	Atoms, molecules, solids, and surfaces:...
1992	1445	Accurate and simple analytic representation...



curve, but will demonstrate through several case studies that the proposed method is able to identify examples of transformative research. We use the APS and DBLP datasets described in Section 2.3.1 to identify examples of transformative research in physics and computer science, respectively. We compare our method to two *baselines*, one of them is to order the papers within a cascade by their popularity, *i.e.*, the number of citations they receive, and show that our method identifies more relevant challengers. The other, which we called *cover ratio*, is the ratio of the cascade sizes of the seed and the complement created by a candidate challenger.

2.5.1 Physics

We chose several papers with the most citations in our dataset, which came from different subfields of physics. We identified the most disruptive challengers of these papers and carried out quantitative analysis of their topics.

Case Study 1 In 1957 Bardeen, Cooper and Schrieffer published a seminal paper titled “Theory of Superconductivity” which explained the mechanism by which some metals became perfect electrical conductors (*i.e.*, they lost their electrical resistance) at low temperatures. The authors were awarded a Nobel prize for this discovery in 1972. This paper is one of the ten most cited papers in the APS dataset. Table 2.2 lists the ten top-ranked challengers identified by the proposed method and the baseline. The disruption score of challengers was computed for a ten-year period ($\tau = 10$). Compared to the citations baseline, both our method and the cover ratio baseline identifies papers that are relevant to the topic of superconductivity. All ten of the top challengers identified by baseline are papers dealing with calculations of electronic structure of materials, and include other most-cited papers in the APS dataset. While this is a very important topic, it is only peripherally related to superconductivity, in as much as this phenomenon is a result of correlated electron pairs.

While top-rated challengers discovered by the cover ratio baseline are on the topic of superconductivity, only the proposed method discovered papers on high temperature superconductivity (HTS). The discovery of HTS was an important development in the study of superconductivity, recognized with a Nobel prize in 1987. Although the original paper announcing the discovery is not in our dataset, presence of several other papers on HTS among the top challengers demonstrates the efficacy of our method to identify disruptive papers. These challengers include “Antiferromagnetism in $\text{La}_2\text{CuO}_{4-y}$ ”, “Two-dimensional antiferromagnetic quantum spin-fluid state in La_2CuO_4 ”, “ $\text{Ba}_2\text{YCu}_3\text{O}_7$: Electrodynamics of Crystals with High Reflectivity” and “Momentum Dependence of the Superconducting $\text{Sr}_2\text{CaCu}_2\text{O}_8$ ”.

We verify that our method identifies more relevant challengers through the analysis of their PACS numbers. The Physics and Astronomy Classification Scheme (PACS) was introduced in 1975 to allow authors to identify the field and subfields of their papers. Fig. 2.5 shows the frequency distribution of PACS categories of the 30 top-ranked challengers (with PACS numbers) identified by our method and the citations baseline, weighted by the score assigned to the paper by the method. We aggregated the numbers by their top-

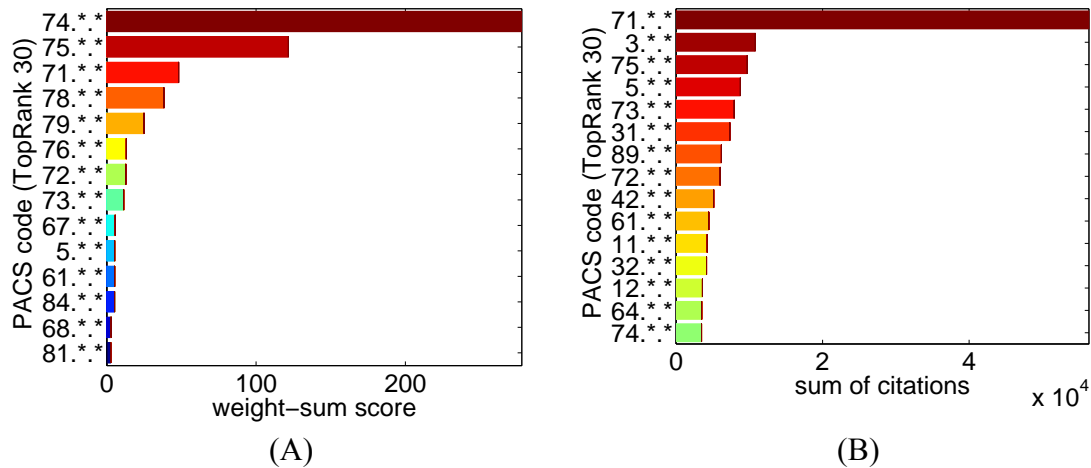


Figure 2.5: Distribution of PACS numbers of challenger papers identified by (A) our method and (B) baseline method.

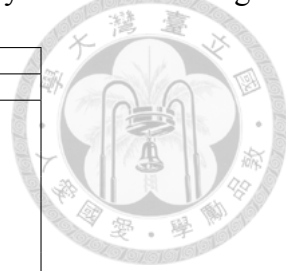
level category. Our method finds many more papers about “Superconductivity” (category 74) and “Magnetic properties and materials” (75) than baseline, which finds papers about “Electronic structure of bulk materials” (71), “Quantum mechanics” (3) and “Statistical physics” (5), while “Superconductivity” (topic 74) is 15th most frequent topic among these challengers. In fact, topics (71, 75, 5) are the most common PACS numbers in the entire dataset, suggesting that baseline method picks out globally popular papers, even though it considers only the papers in the cascade created by the seed node.

Case Study 2 We used our method to rank challengers of the most cited paper in particle physics¹. This is the 1967 paper by Steven Weinberg titled “A Model of Leptons.” This seminal work unified weak and electromagnetic interactions within a single theory of electroweak interactions. It won its authors a Nobel prize in 1979. Table 2.3 lists the ten challengers to this paper with highest disruption score, which was computed for $\tau = 10$ years. The first and second challengers are papers by David Gross and Frank Wilczek, and David Politzer respectively. These three physicists shared a Nobel prize in 2004 for elucidating the theory of strong interactions, which along with gravity, electromagnetic, and weak interactions forms the four fundamental forces of nature. Though these papers received a nod from the Nobel committee 30 years after their publication, our method identifies them as important already ten years after publication.

¹<http://www.slac.stanford.edu/spires/topcites/2010/alltime.shtml>

Table 2.3: Top ten challengers to the 1967 “A Model of Leptons” by Steven Weinberg identified by (a) proposed method and (b) baseline method.

Year	Cites	Title
(a) our method: sorted by disruption score		
1974	190	Asymptotically free gauge theories. II
1974	123	Electroproduction scaling in an ... of strong interactions
1974	309	Hierarchy of Interactions in Unified Gauge Theories
1974	696	Confinement of quarks
1973	162	New Approach to the Renormalization Group
1972	46	Spontaneous Breakdown and Hadronic Symmetries
1973	44	Unified Gauge Theories of Hadrons and Leptons
1972	208	Effects of a Neutral Intermediate Boson in Semilep...
1974	361	Experimental Observation of a Heavy Particle J
1973	59	Current Algebra and Gauge Theories. I
(b) baseline: sorted cover ratio		
1971	236	Physical Processes in a Convergent Theory ...
1970	35	Spontaneous Breakdown ... Interaction Symmetry
1972	95	Renormalizable Massive Vector-Meson Theory-Perturbation ...
1972	71	Short-Distance Behavior of Quantum Electrodynamics ...
1973	742	Radiative Corrections ... Spontaneous Symmetry Breaking
1972	157	Spontaneously Broken Gauge Symmetries. I. Preliminaries
1972	109	Spontaneously Broken Gauge Symmetries. II. ...
1972	68	Approximate Symmetries and Pseudo-Goldstone Bosons
1972	94	Spontaneously Broken Gauge Symmetries. III. Equivalence
1972	58	Mixing Angle in Renormalizable Theories of ...
(c) baseline: sorted by citations		
1981	3191	Self-interaction correction to density-functional ...
1996	3088	Generalized Gradient Approximation Made Simple
1996	2387	Efficient iterative schemes for ab initio total-energy ...
1990	1951	Soft self-consistent pseudopotentials in a generalized ...
1991	1950	Efficient pseudopotentials for plane-wave calculations
1992	1567	Atoms, molecules, solids, and surfaces: Applications ...
1992	1445	Accurate and simple analytic representation of the ...
1994	1430	Projector augmented-wave method
1999	1424	From ultrasoft pseudopotentials to the projector ...
1993	1345	Ab initio molecular dynamics for liquid metals

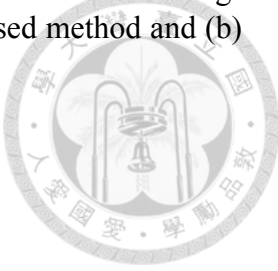


The top-ten challengers include four more papers by Steven Weinberg (3, 5, 8, and 10), and papers by Nobel laureates Kenneth G. Wilson (paper 4 “Confinement of quarks”), and Samuel C. Ting (paper 9 “Experimental Observation of a Heavy Particle J”). Though these papers are on slightly different topics than the seed, they are all important works within the particle physics community, demonstrating that our method is able to capture how the community shifts its attention between different topics.

The cover ratio baseline, on the other hand, identifies papers on several topics of particle physics, most notably gauge symmetry breaking. This is an important research area in theoretical particle physics, but one that generalizes to all forces. In contrast, our method found more challengers relevant to the topic of electroweak interactions.

Of the top ten challengers identified by citations baseline, seven are the same as the baseline challengers of the first case study. These are papers in a popular topic of using density functional theory, or its variants, for electronic structure calculations, and are not

Table 2.4: Top ten challengers (published after 1994) to the 1982 “Two-Dimensional Magnetotransport in the Extreme Quantum Limit” identified by (a) proposed method and (b) baseline method.



Year	Cites	Title
(a) our method: sorted by disruption score		
1995	246	Spontaneous interlayer ... double-layer quantum Hall ...
2005	179	Unconventional Integer Quantum Hall Effect in Graphene
2005	65	Electric Field Modulation ... Mesoscopic Graphite
2005	178	Quantum Spin Hall Effect in Graphene
1995	199	Optically Pumped NMR Evidence ... Skyrmions GaAs ...
2005	42	Coulomb interactions and ferromagnetism ... graphene
2005	21	Disorder and interaction ... two-dimensional graphene ...
2005	14	Coexistence of sharp quasiparticle ... in graphite
2005	45	Local defects and ferromagnetism in graphene layers
2005	121	Z_2 Topological Order and the Quantum Spin Hall Effect
(b) baseline: sorted by cover ratio		
1995	85	Updated analysis ... baryon spectrum
1996	857	Review of Particle Physics
1995	50	... partial-wave T matrices in a ...
1995	51	Baryon current matrix elements in a light-front framework
1995	22	Kinematic evidence for top quark pair production ...
1995	18	Search for High Mass Top Quark Production in pp ...
1995	269	Observation of the Top Quark
1995	337	Observation of Top Quark Production in p...
1995	11	... in a three-coupled-channel, multiresonance, unitary model
1995	72	Static Response and Local Field Factor of the Electron Gas
(c) baseline: sorted by citations		
1996	3088	Generalized Gradient Approximation Made Simple
1996	2387	Efficient iterative schemes for ab initio total-energy ...
1999	1424	From ultrasoft pseudopotentials to the projector ...
1998	1003	Quantum computation with quantum dots
1996	857	Review of Particle Physics
1998	845	Entanglement of Formation of an Arbitrary State of ...
1996	795	Mixed-state entanglement and quantum error correction
1998	748	Evidence for Oscillation of Atmospheric Neutrinos
1998	737	Cold Bosonic Atoms in Optical Lattices
1995	664	Double Exchange Alone Does Not Explain the Resist...

relevant to the topic of high energy physics. This case study further highlights the ability of our method to identify important and relevant challengers.

Case Study 3 Our final study considers the fractional quantum Hall effect, a phenomenon in which the conductance of 2-dimensional electrons is quantized at certain levels. This effect was first reported in a 1982 paper titled “Two-Dimensional Magnetotransport in the Extreme Quantum Limit.” The discovery and explanation of this effect was recognized with a Nobel prize in 1998. Table 2.4 shows the top ten challengers identified by our method and baseline. Since the APS dataset ends in 2009, the disruption score for 2005 papers were computed for the four year period. To mitigate the bias that incomplete data introduces, we show only the challengers published after 1994.

Several of the challengers with highest disruption score are about graphene, a one-atom-thick layer of graphite, whose discovery has facilitated new investigations of the

properties of matter and electrons confined to 2-dimensional surfaces, and resulted in a Nobel prize in 2010. In comparison, both baseline methods identify irrelevant challengers, including those dealing with the top quark (papers (b)6–8), calculations of electronic structure of bulk materials (papers (c)1–3), quantum computing (papers (c)4, 7) and high energy physics ((b)2, (c)5, 8, 9).

Quantitative Analysis We validate quantitatively that the proposed method identifies more relevant challengers than baseline by performing PACS number analysis of the challengers for the ten most-cited papers in the APS dataset. We compared the PACS number distribution of the 30 top-ranked challengers identified by each method with the distribution of PACS numbers of all papers in the APS dataset (with PACS numbers). The mean correlation of the distributions of PACS numbers of challengers of the 10 top-ranked papers identified by our method with the global PACS number distribution is 0.4611 ± 0.0048 . The mean correlation of PACS number distribution of challengers for the 10 top-ranked papers found by baseline with the global PACS number distribution is 0.5800 ± 0.0033 . Higher correlation of the baseline method indicates that it tends to identify challengers on globally popular topics, compared to the proposed method, which tends to identify challengers that are topically relevant to the seed.

2.5.2 Computer Science

We report results of two case studies of high interest to the data mining community, using the most highly cited papers in the DBLP dataset as seeds. Due to the fast pace of computer science research, we set $\tau = 4$ years to compute the disruption score $\delta(\tau)$.

Case Study 1 Ross Quinlan’s 1986 paper on ID3 is one of the most influential papers in computer science that laid the foundation of the field of classifier learning. One may expect that papers about new algorithms of classifier learning are top challengers that transform the field, but surprisingly the results given in Table 2.5 shows that it is papers about association rule mining. Research in association rule mining led to a whole new field of data mining, while in comparison, new research in classifier learning is still within the

Table 2.5: Top challengers to the 1986 “Induction of Decision Trees” paper identified by (a) proposed method and (b) baseline method.

Year	Cites	Title
(a) our method: sorted by disruption score		
1995	189	Discovery of Multiple-Level Association Rules ...
1995	227	An Effective Hash Based Algorithm ... Association Rules
1996	211	Sampling Large Databases for Association Rules
1995	254	An Efficient Algorithm for Mining Association Rules...
1997	191	Beyond Market Baskets: Generalizing Association Rules
(b) baseline: sorted by cover ratio		
1992	53	Querying in Highly Mobile Distributed Environments
1993	33	Relevance Feedback and Inference Networks
1992	64	An Interval Classifier for Database Mining Applications
1993	143	Database Mining: A Performance Perspective
1993	1372	Mining Association Rules between Sets of Items ...
(c) baseline: sorted by citations		
1994	1592	Fast Algorithms for Mining Association Rules...
1993	1372	Mining Association Rules between Sets of Items ...
2000	647	Directed diffusion: a scalable and robust ...
2002	602	Wireless sensor networks: a survey.
2000	523	Content-Based Image Retrieval at the End of ...



realm laid out by Quinlan’s ID3. In this sense, our result is more reasonable. The top challenger is perhaps the most related to ID3 among papers on association rule mining because a decision tree can be considered as a set of multiple-level rules. Our results are also more reasonable than those found by both baselines. Though two association rule mining papers appear in the top-5 lists found by the challengers, the remaining papers are irrelevant to decision tree learning.

Case Study 2 Next, we asked what challenges association rule mining. Our seed selection is Agrawal and Srikant’s 1994 seminal paper, which is the third most-cited paper in DBLP. The results, shown in Table 2.6, suggest that it remains dominant in data mining, as top five challengers are all follow-up papers with relatively low disruption scores (data not shown). Here, cover ratio baseline identifies similar challengers as those found by our method. The citations baseline selects mostly irrelevant papers.

