Axiomatic Analysis of Unsupervised Diversity on Large-Scale High-dimensional Data

by

Shiyan Yan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2021

Doctoral Committee:

Professor Qiaozhu Mei, Chair
Associate Professor Daniel Romero
Assistant Professor Misha Teplitskiy
Associate Professor VG Vinod Vydiswaran

Shiyan Yan

shiyansi@umich.edu

ORCID iD:0000-0002-3264-149X

To my dearest family and friends

ACKNOWLEDGEMENTS

There is no doubt that this very little piece of work written by Shiyan Yan cannot occupy the entire universe. It only contributes to a tiny nudge in the society, hopefully. However, to myself, this piece of work is my precious, accumulated with my right and wrong paths over the past eight years of career. Although I make tons of mistakes in small and significant decisions, I finally reach to this stage with great help from many people.

"When it rains, it pours." I cannot list all their names here. So dear reader, please forgive me if I miss your name.

Firstly, I want to say thank you to my doctoral committee members. Qiaozhu has offered me tremendous help when the challenges, academic and non-academic, come to me. It's my honor to defend this piece of work under his supervision. I have learned too much from him. Daniel, Misha, and Vinod accept my invitation kindly. Their suggestions make my piece of work solid. Their comments are really helpful to improve my mindset.

My peers and collaborators are very awesome. In every stage of my PhD career, they support my decisions and answers all the questions I raised up. Wei and I work together and his unforgettable help appear everyday in Michigan. So many other colleagues act like my siblings in these years. A long list of their names cannot cover their kindness to me: Tao Dong, Zhe Zhao, Xuan Zhao, Youyang Hou, Yue Wang, Cheng Li, Yang Liu, Ning Jiang, Fengmin Hu, Jiaqi Ma, Kan Yu, Sam Carton, Cristina Garbacea, Ark Zhang, Yingzhi Liang, Teng Ye, Xuan Lu, Theresa Velden, Yan Chen, Xinda Zeng, Linfeng Li, Licia He, Zhuofeng Wu, Xinyan Zhao, Yixin Zou, Xuedong Li, Huoran Li, Yutong Xie, Xiaolin Shi, Weinan Wang, Lisa Yu, Alessandro Oltramari, Xieyang Liu.

My friends give me hands when I encounter hardness over and over again. Special thanks to Fangyang Wang. She tried her best to persuade me to conquer something impossible. It is so amazing. I also want to thank Yuyan Han, Hongjiu Zhang, Zhenting Zou, Shaowu Pan, Yao Shi, Tianyi Liu. They brought the thoughts and happiness to my life.

There are another two names I cannot detour in any narrative of my story. Carl Lagoze, my first advisor, trust me and brought me to Michigan. It is hard to imagine that he gave me an offer when I nearly knew nothing about research and life in America. He gave me the start point for everything after my undergraduate years. Xin Rong, my great roommate, discussed with me about everything in our small apartment. His wisdom and pathetic death teaches me a lot, far beyond the books and papers I read in this eight years.

My parents support always and forever. I hope my work can make them to be proud of their son. It is such a tough process to turn a boy to a man. Loving them.

"Star and star meet dramatically" This dissertation and I meet with each other in this special year. I really appreciate the serendipity in my fate.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Diversity is a concept widely used in every corner of our society. It represents the "breadth" of a set of objects, which needs to be promoted or reduced in different scenarios. Though many people have discussed it, how to define diversity in a reliable way is still a non-trivial task. In particular, when we are facing large-scale high-dimensional data, it is impossible to use pre-defined classifications to divide each object into categories and utilize diversity measurements in downstream tasks. An unsupervised methodology is necessary to handle this challenge.

In this dissertation, I explore different methods to address the research question: how to measure diversity in an unsupervised manner based on large-scale high-dimensional data. I leverage representation learning algorithms to project objects into a discrete or continuous space and design several metrics to measure diversity in real-world applications. Furthermore, I introduce an axiomatic analysis method to help us choose and evaluate diversity metrics in both discrete and continuous settings.

Following the guidelines derived from the axiomatic analysis, I define diversity in terms of metrics to map distributions of topics to real numbers in discrete space. I also find a simple and intuitive metric to measure diversity, which is defined in continuous space, that performs surprisingly well to satisfy different axioms.

The sound and reliable metrics motivate me to focus on some controversial research topics in real applications. I explore the effect of research diversity i.e., how broad researchers' research interests are. I conduct several studies to figure out whether publishing papers with high diversity results in greater research impact. Furthermore, I track trajectories of researchers' careers and try to find the effects of research diversity at different stages.

Another real-world application appears in online social networks. Structural diversity, the closeness of users' friends, has a substantial influence on users' behavior from many perspectives. I define users' structural diversity using the results of axiomatic analysis. I track the pattern within the variation in structural diversity in both static and dynamic networks and simulate it with an intuitive graph generation algorithm. An interesting pattern of structural diversity and user engagement in online social media is illustrated.

CHAPTER I

Introduction

Diversity has been the buzzword at the center of both public and academic discussions for a long time. It is so important to our society. Researchers keep seeking fair and accurate ways to measure it in different domains and evaluate its social implications. Sociologists want to measure the ethnic diversity of certain groups of people. Biologists are interested in tracking the variations of diversity in animal species [75]. In the online community, scholars dive into the relationship between users' social diversity and social capital [32]. In the research of science of science, researchers explore the effect of diversity in publications on authors' research impact [61]. The idea of measuring diversity resides in all corners of research communities.

Long-term research on diversity has produced some results. Researchers have found effects of different types of diversity in various domains. For instance, greater ethnic diversity can improve people's happiness in a work environment [29]. Richer diversity in animal species can provide more robustness in an environmental system [75]. Scientists have conducted research on the effects of diversity and provided actionable suggestions to the public.

However, before drawing any conclusions about the complicated effects of diversity in different domains, there is a single fundamental question that interests many researchers: How can a reliable metric be designed to measure the diversity of a given

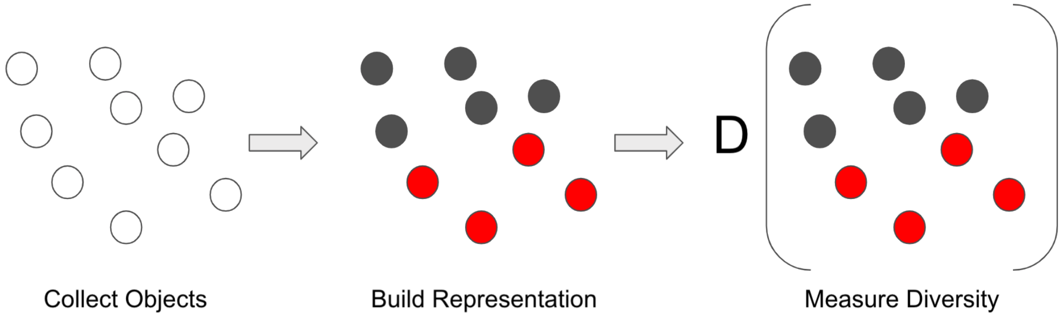


Figure 1.1: Diversity measurement pipeline

set of objects? Without a reliable method to measure diversity, we cannot build a solid foundation for diversity-related research.

To measure the diversity of a certain set of objects, such as the diversity of ethnic groups or animal species in certain areas, the most common practice follows three steps (as shown in Figure 1.1):

1. Collect information about a set of objects;
2. Build a low-dimensional representation of these objects e.g., assign objects to some categories;
3. Measure diversity using designed diversity metrics based on this representation.

For example, if we want to measure the diversity of races in a group of people, we need to collect information from people through surveys or various tests. We implement some criteria like the one-drop rule to categorize people into different races such as African American. Finally we design a metric like variance or the Gini index to test whether the distribution of people across different races is even or not.

This procedure is commonly used on a set of objects. If people intend to measure the inner diversity of a single object, it is better to render different properties of this object into a representation and follow step 2 and step 3 accordingly.

People have followed these common practices for a long time, but these practices have some significant drawbacks. First, building the low-dimensional representation is costly, which makes it hard to scale up. There are millions of species of animals around world. Scientists have spent a very long time testing the genes of animals and deciding their position in the gigantic species classification. Second, assigning objects into different categories requires lots of human effort and domain knowledge. For example, librarians need to assign books into different classes in the Dewey Decimal Classification (DDC), which contains many classes. This task requires librarians to obtain substantial training through school and real working scenarios. Last but not least, classifications themselves change over time. Our recognition of the world keeps changing rapidly, which makes some classifications unsuitable after a certain period. The first international classification of diseases (ICD) can be dated back to the year of 1893. Researchers in medicine have updated the diseases over and over again since new diseases continue to arise and the existing disease classification may not adequately reflect their nature.

In the big data era, we face more complicated diversity-related tasks which can not be resolved solely with supervised information. We usually do not even have any prior knowledge of applications and cannot generate its underlying data distribution as ground truth. Given a large set of images and texts generated by users, how can we build a classification and represent it within a few dimensions? When users change their data or add completely fresh content, how can the representation respond to it immediately? It is hard to imagine a well-defined classification that can handle a continuous stream of high-dimensional messy data easily. To tackle the drawbacks of traditional diversity measurement, we have come across a specific challenge which I try to address in this dissertation: How can diversity be measured without the supervised information of a large dataset?

In recent years, the rapid development of representation learning techniques empowered by growing computational resources paved the way to the possibility of learning accurate representations for a considerable number of objects in an unsupervised manner. We can learn reliable representations of objects, like images and natural language, with less supervision information. This gives us a chance to solve the problem of scalability and lack of supervision information.

In this dissertation, as an attempt to handle the scalability and lack of supervision, I have changed the second and third steps in the common practice of diversity measurement: learning representations of objects and designing diversity metrics based on representations, in which modern representation learning are incorporated. To avoid over general, I focus on two research domains: science of science and online social media, and demonstrate some ways to change the two steps.

Regarding the second step (representation learning), I have proposed a method in a discrete space to extract concepts and topics, and assign publications to topics in our science of science research. I also implement text embedding techniques to learn the embedding of publications and authors, which occurs in a continuous space. In addition, I implement several network embedding techniques in the research on online social media to represent users in a continuous embedding space.

Regarding the third step (diversity metric design), I summarize existing work on diversity metric design and propose a new metric in discrete space. The proposed metric is compared with existing ones, and it is found that each has advantages in different aspects. In continuous space, I propose a simple, intuitive but effective metric to represent the diversity of objects and provide some theoretic analysis of its good properties.

Designing a diversity metric is just the important first step for diversity research. We cannot accept every metric proposed without conditions. We still need to take a further step to make the diversity measurement sound and reliable. However, although we can propose many metrics to measure diversity in discrete and continuous spaces, it is hard for us to judge which diversity metric should be used in different scenarios. If we can build a general testbed with multiple reliable criteria, we can provide clear guidelines when implementing diversity metrics. This leads to another challenge that arises in the diversity metric design: How can we evaluate the reliability of a diversity measurement?

In the real world, there are very few datasets that can be used as gold standard to indicate the diversity of objects, especially in domains such as the science of science and social media research. We, unfortunately, do not have solid methods to test the performance of different metrics using ground truth, such as calculating precision and recall in a classification task or mean square error in regression. As an alternative, I propose a series of axiomatic analyses of proposed and existing diversity metrics in this dissertation.

In my axiomatic analysis, I propose some axioms that a good diversity metric should follow. If a diversity metric can satisfy most axioms, we can argue the soundness of the metric to some extent. The diversity metrics are defined in both discrete and continuous spaces. So are the axioms. In discrete space, I propose five axioms, such as "if adding a new object into a new category, the diversity metric should increase." I have run several simulation experiments to test existing and proposed diversity metrics. In continuous space, I also propose five axioms which are similar to those in discrete space. Additionally, I provide some mathematical proofs to illustrate why some metrics can meet all the axioms when others fail on some axiomatic testing.