2.6 Summary

Transformative research shifts attention of the scientific community from the established paradigms that represent theories and methods accepted and practiced by the community. The degree to which the paradigm is accepted by the community is reflected in the citations received by papers that first describe it, and citations received by these papers,

Table 2.6: Top challengers to the 1994 “Fast Algorithms for Mining Association Rules in Large Databases” paper identified by (a) proposed method and (b) baseline method.

Year	Cites	Title
(a) sorted by disruption score		
1995	227	An Effective Hash Based Algorithm ... Association Rules
1995	189	Discovery of Multiple-Level Association Rules...
1996	211	Sampling Large Databases for Association Rules
1995	254	An Efficient Algorithm for Mining Association Rules...
1998	170	Exploratory Mining and ... Association Rules.
(b) baseline: sorted by cover ratio		
1995	227	An Effective Hash Based Algorithm ... Association Rules.
1995	189	Discovery of Multiple-Level Association Rules ...
1995	254	An Efficient Algorithm for Mining Association Rules ...
1996	211	Sampling Large Databases for Association Rules
1997	252	Dynamic Itemset Counting and Implication Rules ...
(c) baseline: sorted by citations		
2000	523	Content-Based Image Retrieval at the End of the Early...
2000	492	Mining Frequent Patterns without Candidate Generation
2002	350	Optimizing search engines using clickthrough data
2002	338	Models and Issues in Data Stream Systems
2001	328	Item-based collaborative filtering recommendation...



and so on. By looking at the structure of the citations cascade, we can determine when a new paradigm attracts attention of the scientific community. This happens when citations received by papers advancing the new paradigm can explain most of the new citations received by the old paradigm. These shifts of attention are evident soon after the challengers’ publication, enabling early detection of transformative research. We have proposed a method to identify transformative challengers, i.e., scientific papers that shift attention of the community, by measuring how much they disrupt the growth of citation cascades of papers representing the established paradigm. When applied to citations networks of physics and computer science papers, our method correctly identified several examples of transformative research.

More work needs to be done to elucidate the processes that lead to shifts of attention. We need to identify seeds which simply do not have any significant challengers. Also, we would like to develop scalable methods that take into account a set of seeds and a set of challengers. Another interesting direction is to develop methods to identify which established idea a given paper disrupts. We believe that identifying transformative research by analyzing citations cascades will prove to be a productive line of inquiry.



Chapter 3

Citing the Protein Data Bank and Related Repository

3.1 Introduction

In this Chapter, we focus on analyzing citations to the PDB data repository. We will then investigate citations to individual structures as our next step. PDB users currently have different choices to cite the PDB data repository. They can cite the original debut publication of the RCSB PDB published in 2000 [11] (hereinafter, “the PDB debut paper”), which was highly cited, ranked 92 among the top 100 most-cited research of all time [71] with 12,754 citations. Alternatively, PDB users can cite one of the follow-up update papers of PDB published in the annual Database Special Issue of Nucleic Acids Research (NAR) from year 2002 to 2008 [74, 73, 14, 23, 43, 10, 35] and in other venues [9, 15, 32, 75, 63]. These publications describe the progress of continued enhancement and development of PDB. Citing journal publications represents a traditional method of data citation, with the benefit of being persistent and unambiguous. Alternatively, PDB users can cite PDB by mentioning URLs linking to the PDB home pages on the Web in the text, like “(<http://www.rcsb.org>).” URLs are unique but not persistent. Also, URL mentions are hardly recognized as academic accomplishment. In addition to URL mentions, data usage statistics, such as download counts, is proposed to be considered to measure the impact of

research works [58]. This chapter aims to answer the following questions:

- Does a new PDB publication by any of the wwPDB members attract more new citations and does a new PDB publication decrease the growth of citations and influence of its predecessors?
- Do PDB users refer to PDB URLs more often than citing PDB publications? How many use both? If we consider URLs and PDB publications as independent works, do URLs decrease the growth of citations and influence of PDB publications?
- How does data usage statistics correlate to paper citations and URL mentions?

Our main analysis tool is the citation cascade analysis. Citation cascades are chains of citations between two articles in a citation network. Citation cascades can be quantified by a function that considers both the length of the chain and the number of paths. Previously, we have shown that the growth of citation cascades correlate with the lasting influence of research articles better than citation counts [29], which usually favor an old paper because it takes long to accumulate citations for a new paper to be considered more influential than an old one. In contrast, disruption of citation cascades of an established paradigm can serve as an early indicator of paradigm shift [37].

One of the technical challenges is how to quantify and compare the influence of the PDB publications and URL mentions. Our approach to quantifying influence allows us to overcome this challenge by constructing citation cascades originated from papers with URL mentions. In this way, influence of URL mentions and PDB publications can be normalized and comparable, though cares must be taken in matching PMC full-text data, where URL mentions can be observed, with the PubMed citation network data set, where only abstracts are available.

3.2 Related Work

Data citation is receiving increasing attention in all disciplines of science as data become essential and ubiquitous in research. CODATA/ITSCI Task Force on Data Citation published a report on the current state of data citation in 2013 [69]. FORCE 11

(<http://www.force11.org>) has its final release of *Joint Declaration of Data Citation Principles* in 2014 [28], which identifies six principles as the guideline for the design of data citation standards and practices. A few studies have focused on automatically connecting the citation patterns that are resident in the literature data to the biomedical databases. BioLit [59] provided a comprehensive view on the literature data that links to biomedical databases by integrating the content of PubMed Central (PMC) with that of the PDB repository, based on the text-mining approach. Senay [40] characterized the patterns of how PDB entries are cited in research articles, based on analysis of the full text literature data available from Europe PubMed Central. Aurélie [53] developed a framework that improves links between literature data and various biomedical databases.

Much of bibliometric analysis uses traditional academic citations to measure a paper's quality or scientist's productivity [36]. Beyond simple citations counts, researchers have explored methods that analyze the structure of citation networks to identify important papers [19, 29] or predict which papers will be important in the future [62]. Moreover, Lovro implemented a network-based statistical comparison of the citation topology for analyzing the consistency of various bibliographic databases [64]. Our analysis method differs from related work in that we consider cascades, which take chains of citations, into account. It is well known that citation counts decay over time even for a highly influential work [3]. Therefore, it is important to consider its continuing influence of cascades, which provide indirect exposure to the work. Ghosh and Lerman [30] developed a function to quantify the structure of a growing cascade of information spreading in social media, which we use to measure the size of evolving cascades. We have developed a preliminary approach to quantifying transformative research with a disruption score that based on this model.

3.3 Materials and Methods

3.3.1 Paper Citation Data

The citation data used in our study were collected from MEDLINE \ PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed>) through the Entrez system and the XML for-

mat files from the NLM's FTP sever (http://www.nlm.nih.gov/bsd/licensee/access/medline_pubmed.html). Each record contains XML elements <CommentsCorrections>. The attribute RefType="Cites" of the element lists references or the bibliography of an article, from which we can obtain the citation information (see http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html). Our data set contains totally 22,732,343 articles and 102,783,011 pairs of cited-citing relation from PubMed, obtained in August 2015.

3.3.2 Mining URL Mentions

We extracted and counted articles containing mentions of PDB URLs from the full-text article data available from PMC. The data is available for download from (<http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/>), in either NXML markup language or plain text. We obtained 782,890 articles in NXML format as of October 2014, and 967,022 articles in plain text format as of February 2015. Removing duplicate PMC IDs yielded a total of 972,725 articles.

We extracted mentions of URLs linking to the home pages of the wwPDB partners, including RCSB PDB, PDBe (PDB Europe) and PDBj (PDB Japan), and wwPDB (world-wide PDB). Table 3.1 shows the patterns that we used to extract URL mentions from the text. URLs that link directly to a landing page of a protein structure in PDB are excluded. These can be recognized by certain suffix patterns in the URLs, as given in Table 3.1. Formal URL citations, that is, citing PDBs as a paper citation and listing a URL in the bibliography section, were not considered. URLs that are DOIs (digital object identifiers) (<http://www.doi.org>) [22] were not included here, either.

Table 3.1: Text patterns considered as PDB URLs.

PDB site	URL	Inclusion Prefix	Exclusion Suffix
RCSB PDB	http://rcsb.org , http://www.rcsb.org , http://www.pdb.org	"*rcsb.org", "*pdb.org"	"structureId=*"
wwPDB	http://www.wwpdb.org	"*www.wwpdb.org"	
PDBe	http://pdbe.org , http://www.ebi.ac.uk/pdbe http://www.ebi.ac.uk/msd	"*pdbe.org", "*www.ebi.ac.uk/pdbe", "*www.ebi.ac.uk/msd"	"/entry*"
PDBj	http://pdbj.org	"*pdbj.org"	"/mine*"

3.3.3 PDB Usage Statistics

The wwPDB provides monthly statistics of *FTP, Archive and Website Downloads*, and *Views* for each PDB structure from 2007 to present, available at (<http://www.wwpdb.org/stats/download.php>).



3.3.4 Calibrated Disruption Score

Previously, we have developed a method for quantifying the disruption of citation cascades of an established paradigm of scientific papers [37]. The disruption can be measured by comparing the growth of the average ϕ over time for all papers in the cascade and the papers in the complement of the cascade. C is the entire cascade rooted by the seed paper. Let $C^{(h)}$ denote the cascade originating from the challenger, h . We define the *residue cascade*, denoted by \tilde{C} , as the complement subgraph of C obtained by subtracting $C^{(h)}$ from C , *i.e.*,

$$\tilde{C} := C - (C \cap C^{(h)}) = C \setminus C^{(h)}. \quad (3.1)$$

By definition, references of papers in \tilde{C} can only be traced back to the seed papers, and note that it is not necessary for the challenger to be in C . Let t_0 be the publication time of the challenger paper, and C_t is the set of papers published at time t . Here, we suppose that we could investigate the complete citation network instead of the sampling network. The *calibrated disruption score* is defined as

$$\delta(\tau) := 1 - \frac{1}{\tau} \sum_{t=t_0+1}^{t_0+\tau} \frac{\sum_{j \in \tilde{C}_t} \phi(j)}{\sum_{j \in C_t} \phi(j)}. \quad (3.2)$$

The calibrated disruption score is a revision of the disruption score of Chapter 2 to normalize the range between 0 and 1 and ensure that scores of challengers published in different years are comparable when τ is set to the same value. Intuitively, a 5-year ($\tau = 5$) calibrated disruption score greater than 0.7 amounts to a large portion of the new influence of the seed paper is indeed due to the challenger, suggesting that its influence has been disrupted by the challenger..

3.4 Results and Discussion

3.4.1 Paper Citations

We start by investigating whether authors choose to cite new PDB follow-up update papers instead of the RCSB PDB debut paper. We consider only those published before 2008 so that for every paper we can observe the growth of its citation counts for at least five years (up to 2013). Moreover, uniprot is another good data warehouse for the comparison of PDB, and we will provide some studies of similarities and differences between these two resources. Fig. 3.1(A) shows that the annual citation counts of these PDB publications are much less than that of the highly cited PDB debut paper. The paper citation result seems to match the well-documented *Matthew effect* in science, which states that *the rich get richer and the poor get poorer* in terms of citations [49, 50]. The Fig. 3.1(B) show the result of UniProt that authors cite the latest core publications more often every year than core publications published in previous years. This is drastically different from PDB, for which authors prefer to cite the original debut paper. This may be because authors mainly follow the "how to cite" instructions given by the respective data repositories.

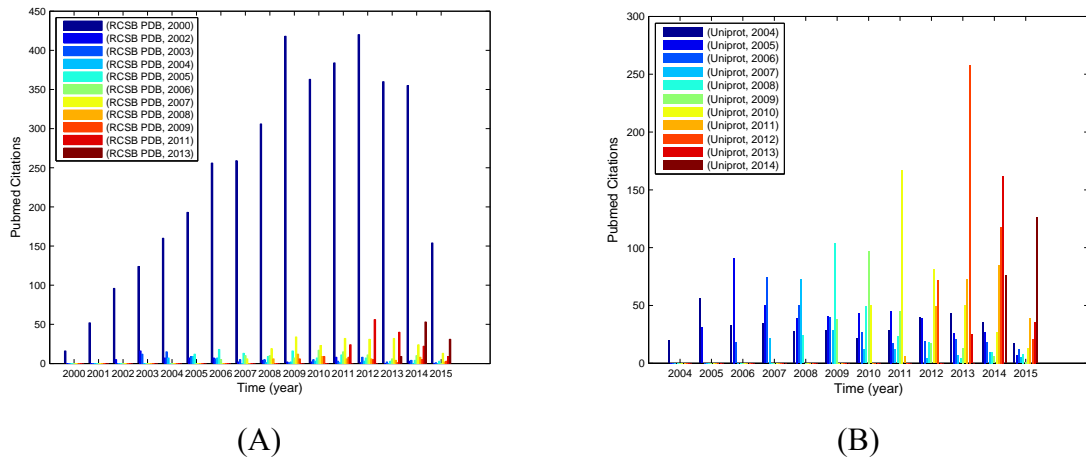


Figure 3.1: Citation growth of the (A) PDB and (B) UniProt debut article and their follow-up articles.

Though the citation counts of the follow-up update papers are not as large as the original debut paper, they may still disrupt the growth of the citation cascade of the PDB debut paper if they were cited by highly influential papers. To visualize if this is the case, we plot

two graphs similar to Fig. 2.1(B) to show the growth of the influence of the PDB debut paper and the growth of the residue cascades by the seven follow-up articles published in the Database Special Issue of NAR. Fig. 3.2(A) shows that the growth of the residue cascade curves are close to the curve of the PDB debut paper after 5 years ($\tau = 5$), suggesting that the follow-up articles hardly disrupt the growth of the cascade and thus the influence of the original PDB debut paper. Besides, we also check if the update papers disrupt the growth of the cascade of the earlier papers among the core publications of UniProt. Fig. 3.2(B) shows the curves of the cascade growth of NAR 2004 as the root and its follow-up papers as challengers, and reveals that the latest core publications are more influential than core publications published in previous years.

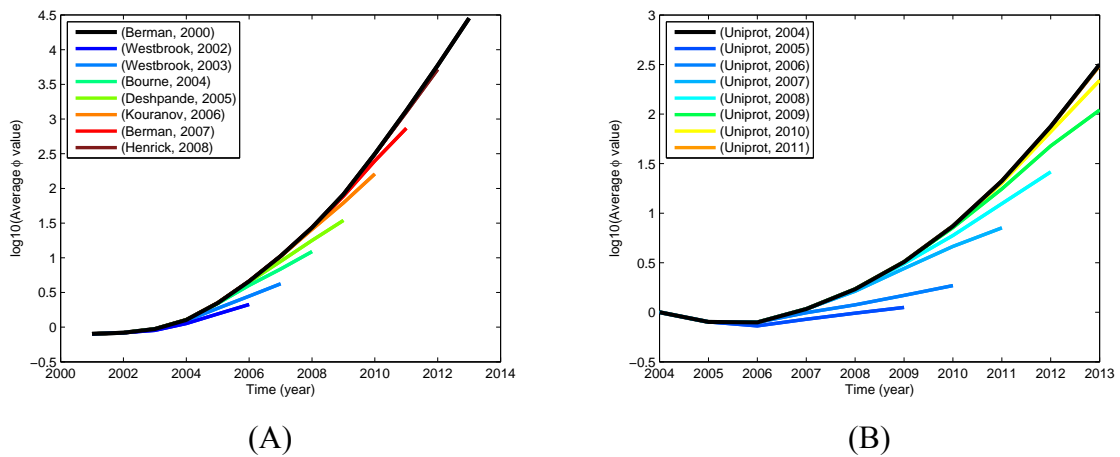


Figure 3.2: Compare the growth of the (A) PDB and (B) UniProt debut paper's cascade with all the residue cascades created by its follow-up articles in 5 years ($\tau = 5$). The y-axis of both panels shows the logarithm of the annual average cascade function values Φ , defined in Eq. 2.3.

Fig. 3.3 compares long-term disruptions of three follow-up articles published in the same year (2003). The figure shows that the growth of these residue cascades start to open large gaps from the black curve but these curves of the residue cascades fail to drop downward, suggesting limited disruption to the influence of the original debut paper. Table 3.2 shows the calibrated disruption scores of all PDB follow-up articles published between 2002 to 2008. The first seven articles are those published in the Database Special Issue of NAR.

The last column of Table 3.2 shows the average scores of five randomly selected arti-

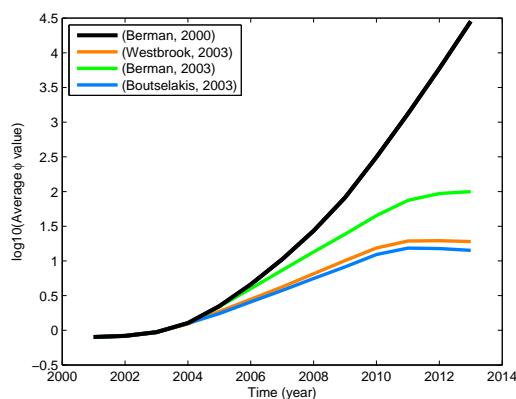


Figure 3.3: The residue cascades created by three 2003 follow-up articles.

Table 3.2: 5-year calibrated disruption scores of the PDB follow-up articles. The last column shows the average scores of randomly selected papers published in the same issue.

Author	Year	Title	Calibrated Disruption Score	Avg of Random 5
Westbrook	2002	The protein data bank: unifying the archive[74]	0.34	0.01
Westbrook	2003	The protein data bank and structural genomics[73]	0.35	0.00
Bourne	2004	The distribution and query systems of the RCSB Protein Data...[14]	0.32	0.00
Deshpande	2005	The RCSB Protein Data Bank: a redesigned query system and ...[23]	0.33	0.00
Kouranov	2006	The RCSB PDB information portal for structural genomics[43]	0.27	0.00
Berman	2007	The worldwide Protein Data Bank (wwPDB): ensuring a single...[10]	0.26	0.00
Henrick	2008	Remediation of the protein data bank archive[35]	0.10	0.00
Berman	2003	Announcing the worldwide Protein Data Bank.[9]	0.17	0.00
Boutselakis	2003	E-MSD: the European Bioinformatics Institute...[15]	0.39	0.00
Golovin	2004	E-MSD: an integrated data resource for...[32]	0.06	0.00
Westbrook	2005	PDBML: the representation of archival macromolecular...[75]	0.08	0.00
Standley	2008	Protein structure databases with new web...[63]	0.00	0.00

cles published in the same issue. The scores show that the follow-up articles still impact on the influence of the original debut papers much higher than other less related articles. We further compute the scores of the most highly cited articles in the Database Special Issues of NAR in each year and show the results in Table 3.3. Again, none of them score very high but three articles related to protein and thus PDB [12, 7, 4] score higher than 0.4, which is higher than the scores of any follow-up papers of PDB, suggesting that these articles impose influence disruption to the PDB debut paper more than the PDB follow-up papers.

3.4.2 URL Mentions

We investigate the trend that authors mention PDB URL(s) in the text as data citation practice. Fig. 3.4(A) shows that the annual citations to the PDB debut paper are higher than the annual counts of mentions of different PDB URLs. Note that since the annual counts

Table 3.3: 5-year calibrated disruption scores of the most highly cited articles in the database special issue of NAR.

Year	Title	Calibrated Disruption Score
2002	Gene Expression Omnibus: NCBI gene expression and hybridization ...[24]	0.28
2003	The SWISS-PROT protein knowledgebase and its supplement TrEMBL ...[12]	0.53
2004	The Pfam protein families database. [7]	0.50
2005	The Universal Protein Resource (UniProt). [4]	0.42
2006	miRBase: microRNA sequences, targets and gene nomenclature. [33]	0.29
2007	NCBI reference sequences (RefSeq): a curated non-redundant sequence...[60]	0.39
2008	The Pfam protein families database. [27]	0.30

were obtained from full-text articles in PubMed Central, we only counted the citations from papers in PubMed Central too for the PDB debut paper here so that the numbers are comparable. Though the annual counts of URL mentions are low, they grow as fast as the citations, which drop in 2013 while the counts of URL mentions continue growing. Fig. 3.4(B) shows that the sum of the annual counts of mentions grows steadily and in 2013 surpasses the citations to the PDB debut paper in that year. The figure also shows the annual counts of the papers that not only cite the PDB debut paper but also mention one of the PDB URLs. Nearly all authors who cited the PDB debut paper did not mention any PDB URL (94%), while authors who chose to directly mention the PDB URLs rarely cite the PDB debut paper (87%). In other words, authors chose to either cite the PDB debut paper or mention URL but rarely do both.

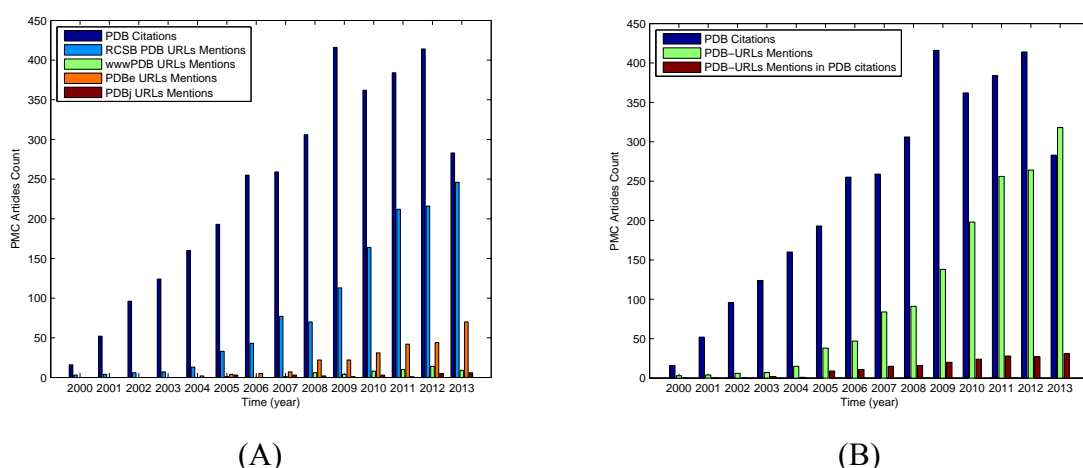


Figure 3.4: (A) Annual growth of the citations to the PDB debut paper and the counts of the different PDB URL mentions. (B) Annual growth of the citations to the PDB debut paper (blue bar), sum of all PDB URL mentions (green bar) and the count of the articles that not only directly cite the PDB debut paper but also mention PDB URLs (red bar).

We next consider mentioning of URL as a challenger and investigate whether it dis-

rupts the influence of the PDB debut paper. Here, the citation cascade of the URL mentioning is different from a paper citation cascade only in that its roots are those papers with PDB URL mentions. Then the cascade expands with papers citing these roots and papers citing those citing roots and so on to constitute the citation cascade. We also consider the seven PDB follow-up papers published in the Database Special Issue of NAR between 2002 to 2008 shown in Table 3.2 collectively as a challenger to compare their disruption impact with the URL mentioning.

Fig. 3.5 plots the growth of the cascades of the PDB debut paper, NAR follow-up papers, and URL mentioning, as well as the growth of the residue cascades by the follow-up NAR papers and URL mentioning. Again, the wider the gap between the curve for the PDB debut paper and the curve of a residue cascade, the higher the disruption of the influence. The figure shows that the gap of the residue cascade of the NAR follow-up papers is also taller than that of the URL mentioning, suggesting that the NAR follow-up papers collectively pose a higher disruption impact to the PDB debut paper than the URL mentioning, though individually, their impact is not apparent. Meanwhile, the growth curve of the NAR follow-up papers rises faster than the curve of the URL mentioning, but the latter is catching up rapidly after 2010.

3.4.3 Data Usage Statistics

Various data usage statistics may provide alternatives to citation counts as metrics of impact of a data repository. Yet it is not clear whether these statistics and citation counts are correlated or not. Fig. 3.6 shows that the annual counts of PDB FTP archive access and the citations to the PDB debut paper appear highly correlated before 2013, when the count of citations to the PDB debut paper drops, while the counts of PDB website downloads and views and the counts of the PDB URL mentions appear highly correlated as they grow at a similar rate. Other pairs appear uncorrelated.

We fit linear models to confirm and quantify the observed correlations. Table 3.4 shows the results of pairing data citations (including both paper citations and URL mentions) and data usage statistics (including both website and FTP access) as either dependent

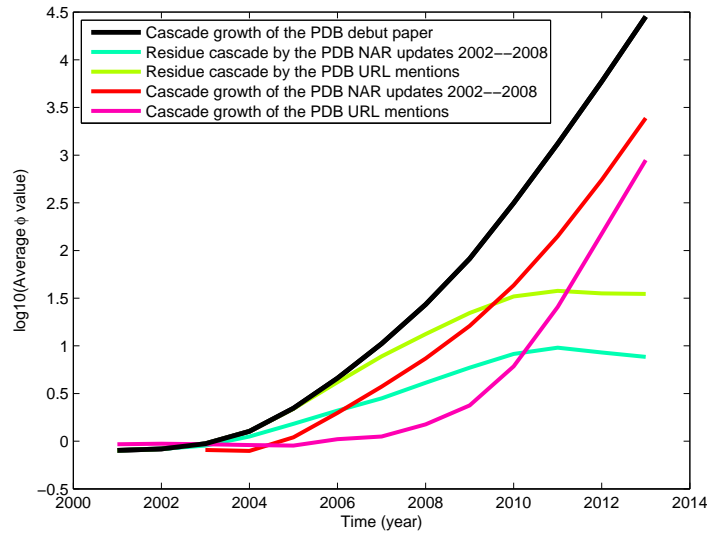


Figure 3.5: Growth of the cascade of the PDB debut article (black curve), the collection of PDB NAR update articles from 2002 to 2008, the PDB URL mentions articles, and their corresponding residue cascades. Notice the split between the black curve and green curve, indicating the cascade disruption.

variable or independent variable with different time frames. For example, row No. 19 in the table shows the result of fitting the linear model:

$$c(t) + c(t + 1) = w \cdot (u(t - 1) + u(t)) + \beta,$$

where $c(t) + c(t + 1)$ is the sum of the counts of data citations by PDB URL mentions of the current and next year and serves as the dependent variable in the model, $u(t - 1) + u(t)$ is the sum of the access counts of the website downloads and views of the previous year and this year and serves as the independent variable to predict the dependent variable, and w and β are the model parameters that we fit from the data. We quantify the fitness of all results with the R^2 value. The results show that regardless of the settings the PDB URL mentions and the website downloads and views are highly correlated with $R^2 > 0.9$ (in bold fonts). The best fit was found between the two-year sum of the counts of the website downloads and views and the URL mentions (row No. 11). Fig. 3.7 shows the fit of these four cases.

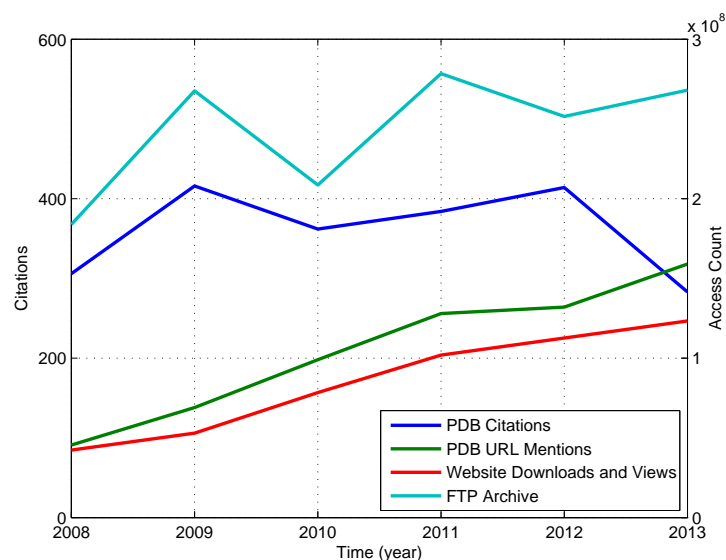


Figure 3.6: The growth of citations of the PDB debut paper, PDB URL mentions, website downloads and views, and FTP archive access from 2008 to 2013. This analysis only considers citations and mentions available from the PubMedCentral archive.

3.5 Summary

In this study, we compare data citations to a data repository by citing original and follow-up publications and URL mentioning by applying an approach using disruptions of citation cascades and correlate data citations with data usage statistics for PDB, one of the most widely used biomedical data repositories. Our findings include that

1. Authors still prefer citing the original PDB debut paper to citing follow-up papers.
2. The number of authors citing PDB by URL mentioning is growing rapidly.
3. The impact of PDB URL mentioning, however, is still lower than that of PDB follow-up papers collectively.
4. PDB website access statistics and URL mentions are highly correlated.
5. Correlations between PDB data usage statistics and PDB paper citations are not as high, though PDB FTP access seems to correlate with paper citations in early years.

These trends may be in part the result of the citation policy of the RCSB PDB, which recommends the original PDB debut paper and the URL <http://www.rcsb.org> as the data

Table 3.4: The correlations between PDB data citations and PDB data usage statistics by linear modeling.

No.	Dependent Variable (y)	Time Frame	Independent Variable (x)	Time Frame	R^2
1	Website Downloads and Views	$u(t)$	PDB Citations	$c(t)$	0.01
2	FTP Archive	$u(t)$	PDB Citations	$c(t)$	0.13
3	Website Downloads and Views	$u(t)$	PDB URL Mentions	$c(t)$	0.98
4	FTP Archive	$u(t)$	PDB URL Mentions	$c(t)$	0.41
5	Website Downloads and Views	$u(t)$	PDB Citations	$c(t-1)$	0.67
6	FTP Archive	$u(t)$	PDB Citations	$c(t-1)$	0.13
7	Website Downloads and Views	$u(t)$	PDB URL Mentions	$c(t-1)$	0.97
8	FTP Archive	$u(t)$	PDB URL Mentions	$c(t-1)$	0.30
9	Website Downloads and Views	$u(t) + u(t+1)$	PDB Citations	$c(t) + c(t+1)$	0.49
10	FTP Archive	$u(t) + u(t+1)$	PDB Citations	$c(t) + c(t+1)$	0.71
11	Website Downloads and Views	$u(t) + u(t+1)$	PDB URL Mentions	$c(t) + c(t+1)$	0.99
12	FTP Archive	$u(t) + u(t+1)$	PDB URL Mentions	$c(t) + c(t+1)$	0.88
13	PDB Citations	$c(t)$	Website Downloads and Views	$u(t-1)$	0.26
14	PDB Citations	$c(t)$	FTP Archive	$u(t-1)$	0.08
15	PDB URL Mentions	$c(t)$	Website Downloads and Views	$u(t-1)$	0.91
16	PDB URL Mentions	$c(t)$	FTP Archive	$u(t-1)$	0.28
17	PDB Citations	$c(t) + c(t+1)$	Website Downloads and Views	$u(t-1) + u(t)$	0.26
18	PDB Citations	$c(t) + c(t+1)$	FTP Archive	$u(t-1) + u(t)$	0.55
19	PDB URL Mentions	$c(t) + c(t+1)$	Website Downloads and Views	$u(t-1) + u(t)$	0.96
20	PDB URL Mentions	$c(t) + c(t+1)$	FTP Archive	$u(t-1) + u(t)$	0.89

resource reference. Since the citation network could be pretty large and could be obtained from different data source, the major technical challenge is to collect a complete set of citation network. Also it can be challenging to integrate the Pubmed citation data with the PMC full-text data for comparing the citing or mention behaviors of PDB users. The analysis of citation trends of other biological data resources with different citation policies will be analyzed in the future to explore this effect and to develop recommendations for data citation practices.

Our analysis methodology is applicable to analyzing citations of Web servers as long as a web server has primary publications that can be used as the root nodes of citation cascades and maintains Web access logs to correlate with citation counts and/or URL mentions.