The axiomatic analysis of diversity metrics sheds light on the pros and cons of different diversity metrics, but the results are only theoretical. We then encounter another challenge in diversity research in the context of real application: whether the diversity defined through this theoretical analysis can make some difference in real-world applications. Beyond the solid theory, we need more evidence from empirical studies to validate the effect of diversity metrics.

In this dissertation, I have conducted several studies in the domain of science of science and online social media to verify the implications of diversity. Some empirical results have revealed the predictive power of diversity metrics. I illustrate some interesting findings regarding the effect of research diversity in science of science and structural diversity in online social media.

Research diversity is one of the pivotal topics in the research in science of science. There is a very furious debate on whether publishing papers in multiple disciplines will bring researchers more benefits or not. The diversity metrics referred to in the debate are not consistent. I, instead, utilize the metric I choose based on the results of axiomatic analysis to explore the influence of research diversity. The results show that research diversity has strong predictive power regarding the increase in research impact. Diversity variation patterns, which indicate researchers' publishing strategies, differ a lot among researchers and lead to variation in research reputations in the end.

Structural diversity for users in online social media indicates the bandwidth of friends from different sources. It has proved to be an important factor in research on information propagation and user communication [108]. In my research on online social networks, I demonstrate the complicated relationship between online users' structural diversity and their engagements in social media: high friendship structural diversity leads to more broadcasting behavior and decreases narrowcasting behavior. A novel graph generation model is proposed to simulate the growth pattern of online social networks, which captures the variation in structural diversity for social media users. It is found that growth in structural diversity follows a "rise-fall-rise" pattern, which is named "open-closed-open." The pattern appears in both static and dynamic structural diversity in one of the leading online social media platforms, Snapchat.

The dissertation covers the discussion of diversity metric design in both discrete and continuous space. The axiomatic analysis of these metrics shows that no single metric can outperform others under all constraints in discrete space. The average distance, however, can meet all of the proposed axioms in the continuous space. The

	Chapter III	Chapter IV	Chapter V
Axiomatic Analysis	✓	✓	
Representation Space	discrete	continuous	
Social Implications	research diversity		structural diversity

Table 1.1: Coverage of chapters

real world applications, on the other hand, illustrate the complexity of the social implications of diversity metrics. Research diversity has a positive correlation with the research impact and can be used in some prediction tasks. The structural diversity variation pattern indicates online users' friending strategies and it has opposite relationships with online broadcasting and narrowcasting behaviors. In general, the theoretical and empirical analyses within this dissertation shed light on the importance of understanding diversity and utilizing it in reasonable ways.

1.1 Dissertation Outline

This dissertation contains both theoretical and empirical studies with different types of diversity in two applications. Table 1.1 summarizes the scope of the three studies, which are included in three chapters.

Chapter II offers a general literature review that introduces related work in axiomatic methodology and research regarding computational diversity measurements. I will analyze the advantages and limits of axiomatic analysis to show why it is important in diversity measurement design. I also distinguish the difference between supervised and unsupervised settings and illustrate the challenge we face in measurement design in large-scale high-dimensional data.

Chapter III introduces my work on diversity measurement in discrete space and its axiomatic analysis. It compares different metrics within the same axiomatic framework and demonstrates the pros and cons of the metrics. I also disclose some initial empirical results of research diversity in real world.

In Chapter IV, I move on to diversity measurement in continuous space and its implementation to find diversity variation in scientific communities. The axiomatic analysis helps us select one simple but effective diversity metric defined in continuous space. In this chapter, I also explore the effect of research diversity on research impact and analyze the research trajectories for different researchers.

Finally, I extend the diversity research from science of science to online social media in Chapter V. I track the variation in social structural diversity for online users and propose generative models to simulate users' diversity-aware friending strategies. The variation is validated in both static and dynamic networks and a graph generation model is proposed to capture this variation. I also unveil some relationships between diversity and user engagements to illustrate the real application of diversity metrics.

I summarize my major findings and propose several research directions for the future within the last chapter.

CHAPTER II

Preliminaries

Diversity and its effects have been studied thoroughly by researchers. In this dissertation, I will not include all of the domains that are related to the study of diversity measurement and evaluation. Instead, I will primarily summarize existing research regarding axiomatic analysis as a research methodology and distinguish between supervised and unsupervised diversity measurements in data science. Other helpful preliminaries and related work will be covered in other chapters. Research related to diversity metric design will be covered in Chapter III and Chapter V. Unsupervised representation techniques, such topic modeling, text embedding, and graph representation learning will be discussed in Chapter III, Chapter IV, and Chapter V separately. I will also summarize the effect of diversity in the domain of science of science and online social network research in the remaining chapters.

2.1 Axiomatic Methodology

Axiomatic analysis (or the axiomatic approach) is a research method used to evaluate the reliability of measurements. Researchers usually propose several axioms that an ideal metric should follow. The axioms usually come from common sense in real applications and findings from existing research. Both mathematical proofs and simulation experiments can evaluate how likely it is that a metric can follow the axioms. Comparing different metrics on multiple axioms can shed light on how reliable the metrics are.

Axiomatic analysis and the axiomatic approach have a long history of use. One of the earliest implementations of axiomatic analysis is the axiomatic system of utility functions developed in 1950s, which was published in *Econometrica* [46]. The authors propose a series of axioms, which are combined with axioms proposed by other economists, to argue that a good utility function must meet certain criteria. Another important axiomatic analysis in economics resides in the measurement of accessibility. Weibull proposes a new accessibility measurement which meets some mathematical constraints in different regions [120].

The axiomatic method has been used extensively in later studies in many disciplines, especially in data science. A well-known implementation of axiomatic analysis is Kleinberg's impossibility theorem of clustering algorithms [58]. Kleinberg proves that it is impossible to create a perfect clustering algorithm to meet three constraints at the same time. Researchers also leverage the method in various tasks like image interpolation [22], capital allocation [53], cohesiveness measurements [2], and recommender system [109].

Axiomatic analysis is suitable for metric design tasks, especially the task of information retrieval. Researchers in information retrieval utilize this method to design solid retrieval functions and evaluation metrics. Fang et al. have argued for a few axioms in retrieval function design, such as "adding a query into a document should improve the relevance," and proposed a justified retrieval function based on existing ones which outperform other metrics in retrieval tasks [37]. Gollapudi et al. have proposed an axiomatic framework to evaluate the diversification of search results and claim that there is no single retrieval function that can satisfy all of the axioms [40]. Researchers have also extended the axiomatic approach to the evaluation of retrieval functions. They have built an axiomatic framework to evaluate a retrieval function in terms of both novelty and diversity [3].

Axiomatic analysis methods are also incorporated in the domains covered by this dissertation, science of science and online social media research. Researchers in science of science also embrace axiomatic analysis when they design metrics to quantify authors' research impact. Since the H-index only counts the ranking of citations and ignores the numbers of citations, researchers have proposed an alternative index, the g-index [123]. The new metric can satisfy some simple axioms like "adding a paper with 0 citations should not improve the value of the index." It performs better than the H-index from many perspectives. The axiomatic approach also extends to research on online social media. Nilforoshan et al. propose a method to measure suspicious behavior in a social network, which meets many criteria in graph mining [81].

Some previous studies have illustrated how to design a metric for evaluating diversity. Researchers have defined diversity metrics based on different genres of data: sets [33, 17, 13], ranking [3, 40], and information network [79]. The metrics are used in different applications, such as information retrieval [3, 40] and option similarity measurement [17]. The researchers have proposed many axioms that a good diversity metric should satisfy: symmetry[79], redundancy [3], scale invariance [40], monotonicity [17], scaling [33], recursive property [63], etc. Some of the axioms mentioned in these studies are adopt in Chapter III and Chapter IV. I will describe the rationale to pick appropriate axioms based on those metrics in the next two chapters.

Axiomatic analysis is very helpful when evaluating metrics and exploring the effectiveness of novel algorithms. However, it is worth noting that axiomatic analysis only provides a "lower bound" for a good metric. A metric could still be suboptimal even if it satisfies all of the proposed axioms perfectly.

2.2 Supervised Diversity Measurement and Applications

Evaluating diversity and exploring the effect of diversity are very important in different scenarios of our social life. As I mentioned in the introduction, this usually follows a paradigm consisting of three steps: collecting information about objects, representing data objects, and designing and calculating diversity metrics.

This paradigm has incubated many influential research and social implications that have changed our society deeply. In biology, scientists invented the Linnaean taxonomy in the 18th century and defined biodiversity based on a set of taxonomies [75]. Biodiversity measurement helps us understand the state of an ecosystem and build environmentally friendly living spaces. Racial classification evolved over many years [90]. Sociologists defined diversity in society and organizations and found that it has a huge effect on efficiency, justice, and fairness [28].

The supervised method of defining and calculating diversity follows a top-down paradigm. It requires domain knowledge from experts to design categories and assign objects into categories. This method is not doable in many scenarios in the big data era.

When users generate large amounts of data, such as posting many tweets or uploading many photos to their social media accounts, it is nearly impossible to organize all of them through designing classifications and recruiting people to assign the resources to categories in classifications. We need to leverage some unsupervised method to manage the objects and compute their diversity. We need to either extract categories from data and design a method to assign objects into different categories or represent each object in a shared space and calculate the diversity based on the representation directly. I propose methods that follow these new paradigms in this dissertation that will save costs in the process of codebook design and object annotations when we face large-scale data.

2.3 Unsupervised Diversity Measurement and Applications

Different from the traditional supervised diversity measurement, unsupervised diversity measurement does not apply the top-down paradigm. Instead of utilizing existing classifications, researchers prefer to follow a bottom-up paradigm to extract important information for data directly. I will primarily focus on research about measuring diversity in unsupervised manner.

There is some research that focuses on defining and promoting diversity in different tasks, especially generation tasks. Researchers have added diversity-aware terms into objective functions and algorithms, like mutual information between words and pixel difference between pictures [72, 98]. But the actual meaning of this diversity metric is usually vague and hard to interpret. These terms usually aim toward a general idea of generating "different" objects in a particular space.

Besides embedding diversity as a part of an objective function, scientists define unsupervised diversity directly in many tasks. However, the definition and implementation of unsupervised diversity measurement varies greatly in data science. For example, in a QA system, researchers want the system to provide diverse replies to end users since the answer "I don't know" usually has a high probability of being selected. Although this answer is "correct" and safe, it is useless for solving the real problem. Therefore, a metric like the number of different answers generated by the system is sometimes defined as diversity[54].

In tasks like image captioning, researchers want to generate diverse captions for given images. Therefore, the number of objects described or number of unique words mentioned in the generated captions has become a method to evaluate the diversity of generated captions [49]. Similar metrics appear in other tasks. For example, counting the number of distinct n-grams when evaluating the diversity of generated texts [72, 114, 98, 73], counting the number of modes that appeared in generated images [39], and counting the number of groups joined by a particular item [89].

There are more advanced metrics to evaluate diversity based on these simple metrics, such as estimated parameters of a Zipf distribution fitted to generated text [20], variances of captions from different models across images [71], and KL-divergence between the model and ground truth distribution [73].

However, there is no study that thoroughly analyzes why we should use these diversity metrics when we have a collection of generated data. Instead, some researchers ask human annotators to rate the diversity directly based on the generated text or images [124, 43], which is costly to scale up. The lack of detailed analysis of the reliability of diversity metrics motivates us to incorporate other methods, like axiomatic analysis, into our measurement design. We can only draw conclusions about the effect of diversity based on reliable metrics verified by reasonable requirements. The various social implications of diversity can then rest on solid common ground.

CHAPTER III

Axiomatically Measuring Research Diversity in Scientific Communities

The preliminaries section shed light on the importance of defining a sound diversity metric based on large-scale high-dimensional data without much supervised information. In this chapter, I focus on a real application, measuring research diversity for researchers, as an initial exploration within this complicated research topic.