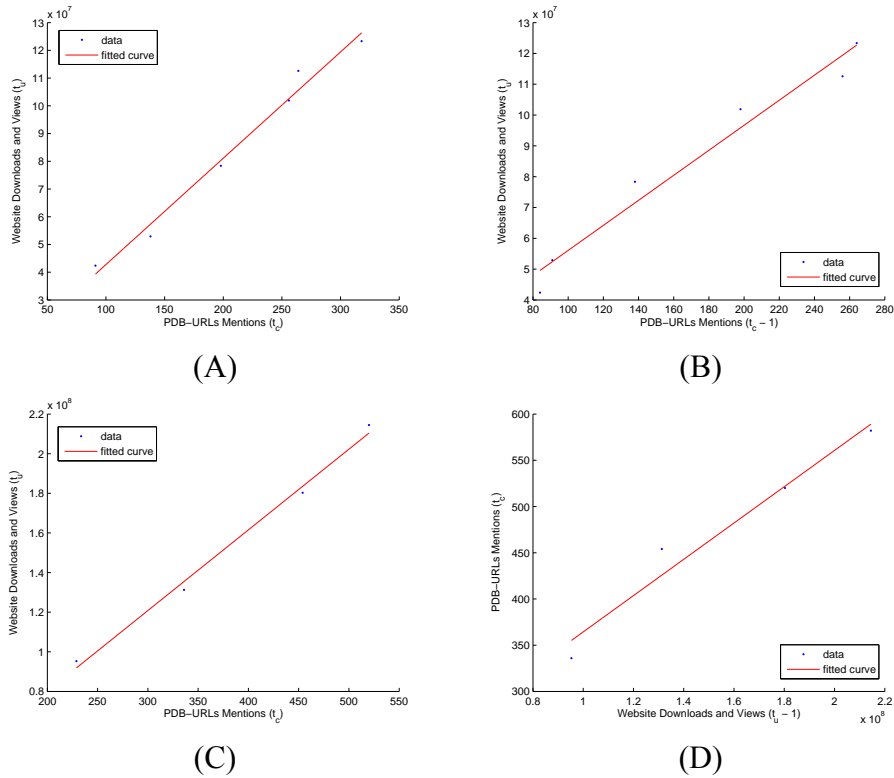


Figure 3.7: The plots of the fitting of linear models between the PDB URL mentions c and the website downloads and views u , referred to by their case No.'s in Table 3.4: (A) Case No. 3, $y = u(t)$, and $x = c(t)$, (B) Case No. 7, $y = u(t)$, and $x = c(t - 1)$, (C) Case No. 11, $y = u(t) + u(t + 1)$, and $x = c(t) + c(t + 1)$, (D) Case No. 19, $y = c(t) + c(t + 1)$, and $x = u(t - 1) + u(t)$.



Chapter 4

Data Citation to the Protein Data Bank

4.1 Introduction

In this chapter, we focus on analyzing the various of data citation to the temporal patterns. An appropriate data citation will benefit the data reused, experiments reproduced, and even provide machine readability for tracing the data usage. Temporal patterns could be considered as the simply annual growth of data citation, or concerned with the changes in occurrence frequency over time of keywords in research articles. We will apply this methodology to study the temporal patterns of PDB data, it will help us to know the trends of protein structure researches.

The major data of The Protein Data Bank (PDB) [11, 74, 73, 14, 23, 43, 10, 35] are the experimentally determined structures of protein. The PDB provides unique identifiers (PDB IDs) and digital object identifiers (DOIs) that make the data are accessible and persistence for researchers to use it as the referenced data. For a PDB entry, the primary citation papers is the study of crystallography process for a specific protein, and the primary citation should be declared when it was deposited to repository that have it be seen as legitimate, citable products of research. Hence, the data are easily to be given scholarly credit to all contributors to the data. All the characteristic make the PDB data be a good practice model to help us study the behaviors that how the protein structure data being used by the researchers. There are two major ways to cite these data items: citing the primary citation paper (citation), or mention the PDB ID (mention). Although the DOI or URL of

data are trackable, the usage of other data citation practices include URL mentions, DOI mentions is still low, so we focus on the two major ways. We believe that if users could mention the PDB IDs or citing the primary citation papers in the article, which can be great benefit to both sides of data provider and repository developer.

Another aspect is to consider the co-cited relationships between articles. Co-citation links two articles that are cited together by another article. To study the co-citation network may help address a problem of citation counts, which usually take too long to accumulate for a new paper to be considered more influential than an old one. Similar to the co-citation, we also try to investigate another pattern that is the co-mention. Co-mention is defined as the frequency with which two PDB IDs are mentioned together in a research article. The higher co-citations or co-mentions two articles receive could assert that the more likely they are related. Analysis of the co-citation and co-mention patterns will not only highlight behaviors of how the PDB data being used, but also help to establish the quantitative methods for measuring the similarity of two PDB entries. We want to compare them so that we can see how citing primary citation and PDB ID mention lead to different results as a metric of influence.

In the previous work, we have studied PDB as a data repository [38], now we want to study its' data items, which are the protein structure data. We aims to answer the following questions:

- Do PDB users mention to PDB IDs in their paper more often than citing the primary citation papers? How many use both?
- How is the PDB entry statistically dependent to the data citation frequency?
- For each PDB entry, how does their citation count statistically correlate to mention count?
- What are the co-citation and co-mention patterns of the PDB entries? Are these two kinds of patterns consistent to each other?
- If the authors clearly cite data sources will also help improve impact of their own papers?

In practice, we organize two of the questions as the null hypothesis statements involves the variables, including the PDB IDs, citation count and mention count. And we apply statistical hypothesis test to verify the difference of these variables. Specifically, we consider the null hypothesis that the PDB IDs are independent of the corresponding data citation frequencies, and the other is that citation frequencies and the mention frequencies of the PDB entry are linearly independent. We then perform the G-test of independence and Pearson correlation test to verify these statements. Moreover, we illustrate the analysis of co-citation and co-mention patterns from the view of data citation network. We also try to identify the influential studies of protein structure by previously proposed model, the calibrated disruption score [38, 37].

4.2 Materials and Methods

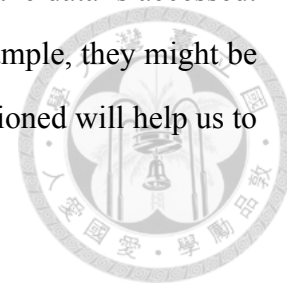
4.2.1 Citation data

The citation data used in our study were collected from MEDLINE \ PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed>) through the XML format files from the NLM's FTP server (http://www.nlm.nih.gov/bsd/licensee/access/medline_pubmed.html), given in the XML structure of `<CommentsCorrections RefType="Cites">`. Our data set contains totally 22,732,343 articles and 102,783,011 pairs of cited-citing relation from PubMed, obtained in August 2015.

4.2.2 Mention data

We extracted and counted articles containing mentions of PDB ID from the full-text article data available from PubMed Central (PMC). The data is available for download from (<http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/>), in either NXML markup language or plain text. We obtained 1,015,179 articles in NXML format, and 1,093,980 articles in plain text format as of August 2015. Removing duplicate PMC IDs yielded a total of 1,015,233 articles. Implementing the full-text mining will help to the statistic of the PDB IDs mentioned in the research articles, and it will also reflect the truly data usage. wwwPDB

provide data download statistics, which directly measure how often the data is accessed. However, some downloads are not reflected to actual usage. For example, they might be performed by mirroring software. The statistic of the PDB IDs mentioned will help us to distinguish "legitimate" data usage from downloads count.



4.2.3 Mentions of issued PDB IDs

Each PDB entry has an unique identification code, and these codes are recorded as 4 characters in length. The first character is a numeral in the range 1-9, while the other three characters can be mixed with either numerals or letters. Table 4.1 shows all the issued PDB IDs presented in full-text format articles. In free text, the PDB IDs sometimes will be confused with other abbreviations in the text mining process, i.e., false positives of PDB IDs. For example, the PDB ID "3AUT" will be confused with the the abbreviation of postal code "385 Euston Road, London, NW1 3AUT, UK", the PDB ID "2NO3" will be confused with the the abbreviation of chemical formulas " $Zn(H_2O)_2(C_5H_5N_3O_2)_2$ 2NO3 ..", and the PDB ID "3DEE" will be confused with the the abbreviation of software "domain definitions from SCOP, CATH, DALI, 3DEE, and MMDB are ..". In order to solve this problem, we develop a machine learning based approach for recognizing the PDB IDs mentioned in the research articles that incorporated with the prefix information to minimize ambiguities, and it greatly decreasing the false positive rate of identifier.

Table 4.1: Mentions of Issued PDB IDs.

Identifier	Example	Machine Readable	Mentions	%
PDB ID	PDB ID: 1STP	Y	14,888	4.8
PDB DOI	http://dx.doi.org/10.2210/pdb1stp/pdb	Y	155	0.05
External Link Tag	<ext-link .. ext-link-type="pdb" xlink:href="1STP">	Y	32,108	10
PDB File Name	1stp.pdb	Y	895	0.03
PDB URL	http://www.rcsb.org/././structureId=1stp	Y, but URL may change	657	0.2
Non-standard PDB ID	PDB code: 1STP , PDB reference 1STP , PDB accession number 1STP , Many variations..	Y/N	22,081	7.1
PDB in Context	We employed the following PDB coordinates: glycogen phosphorylase , 1gpy ..	Y/N with NLP or ML	16,726	5.4
Free Text	We first placed S2 bound to human PI3KC; (3ene) into the reference coordinates..	Y/N with NLP or ML	221,287	72

4.2.4 G-test of Independence

P is the set of all the entries deposited to PDB. For a PDB entry, $p \in P$, the citation of p , $cite(p)$ is the set of articles that citing on the primary citation of p . The mention of p , $ment(p)$ is the set of articles that have mentioned the ID of p anywhere in the text area. G-test is a good method to see whether the observations of distribution fits to a theoretical expectation. The null hypothesis is that the PDB IDs are independent of the corresponding data citation frequencies. The observed matrix includes the data citation frequency of each PDB entry. We let each row cells of the observed matrix be the pair of values, $(|cite(p)|, |ment(p)|)$. The expected matrix is constructed by random sampling from a distribution with the given expected frequencies. We then want compute each row cells of expected matrix, denoted as the pair of values, $(excite(p), exment(p))$, as the following values,

$$excite(p) = \left(\sum_{p \in P} |cite(p)| + |ment(p)| \right) \times \frac{|cite(p)|}{|cite(p)| + |ment(p)|} \times \frac{|cite(p)|}{\sum_{p \in P} |cite(p)|}, \quad (4.1)$$

$$exment(p) = \left(\sum_{p \in P} |cite(p)| + |ment(p)| \right) \times \frac{|ment(p)|}{|cite(p)| + |ment(p)|} \times \frac{|ment(p)|}{\sum_{p \in P} |ment(p)|}, \quad (4.2)$$

which are the expected citation and expected mention of p . We define the G-test statistic, G , as the following value,

$$G = 2 \sum_{p \in P} \left(|cite(p)| \ln \frac{|cite(p)|}{excite(p)} + |ment(p)| \ln \frac{|ment(p)|}{exment(p)} \right). \quad (4.3)$$

We use the degrees of freedom, which is the size of total PDB IDs to decide the G-test distribution function, and we can use the function to calculate the p-value by the given G .

4.2.5 Pearson Correlation Coefficient

Pearson correlation coefficient is used to quantify the dependence of two variables. We use the Pearson correlation coefficient for calculating the dependence between the distributions of citation and mention for the whole data. The null hypothesis is that the citation

frequencies and the mention frequencies of each p are linearly independent. We let $mcite$ be the value, $\frac{1}{|P|} \sum_{p \in P} |cite(p)|$ and $mment$ be the value, $\frac{1}{|P|} \sum_{p \in P} |ment(p)|$. We define the Pearson correlation coefficient statistic as the following value,

$$\frac{\sum_{p \in P} (|cite(p)| - mcite)(|ment(p)| - mment)}{\sqrt{\sum_{p \in P} (|cite(p)| - mcite)^2} \sqrt{\sum_{p \in P} (|ment(p)| - mment)^2}} \quad (4.4)$$

The p-value here is the probability that the correlation coefficient between citation frequencies and the mention frequencies were in zero, which is the null hypothesis. We can compute the p-value by the Pearson's correlation coefficient distribution function.

4.2.6 Co-citations/mentions between PDB Entries

The co-citation, **co-cite**($p1, p2$) of two PDB entries $p1$ and $p2$ is the set of papers that both citing on the primary citations of $p1$ and $p2$, it can be defined as,

$$\mathbf{co-cite}(p1, p2) = cite(p1) \cap cite(p2), \quad (4.5)$$

and we call $|\mathbf{co-cite}(p1, p2)|$ as the **co-citation degree** of $p1$ and $p2$. Similarly, the co-mention papers, **co-mention set** of two PDB entries $p1$ and $p2$ can be also defined as,

$$\mathbf{co-ment}(p1, p2) = ment(p1) \cap ment(p2), \quad (4.6)$$

and the size of **co-ment**($p1, p2$) is the **co-mention degree** of $p1$ and $p2$.

4.2.7 Jaccard Index

The **Jaccard Index** is used to quantify the difference of the identified neighborhoods for a PDB entry, $p \in P$, according to the corresponding co-citation or co-mention sets. The top k ranking neighborhoods of p ordered by the co-citation degree is defined as the set, $nc_k(p)$, and the top k ranking neighborhoods of p ordered by the co-mention degree is

defined as the set, $nm_k(p)$, then

$$intersection_k = nc_k(p) \cap nm_k(p), \quad (4.7)$$

and

$$union_k = nc_k(p) \cup nm_k(p), \quad (4.8)$$

and then we define the *Jaccard Index* $_k$ as,

$$Jaccard\ Index_k = \sum_{i \in k} \frac{intersection_i}{union_i}. \quad (4.9)$$

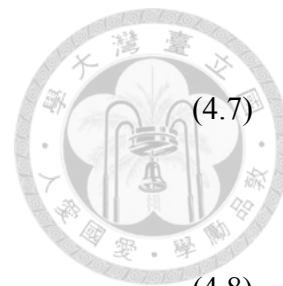
4.3 Results and Discussion

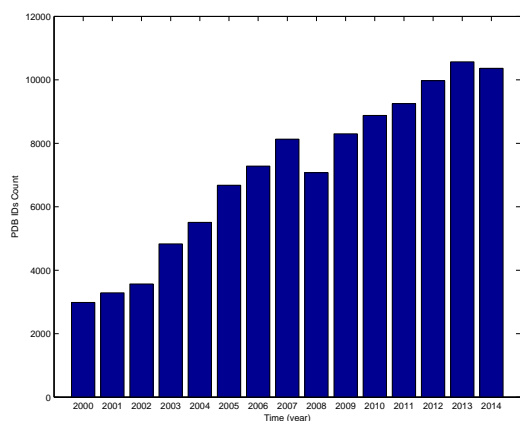
4.3.1 User Tendency to the PDB Data Citation

We start by investigating the tendency that the authors tend to cite primary citation papers or mention the PDB IDs in the text as data citation practice. Fig. 4.1(A) shows that the annual growth to the count of entries depositing to the PDB repository. The number grows very fast, and there are totally 110,790 entries as of 4 August 2015. Most of them are crystalized by the X-ray diffraction. Fig. 4.1(B) shows that the annual growth of the citation to the primary citation papers and the mention frequency of PDB data. Note that since the annual counts were obtained from full-text articles in PubMed Central, we only counted the citations from the papers in PubMed Central to make the numbers are comparable. The result of Fig. 4.1(B) shows that the growth rate of the annual counts of data citation is higher than the growth rate of depositing PDB entries. Most of the authors tend to directly cite on the primary citation papers of the used PDB data instead of mention the PDB IDs in the papers.