Since many existing metrics to evaluate scholars consider their scientific impact without considering the importance of diversity of researchers' work, I define a new metric for research diversity, based on the existing generalized Stirling metric, in discrete space that considers multiple aspects. I extract research topics in computer science using concept extraction and clustering from the literature in the ACM dataset. I then assign authors a distribution over these research topics, from which I calculate scores of research diversity for each author. To evaluate the reliability of diversity metrics, I propose five axioms that a sound diversity metric should satisfy and design corresponding simulation experiments to evaluate the ability of diversity metrics to follow the axioms. The results show how these metrics perform in different experiments, concluding that no metric consistently outperforms the others. I briefly test the relationship between our proposed metric and scientific impact and find a weak correlation between them. Finally, I demonstrate that the variation in the metric over time illustrates a possible publication pattern for scholars.

3.1 Introduction

With the development of research communities, different disciplines of research arose in the past century. Researchers in their own domains propose ideas and publish papers to advance the human knowledge. Meanwhile, an increasing number of scholars are engaged in interdisciplinary research [88, 115]. Some of this is due to the emergence of new scholarly "disciplines" that are inherently multidisciplinary such as information science, while some arises from scientific problems such as climate change that require expertise from multiple fields.

At the same time, scholarly impact and influence continue, by and large, to be measured by indices that ignore research diversity and may even penalize scholars who diversify their research portfolios. For example, the H-index, which is used extensively to measure scholarly impact, and which has been criticized for its limited focus [121], may be unfair when comparing scholars with different diversity of research interests. Researchers may not publish many papers within a single track but build bridges between research communities. Beyond just counting publications and citations, a metric or a set of metrics is needed that accounts for research diversity, so that research diversity can be measured and be included in an evaluation system of scholars' scientific influence.

In this chapter, I describe my research that explores the area of scholarly impact metrics and research diversity. The contributions of this work are as follows. I design a new metric to measure scholars' research diversity, called breadth of research, that builds on traditional metrics. I develop a multi-stage method for extracting topics from a corpus (in our case computer science papers) and calculate the scores of research diversity for authors who have published computer science conference papers. I design five simulation experiments that compare the relative performance of existing metrics and my new metric for measuring research diversity. I measure the relationship between research diversity and the H-index for scholars who are authors in the corpus. Finally, I explore the variation in research diversity for scholars over time to observe their paper publication behavior over their careers.

3.2 Related Work

There is a variety of existing literature relevant to the area of research diversity or degree of interdisciplinarity. The areas covered by this literature include topic extraction, topic relationship extraction, metrics design and the relationships between different aspects of research evaluation systems.

3.2.1 Topic Modeling for Scientific Literature

There are many methods that can be used to associate topics with publications. The simplest one is to use the classification codes in a dataset, such as ISI subject categories in Web of Science, as the set of topics. But these categories are too coarse-grained and hide intra-disciplinary variability. Another method is to use unsupervised learning algorithms to extract some topics according to the content of papers or the citation network of papers. Topic modeling [14] is one of the popular unsupervised learning algorithms based on content of papers. This model has been used to identify the disciplines that comprise interdisciplinary work funded by the NSF [80]. Researchers also adapted topic models as the ACT model (author-conference-topic) [70] for academic literature clustering purposes. Another family of approaches is to use community detection in networks as a basis for finding topics. One example is the use of two-round clustering [94] over the citation network to extract topic-associated communities [112]. There are other methods to combine both the citation network and the word distribution of abstracts [51] to find temporally ordered topics from a corpus of scientific literature, such as the ACM dataset.

Understanding the relationship between topics is also an important step after topic extraction, because calculating the similarity of topics is necessary for understanding research diversity. Some researchers have extracted relationships and used information visualization techniques to represent the relationship between different topics. For example, Yan detects the path between different disciplines to find the evolution of some areas [125]]. Another paper describes a new method to find the diversity subgraph in a multidisciplinary scientific collaboration network [45]. An interesting visualization method leverages the circle of science to visualize the relationship between disciplines in one dimension [18].

3.2.2 Measuring Research Impact and Research Diversity

Many metrics have been designed to measure factors related to scientific impact. The most common metrics are the impact factor and the H-index, which measure the number of citations of scholars' papers. Although these metrics have many problems, such as lack of universality between different disciplines [56], they are still widely used in systems like Google Scholar. Some alternative metrics also use the number of citations to measure the scientific influence of scholars [95]. They offer advantages over simple metrics such as the H-index, but they also focus solely on the citation count of papers. Other metrics based on the centrality of scholars in a network (e.g., co-authorship) like PageRank and betweeness centrality [16] are also widely used. However, the correspondence of centrality to actual influence is unknown.

As mentioned earlier, commonly used metrics of scholarly influence fail to consider the breadth of scholars' research. In response, a number of researchers have created some metrics for the degree of interdisciplinarity and more generally research diversity. The report of quantitative metrics and context in interdisciplinary scientific research [115] is a good survey for metrics for interdisciplinarity. Specialization and integration [92, 91] relate to diversity, coherence and intermediation. They define diversity as a combination of variety, balance and disparity. Coherence means the strength of links between different disciplines. Intermediation is based on network structure and is measured by betweeness centrality, clustering coefficient and average similarity. Other papers describe metrics based on these dimensions. Cassi et al. [23] divide the Stirling metric into a “within component” and a “between component” to measure the diversity of articles. Jensen et al. [50] propose six indicators based on the dimensions and measure the research diversity at two levels (article and laboratory). Karlovcec at al. [55] defines a new diversity metric based on Generalized Stirling. This metric incorporates connectedness of the citation graph into the original metric and applies it in exploratory analysis of the research community. Roessner et al. [93] validate the interdisciplinarity metrics with ethnographic materials (field observations and unstructured interviews).

3.2.3 Influence of Research Diversity

Some research has focused on the relationship between research diversity and other factors considered in scientometrics (not just scientific influence). One interesting paper finds that papers with an average degree of interdisciplinarity will get a higher impact rating than papers with too high or too low a degree of interdisciplinarity [103]. The results are convincing but the metrics used in this paper are quite simple (Jaccard similarity and cosine similarity). Two papers find that interdisciplinary papers have potentially lower impact than more focused papers. One of them finds that multidisciplinary papers are not frequently cited, in contrast to single-discipline papers [69]. The other explains how high-ranked journals suppress interdisciplinary research [92]. Other papers describe some factors that can encourage researchers to be involved in interdisciplinary research[21, 110]. They provide some theories to explain why scholars choose interdisciplinary projects. Some findings support that there are no correlations between citation ranks and ranked interdisciplinarity indices [87]. In contrast, other researchers confirm that the degree of interdisciplinarity is strongly correlated with the impact factor [100].

3.3 Diversity Measurement

The key question in the research on research diversity is how to measure diversity based on various scientific literatures. As mentioned in the section describing related work, many metrics have been used to measure the "degree of interdisplinarity." Compared to previous metrics to measure research diversity, we design a new metric that considers the topic distribution, similarity distribution and coherence within research topics.

3.3.1 Summary of Existing Measurements

There are many measurements of diversity or interdisciplinary, like entropy [119], Simpson's index [101] and generalized Stirling [104]. Each of these is computed as follows. Denote p_i as the probability of topic distribution for an author over topic i, and d_{ij} as the distance between topic i and topic j.

$$\begin{aligned} Entropy &= \sum_{i=1}^{n} -p_i \times \log p_i \\ Simpson &= 1 - \sum_{i=1}^{n} p_i^2 \\ GS &= \sum_{i,j} d_{ij}^{\alpha} \times (p_i p_j)^{\beta} \end{aligned} \tag{3.1}$$

Comparing them, only generalized Stirling considers not only the distribution of topics but also the similarity between topics. The further the distance between topics about which an author publishes papers, the more diverse will the author's research interest be. However, the traditional metrics do not consider the notion of differing coherence between different research topics. The degrees of influence of topics with small proportions are very limited. I propose a modified version of the generalized Stirling metric that incorporates the coherence of topics and value of minor topics (topics with small proportions).

3.3.2 Proposed Measurement

The new metric for research diversity, called BOR (breadth of research), is defined as follows: Denote d_{ij} as the distance between two topics, which is defined as the average distance (inverse of similarity defined above) between terms in the two topics, p_i as the probability of an author's paper belong to topic i, and coh_i as the coherence of topic i. The coherence of each topic is based on the proportion of authors for whom the respective topic is their major research topic.

$$BOR = \sum_{i,j} d_{ij}^{\alpha} \times (p_i + p_j)^{\beta} \times (Coh_i \times Coh_j)^{\gamma} \tag{3.2}$$

I modify the product of p_i and p_j in generalized Stirling to the summation of p_i and p_j because the summation will give minor topics more chances to be counted into the measurement of research diversity.

I add the coherence term into the metric because different topics have different "density" within themselves. For example, some topics, like digital library, are less coherent topics because there are many diverse subtopics within them. But for topics like operation systems, researchers concentrate on several narrow subtopics. A researcher focusing on digital library should have greater research diversity than operating systems researchers if other variables are controlled (so the gamma should have a negative value).

The new metric leverages properties of papers (topic distribution), properties of topics (coherence) and properties of relationships (topic similarity). The tuneable parameters give the metric more flexibility to balance different aspects of research diversity.

3.4 Axiomatic System

There is no established standard for determining the quality of metrics of research diversity. Furthermore, there is no ground truth to show the rankings of scholars' research diversity with which to validate the various metrics. I design an alternative evaluation method based on a set of axioms concerning research diversity and then test how the metrics perform according to these axioms. I propose five axioms that a good metric of research diversity should follow.

In addition to the definition of d_{ij} and coh_i, defined in the previous section, the following definitions relate to the axioms.

Notation:

- A_i: article A; $C = \{A_1, A_2, ..\}$: a collection of articles. N_C: number of articles in collection C
- t_i: topic i; $D_A(t)$: topic distribution of article A over topic t. $\sum_t D_A(t) = 1$
- $D_C(t)$: topic distribution of collection C over topic t. $D_C(t) = \frac{1}{N_C} \sum_{A_i \in C} D_{A_i}(t)$ $\sum_t D_C(t) = 1$
- d_{ij}: distance between topic i and topic j
- coh_i: the degree of cohesiveness in topic i
- $score(C)$: diversity score of the collection C

Axiom 3.1. ***Add to Old Topic***: *If an author publishes a paper on a topic on which she has published many papers before, her research diversity should decrease.*

Choose t, s.t. $t = \arg\max_t D_C(t)$, construct a new article A_n, s.t. $D_{A_n}(t) = 1$. $C' = C \cup \{A_n\}$. $score(C') < score(C)$

Some researchers suggest that diversity should increase with new publications no matter how close their new study is to their current research agenda [17]. In contrast, I propose an axiom to test the possibility of the decrease of diversity. Since every new publication costs some "capitals", I plan to give some penalty in terms of research diversity if researchers only publish in their familiar topics. The decrease of diversity is not necessarily a bad practice, but it can illustrate the tradeoff between depth and breadth when choosing research topics.

Axiom 3.2. ***Add to New Topics***: *If an author publishes a paper on a new topic on which she has never published, her research diversity should increase. Choose* t*, s.t.* $D_C(t) = 0$*, construct a new article* A_n*, s.t.* $D_{A_n}(t) = 1$*.* $C' = C \cup \{A_n\}$*.* $score(C') > score(C)$

Similar to some "monotonicity" proposed in previous literature [17], I expect a new publication in a fresh new topic will increase the scores of diversity metrics, which reflect the nature of research diversity.