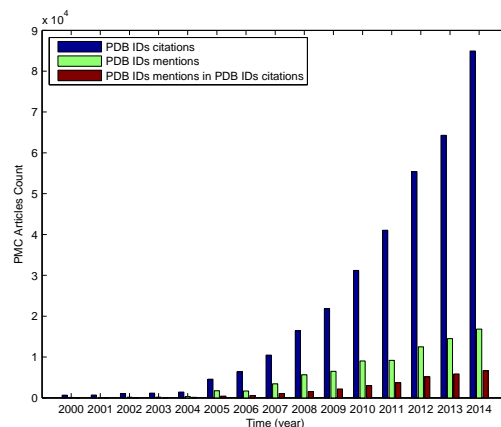
4.3.2 Trends of Protein Structure Researches

We then investigate which kinds of PDB data are the most cited and mentioned on protein structure researches. We list the top ten PDB entries which are sorted in the order of data





(A)



(B)

Figure 4.1: (A) Growth of the depositions of new PDB entries. (B) Annual growth of the citations to the PDB entries' primary citation papers (blue bar), sum of all the PDB IDs' mention (green bar) and the count of the articles that not only directly cite the PDB entries' primary citation papers but also mention the PDB IDs (red bar).

citation and mention frequency, shown as the Table 4.2 and Table 4.3. These PDB entries are also annotated with their properties, the category of the protein structure and the source of organism from which the protein structure is crystallized.

Table 4.2: Top 10 PDB Entries. (Sorted by citation frequency)

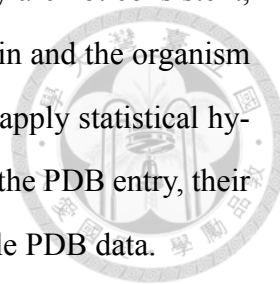
PDB ID	Year	Citations	Citation Rank	Mentions	Mention Rank	Category	Source
1AOI	1997	1527	1	31	37	DNA Binding Protein	XENOPUS LAEVIS
1BL8	1998	1234	2	35	24	Membrane Protein	STREPTOMYCES LIVIDANS
1F88	2000	957	3	44	16	Signaling Protein	BOS TAURUS
1GC1	1998	852	4	26	57	Viral Protein; Receptor; Immune System	HOMO SAPIENS; HIV 1
1RV1	2004	747	5	11	488	Ligase	HOMO SAPIENS
1FFK	2000	746	6	31	34	Ribosome	HALOARCUA MARISMORTUI
2RH1	2007	682	7	124	1	Membrane Protein	HOMO SAPIENS
1YSG	2005	650	8	6	1984	Apoptosis	HOMO SAPIENS
2A79	2005	635	9	49	10	Membrane Protein	RATTUS NORVEGICUS
1AIK	1997	561	10	12	403	Viral Protein	HIV-1 M:B HXB2R

Table 4.3: Top 10 PDB Entries. (Sorted by mention frequency)

PDB ID	Year	Mentions	Mention Rank	Citations	Citation Rank	Category	Source
2RH1	2007	124	1	682	7	Membrane Protein	HOMO SAPIENS
1UBQ	1987	96	2	222	142	Chromosomal Protein	HOMO SAPIENS
1KX5	2002	69	3	272	87	Structural Protein	HOMO SAPIENS; XENOPUS LAEVIS
2R9R	2007	65	4	433	20	Membrane Protein	RATTUS NORVEGICUS
3EML	2008	65	5	408	24	Membrane Protein; Receptor	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
1U19	2004	64	6	227	134	Signaling Protein	BOS TAURUS
1K4C	2001	59	7	454	18	Membrane Protein	STREPTOMYCES LIVIDANS; MUS MUSCULUS
2VT4	2008	55	8	356	38	Receptor	MELEAGRIS GALLOPAVO
2B4C	2005	55	9	289	71	Viral Protein	HOMO SAPIENS; SYNTHETIC CONSTRUCT; HIV 1
2A79	2005	49	10	635	9	Membrane Protein	RATTUS NORVEGICUS

From the results of Table 4.2 and the Table 4.3, we could find that the selected top

ten PDB entries ordered by citation frequency and mention frequency are not consistent, although most of them are belong to the category of membrane protein and the organism source of them are crystallized from homo sapiens. We also want to apply statistical hypothesis test to verify the difference between the variables, including the PDB entry, their corresponding citation frequency and mention frequency for the whole PDB data.



4.3.3 Statistic Test to the Data Citation

We use the G-test of independence to verify the hypothesis that whether the observations of PDB IDs is dependent of the distribution of data citation. Besides, the Pearson correlation coefficient is used for calculating the linear dependence between the citations count and mentions count for the whole PDB entries, and try to test the hypothesis that distribution of citation frequency are dependent of the mention frequency. We observe the p-value of G-test of independence and Pearson correlation coefficient depend on selecting from the front of top highly cited PDB entries to the whole PDB data, shown as Fig. 4.2(A) and (B).

Fig. 4.2(A) shows that the p-value of G-test of independence drops to $4.28e-8$ when the selected k equals to 4, and we get the p-value closed to zero for the whole PDB data. It indicates that the PDB IDs are dependent of the corresponding data citation frequencies. Additionally, Fig. 4.2(B) shows that the p-value of Pearson correlation coefficient grows in oscillation amplitude, but suddenly drops to close to zero when the selected k reaches to 135, revealing that the probability of citation frequencies and the mention frequencies were linearly independent is very low. Fig. 4.2(C) shows that growth of Pearson correlation coefficient. The change of coefficient value is very slow, and increased to 0.48 for the whole PDB data, suggesting that the citation frequencies and the mention frequencies are in moderate linear relationship. We also use the Q-Q plot, which is a probability plot used for comparing two distribution variables by plotting their quantiles against each other, shown as Fig. 4.2(D). We could find that the distribution of data citation and mention frequency are almost in the linear growth. It indicates that the probability distribution of these two variables are similar to each other.

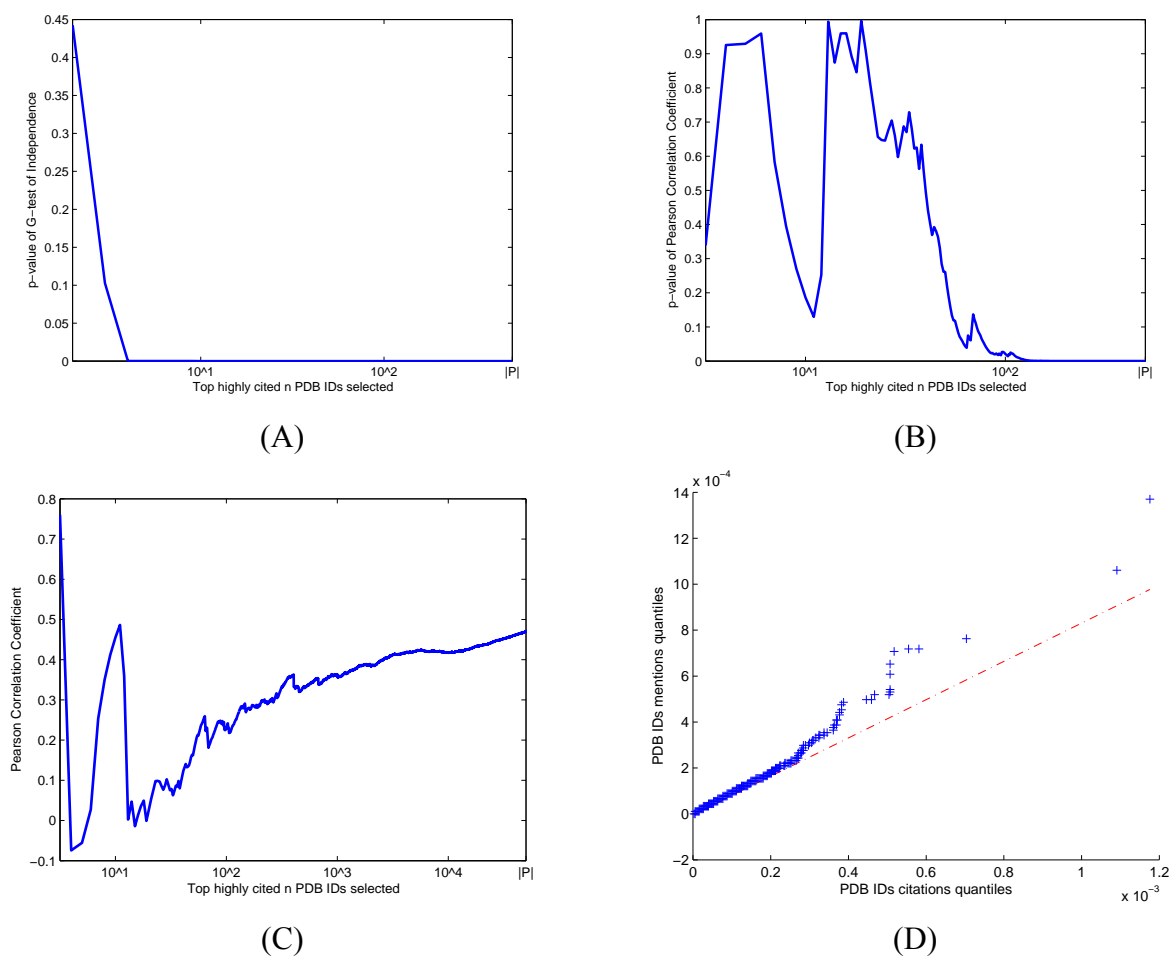


Figure 4.2: (A) P-value of G-test of independence. (B) P-value of Pearson correlation coefficient. (C) The growth of Pearson correlation coefficient. (D) Q-Q plot between the distributions of citation and mention.

4.3.4 Analysis of the Co-citation/mention Patterns

Analysis of the co-citation and co-mention patterns will reveal that how the PDB data being used. We try to use the **Jaccard index** to quantify the difference of the identified neighborhoods for a PDB entry, p , according to the corresponding co-citation or co-mention degree. We only consider those PDB entries possess both the co-cited and co-mentioned neighborhoods. The PDB entries will be sorted in their corresponding Jaccard index that compares the top 3 co-cited or co-mentioned ranked selected neighborhoods, shown as Fig. 4.3(A). The Jaccard index drop a little bit faster than linear. We normalize the area of Fig. 4.3(A) as 1, and calculate ratio of the area under the curve, and get the value is 0.1123, suggesting that it is inconsistent between co-citation and co-mention neighbors for most PDB IDs.

From the observation of the annual growth of the citation and mention to the PDB entries on Fig. 4.1(B), it may suggest that the deposited time is correlated the consistency between co-citation and co-mention neighbors. Fig. 4.3(B) shows the average **Jaccard index** ordered by the deposited time, and support our hypothesis.

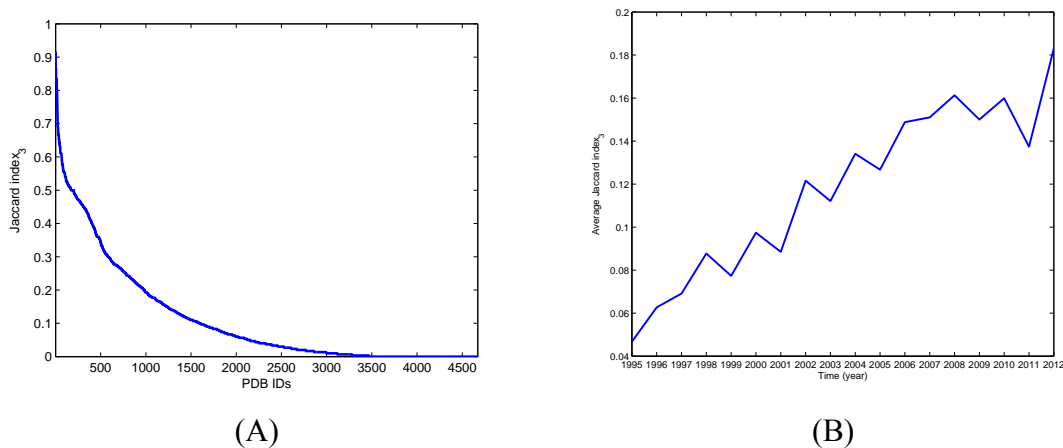


Figure 4.3: (A) Distribution of $Jaccard\ index_3$ of PDB IDs. (B) The average $Jaccard\ index_3$ ordered by the deposited time.

We then try to analysis of the co-citation/mention patterns to PDB categories. We denote the set of PDB categories as Cat , and a PDB entry p will be classified into at least one category, $cat \in Cat$. The top highly cited category of PDB data is selected according to the order of average citations, which can be calculated by,

$$AverageCitation(cat) = \frac{1}{|p \in cat|} \sum_{p \in cat} \|cite(p)\|, \quad (4.10)$$

Table 4.4 shows the selected five category of protein structure. The **co-citation set** and the **co-mention set** of two categories $cat1$ and $cat2$ are defined as, $\bigcup_{p1 \in cat1, p2 \in cat2} \mathbf{co-cite}(p1, p2)$, and $\bigcup_{p1 \in cat1, p2 \in cat2} \mathbf{co-ment}(p1, p2)$. The **co-citation degree** and **co-mention degree** are the length of co-citation set and co-mention set separately. Fig. 4.4(A) and (B) shows the results of co-citation degree and co-mention degree among the selected five categories. The value of heatmap is defined as the normalized co-citation or co-mention degree between pairs of categories. Analysis of the co-citation and co-mention patterns will help to highlight behaviors of how the PDB data being used across the different categories of protein structure. We could find the receptor, membrane protein and viral protein are highly

co-cited and co-mentioned to each other. The results of these two figure are very similar. However the scale of the co-citation degree is large than the co-mention degree. In the Table 4.4, the order of average citation consistent with the order of the average mention.

Table 4.4: Highly cited category of PDB data.

Category	PDB IDs count	Total citations	Average citation	Total mentions	Average mention
Receptor	102	5115	50.15	403	3.95
Ribosome	142	6825	48.06	364	2.56
Membrane Protein	640	21475	33.55	1467	2.29
Gene Regulation	208	6752	32.46	462	2.22
Viral Protein	849	27452	32.33	1530	1.80

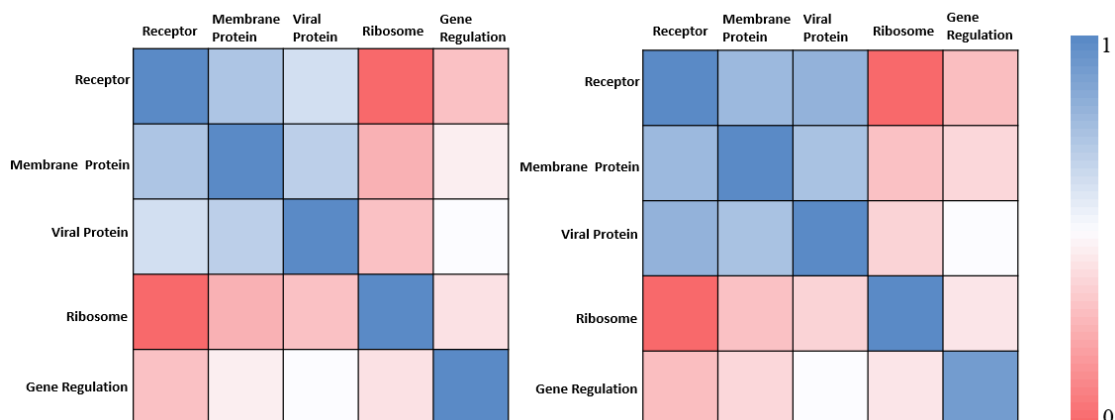


Figure 4.4: Heatmap of (A) co-citation degree between top cited categories of PDB IDs.(B) co-mention degree between top cited categories of PDB IDs.