Axiom 3.3. ***Submodularity***: *If an author publishes papers on two new topics in a sequence, the increase in research diversity the second time should be smaller than the increase the first time. Choose* t_1 *and* t_2*, s.t.* $D_C(t_1) = 0$*,* $D_C(t2) = 0$ *,construct two new articles* A_{n1} *and* A_{n2}*, s.t.* $D_{A_n1}(t_1) = 1$*,* $D_{A_n2}(t_2) = 1$*.* $C' = C \cup \{A_{n1}\}$*,* $C'' = C' \cup \{A_{n2}\}$*.* $score(C') - score(C) > score(C'') - score(C')$

Inspired by [36], I propose this axiom to control the increase of diversity. Intuitively, the increase of diversity caused by the first attempt should be the highest since it breaks the comfort zone and explores a new domain. The publications in the same domain later may still result in increases of diversity but the increases should not be as high as that in previous attempts.

Axiom 3.4. ***Add to Close Topics***: *If an author publishes a paper on a new topic close to the author's research interest, the increase in her research diversity should be* *less than that of publishing a new paper on a randomly chosen new topic. Randomly Choose* t_1 *s.t.* $D_C(t_1) = 0$, *construct a new article* A_{n1}, *s.t.* $D_{A_{n1}}(t_1) = 1$. $C' = C \cup \{A_{n1}\}$. *Choose* t_2 *s.t.* $D_C(t2) = 0$ *and* $t_2 = \arg\min_t \inf_{t_0 \in \{t | D_c(t) > 0\}} d_{t_0 t_1}$ *Construct a new article* A_{n2}, *s.t.* $D_{A_{n2}}(t_2) = 1$, $C'' = C' \cup \{A_{n2}\}$. *Then score*$(C'') <$ *score*(C')

Axiom 4 intuitively tests whether a diversity metric is sensitive to the distance of topics. Research in topics that are far from each other is definitely more diverse than research in close topics.

Axiom 3.5. ***Add to Coherent Topics****: If an author publishes a paper on a new topic with high coherence, the improvement in her research diversity should be less than that of publishing a new paper on a randomly chosen topic. Randomly Choose* t_1 *s.t.* $D_C(t_1) = 0$, *construct a new article* A_{n1}, *s.t.* $D_(A_{n1})(t_1) = 1$. $C' = C \cup \{A_{n1}\}$. *Choose* t_2 *s.t.* $D_C(t2) = 0$ *and* $t_2 = \arg\max_t(coh_t)$. *Construct a new article* A_{n2}, *s.t.* $D_{A_{n2}}(t_2) = 1$, $C'' = C' \cup A_{n2}$. *Then score*$(C'') <$ *score*(C')

The relationship between topics captures the difference between research domains but it cannot cover the heterogeneity within topics. I introduce the concept of coherence, which is proposed in [23]. It indicates the relationship within research topics. High coherence in a topic means its within-diversity is low. The fifth axiom is designed according to this principle.

I implemented five simulation experiments based on the original dataset with 8,911 authors to test how the traditional metrics and our new metric conform to the axioms. The experimental settings and results are presented in the next section.

3.5 Experiment

To verify the soundness of the original and proposed metrics for research diversity, I propose methods to extract research topics in an unsupervised way from a scientific literature dataset. I extract topics and assign topics to different papers and authors, which provides a testbed for simulation experiments to test the soundness of different diversity metrics.

3.5.1 Dataset

I extract abstracts, full text, and other metadata from the ACM digital library for proceedings of major conferences in computer science. From these proceedings I select authors whose names are unambiguous and who have published at least five papers. The standard for unambiguity is whether using the full name as the query sent to Google Scholar returns only one researcher profile with the same name. I extract the citation numbers and H-indexes by crawling Google Scholar. Overall, I crawled H-indexes and citation numbers for 8,911 authors from Google Scholar in August 2014. I also used the Wikipedia dataset to extract important terms in computer science.

Both traditional metrics and the new metric designed in this paper require a distribution over different topics or areas for authors. In order to generate topic distributions, I leverage the text data in the papers of the ACM digital library and implement three steps to form distributions: dictionary extraction, topic extraction and author assignment.

3.5.2 Dictionary Extraction

How to define topics is the first problem to be solved in the topic extraction and assignment. In this work, I extract a dictionary of n-grams in computer science and cluster them into topics using the Affinity Propagation algorithm [38]. Three different sources of dictionaries are used in this chapter: n-grams that are frequently used in papers, n-grams that can be matched to their abbreviations in the papers, and entries in Wikipedia.

Dictionary extraction follows these steps:

1. Extract bigrams and trigrams that occur frequently in papers using a threshold

of more than 10 times for bigrams and more than 5 times for trigrams. The threshold helps to eliminate noisy grams with low frequency.

2. Extract grams from papers that conform to the pattern “n-grams (abbreviation),” e.g., machine learning (ML).

3. Get an intersection between the results of step 1 and step 2 (3,816 terms in total).

4. Build a network of entries in Wikipedia according to hyperlinks between them in the website.

5. Make use of grams in step 3 and search their neighbours in the network of Wikipedia terms. If their neighbours also occur frequently in papers (with frequency higher than the thresholds mentioned above), add the terms into the final dictionary (6,100 terms).

The top 5 bigrams and top 5 trigrams in the final dictionary are shown in Table 3.1.

Table 3.1: N-grams with top frequency

N-grams	Frequency
User interface	2372
Software development	2102
Programming language	2042
Software engineering	1988
Operating system	1761
Wireless sensor network	586
World wide web	467
Graphical user interface	305
Support vector machine	300
Discrete event simulation	287

3.5.3 Topic Extraction and Assignment

After extracting the dictionary, I count the co-occurrence measure for every pair of terms. I then calculate the similarity between different terms by:

$$Sim_{ij} = \log \frac{Cooccur_{ij} + 1}{Max(Coocur_{ij}) + 2} \tag{3.3}$$

The logarithm calculation makes the distribution of similarity more uniform and avoids the influence of outliers of co-occurrence numbers. I weight co-occurrences of terms in abstracts of papers more than those in full text based on the intuition that abstracts generally have a stronger "topic signal." Using the computed similarity matrix of terms, I then run Affinity Propagation to cluster together similar terms and choose an exemplar for every cluster. The benefits of Affinity Propagation are that the exemplars for every cluster provide a straightforward explanation of what these clusters are about. More than two hundred clusters, or topics, are generated. Here are two examples of the clustering results:

- **Exemplar**: digital library
  - Terms: citation analysis, citation index, community building, digital earth, digital library, digital library software, digital preservation, digital reference, discourse analysis, Dublin core.
- **Exemplar**: machine learning
  - Terms: active learning, adaptive control, Bayes classifier, belief propagation, clinical trial, computational learning theory, concept learning, conditional random field.

I then assign each paper a probabilistic assignment to the different topics according to their respective frequency of n-grams associated with the particular topic. Therefore, each paper will have a distribution over topics.

3.5.4 Author Assignment

Using the clusters of n-grams in computer science and the topic distributions for every paper, I assign authors into different topics according to their papers. Every author is represented by a distribution over topics, which are used to calculate scores of metrics. There does not exist a "gold standard" list of researchers that ranks breadth of research that we can use to evaluate how reasonable our topic assignments are. I list below some topic distributions for well-known computer scientists to demonstrate our assignments.

- John Koza

    1. genetic programming 0.567
    2. programming language 0.083
    3. knowledge base 0.063

- Peter Denning

    1. memory management 0.107
    2. computer systems 0.093
    3. information systems 0.050

- Eric Horvitz

    1. user interface 0.082
    2. information retrieval 0.067
    3. machine learning 0.051
    4. speech recognition 0.047

3.5.5 Simulation Results

I implemented five simulation experiments based on the original dataset with 8,911 authors to test how the traditional metrics and our new metric conform to the axioms. The simulation exactly follows what I said in the description of the axioms. For example, to test the performance of "Add a new Topic," I will create a new topic for each individual research profile and compare the values of diversity metrics before adding a new paper to the new topic. The percentages of the values that satisfy axioms are reported in Table 3.2.

The results of simulations (Table 3.2) show that entropy and Simpson's perform well in the first three axioms because they don't consider distances between topics and introduce less noise. Because every new topic will be regarded equally for these metrics, they cannot follow Axiom 4 and Axiom 5. Generalized Stirling and my metric perform reasonably well in Axiom 1 and Axiom 2, but worse than entropy and Simpson's. They perform relatively badly in Axiom 3 because relatively bad performance on publishing a paper on a new topic (Axiom 2) will aggregate when testing the performance of publishing two papers on two new topics. But they perform well in Axiom 4 because of the consideration of distances. In addition, I find that our metric performs better than generalized Stirling in Axiom 5, which means coherence of topics and greater weights on minor topics are beneficial when we consider variation of metrics when people publish on topics with different coherence levels.

	entropy	Simpson	GS	BOR
Old	0.99	0.99	0.97	0.88
New	0.89	0.97	0.86	0.86
Submodularity	0.97	0.94	0.50	0.50
Close	N/A	N/A	0.76	0.70
Coherence	N/A	N/A	0.54	0.62

Table 3.2: Probability that metrics satisfy the axioms

Table 3.3: Average probability of satisfying the axioms with different α

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
Axiom1	0.40	0.42	0.48	0.62
Axiom2	0.33	0.38	0.44	0.55
Axiom3	0.34	0.32	0.24	0.22
Axiom4	0.38	0.57	0.66	0.64
Axiom5	0.63	0.61	0.57	0.52

Table 3.4: Average probability of satisfying the axioms with different β

	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$
Axiom1	0.86	0.67	0.30	0.08
Axiom2	0.69	0.57	0.24	0.16
Axiom3	0.40	0.40	0.29	0.05
Axiom4	0.57	0.57	0.59	0.53
Axiom5	0.61	0.61	0.59	0.52

3.5.6 Parameter Sensitivity

The performance of the new metric is influenced by the value of parameters α, β and γ. I tested the performance of the new metric with different settings. The results are shown in Table 3.3, Table 3.4, and Table 3.5.

The tables show that the metric is very sensitive to the α, β and γ. In order to find the best parameter setting, I calculated the average performance over five different simulation experiments for every parameter setting. I selected the settings with the highest average performance and a minimum threshold of at least 0.5 in every experiment. The best setting for Generalized Stirling is α=2,β=0.3. The best setting for the new metric is α=1, β=0.5 and γ=-0.5. These are used in the comparison of metrics in Table 3.2.

Table 3.5: Average probability of satisfying the axioms with different γ

	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 10$	$\gamma = 100$
Axiom1	0.58	0.47	0.45	0.45
Axiom2	0.24	0.39	0.47	0.48
Axiom3	0.09	0.26	0.34	0.38
Axiom4	0.49	0.57	0.59	0.59
Axiom5	0.62	0.66	0.58	0.53

3.5.7 Analysis of Proposed Metric

One important modification of the metric is the replacement of product with summation in the second term of the metric. I test the effect of this. If we control the distance term and coherence term in the metric to be the same for every topic and set β=1, the metric using summation will definitely follow Axiom 2 but not follow Axiom 1 and Axiom 3, theoretically.

Let n represent the number of topics.

Axiom1: Add to Old Topics

Proof.

$$\begin{aligned} score(C) &= \sum_{i,j} d^{\alpha}(p_i + p_j)(coh \times coh)^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma} \\ &= \sum_{i,j} d^{\alpha}(p'_i + p'_j)(coh \times coh)^{\gamma} = score(C') \end{aligned} \tag{3.4}$$

□

Axiom2: Add to New Topics

Proof.

$$\begin{aligned} score(C) &= \sum_{i,j} d^{\alpha}(p_i + p_j)(coh \times coh)^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma} \\ &< \sum_{i,j} d^{\alpha}(p'_i + p'_j)(coh \times coh)^{\gamma} = (n)d^{\alpha}(coh)^{2\gamma} = score(C') \end{aligned} \tag{3.5}$$

□

Axiom3: Submodularity

Table 3.6: Comparison between metric with summation and production

Metric	Parameter setting	Old	New	Submodularity	Close	Coherence
BOR_m	α=0.1, β= 0.1, γ= -0.1	0.99	0.85	0.45	0.22	0.59
	α=100, β= 1, γ= -1	0.82	0.62	0.47	0.69	0.53
	α=1, β= 1, γ=-10	0.83	0.40	0.39	0.55	0.76
BOR	α=0.1, β=0.1 , γ=-0.1	0.97	0.89	0.45	0.22	0.59
	α=100, β=1 , γ=-1	0.69	0.69	0.50	0.69	0.55
	α=1, β=1, γ=-1	0.69	0.47	0.41	0.54	0.77

Proof.

$$\begin{aligned} score(C'') - score(C') &= (n+1)d^{\alpha}(coh)^{2\gamma} - (n)d^{\alpha}(coh)^{2\gamma} \\ &= (n)d^{\alpha}(coh)^{2\gamma} - (n-1)d^{\alpha}(coh)^{2\gamma} = score(C') - score(C) \end{aligned} \tag{3.6}$$

□

From the derivation above, the performance of the new metric in Axiom 1 and Axiom 3 should be worse than the metric using product. The performance of Axiom 2 should be better than the metric using product. So I construct a metric using product in the second term and compare its performance with the new metric's performance in different parameter settings.

$$BOR_m = \sum_{i,j} d_{ij}^{\alpha} \times (p_i \times p_j)^{\beta} \times (Coh_i \times Coh_j)^{\gamma} \tag{3.7}$$

The results in Table 3.6 show that the metric using summation outperforms product in Axiom 2, and the metric using product outperforms the metric using summation in Axiom1, which is consistent with the results of derivation. But the results for the other three axioms are close between the two metrics, which means the interaction between different terms in the metric (distance term, distribution term and coherence term) will influence the results of the simulation.

Table 3.7: Correlation between research diversity and H-index

	Pearson Corr.	Partial Corr.
Entropy v.s. H-index	-0.1722	-0.0769
Simpson v.s. H-index	0.2102	0.0922
GS v.s. H-index	0.4283	0.1820
BOR v.s. H-index	0.4337	0.1832

3.6 Effect of Research Diversity

I tested the Pearson correlation between metrics of research diversity and H-indexes of scholars. My results (Table 3.7) show that some metrics have a positive relationship with the H-index. Others have weak a negative relationship. Because publication numbers may influence the correlation between research diversity and scientific impact, i.e., the increase in numbers of publications may bring an increase in research diversity and an increase in the H-index simultaneously to make them positively correlated with each other, I test the partial correlations between metrics of research diversity and H-index controlling publication numbers (Table 3.7). They are weaker than the results of Pearson correlations, and none of the weak partial correlation scores illustrate a strong correlation between metrics for research diversity and H-index for scholars.

I also conduct a preliminary study to track the variation in research diversity of scientists. I illustrate in Figure 3.1 the average variation of metrics over publication years for scholars. Simpson's, generalized Stirling and the new metric initially increase and then level off, which explains a possible publication pattern of scholars: scholars' research diversity may increase with the increase of publications in the early stage of their career. But because of accumulation of publications, their accumulative research diversity will not change dramatically in later years.

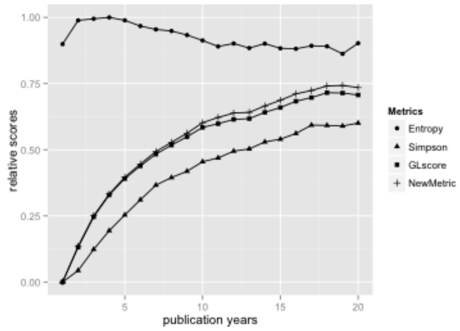


Figure 3.1: Variation of metrics over publication years

3.7 Summary

In this chapter, I describe a new metric based on generalized Stirling to evaluate research diversity for scholars in computer science. The metric makes use of topic distributions, similarity between topics, and coherence of topics, and it can capture research diversity. All of the information used to calculate research diversity is extracted automatically from the ACM Digital Library. The simulation experiments show that traditional metrics can perform well in some situations, but they do not perform well when coherence within topics and similarity between topics are considered. In contrast, the Generalized Stirling metric and the new metric, breadth of research, work better in the simulation related to similarity between topics and coherence but perform worse in the experiments involving adding new topics.

With the new metric for research diversity, I find that the correlations between research diversity and scientific metrics are weak, especially when I control publication numbers. From this study, there is no evidence to show whether the increase in research diversity will influence the impact of scholars' publications. Also, after testing the variation in the new metric over many years, I find a possible publication pattern of scholars: Research diversity increases in the beginning with the increase of publications. But it increases slowly once publications have accumulated.

CHAPTER IV

Research Diversity Measurement and Application in Continuous Space

The axiomatic analysis in the previous chapter shed light on the importance of a reliable definition of research diversity. In this chapter, I extend the axiomatic analysis to continuous space, in which modern representation learning can capture the relationship between objects accurately and efficiently.

The results of the axiomatic analysis yield a surprising result: the simple and intuitive diversity metric, average distance, performs well in axioms. I provide a set of proofs for this metric and conduct a study to leverage it to explore the complexity of the effects of research diversity. The time series analysis of research diversity trajectories reveals the heterogeneous variation in research diversity in researchers' early careers. I also go beyond the simple analysis of research diversity in the previous chapter and conduct a deep author-level regression analysis of the influence of research diversity on research impact. These studies illustrate a strong predictive power of researcher diversity in the early stage of careers for greater research reputation later.

4.1 Introduction

In the previous chapter, I explored metric design for research diversity in discrete space. I represent each object as a distribution of topics learned from existing data in an unsupervised manner. The axiomatic analysis of the metric design has revealed many important concerns to consider when measuring research diversity. The analysis has given us some insights into how to design a solid diversity metric in a discrete space.

The discrete space is easy to understand and interpret. However, the representation ability is limited in some circumstances. It is non-trivial to build connections between fine-granular classes in the discrete space, and it is hard to depict the relationship between different objects and the whole corpus because of its hard-coded nature.

Recent developments in representation learning in continuous space pave the way to a better mapping from objects to an embedded space, which provides a variety of possibilities for data exploration methods as future steps. Different from learning classes or leveraging existing classifications, representation learning, especially deep learning, can map objects to a high-dimensional embedding space based on the information derived from the objects, such as image, sound, and text. The success of learning representations results in good performance of downstream tasks using these representations, such as image classification [59], text generation [30], and recommender systems [27]. These representation techniques provide the possibility of representing publications in scientific literature in a continuous embedding space. I can find the relationships between publications in high-dimensional spaces using unsupervised methods and measure their relationships with a number of metrics designed for the continuous space.

With the representation of papers and authors in the continuous space, we face the challenge of designing a solid diversity metric once again. How to design a sound diversity metric in a continuous space is one of our important research questions. Similar to the practice I introduced in the previous chapter, I design another set of axioms to limit the possible metric design choices. However, I do not implement the simulation experiments like I did in the previous chapter. I provide some mathematical proofs for various metrics to explore whether they can satisfy the axioms correctly. One metric, as a result, can meet all of the defined axioms. This is the average distance between objects, which is impressively simple and intuitive in terms of its form. Compared with other metrics that are commonly used to measure diversity, average distance demonstrates good properties regarding the defined axioms.

Using the new metric designed in continuous space, I want to take a step further into two social implications in the science of science, which were simply studied in the previous chapter. The first is the relationship between research impact and research diversity. The existing discussion about these two factors has shed light on the complexity of the dynamics. There is evidence to support very contradictory results: a positive relationship [68, 102], a negative relationship [69], a neutral relationship [87, 97, 1], and a reversed-U-shaped relationship [61]. I control many variables within the scientific literature and find that the research diversity can be useful for predicting research impacts through a regression analysis.

The other research topic I am interested in is the variations in scholars' research diversity. In the previous chapter, I found that research diversity increases when researchers publish more and more papers. I am interested in more complex research questions such as when people will broaden or narrow down their research interests, and what will happen to research diversity when researchers graduate from doctoral programs, move to different research communities, or get tenure.

Based on the findings from the axiomatic analysis of research diversity in continuous space, I define the diversity metric and leverage it to measure research diversity. Researcher's diversity is represented through time-series for each author and conference. A detailed analysis of these time-series illustrate very different choices for authors and different trends in conferences. The results shows that on increase in diversity in researchers' early stages can serve as a good indicator for their success later.

In the next several sections, I will summarize some existing findings about representation learning and the effects of research diversity, followed by an extensive axiomatic analysis. The author-level regression and trajectory clustering analysis are described based on the results of the axiomatic analysis.

4.2 Related Work

Important research work regarding the axiomatic analysis of metrics was elaborated in Chapter II. In addition, I have summarized most research regarding the design of metrics and effects of research diversity in Chapter III. I will focus on techniques for learning continuous representation of researchers' profiles, especially the text embedding and language models invented in recent years. I will implement these embedding techniques in our proposed studies and calculate the research diversity for individual researchers. I also unfold a substantial amount of research regarding the effect of research diversity on research impact, which is a critical social implication that I focus on in this dissertation.

4.2.1 Text Representation Learning

The amount of data has been increasing explosively in the last two decades. This results in difficulties in getting large-scale labeled data at low costs. It is also not possible to design a pre-defined classification to manage all of the emerging user-generated data.

Many topic modeling techniques have been invented by researchers to cluster documents and represent documents as distributions over learned topics, such as pLSA [47] and LDA [14]. I implemented some techniques within this line of research in the previous chapter.

Researchers are not satisfied with the discrete nature of topic modeling. An unsupervised learning method that can represent documents in continuous space is more ideal for many downstream tasks. Since 2013, scholars have invented a series of techniques called "text embedding." Based on the co-occurrence relationships between words, word embedding techniques can map texts (which could include words, phrases, sentences, or documents) into a high-dimensional continuous space. Word2vec [78] and GloVe [84] are the earliest and most widely used models in this line of research. Many subsequent studies have improved the performance and training efficiency of word embedding, such as fasttext [15]. Others extend the techniques to other objects, such as embedding a whole document into a continuous space, like doc2vec [64].

Another important line of representation learning research is the deep language model. Different from modeling embedding directly, language models can adapt various objective functions and model the probabilistic relationship between tokens. Among all of the language models, the models based on bi-directional LSTM [86], attention mechanism [111], and masked token prediction tasks [30] achieve the best performance. Much other follow-up work, such as XLNet [127], has made the model more complicated and accurate. In particular, researchers have also leveraged more than one million scientific publications and obtained a pre-trained language model for scientific literature [11]. I will utilize this large-scale language model in my following analysis in the domain of science of science.

The development of representation techniques has progressed rapidly in recent years. I adopted the widely accepted technique, doc2vec, in this study. I will see whether better representation can lead to novel findings regarding the effect of research diversity.

4.2.2 Effect of Research Diversity

When scholars choose the venues in which they will publish, they face a common choice: publish papers in different domains and seek broad research impact or focus on a single research domain and make more hard-core contributions. Researchers are interested in whether different strategies result in different research impacts on scientific communities.

There is a considerable amount of research regarding the complicated dynamics between research diversity and research impact. No consensus has been reached among researchers. The effect of research diversity varies in different domains and at different stages of an academic career.

Several studies have supported the positive effects of research diversity or degree of interdisciplinarity on research impact, which is usually represented as citations. Levitt et al.'s research reveals some corroborative evidence to show that interdisciplinary research has higher citation ratings than research that focuses on single disciplines [68]. Researchers also found a positive correlation between research diversity or degree of interdisciplinarity and research impact for both single papers and journals [102, 100]. Larivinere et al. explored relationships between papers with high research diversity. They found that there usually exists a win-win relationship between interdisciplinary paper pairs [62]. Chen et al. even extend this line of research to broader domains and draw a strong conclusion regarding the positive effect of interdisciplinary research in science development, especially in natural sciences and engineering [26]. Additionally, they find that these positive effects are more common in highly cited papers (like the top 1% of highly cited papers) [25].