4.3.5 Identification of the Influential PDB Entries

Based on the co-citation and co-mention metrics, we could list the co-cited and co-mentioned studies of protein structure for the entries of PDB repository. A citation cascade is constructed by a series of citations between two articles. In Previous chapter, we have developed a method that can be used to quantify the disruption of citation cascades of an established paradigm, and it can serve as an early indicator of paradigm shift [38, 37]. Intuitively, a 5-year ($\tau = 5$) calibrated disruption score greater than 0.7 amounts to a large portion of the new influence of the seed paper is indeed due to the challenger, suggesting that its influence has been disrupted by the challenger. In this section, it can be implemented on all the pairs of PDB entries' primary citation papers to identify related influential study of PDB entries. We take two of the highly cited PDB entries, **1AIK** [16]

and **1F88** [57] from the two major PDB categories, membrane protein and viral protein for examples. Glycoprotein 41 (gp41) is a well-known subunit of the envelope protein complex of retroviruses, and the primary citation paper of **1AIK** is the pioneer protein structure study of gp41. In order to identify the influential succeeding PDB entries, Table 4.5 and Table 4.6 list all the related PDB entries as the co-cited and co-mentioned neighborhoods for **1AIK**. We could find all of them are related to the viral protein, immune system or inhibitor and most of them are crystalized from the HIV-1. Based on the results of calibrated disruption score, PDB entry **1ENV** [72] is identified as the influential one to **1AIK**. The study of **1ENV** provide a X-ray crystallography to determine the structure of gp41 ectodomain. On the other hand, the primary citation paper of **1F88** is an important study of crystal structure of G protein-coupled receptor. Table 4.7 and Table 4.8 list all the related PDB entries as the co-cited and co-mentioned neighborhoods for **1F88**. Most of the identified co-cited and co-mentioned neighborhoods are the same, and are belong to signaling or membrane protein. However, the source organism from where they are crystalized are diverse. Based on the results of calibrated disruption score, **1L9H** [56] is identified as the influential one to **1F88**, it is the same authors' consequent study to the primary citation paper of **1F88**. PDB entry **2RH1** [20] is identified as the second influential one. It is consistent to the development of G protein-coupled receptor researches, the work of **2R4R**, **2R4S** [25] is the first time to successful crystallized the G protein-coupled receptor structure from homo sapiens, however, there was immediately another higher resolution of same crystal structure, **2RH1** that delivered by the same authors.

Table 4.5: Co-cited neighbor entries of the PDB entry-*1AIK*

PDB ID	Year	co-citation degree	Calibrated disruption score	Category	Source
1ENV	1997	412	0.74	Viral Protein	SACCHAROMYCES CEREVISIAE
1SZT	1997	178	0.43	Viral Protein	HIV 1
1GC1	1998	163	0.48	Viral Protein	HOMO SAPIENS; HIV 1
1HTM	1994	109	0.29	Viral Protein	UBPU-608
2EZO,2EZQ,2EZP,2EZR,2EZR	1998	99	0.43	Viral Protein	SIMIAN IMMUNODEFICIENCY VIRUS
2Q7C,2Q3I,2Q5U	1999	69	0.38	Viral Protein	HIV 1
1CZQ	1999	69	0.38	Viral Protein; Inhibitor	SACCHAROMYCES CEREVISIAE; HIV 1 ;SYNTHETIC CONSTRUCT
1MOF	1996	68	0.46	Viral Protein	MOLONEY MURINE LEUKEMIA VIRUS
1EBO	1998	54	0.45	Viral Protein	EBOLA VIRUS SP.
2BF1	2005	53	0.32	Viral Protein	SIMIAN IMMUNODEFICIENCY VIRUS

Table 4.6: Co-mentioned neighbor entries of the PDB entry-1AIK

PDB ID	Year	co-mention degree	Calibrated disruption score	Category	Source
1ENV	1997	3	0.74	Viral Protein	SACCHAROMYCES CEREVISIAE
1F23	2001	2	0.16	Viral Protein	HIV 1
1GC1	1998	2	0.48	Viral Protein	HOMO SAPIENS; HIV 1
2NY7	2007	2	0.37	Viral Protein	HIV 1; HOMO SAPIENS
3DNN	2008	2	0.35	Viral Protein; Immune System	HIV-1 M:B-HXB2R
2B4C	2005	2	0.28	Viral Protein; Immune System	HIV 1; HOMO SAPIENS; SYNTHETIC CONSTRUCT
3NGB	2010	2	0.21	Viral Protein; Immune System	HIV 1; HOMO SAPIENS
3MA9	2010	2	0.07	Immune System	HIV 1; HOMO SAPIENS
3MAC	2010	2	0.07	Immune System	HIV 1; HOMO SAPIENS
2X7R	2010	2	0.05	Viral Protein	HIV 1 LW12.3 ISOLATE

Table 4.7: Co-cited neighbor entries of the PDB entry-1F88

PDB ID	Year	co-citation degree	Calibrated disruption score	Category	Source
2RH1	2007	313	0.41	Membrane Protein	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
2R4R,2R4S	2007	222	0.34	Signaling Protein	HOMO SAPIENS; MUS MUSCULUS
3EML	2008	217	0.33	Membrane Protein; Receptor	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
2VT4	2008	210	0.33	Receptor	MELEAGRIS GALLOPAVO
3DQB	2008	166	0.37	Signaling Protein	BOS TAURUS; SYNTHETIC CONSTRUCT
3CAP	2008	155	0.31	Signaling Protein	BOS TAURUS
1U19	2004	153	0.32	Signaling Protein	BOS TAURUS
1GZM	2004	136	0.34	Signaling Protein	BOS TAURUS
3OE0,3OE9,3OE8,3OE6,3ODU	2010	125	0.13	Signaling Protein	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4 ; SYNTHETIC CONSTRUCT
2I36,2I37,2I35	2006	107	0.38	Membrane Protein	BOS TAURUS

Table 4.8: Co-mentioned neighbor entries of the PDB entry-1F88

PDB ID	Year	co-mention degree	Calibrated disruption score	Category	Source
2RH1	2007	10	0.41	Membrane Protein	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
3EML	2008	8	0.33	Membrane Protein; Receptor	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
1GZM	2004	7	0.34	Signaling Protein	BOS TAURUS
2VT4	2008	8	0.33	Receptor	MELEAGRIS GALLOPAVO
1U19	2004	7	0.32	Signaling Protein	BOS TAURUS
3CAP	2008	7	0.31	Signaling Protein	BOS TAURUS
3ODU	2010	7	0.13	Signaling Protein	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4
3PBL	2010	6	0.12	Hydrolase	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4
1L9H	2002	5	0.44	Signaling Protein	BOS TAURUS
3NY9	2010	5	0.10	Membrane Protein	HOMO SAPIENS; ENTEROBACTERIA PHAGE T4

4.3.6 If the Authors Clearly Cite Data Sources Will Also Help Improve Impact of Their Own Papers?

Based on the results of Pearson correlation coefficient, we could find that the growth of citation and mention frequency are in moderate linear relationship. It may suggest that the PDB users are encouraged to do the clear data mention in their papers will also help to increasing the citation. However, we should point out the problem that do we need to suggest authors that clearly citing data sources help improve impact of their own papers. We try to answer this question through dividing the PDB entries' citing or ID mentioning articles into some groups. For a PDB entry, $p \in P$, a specific journal j of published

year t , those papers citing to p 's primary citation is denoted as $cite_{jt}(p)$ and those papers mentioning p 's is denoted $ment_{jt}(p)$. We consider five journals that all related the protein structure researches, including PLOS Pathogens (PLoS Pathog.), Acta Crystallographica Section D (Acta Crystallogr. D), The Journal of Biological Chemistry (J. Biol. Chem.), BMC Structural Biology (BMC Struct. Biol.), and Nature Structural and Molecular Biology (Nat. Struct. Mol. Biol.). The papers are divided into as the following four patterns for further discussion.

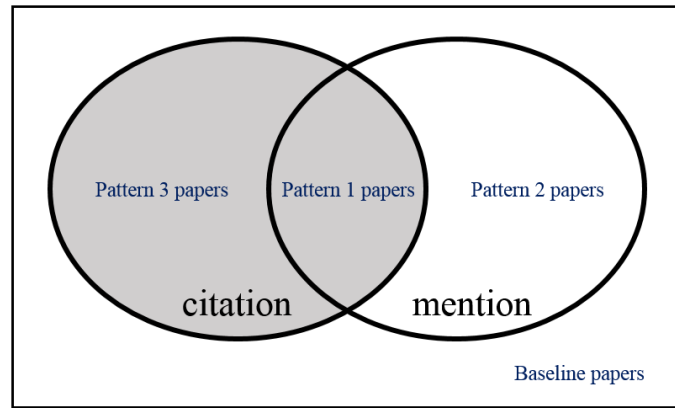


Figure 4.5: Venn Diagram of the selected papers.

- Pattern 1 papers: Those PDB ID mentioning articles that also cites to the corresponding PDB primary citation, which is denoted as the set, $ment_{jt}(p) \cap cite_{jt}(p)$, $p \in P$.
- Pattern 2 papers: Those PDB ID mentioning articles that do not cite to the corresponding PDB primary citation, which is denoted as the set, $ment_{jt}(p) \setminus (ment_{jt}(p) \cap cite_{jt}(p))$, $p \in P$.
- Pattern 3 papers: Those articles citing to the PDB entries' primary citation, but do not mention the corresponding PDB ID, which is denoted as the set, $cite_{jt}(p) \setminus (ment_{jt}(p) \cap cite_{jt}(p))$, $p \in P$.
- Baseline papers: Those articles do not cite any of the PDB entries' primary citation, nor mention the PDB IDs, which is denoted as the set, $J_t \setminus (ment_{jt}(p) \cup cite_{jt}(p))$, J_t is the set of articles of the corresponding journal published at time t and $p \in P$.

The difference of citations between pattern i articles of p , $ptn_i(p)$, and the pattern j articles of p , $ptn_j(p)$ is given by,

$$diff(p) = \frac{1}{\|ptn_i(p)\|} \sum_{k \in ptn_i(p)} \|cite(k)\| - \frac{1}{\|ptn_j(p)\|} \sum_{k \in ptn_j(p)} \|cite(k)\|, \quad (4.11)$$

then we summarized the difference results for a subset, $P' \subset P$, where the element $p \in P'$ both contains pattern i and pattern j articles, and the difference can be calculated by,

$$diff(P') = \frac{1}{|P'|} \sum_{p \in P'} diff(p). \quad (4.12)$$

Fig. 4.6 help us to know that if the authors both clearly citing the data sources and mention the IDs of used data will get more citations than those only citing the data sources or mention the IDs, all the pairs from pattern 1 papers to pattern 3 papers are considered to answer this question. However, the result shows that it is not clear if the authors clearly citing data sources or mention the IDs of used data on the papers will help improve impact of their own papers. Moreover, Fig. 4.7 answer the question that if the authors both clearly citing the PDB data or mention the PDB IDs will get more citations than those papers nor citing PDB data or mention the PDB IDs. In comparison with baseline papers, the results of PLoS Pathog. and Nat. Struct. Mol. Biol. show that those authors citing PDB primary citation papers or mention the PDB IDs have more citations.

4.4 Summary

In this chapter, we consider the issues surrounding the various of data citation to the PDB data. These analyses offer insights into the investigating of data citation behavior patterns of the users and help us to know the trends of protein structure researches. And this understanding can then hopefully help us to figure out what the properties of protein structure be studied as the popular research topics over the past decade. Our findings include that

1. The user prefer to only cite the primary citation of PDB data, instead of mentioning the IDs of PDB data.

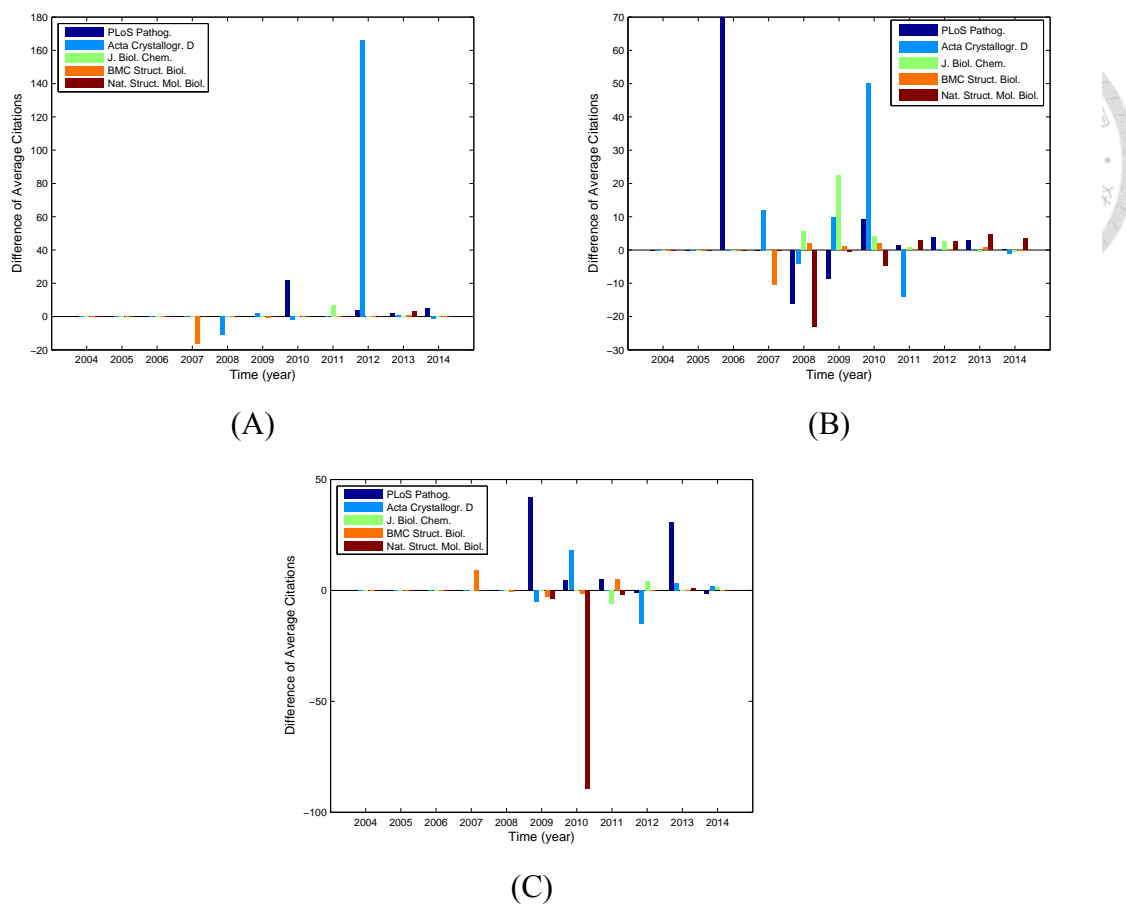
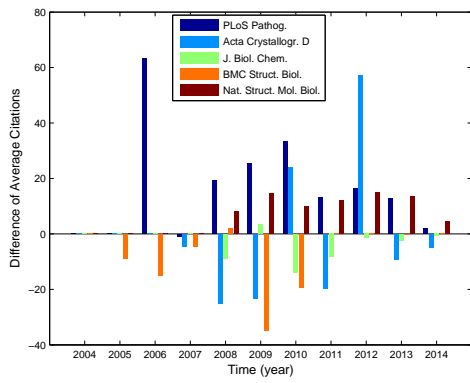


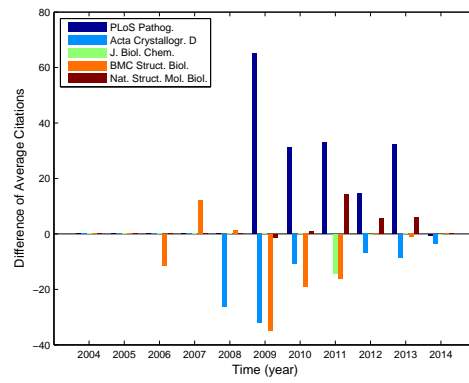
Figure 4.6: The difference of two pattern articles in (A) Case 1: Both mention & citing (pattern 1) vs. Only mention (pattern 2), (B) Case 2: Both mention & citing (pattern 1) vs. Only citing (pattern 3), and (C) Case 3: Only mention (pattern 2) vs. Only citing (pattern 3).

2. The PDB entries are dependent of their data citation frequencies.
3. The citation frequencies and the mention frequencies are in moderate linear relationship.
4. A comparison of co-citation and co-mention shows that the similar protein structures researches tend to potentially be clustered together.