Although the existing evidence has shown a positive relationship, many researchers also find a negative effect of interdisciplinarity within some particular domains. For example, within one study of papers in Web of Science and Scopus, researchers find that the average citation counts of papers are very similar in Mono and Multi discipline scenarios. Furthermore, in life science, health science and physical science, the monodisciplinary papers have roughly twice the average citation counts as the multidisciplinary cases [69]. Researchers also claim that research in Dutch physics gets more citations when it is not in interdisciplinary programs. These studies focus on some natural science disciplines and prove that the interdisciplinarity is not always helpful across all domains.

Besides the mixture of positive and negative effects, some researchers also hold the idea that there is no significant relationship between research diversity and research impact. Studies like [87, 97, 1] find no significant effect of research diversity using various metrics, like the Simpson index and Shannon index. Wang et al. and Yegros et al. argued that some diversity metrics have a negative effect on citations whereas some other metrics have an opposite effect [117, 128].

Beyond the positive, negative, and neutral effects, the complexity of research diversity is further revealed by a few researchers. They find that the relationship between research diversity and research impact has a reversed-U shape i.e., highly cited papers are usually not very diverse or very narrow. Papers with a medium level of interdisciplinarity have the greatest advantage in getting cited [61].

All of these studies demonstrate the complexity of the dynamics between research diversity and research impact. Researchers have explored effects of different metrics in different datasets. My study goes beyond their choice of metrics to figure out whether a better representation of publications and metrics defined based on continuous representations provides more insights about the effects on research diversity. I also explore the dynamics of this research at different stages of researchers' careers.

4.3 Axiomatic System

Similar to what I implemented in the previous chapter, I propose five different axioms that a good diversity metric should follow. I limit the discussion of these axioms to continuous spaces, which is different from the discrete setting of the axioms in the last section. Some of the axioms are based on intuitions about the science of science, while others are discussed in previous studies.

4.3.1 Notations

In continuous space, objects are usually represented as data points within the space. I can represent the relationship between nodes using the distance between nodes. Ideally, a reliable metric should satisfy several axioms no matter what kind of distances are deployed in the system.

I define some preliminary variables using notations as follows:

- N: a set of nodes.
- $div(N)$: Diversity of N
- $n = |N|$: number of nodes in N
- $d(u, v)$: distance between u and v

I also define several metrics that are usually used in the continuous space:

- $avg(N) = \left(\sum_{u,v \in N} d(u, v)\right) \Big/ \binom{|N|}{2}$: average distance of pairs of nodes in N
- $L_{dis} = \max_{u,v} d(u, v)$: largest distance between all pairs of node s
- entropy $= -\int_x p(x) \ln p(x) dx$: the entropy of data points, where $p(x)$ is the probabilistic distribution of each node in the space.
- Gini-Simpson Index $= \int_x (1 - p(x)^2) dx$: gini index of data points, where $p(x)$ is the probabilistic distribution of each node in the space.
- Variance $= \sum_{i=1}^{n} \frac{1}{n-1} d(i, avg)^2$, the variation of nodes in the continuous space, where avg indicates the average representation of nodes

- $L_{disC} = \max_u d(u, C)$: max distance to the center of the nodes, where C represents the center.

4.3.2 Axioms and Metric Comparison

Based on our understanding of research diversity, considering the axioms I designed in the discrete setting and the axioms proposed by other diversity axiom studies[40, 33, 63], I propose five axioms that a diversity metric defined in continuous space should follow:

- Add to new topic: adding an object that is farther than any other objects, the diversity should increase
- Add to old topic: adding an object that is closer than any other objects, the diversity should decrease
- Add to closer topic: adding an object that is far from existing objects will increase the metric more than another object that is close to existing objects.
- Duplicate: copying all the nodes, the doubled set of objects should have lower diversity
- Submodularity: adding an object twice, the increase of metrics for the second addition will be smaller than the increase for the first one.

The strict mathematical definitions of these axioms are described in the next section.

These axioms are designed based on our application: measuring research diversity in research communities. Four of these axioms are aligned with the axioms proposed in the discrete space. I adapt them into continuous form. Since there is no clear definition of “topic” in the continuous space, I can not measure “coherence” of topics.

Instead, I propose an alternative axiom: “duplicate”. The rationale behind “duplicate” is similar to the axiom of “add to old topic”. If researchers have the access to more resources, duplicating the existing publications should not be encouraged from the perspective of diversity promotion. The researchers can expand their research to other places within the continuous space to improve their research diversity. Thus, I suggest that it is an appropriate practice to decrease the diversity when research profiles are self-duplicated.

The list of axioms is not exhaustive compared to axioms proposed by the few previous studies. However, not all the axioms proposed for diversity are suitable in this application. They are designed for other applications like evaluating the results of a retrieval system [40, 3] and similarity of options [17]. Some axioms are not very important for research diversity metrics since nearly all of the metrics will meet them. For example, the symmetry axiom proposed by Laxion[63] i.e., $D(a, b) = D(b, a)$, can be naturally satisfied by nearly all of the metrics. Continuity and Scaling, proposed in [33], have a similar problem. They cannot help decision makers to filter out good metrics.

A few other axioms may not fit the context of a diversity metric, such as Monotonicity [63, 13, 40]. Monotonicity means no matter what new object is added into the collection, the diversity should always increase. It is not desirable for research diversity. The diversity metric will not be able to measure the “narrowing down” of research topics if monotonicity is one of the constraints. In contrast, I propose the new topic and old topic axioms to depict the requirement of variation.

Another controversial choice of axiom appears in scale invariance v.s. the duplicate axiom. Some researchers insist that diversity should be scale-free [40], which means self-duplicating an author’s publication will not decrease its diversity. It ignores the effect of the number of publications but only considers the true “distribution” of research topics. However, I have seen the shortcoming of the discrete space represen-

Table 4.1: Axiom satisfaction for metrics

	Old	New	Close	Duplicate	Submodularity
Average Distance	✓	✓	✓	✓	✓
Largest Distance	✗	✓	✓	✗	✓
Continuous Entropy	✗	✓	✗	✗	✓
Gini-Simpson Index	✗	✓	✗	✗	✓
Variance	✗	✓	✗	✗	✓
L-Distance to Center	✗	✓	✗	✗	✓

tation and try to go beyond the simple idea of "distribution" of research topics. The scaling effect of the number of objects should be taken into account. Policy should not encourage researchers to repeat themselves when considering their research diversity.

With the carefully designed axioms, I evaluate whether the metrics mentioned in the previous section will satisfy these axioms. The results in Table 4.1 show that average distance is the only metric that satisfies all of the axioms. This metric has been used extensively in diversity measurement in domains like drug discovery [12, 34] and social network analysis [105]. It is impressive that the most intuitive and simple metric can actually beat other metrics. I will provide the proof for average distance in the next section.

The reasons why other metrics cannot meet the constraints vary. The largest distance is not very stable for extreme data points. The Gini-Simpson index and entropy ignore information about distance sometimes. As a result, when the distribution of data points is very sharp, it cannot satisfy some constraints. Variance is actually equal to average distance when the distance has a special form. However, they are not consistently equal to each other. Average Distance is invariant of distance choice so it can meet the constraints easily.

4.3.3 Proof of Axioms

In this section, I provide a series of proofs to validate that the simple and intuitive metric, average distance, can meet all of the constraints. In our proof, $div(N)$ is defined as $avg(N)$.

Axiom 4.1. *Add to new topic*

For v, $\forall u \in N, d(v,u) >= \max_{u' \in N} d(u',u) \implies div(N \cup \{v\}) >= div(N)$

Proof.

$$\begin{aligned} div(N) &= \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u')}{\binom{n}{2}} = \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u') + n \times avg(N)}{\frac{1}{2}n(n+1)} \\ &<= \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u') + \sum\limits_{u\in N} d(v,u)}{\frac{1}{2}n(n+1)} = div(N \cup \{v\}) \end{aligned} \tag{4.1}$$

□

Axiom 4.2. *Add to old topic*

For v, $\forall u \in N, d(v,u) <= \min_{u' \in N} d(u',u) \implies div(N \cup \{v\}) <= div(N)$

Proof.

$$\begin{aligned} div(N) &= \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u')}{\binom{n}{2}} = \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u') + n \times avg(N)}{\frac{1}{2}n(n+1)} \\ &>= \frac{\frac{1}{2}\sum_{u\in N}\sum_{u'\in N\backslash\{u\}} d(u,u') + \sum\limits_{u\in N} d(v,u)}{\frac{1}{2}n(n+1)} = div(N \cup \{v\}) \end{aligned} \tag{4.2}$$

□

Axiom 4.3. *Add to closer topic*

For node v *and* v‘, $\forall u \in N, d(v,u) <= d(v',u) \implies div(N \cup \{v\}) <= div(N \cup \{v'\})$

Proof.

$$\begin{aligned} div(N \cup \{v\}) &= \frac{\frac{1}{2}\sum_{u \in N}\sum_{u' \in N \setminus \{u\}} d(u,u') + \sum\limits_{u \in N} d(v,u)}{\frac{1}{2}n(n+1)} \\ &<= \frac{\frac{1}{2}\sum_{u \in N}\sum_{u' \in N \setminus \{u\}} d(u,u') + \sum\limits_{u \in N} d(v',u)}{\frac{1}{2}n(n+1)} = div(N \cup \{v'\}) \end{aligned} \tag{4.3}$$

□

Axiom 4.4. *Duplicate*

Create a set N' *as a copy of* N*,* $div(N \cup N') < div(N)$*, and* $\exists(u,v), d(u,v) > 0$

Proof.

$$div(N \cup N') = \frac{\frac{1}{2}\sum_{u \in N \cup N'}\sum_{u' \in N \cup N' \setminus \{u\}} d(u,u')}{\binom{2n}{2}} = \frac{\sum_{u \in N}\sum_{u' \in N \cup N' \setminus \{u\}} d(u,u')}{\binom{2n}{2}} \tag{4.4}$$

Since for each $u \in N$, there is a corresponding node $u' \in N'$ that takes the same place as u in the space, $d(u,u') = 0$.

$$\begin{aligned} div(N \cup N') &= \frac{\sum_{u \in N}\sum_{u' \in N \cup N' \setminus \{u\}} d(u,u')}{\binom{2n}{2}} = \frac{2\sum_{u \in N}\sum_{u' \in N \setminus \{u\}} d(u,u')}{\frac{1}{2}2n(2n-1)} \\ &< \frac{2\sum_{u \in N}\sum_{u' \in N \setminus \{u\}} d(u,u')}{\frac{1}{2}2n(2n-2)} = \frac{\frac{1}{2}\sum_{u \in N}\sum_{u' \in N \setminus \{u\}} d(u,u')}{\frac{1}{2}n(n-1)} = div(N) \end{aligned} \tag{4.5}$$

□

Axiom 4.5. *Submodularity*

For node v *and* v'*,* $\forall u \in N, d(v,u) = d(v',u)$*,* $\forall u \in N, d(v,u) >= \max\limits_{u' \in N} d(u',u) \implies div(N \cup \{v\}) - div(N) > div(N \cup \{v,v'\}) - div(N \cup \{v\})$

Proof. denote $|D'| = \sum\limits_{u \in N} d(v,u) > n(avgN)$

$$\begin{aligned}
&div(N \cup \{v\}) - div(N) - div(N \cup \{v, v'\}) + div(N \cup \{v\}) \\
&= 2\frac{\frac{1}{2}(n-1)n(avg(N)) + |D'|}{\frac{1}{2}(n+1)n} - \frac{\frac{1}{2}(n-1)n(avg(N)) + 2|D'|}{\frac{1}{2}(n+2)(n+1)} - avg(N) \\
&= \frac{(n+2)n(n-1)avg(N) + 2(n+2)|D'| - \frac{1}{2}n^2(n-1)avg(N) - 2n|D'|}{\frac{1}{2}(n+2)(n+1)n} \\
&- \frac{\frac{1}{2}(n+2)(n+1)n(avg(N))}{\frac{1}{2}(n+2)(n+1)n} \\
&> \frac{(n^2+n-2) + 4 - \frac{1}{2}(n^2-n) - \frac{1}{2}(n^2+3n+2)}{\frac{1}{2}(n+2)(n+1)} avg(N) \\
&= \frac{1}{\frac{1}{2}(n+2)(n+1)} avg(N) > 0
\end{aligned} \tag{4.6}$$

□

4.4 Author-Level Regression Analysis on Research Diversity

With the accurate representation of documents and the solid research diversity metric, I have the ability to extend the studies in Chapter III regarding the effect of research diversity. I have a chance to disentangle the complex relationship between research diversity and research impact. A causality analysis on the existing dataset would be very hard to conduct. I, instead, design a regression task to evaluate the predictive power of diversity. If research diversity can help predict variation in research reputation, it can be an important indicator to predict whether a researcher will be successful in advance.

4.4.1 Data Collection and Representation Learning

I have chose and processed dblp data from the Aminer website [1]. This dataset contains all the important publications in computer science conferences and journals until early 2018. This dataset includes the metadata of more than 3 million papers. I have constructed a collection of statistics for more than 392K authors after some

[1] https://www.aminer.cn/dataCitation

simple data cleaning steps. The major reason to select this dataset is that I have rich domain knowledge in computer science and I am able to conduct sanity checks easily.

With the collected metadata for papers, I train and calculate the research diversity based on continuous embedding learned from the data. The abstracts of papers are selected to train their document embedding. I adopt the doc2vec model [64], which is widely used and adopted in text representation learning, to train the representations for each document with the dimension as 300 and window size as 10.

4.4.2 Experiment Setting

The goal of this study is to test the predictive power of research diversity for academic impact, which can be influenced by many factors. Based on the accessibility of data within the AMiner dataset, I carefully choose the variables which have been tested effectively in previous literature [31, 83, 126]. This includes variables at the author level, coauthor level, and venue level. Author-level variables depict the reputations of authors, which include career age, paper number, citation number, and H-index. In addition, network-based metrics such as page rank for authors are included as well. Coauthor-level variables describe the reputation of authors' coauthors, which includes coauthors' paper number, citation number and H-index. Venue-level variables summarizes the citation patterns in the venues where authors publish their papers, including citation numbers and paper numbers. The details of these variables are summarized in Table 4.2

I calculated the values of these controlled variables and research diversity in the year of 2013. The dependent variable I want to predict is the difference in H-index between the year 2013 and the year 2017, i.e. $delta(H) = H_{2017} - H_{2013}$. This variable can measure the increase researchers' reputations over the next several years.

Table 4.2: List of controlled variables for the regression model

	Controlled Variables	Variable Meaning
author level	career_age	number of years since the first publication
	page_rank	page rank score for author in coauthor network
	paper_number	paper number
	citation_number_cu	citation number in last year
	citation_number_tot	total citation number
	h-index	h-index score for the author
	average_cit_num_year	average citation number per year
	average_paper_num_year	average paper number per year
coauthor level	avg_author_num	average author number of papers published by the author
	avg_author_paper_num_max	average maxium paper number of coauthors
	avg_author_avg_cit_num_max	average maxium citation number of coauthors
	avg_author_h_index_max	average h-index of coauthors
venue level	avg_venue_avg_citation	average citation number in the published venues
	avg_venue_avg_paper_num	average paper number in the published venues

Table 4.3: OLS regression results with Doc2Vec paper representation

	coef	**std err**	**P> |t|**	**[0.025**	**0.975]**
research_diversity	0.8647	0.017	0.000	0.831	0.899
page_rank	2.045e+04	1328.929	0.000	1.78e+04	2.31e+04
career_age	0.0085	0.000	0.000	0.008	0.009
paper_number	-0.0114	0.000	0.000	-0.012	-0.011
citation_number_cu	0.0754	0.001	0.000	0.074	0.077
citation_number_tot	-0.0137	0.000	0.000	-0.014	-0.013
h-index	-0.0115	0.001	0.000	-0.014	-0.009
average_cit_num_year	0.0080	0.000	0.000	0.008	0.008
average_paper_num_year	0.7669	0.003	0.000	0.761	0.773
avg_author_num	-0.0145	0.001	0.000	-0.016	-0.013
avg_author_paper_num_max	0.0005	9.41e-05	0.000	0.000	0.001
avg_author_avg_cit_num_max	0.0035	0.000	0.000	0.003	0.004
avg_author_h_index_max	0.0145	0.001	0.000	0.013	0.016
avg_venue_avg_citation	0.0187	0.000	0.000	0.018	0.019
avg_venue_avg_paper_num	0.0009	3.04e-05	0.000	0.001	0.001
intercept	-0.5671	0.011	0.000	-0.589	-0.545

R-squared:	0.438	**Adj. R-squared:**	0.438
F-statistic:	2.037e+04	**Prob (F-statistic):**	0.00
Log-Likelihood:	-5.5897e+05	**AIC:**	1.118e+06
BIC:	1.118e+06	**Kurtosis:**	20.951
Omnibus:	192058.554	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5475044.868
Skew:	1.784	**Prob(JB):**	0.00

4.4.3 Results

Table 4.3 illustrates the detailed results of regression analysis. The regression has a moderate adj. R^2 value, which indicates the fairly good fitness of this model. Different variables have heterogeneous impacts on the increase in H-index. Some accumulative controlled variables, such as paper number, citation number, and H-index have negative relationships since people who already have strong reputations may have difficulty increasing their H-index further. In contrast, other factors like coauthor-level factors and venue-level factors have positive relationships to the increase in H-index.

Research diversity has a positive coefficient with a p-value less than 0.001, which indicates the positive relationship between research diversity and the increase in H-index. In general, researchers with greater research diversity tends to gain stronger reputations, measured by the increase in H-index, in computer science..

I also conduct an F-test for comparing models with and without research diversity as a variable. The F-value is equal to 2,494, and p-value is less than 0.001. This again proves the effect of research diversity based on other controlled variables.

4.5 Research Diversity Trajectory Analysis

4.5.1 Trajectory Analysis of Research Diversity

Researchers have very different publishing strategies when they choose which venue to publish their research findings in. I track the trajectory of research diversity along the way since I am interested in how research diversity will change over time. I want to shed light on the changes in researchers' interests and figure out when people will broaden or narrow down their research interests, what will happen when researchers graduate from doctoral programs, move to different research communities, or get tenure. The accurate depiction of researchers' trajectories will be beneficial to researchers when they choose their publishing venues and will motivate policy makers to encourage different styles of research.

I first conduct some research trajectory analysis on individual authors. I define their annual research diversity as the average distance between papers every year and draw the variation in research diversity over time. As shown in Figure 4.1, even for famous researchers, variations in research diversity are very diverse. Some researchers like Jon Kleinberg (right) have increasing research diversity in their early career and keep the diversity high over time. For some other researchers, on the other hand, research diversity fluctuates over time, and it even decreases in Jure Leskovec's curve (left).

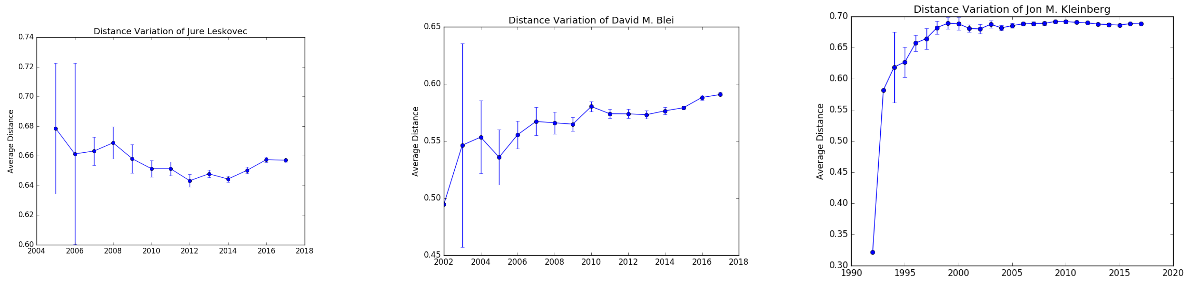


Figure 4.1: Variation of research diversity for researchers

I also have done some preliminary studies of the diversity at the conference level. I treat each conference just as an author with all of the papers in the corresponding domain. I calculate the variation in the research diversity of these research communities and find that the variations are quite different from each other. As shown in Figure 4.2, the research diversity of some data mining conferences like KDD (left) and the Web Conference (middle) keeps increasing over time with constant fluctuation. In contrast, conferences like SIGIR (right) keep their diversity at a constant level. These patterns reveal how the research communities define and change their research topics over time.

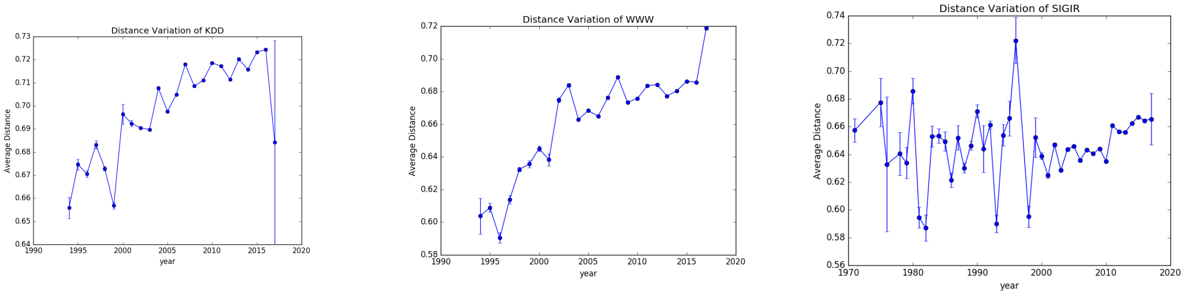


Figure 4.2: Research diversity for research communities

4.5.2 Researcher Trajectory Clustering

The different phenomena in research diversity change motivates us taking the next step to summarize researchers' patterns. Beyond individual researchers or conferences, I represent researchers' trajectories as time series and explore the patterns in the variation of research diversity.

I select authors with at least ten years in their careers for the study and pick 5,000 of them as the training data. The first ten years of variation in research diversity is represented as a set of time series. They are normalized using the Mean Variance method to remove the effect of time-series scales. DTW (Dynamic Time Warping) is used to calculate the distance between time-series. This distance metric can align the shapes of time-series accordingly without the influence of the inconsistency of time windows. A K-means clustering algorithm with $[n_{init} = 10, k = 10]$ is implemented and the clustering results are listed in Figure 4.3.

The clusters are represented by a bunch of time series, drawn as black lines, in the cluster plots, along with a red line to indicate the center of that cluster. We can observe very heterogeneous variation patterns in the time series in Figure 4.3. Clusters 2, 3, and 6 show continuous increase in diversity while all of them have plateaus either in the earlier stage or later ones. In contrast, Clusters 1 and 5 illustrate a decreasing tendency over time.