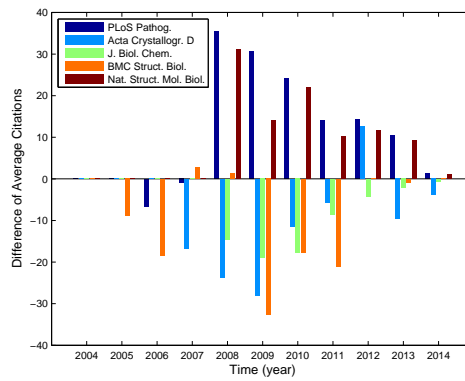
Additionally, we do a complete data usage study to the PDB repository that incorporated with the co-citation/mention metrics and the disruption quantization, that will mature the process for helping PDB users to find out those concerned and needed protein structure data, and will also help to facilitate data sharing and reusing. Finally, we believe that if users could cite the data and mention the IDs (or DOIs) in the article, that can be benefit



(A)



(B)



(C)

Figure 4.7: The difference of two pattern articles in (A) Case 4: Both mention & citing (pattern 1) vs. Nor mention & citing (baseline), (B) Case 5: Only mention (pattern 2) vs. Nor mention & citing (baseline), and (C) Case 6: Only citing (pattern 3) vs. Nor mention & citing (baseline).

to both sides of data provider and repository developer.



Chapter 5

Summaries and Future Work

5.1 Summary of the results

Encouraging the practice of data citation that contributes to data reused, experiments reproduced, and provide machine readability for tracing the data usage, and make the data are easily to be given scholarly credit to all contributors to the data. Moreover, it help to be sufficiently flexible to accommodate the variant data interpretability among different database. In this thesis, we studied the data citation patterns of the PDB repository. From the results of these analyses, we recommended data citation and provenance practices, approaches to discover data citations, ways of linking citations and data, and data access metrics. Here, we summarize our analytical methodology and review the results.

In Chapter 2, we proposed a method to identify transformative challengers by measuring how much they disrupt the growth of citation cascades of papers representing the established paradigm. We studied citations records of physics and computer science papers. Our method can efficiently calculate the disruption score of challenger papers in these large citation datasets. For each case study, our method found challengers that were more relevant to the seed and more important, as judged by later citation by the Nobel prize committee.

In Chapter 3, we applied an approach using disruptions of citation cascades and correlate data citations with data usage statistics to compare the citations to PDB repository by citing original and followup publications and URL mentioning. From the experimental

results, it revealed that the traditional academic citations is still not sufficient. Especially, we could find that the authors often mention the PDB URL, instead of citing on these PDB publications. They are certainly the latent PDB repository's users, but will not be reflected on the academic citations. Therefore, it will result in an underestimation of the impact of PDB.

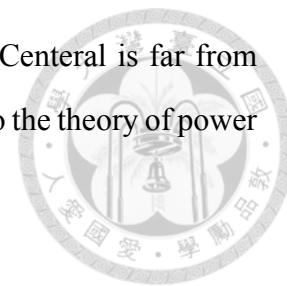
In Chapter 4, we addressed the issues surrounding the various of data citation and access metrics to the PDB data. Meanwhile, our studies focus on the interplay of PDB IDs mentions recognition and references cited of the literature, and the relative importance of these two mechanisms can be expressed by investigating the data citation patterns. We believe that if users could cite the data and mention the IDs (or DOIs) in the article, that can be great benefit to both sides of data provider and repository developer. However, the results reveal that authors increasingly choose to only cite the primary citation of PDB data instead of mentioning the IDs of PDB data. This chapter described a general approach for visualizing the trend of how authors use PDB data and offered insights into the data citation behavior patterns of the users.

5.2 Limitations

Although our framework tried to reflect scholarly impact in the broader sense, our model often have been limited by accessibility to and scalability of data. By contrast, altmetrics probably take a broad view of visibility in comparison to scholar citation metrics. Several social media platforms have been proposed as alternatively sources for measuring the impact of altmetrics on scholarly publications, such as search engine (like Google) query counts, social media mentions (mentioned in Twitter, Facebook, or Github) that often provide free access to usage data through corresponding APIs, hence, data collection is relatively easier and less cost. In response, more and more electronic journals have turned to altmetrics, which providing citations or mentions count in specific social media services. However, these kinds of citations or mentions tend to increase explosively in the short term that against the characteristic of our framework. Moreover, the coverage of all the social media sources seems to be low except for Twitter, so it is not clear if it is

suitable to be useful in practice.

Another limitation is that the text and citation data in PubMed Central is far from complete, missing important journals like Nature, though according to the theory of power law network this might not affect the conclusions too much.



5.3 Future directions

In the future, we will try to analyze the characteristics of all kinds of PDB related citation networks to see how the data and citations influenced or enabled groundbreaking research and development of drugs. This analysis will use the drug and drug target mapping between the RCSB PDB and DrugBank. The purpose here is to correlate various access metrics with tangible impact indicators to determine empirically which metrics are more informative. Analysis of citation and data cascades of these networks will highlight putative pathways of how data and concepts led to the discovery of drug candidates.





Appendices





Appendix A

Estimation of the Subsampling Error of Disruption Score

For any strictly positive constant ε , with probability greater than $1 - 2e^{-2\varepsilon^2|C'_t|} \frac{1}{4\varepsilon|C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} - 2e^{-2\varepsilon^2|\widetilde{C}'_t|} \frac{1}{4\varepsilon|\widetilde{C}'_t|} \sqrt{\frac{\pi}{2|\widetilde{C}'_t|}}$, if $(\Phi_t(C') - \Phi_t(\widetilde{C}')) > 0$ then $\exists S > 0$, $(\Phi_t(C) - \Phi_t(\widetilde{C})) \in (S + \varepsilon, S - \varepsilon)$. We divide the proof into the following two parts.

- part 1. Prove that let $S = (\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)])$, we have

$$(\Phi_t(C') - \Phi_t(\widetilde{C}')) = (\mathbb{E}_{j \sim C'_t}[\phi_{C'}(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_{C'}(j)]) > 0 \implies S > 0. \quad (\text{A.1})$$

- Part 2. Prove

$$\begin{aligned} \Pr \left(\left| S - (\mathbb{E}_{j \sim C_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}_t}[\phi_C(j)]) \right| > \varepsilon \right) \\ < 2e^{-2\varepsilon^2|C'_t|} \frac{1}{4\varepsilon|C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} + 2e^{-2\varepsilon^2|\widetilde{C}'_t|} \frac{1}{4\varepsilon|\widetilde{C}'_t|} \sqrt{\frac{\pi}{2|\widetilde{C}'_t|}}. \end{aligned}$$

Proof of Part 1.

$$(\mathbb{E}_{j \sim C'_t}[\phi_{C'}(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_{C'}(j)]) > 0 \implies (\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)]) > 0 \quad (\text{A.2})$$

Let the difference between the new ϕ of the nodes in the subsampled cascade C' and

the original ϕ as the following,

$$\begin{aligned}\Delta\phi_C(j) &= \phi_C(j) - \phi_{C'}(j) \\ &= \alpha \left(\sum_{i \in C} \phi_C(i) I(i \in \text{cite}(j) \& i \notin C') + \sum_{i \in C} \Delta\phi_C(i) I(i \in \text{cite}(j) \& i \in C') \right)\end{aligned}$$



and its expectation is

$$\mathbb{E}[\Delta\phi_C(j)] \approx \alpha \mathbb{E}[|\text{cite}(j)|] ((1 - \rho)|C| \mathbb{E}[\phi_C(i)] + \rho|C| \mathbb{E}[\Delta\phi_C(i)]), \quad (\text{A.3})$$

where $\rho \equiv \frac{|C'|}{|C|}$ is the sampling rate. From Eq. (A.3), we can conclude that,

$$\mathbb{E}[\Delta\phi_C(j)] > \mathbb{E}[\Delta\phi_C(i)], \forall i \in \text{cite}(j). \quad (\text{A.4})$$

Furthermore, for those nodes $j \in C_1$ at the initial time $t = 1$, it is held that

$$\Delta\phi_{C_1}(j) \geq 0. \quad (\text{A.5})$$

From (A.4) and (A.5), we have

$$\mathbb{E}[\Delta\phi_C(j)] \geq 0, \forall j \in C, \quad (\text{A.6})$$

and the main statement (A.1) can be proved by

$$\begin{aligned}& (\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)]) \\ &= \alpha \mathbb{E}[|\text{cite}(j)|] |C'_t| \mathbb{E}[\phi_C(i)] - \alpha \mathbb{E}[|\text{cite}(j)|] |\widetilde{C}'_t| \mathbb{E}[\phi_C(i)] \\ &= \alpha \mathbb{E}[|\text{cite}(j)|] (|C'_t| - |\widetilde{C}'_t|) \mathbb{E}[\phi_C(i)] \\ &= \alpha \mathbb{E}[|\text{cite}(j)|] (|C'_t| - |\widetilde{C}'_t|) \mathbb{E}[\phi_{C'_t}(i) + \Delta\phi_C(i)] \\ &= \alpha \mathbb{E}[|\text{cite}(j)|] (|C'_t| - |\widetilde{C}'_t|) (\mathbb{E}[\phi_{C'_t}(i)] + \mathbb{E}[\Delta\phi_C(i)]) \\ &= \alpha \mathbb{E}[|\text{cite}(j)|] (|C'_t| - |\widetilde{C}'_t|) \mathbb{E}[\phi_{C'_t}(i)] + \alpha \mathbb{E}[|\text{cite}(j)|] (|C_t| - |\widetilde{C}'_t|) \mathbb{E}[\Delta\phi_C(i)] \\ &= (\mathbb{E}_{j \sim C'_t}[\phi_{C'_t}(i)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_{C'_t}(i)]) + \alpha \mathbb{E}[|\text{cite}(j)|] (|C_t| - |\widetilde{C}'_t|) \mathbb{E}[\Delta\phi_C(i)].\end{aligned} \quad (\text{A.7})$$

Due to (A.6) and (A.7), if $(\mathbb{E}_{j \sim C'_t}[\phi_{C'_t}(i)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_{C'_t}(i)])$ is positive, $(\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)])$ will be positive, too. Consequently,

$$(\mathbb{E}_{j \sim C'_t}[\phi_{C'}(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_{C'}(j)]) > 0 \implies (\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)]) > 0. \quad \square$$



Proof of Part 2. Our goal is to show that $(\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)]) \simeq (\mathbb{E}_{j \sim C_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}_t}[\phi_C(j)])$ with a probability higher than $1 - q$, or more precisely,

$$\Pr(|(\mathbb{E}_{j \sim C_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}_t}[\phi_C(j)]) - (\mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)])| > \varepsilon) < q. \quad (\text{A.8})$$

The proof is to derive q .

To simplify the notations, let $A \equiv \mathbb{E}_{j \sim C_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}_t}[\phi_C(j)]$ and $B \equiv \mathbb{E}_{j \sim C'_t}[\phi_C(j)] - \mathbb{E}_{j \sim \widetilde{C}'_t}[\phi_C(j)]$. The probability in Inequality (A.8) can be rewritten as

$$\begin{aligned} \Pr(|A - B| > \varepsilon) &= \Pr(((A - B) > \varepsilon) \vee (-(A - B) > \varepsilon)) \\ &= \Pr((A - B) > \varepsilon) + \Pr(-(A - B) > \varepsilon) - \Pr(((A - B) > \varepsilon) \wedge (-(A - B) > \varepsilon)) \\ &= \Pr((A - B) > \varepsilon) + \Pr(-(A - B) > \varepsilon) \\ &< q. \end{aligned}$$

Since the derivations are similar, we only provide that of $\Pr((A - B) > \varepsilon)$. The derivation will involve Hoeffding's inequality. Since C' is a random sample of C , from Hoeffding's inequality,

$$\Pr((\mathbb{E}_{j \sim C_t}[\phi_C(j)] - \mathbb{E}_{j \sim C'_t}[\phi_C(j)]) > \varepsilon) < e^{-2\varepsilon^2|C'_t|}. \quad (\text{A.9})$$

Decomposing the convolution of the density functions of A and B , we have

$$\begin{aligned} \Pr((A - B) > \varepsilon) &= \int_{A-B>\varepsilon} f_A(x) \cdot f_B(y) dx dy \\ &= \int_0^\infty \int_{\varepsilon+y}^\infty f_A(x) \cdot f_B(y) dx dy + \int_{-\infty}^0 \int_{\varepsilon+y}^\infty f_A(x) \cdot f_B(y) dx dy, \end{aligned} \quad (\text{A.10})$$

where $f(\cdot)$ is the probability density function. The first term in (A.10) is

$$\begin{aligned}\int_0^\infty \int_{\varepsilon+y}^\infty f_A(x) \cdot f_B(y) dx dy &= \int_0^\infty f_B(y) \int_{\varepsilon+y}^\infty f_A(x) dx dy \\ &= \int_0^\infty f_B(y) \Pr(A > \varepsilon + y) dy\end{aligned}$$

(Based on Inequality (A.9))

$$\begin{aligned}&< \int_0^\infty f_B(y) e^{-2(\varepsilon+y)^2 |C'_t|} dy \\ &< \int_0^\infty f_B(y) dy \int_0^\infty 2e^{-2(\varepsilon+y)^2 |C'_t|} dy\end{aligned}$$

(By Cauchy-Schwarz inequality)

$$\begin{aligned}&= \Pr(B > 0) \int_0^\infty e^{-2(\varepsilon+y)^2 |C'_t|} dy \tag{A.11} \\ &< e^0 \cdot \int_0^\infty e^{-2(\varepsilon+y)^2 |C'_t|} dy \\ &< 1 \cdot \int_0^\infty e^{-2(\varepsilon+y)^2 |C'_t|} dy \\ &= e^{-2\varepsilon^2 |C'_t|} \int_0^\infty e^{-4\varepsilon y |C'_t|} e^{-2y^2 |C'_t|} dy \\ &< e^{-2\varepsilon^2 |C'_t|} \int_0^\infty e^{-4\varepsilon y |C'_t|} dy \int_0^\infty e^{-2y^2 |C'_t|} dy\end{aligned}$$

(By Cauchy-Schwarz inequality)

$$= e^{-2\varepsilon^2 |C'_t|} \frac{1}{4\varepsilon |C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} \cdot (\text{By Gaussian integral})$$

It is symmetric for the second term:

$$\int_{-\infty}^0 \int_{\varepsilon+y}^\infty f_A(x) \cdot f_B(y) dx dy < e^{-2\varepsilon^2 |C'_t|} \frac{1}{4\varepsilon |C'_t|} \sqrt{\frac{\pi}{2|C'_t|}}. \tag{A.12}$$

Combining (A.11) and (A.12), we have

$$\Pr(A - B > \varepsilon) < 2e^{-2\varepsilon^2 |C'_t|} \frac{1}{4\varepsilon |C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} \tag{A.13}$$

Similarly, it holds that

$$\Pr(-(A - B) > \varepsilon) < 2e^{-2\varepsilon^2 |\widetilde{C}'_t|} \frac{1}{4\varepsilon |\widetilde{C}'_t|} \sqrt{\frac{\pi}{2|\widetilde{C}'_t|}}. \tag{A.14}$$



From (A.13) and (A.14), the lower bound q of the probability in (A.8) is

$$2e^{-2\varepsilon^2|C'_t|} \frac{1}{2\varepsilon|C'_t|} \sqrt{\frac{\pi}{2|C'_t|}} + 2e^{-2\varepsilon^2|\widetilde{C}'_t|} \frac{1}{2\varepsilon|\widetilde{C}'_t|} \sqrt{\frac{\pi}{2|\widetilde{C}'_t|}}.$$







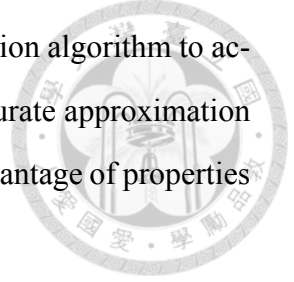
Appendix B

Model-based Approximation for Cascade Generating Function

Computation of the disruption scores must be efficient given ever-growing number of biomedical sciences papers and citations. One challenge of the proposed approach is that the computation of the cascade structure and pairwise comparison can be intractable. We have several strategies to accelerate the computation. One strategy is to avoid exhaustive pairwise comparison by reusing intermediate results. Suppose we would like to rank 100 candidate challengers by their disruption scores. A brute-force approach is to compute the residue cascades for each of the candidates. By sorting these candidates in their topological order in the citation network, the ϕ values computed for the upstream candidates can be reused for the downstream candidates and significantly reduce the computational costs.

Another strategy is by approximation, where we can take advantage of the fact that citations decay exponentially over time to estimate the size of cascades. Computing average cascade Φ can be intractable. A brute-force algorithm to compute Φ is to traverse the citation network and update ϕ for each node visited by a topological sorting algorithm. Such an exhaustive search algorithm slows down as the size of the citation network increases exponentially in recent years. Arbesman [3] shows that, for any paper, the longer away from the citing paper, the less likely that the paper will be cited, and the decay is approximately exponential. Also, modern papers cite more often and the average cita-

tions increase each year. The result suggests that it is possible to model the citation counts accurately and we will take advantage of that to derive an approximation algorithm to accelerate the computation of cascade overtaking. We now present accurate approximation algorithms scalable to very large scale citation networks by taking advantage of properties of Φ .



It has been suggested that citation counts of papers decay exponentially over time and the rate of decay can be estimated accurately. We plotted the curves of the annual average citation count of the papers in the APS citation network dataset as shown in Fig. B.1, which shows that the longer away from the citing paper, the less likely that the paper will be cited, and the decay is approximately exponential. Also, modern papers cite more often and the average citations increase each year. The plot suggests that it is possible to model the citation counts accurately and we will take advantage of that to derive an approximation algorithm to accelerate the computation of cascade overtaking.

Given a cascade C , We model the number of citations by a function $\Gamma_C(t, \tau)$, the average citations from papers published in time t to papers published τ time steps (*i.e.*, year, in this case) early. $\Gamma_C(t, \tau)$ can be estimated and presented as a data point on a curve in the plot.

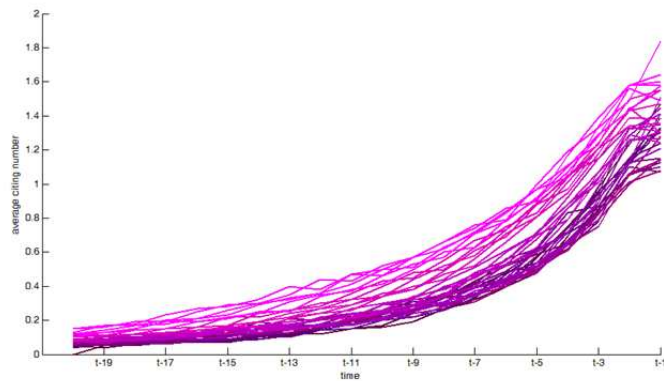


Figure B.1: Average citations of papers in the APS citation network dataset. Each curve is for the papers published in one year from 1970 (darkest) to 2009 (brightest). The curve plots the change of the average citations to the past years. The horizontal-axis indicates how many year from the publication time t .

Consider the problem of computing Φ . If we consider paper $j \in C_t$ as a random variable, then $\Phi_t(C)$ is by definition the expectation of $\phi(j)$. The number of citations

from j to papers published in time $t - \tau$ will also be a random variable and its expectation will be $\Gamma_C(t, \tau)$. Then the expected contribution of these citations to $\phi(j)$ will be $\Gamma_C(t, \tau) \cdot \alpha \cdot \Phi_{t-\tau}(C)$ approximately. Fig. B.2 illustrates this idea.

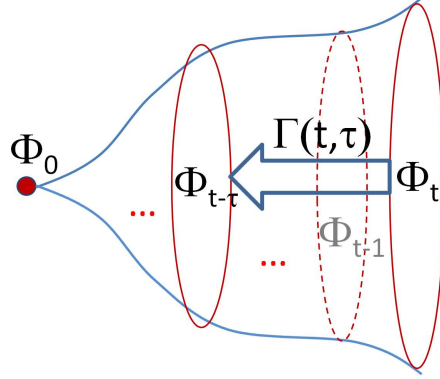


Figure B.2: Estimating average cascade function values Φ by modeling citation counts Γ .

Therefore, we can derive an approximation of Φ as follows:

$$\begin{aligned}
 \Phi_t(C) &= \mathbb{E}_{j \sim C_t}[\phi(j)] \\
 &= \mathbb{E}_{j \sim C_t} \left[\sum_{\tau=1}^t \alpha \sum_{i \in C_{t-\tau}} I[i \in cite(j)] \phi(i) \right] \\
 &= \sum_{\tau=1}^t \alpha (\mathbb{E}_{i \sim C_{t-\tau}}[\sum I[i \in cite(j)]] \mathbb{E}_{i \sim C_{t-\tau}}[\phi(i)] + \text{cov}_{i \sim C_{t-\tau}}[\sum I[i \in cite(j)], \phi(i)])
 \end{aligned} \tag{B.1}$$

$$\approx \sum_{\tau=1}^t \alpha \Gamma_C(t, \tau) \Phi_{t-\tau}(C). \tag{B.2}$$

We note that $I[s]$ is the identity function that returns 1 if the parameter s is true and 0 otherwise. The difference between the approximation (B.2) and (B.1) is the sum of the covariance terms in (B.1). These terms are zero under the assumption that the cascade function value ϕ of a paper, which depends on how many papers it cites, and how often it is cited in the future, are uncorrelated. The assumption is reasonable and can be confirmed empirically. Therefore, we expect that the approximation error will be negligible.

Compared to an exhaustive search algorithm, computing $\Phi_t(C)$ using Eq. (B.2) reduced the complexity from exponential to quadratic. Since the equation is defined recursively, there are plenty of room to optimize its implementation. The preprocessing step

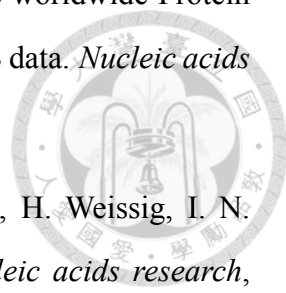
that estimates Γ requires to visit each node once and therefore its time complexity is linear to the size of the citation network.





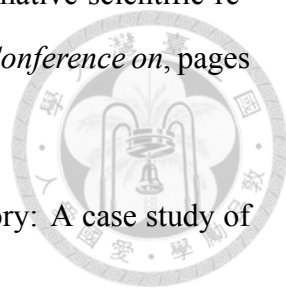
Bibliography

- [1] P. D. Allison. Inequality and Scientific Productivity. *Social Studies of Science*, 10(2):163–179, May 1980.
- [2] P. D. Allison, J. S. Long, and T. K. Kraze. Cumulative advantage and inequality in science. *Ame. Sociological Review*, 47(5):615–625, 1982.
- [3] S. Arbesman. *The Half-life of Facts: Why Everything We Know Has an Expiration Date*. Current Hardcover, first edition edition, Sept. 2012.
- [4] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. The universal protein resource (uniprot). *Nucleic Acids Research*, 33(suppl 1):154–159, 2005.
- [5] J. Bardeen, L. N. Cooper, and J. R. Schrieffer. Microscopic theory of superconductivity. *Phys. Rev.*, 106:162–164, Apr 1957.
- [6] J. Bardeen, L. N. Cooper, and J. R. Schrieffer. Theory of superconductivity. *Phys. Rev.*, 108:1175–1204, Dec 1957.
- [7] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. The pfam protein families database. *Nucleic Acids Research*, 32(suppl 1):138–141, 2004.
- [8] J. Bednorz and K. Müller. Possible high T_c superconductivity in the Ba–La–Cu–O system. *Z. Phys. B*, 64(2):189–193, June 1986.
- [9] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.

- 
- [10] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research*, 35(suppl 1):301–303, 2007.
- [11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [12] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O’Donovan, I. Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.
- [13] P. Bonacich. Power and centrality: a family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.
- [14] P. E. Bourne, K. J. Address, W. F. Bluhm, L. Chen, N. Deshpande, Z. Feng, W. Fleri, R. Green, J. C. Merino-Ott, W. Townsend-Merino, et al. The distribution and query systems of the rcsb protein data bank. *Nucleic acids research*, 32(suppl 1):223–225, 2004.
- [15] H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel, et al. E-MSD: the European bioinformatics institute macromolecular structure database. *Nucleic Acids Research*, 31(1):458–462, 2003.
- [16] D. C. Chan, D. Fass, J. M. Berger, and P. S. Kim. Core structure of gp41 from the hiv envelope glycoprotein. *Cell*, 89(2):263–273, 1997.
- [17] C. Chen. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5303–5310, Apr. 2004.
- [18] P. Chen and S. Redner. Community structure of the physical review citation network, Nov 2009. Comments: 14 pages, 7 figures, 8 tables.

- [19] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, Jan. 2007.
- [20] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka, et al. High-resolution crystal structure of an engineered human β 2-adrenergic g protein–coupled receptor. *science*, 318(5854):1258–1265, 2007.
- [21] U. Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.
- [22] L. A. Davidson and K. Douglas. Digital Object Identifiers: Promise and problems for scholarly publishing. *Journal of Electronic Publishing*, 4(2), 1998.
- [23] N. Deshpande, K. J. Address, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, et al. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mm-CIF schema. *Nucleic acids research*, 33(suppl 1):233–237, 2005.
- [24] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.
- [25] P. Edwards, M. Burghammer, V. Ratnala, R. Sanishvili, R. Fischetti, G. Schertler, W. Weis, and B. Kobilka. Crystal structure of the human beta2 adrenergic g-protein-coupled receptor. *Nature*, 450(7168):383387, 2007.
- [26] V. J. Emery. Theory of high- t_c superconductivity in oxides. *Physical Review Letters*, 58(26):2794–2797, June 1987.
- [27] R. D. Finn, J. Tate, J. Mistry, P. C. Coghill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, et al. The pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):281–288, 2008.

- [28] FORCE11 Data Citation Synthesis Group. Joint Declaration of Data Citation Principles - FINAL. 2014.
- [29] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman. Time-aware ranking in dynamic citation networks. In *COMMPER 2011: Mining Communities and People Recommendations, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 373–380, December 2011.
- [30] R. Ghosh and K. Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of Web Search and Data Mining Conference (WSDM)*, February 2011.
- [31] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012)*, 2012.
- [32] A. Golovin, T. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J. M. Ionides, et al. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 32(suppl 1):211–216, 2004.
- [33] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):140–144, 2006.
- [34] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.
- [35] K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, et al. Remediation of the protein data bank archive. *Nucleic acids research*, 36(suppl 1):426–433, 2008.
- [36] J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, Nov. 2005.

- 
- [37] Y.-H. Huang, C.-N. Hsu, and K. Lerman. Identifying transformative scientific research. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 291–300, 2013.
- [38] Y.-H. Huang, P. W. Rose, and C.-N. Hsu. Citing a data repository: A case study of the protein data bank. *PloS one*, 10(8):e0136631, 2015.
- [39] S. Iijima. Helical microtubules of graphitic carbon. *Nature*, 354:56–58, Nov. 1991.
- [40] Ş. Kafkas, J.-H. Kim, and J. R. McEntyre. Database citation in full text biomedical articles. *PLoS ONE*, 8(5):e63184, 2013.
- [41] A. B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.
- [42] A. Klamer and H. P. Van Dalen. Attention and the art of scientific publishing. *Journal of Economic Methodology*, 9(3):289–315, 2002.
- [43] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P. E. Bourne, and H. M. Berman. The rcsb pdb information portal for structural genomics. *Nucleic acids research*, 34(suppl 1):302–305, 2006.
- [44] T. S. Kuhn. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University Of Chicago Press, fourth edition edition, 2012.
- [45] E. S. Lang, P. C. Wyer, and R. B. Haynes. Knowledge translation: closing the evidence-to-practice gap. *Annals of emergency medicine*, 49(3):355–363, 2007.
- [46] K. Lerman and R. Ghosh. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, May 2010.
- [47] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proceedings of 7th SIAM International Conference on Data Mining (SDM)*, Apr. 2007.

- [48] A. Mazlounian, Y.-H. Eom, D. Helbing, S. Lozano, and S. Fortunato. How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes. *PLoS ONE*, 6(5):e18975+, May 2011.
- [49] R. K. Merton. The Matthew Effect in Science. *Science*, 159(3810):56–63, Jan. 1968.
- [50] R. K. Merton. The matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, 79(4):606–623, 1988.
- [51] Z. S. S. Morris, S. Wooding, and J. Grant. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, 104(12):510–520, 2011.
- [52] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of ICDM*, 2012.
- [53] A. Névéol, W. J. Wilbur, and Z. Lu. Improving links between literature and biological data with text mining: a case study with geo, pdb and medline. *Database*, 2012:bas026, 2012.
- [54] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov. Electric Field Effect in Atomically Thin Carbon Films. *Science*, 306(5696):666–669, 2004.
- [55] K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov, and A. K. Geim. Two-dimensional atomic crystals. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10451–10453, July 2005.
- [56] T. Okada, Y. Fujiyoshi, M. Silow, J. Navarro, E. M. Landau, and Y. Shichida. Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography. *Proceedings of the National Academy of Sciences*, 99(9):5982–5987, 2002.

- [57] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, et al. Crystal structure of rhodopsin: Ag protein-coupled receptor. *science*, 289(5480):739–745, 2000.
- [58] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto. 2010.
- [59] A. Prlić, M. A. Martinez, D. Dimitropoulos, B. Beran, B. T. Yukich, P. W. Rose, P. E. Bourne, and J. L. Fink. Integration of open access literature into the rcsb protein data bank using biolite. *BMC bioinformatics*, 11(1):220, 2010.
- [60] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(suppl 1):61–65, 2007.
- [61] S. Redner. Citation Statistics from 110 Years of Physical Review. *Physics Today*, 58(6):49–54, 2005.
- [62] H. Sayyadi and L. Getoor. Future rank: Ranking scientific articles by predicting their future PageRank. In *2009 SIAM International Conference on Data Mining (SDM09)*, 2009.
- [63] D. M. Standley, A. R. Kinjo, K. Kinoshita, and H. Nakamura. Protein structure databases with new web services for structural biology and biomedical research. *Briefings in bioinformatics*, 9(4):276–285, 2008.
- [64] L. Šubelj, D. Fiala, and M. Bajec. Network-based statistical comparison of citation topology of bibliographic databases. *Scientific reports*, 4, 2014.
- [65] J. Tang, L. Yao, D. Zhang, and J. Zhang. A combination approach to web user profiling. *ACM TKDD*, 5(1):1–44, 2010.
- [66] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In *ICDM'07*, pages 292–301, 2007.

- [67] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic level expertise search over heterogeneous networks. *Machine Learning Journal*, 82(2):211–237, 2011.
- [68] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD '08*, pages 990–998, 2008.
- [69] Task Group on Data Citation Standards and Practices, CODATA-ICSTI. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12(0):1–75, 2013.
- [70] N. S. B. (U.S.). *Enhancing support of transformative research at the National Science Foundation [electronic resource]*. National Science Foundation, Arlington, VA :, 2007.
- [71] R. Van Noorden, B. Maher, and R. Nuzzo. The top 100 papers. *Nature*, 514(7524):550–553, 2014.
- [72] W. Weissenhorn, A. Dessen, S. Harrison, J. Skehel, and D. Wiley. Atomic structure of the ectodomain from hiv-1 gp41. *Nature*, 387(6631):426–430, 1997.
- [73] J. Westbrook, Z. Feng, L. Chen, H. Yang, and H. M. Berman. The protein data bank and structural genomics. *Nucleic acids research*, 31(1):489–491, 2003.
- [74] J. Westbrook, Z. Feng, S. Jain, T. Bhat, N. Thanki, V. Ravichandran, G. L. Gilliland, W. Bluhm, H. Weissig, D. S. Greer, et al. The protein data bank: unifying the archive. *Nucleic Acids Research*, 30(1):245–248, 2002.
- [75] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, and H. M. Berman. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, 21(7):988–992, 2005.
- [76] M. K. Wu, J. R. Ashburn, C. J. Torng, P. H. Hor, R. L. Meng, L. Gao, Z. J. Huang, Y. Q. Wang, and C. W. Chu. Superconductivity at 93 K in a new mixed-phase Y-

Ba-Cu-O compound system at ambient pressure. *Physical Review Letters*, 58(9), 1987.

[77] F. C. Zhang and T. M. Rice. Effective Hamiltonian for the superconducting Cu oxides. *Physical Review B*, 37(7):3759–3761, Mar. 1988.