I, meanwhile, calculate some important scientometric metrics for each cluster and

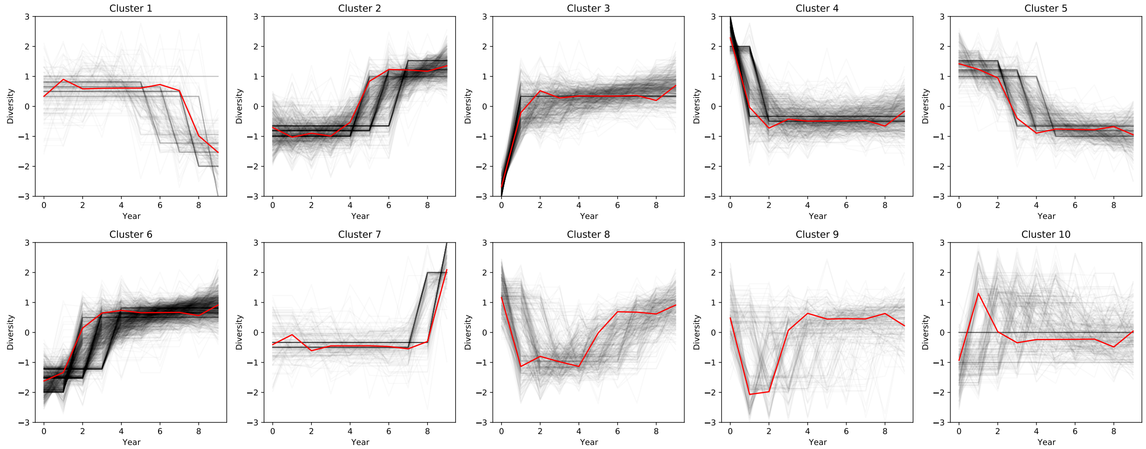


Figure 4.3: Kmeans clustering results of author research trajectories

Table 4.4: Scientometric variables for clusters using K-means

cluster label	h-index	average citation	paper number
1	6.09	11.70	20.07
2	7.77	12.38	29.44
3	8.94	13.96	34.70
4	8.15	13.97	29.74
5	7.27	12.80	25.21
6	8.53	13.08	33.24
7	5.67	11.00	18.55
8	9.20	13.11	36.83
9	8.81	13.91	33.66
10	7.79	12.62	28.26

summarize them in Table 4.4. It is found that authors in Clusters 2, 3, 6, 8, and 9 have larger numbers of H-index, papers, and average citations. These clusters share a pattern of an increase in diversity from year 4 to year 10. Clusters 1 and 5, however, have a lower values of these scientometric variables. An increase in research diversity in the early stage of researchers' careers is a good indicator for the success of their career later.

To get rid of the influence of clustering methods, I also implement the KShape clustering method [82] to cluster the time series and analyze the scientometric variables again. The clustering curves summarized in Figure 4.4 are similar to the curves

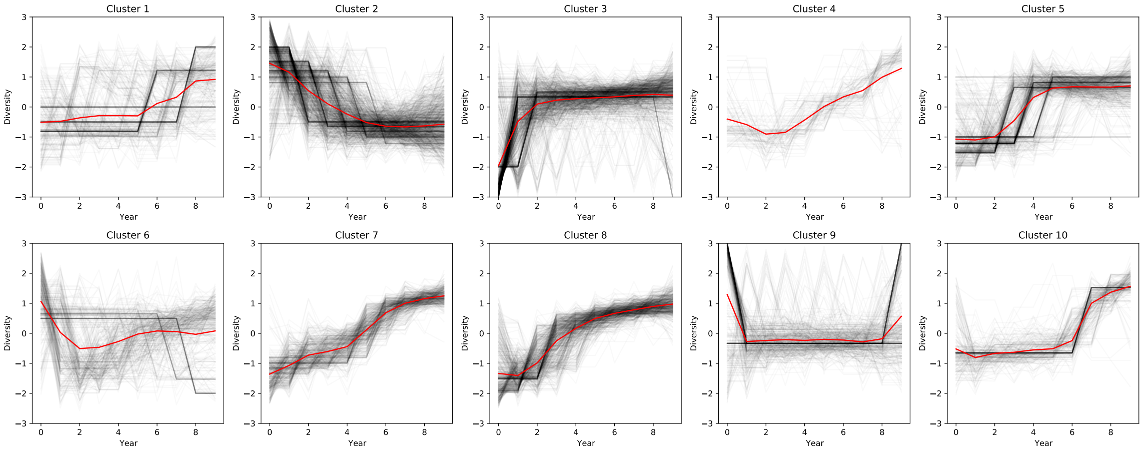


Figure 4.4: KShape clustering results of author research trajectories

Table 4.5: Scientometric variables for clusters using KShape

cluster label	h-index	average citation	paper number
1	6.21	11.74	20.78
2	7.78	13.02	27.98
3	8.51	13.68	32.44
4	9.10	13.27	36.10
5	7.26	12.76	25.99
6	8.29	13.11	31.89
7	9.12	12.77	36.78
8	9.52	13.49	38.76
9	7.13	13.17	24.96
10	7.34	11.47	27.28

in Figure 4.3. There are increasing tendencies in many clusters like Clusters 3, 4, 7, and 8. The H-index of authors in this cluster is also higher, as shown in Table 4.5. The results based on these two clustering algorithms are aligned to some extent.

4.6 Discussion

This chapter discusses how to design a diversity metric in a continuous space. Since continuous space usually can include rich information for data, nearly all of the modern machine learning methods leverage that instead of discrete representations like one-hot dictionaries. Although continuous representation learned by modern algorithms is powerful, discrete space still has its own advantage. Data defined in discrete spaces are usually more interpretable to humans. When the metrics or algorithms encounter errors, it is easier to troubleshoot on the discrete space. In real applications, when there are very clear conceptual boundaries between objects, like species of birds, representation in a discrete space could be a good option as well. In addition, representation space is not the only factor to influence metric design. Other factors, such as the distance choice in our diversity design task, have great impact on the final success of measurement. There is no simple answer between different genres of representations.

This study, especially the axiomatic analysis, suggests average distance to be a good metric. However, is it the "optimal" metric we should always choose? The axioms in this chapter are derived from intuition about research diversity in science of science. Some axioms like "duplicate" may not be desirable in other applications. Furthermore, axiomatic analysis only serves as a "lower bound" analysis for metrics. A metric satisfying the constraints perfectly does not mean it will work well enough in tasks. It is difficult to claim that average distance is the best choice independent of real world concerns, especially when the representation is hard to interpret and reproduce. This study only shed light on one perspective to be considered in the metric choice process.

4.7 Summary

Within this chapter, I focus on the problem of metric design in continuous space. Similar to the analysis in Chapter III, I propose a series of axioms to limit the metric choice in continuous space. I surprisingly find that the simple and intuitive metric, average distance, performs well in the axiomatic analysis. The metric is endorsed by a complete set of proofs of the satisfaction of axioms, which is not equipped by other metrics. On the application side of this study, the results of author-level regression illustrate a positive relationship between diversity and the increase in research impact. With my analysis of researchers' trajectories, I find some interesting variation patterns for famous researchers and conduct a time-series clustering on the research trajectories. It is shown that a continuous increase in diversity can be a good indicator for the success of scholars.

The metric design and the results are limited to this area of science of science. I will explore the usage of a metric derived from axioms in other domains in the next chapter.

CHAPTER V

Measuring Structural Diversity in Online Social Networks

The success of utilizing a diversity metric, endorsed by the results of axiomatic analysis, motivates me to extend this method beyond research diversity. I pay attention to another real application, the structural diversity in online social networks to explore more effects that diversity can bring to our society.

User experience and behaviors in online communities are influenced by structural properties of the social networks. Beyond friend counts and centrality, the structural diversity of a node, or how much its neighbors are different from each other, has been recognized as a surprising and critical factor of social contagion and information diffusion. While it is intrinsically difficult to measure diversity in a discrete topological space, recent developments in large-scale network embedding algorithms have provided a powerful way to project the nodes of a social network into continuous spaces, where their subtle relationships can be captured and computed much more efficiently. In this chapter, I utilize the embedding-based structural diversity metric as discussed in the last chapter and show its advantages over alternative node-level metrics for measuring structural diversity.

Applying this metric to a leading online social network graph (Snapchat), I discover an intriguing pattern in friendship formation: when building their local networks, users start by making friends in local neighborhoods and consuming the known close-by nodes, resulting in an intriguing pattern of first increasing then decreasing diversity; when local friendships are gradually exhausted, they reach out to outer communities, resulting in a sustainable increase in diversity again.

I characterize this “open closed open” (OcO) dynamic of diversity in real world networks and propose an intuitive network generation model (OcOM) which effectively mimics this new network property. I also investigate the relationship between structural diversity and different types of engagement metrics on Snapchat, where I observe heterogeneous correlations between diversity and narrowcasting/broadcasting social behaviors.

5.1 Introduction

Diversity is not only defined in the science of science. When addressing other research questions about diversities, the problem of how to design and evaluate a diversity metric is very critical to researchers and stakeholders. Online social media analysis is one active domain where people are interested in the effect of diversity between people within a small community.

In an era when social media dominates traditional media, people reside in the social contexts of multiple online communities and are shaped by opinions and actions of their friends. Friend counts, centrality in the network, structural holes, and many other local and global properties of social networks can influence a user’s experience and behavior significantly. Among these factors, structural diversity, or how much the neighbors of a node are different from each other, has emerged as a pivotal topic in the research on online social networks [32, 105]. Various evidence has shown that structural diversity of users in social network influences user behavior [108, 6].

In real life, our decisions and behaviors are largely influenced by close friends and family members. We tend to adopt a new product or trust a story due to the process of social reinforcement [77] from friends with high affinities. On the other side, friends that span weak ties are able to disseminate novel information to people and influence their decision making collectively [4, 8, 41]. Information endorsed and spread by friends from different communities could change people's minds in an aggregated way. It is clear that structural diversity, which represents the variety of friendships between friends, plays an important role in information diffusion and decision making in various scenarios. Indeed, existing literature has reported the mixed effect of structural diversity on social contagion, user engagement, retention, and adoption of innovations [108, 122, 32, 105, 4].

Although there have been a few studies regarding the measurement and impact of structural diversity, two fundamental research questions in this direction are still wide open:

First, **how can we efficiently measure the structural diversity in large-scale real-world social networks?** Many existing measurements of structural diversity rely on well-defined community labels. For example, researchers have previously defined structural diversity as the number of connected components or communities [108, 32, 131]. However, connected components do not effectively represent the community structures. Handcrafted labels and fine-tuning are usually necessary to obtain a decent network partition, which is expensive and may not scale to real-world social networks. The algorithms used to detect communities in networks usually face the challenge of choosing the appropriate granularity for graph partitions. Meanwhile, partitions and labels of communities usually fail to represent the complexity of network structures; in a real social network, many users are members of a wide range of communities with different degrees of affinity, which cannot easily be reflected by binary labels or memberships in a predefined number of communities. We need efficient and accurate representation of networks and design a structural diversity metric accordingly.

Second, **what is the influence of structural diversity on a user's engagements in complicated social contexts?** Current evidence of the effect of structural diversity appears to be controversial. Researchers have shown both positive and negative effects of structural diversity on various types of behaviors in different online communities [32, 24, 7]. However, user experience and behavior have a mixed initiative nature, which takes place in a mixture of social contexts. One needs to carefully evaluate the correlation of structural diversity and other social factors within a complicated scenario. Furthermore, the diversity metrics defined in different studies are not consistent. The heterogeneity of results within existing research may be a result of the heterogeneity of diversity metrics used, which illustrates the close connection between these two research questions.

I aim to provide transformative answers to these two salient research questions. To tackle the rarity of community labels, the lack of representation capacity, and the computational inefficiency of metrics in discrete topological spaces, I leverage the recent developments in graph representation learning. Through state-of-the-art node embedding techniques, nodes in a real social network can be projected into high-dimensional continuous spaces. Based on the findings from the previous chapter, a simple and intuitive metric can then be defined in the continuous node embedding spaces to measure the structural diversity of a node.

This metric is not only more informative than alternative network metrics (i.e., node-level centrality measures) in terms of quantifying diversity, but also simple to implement and compute in an embedding space.

With this measurement, I evaluate users' structural diversity in real-world, Web-scale social networks and present a novel discovery: users typically encounter a "friendship saturation" phenomenon when they keep building new friendships. One possible explanation for this phenomenon is: as users tend to connect to friends of friends, they will exhaust the availability of new friends in their local neighborhoods.

As a result, the structural diversity of users will **increase and then decrease** with the accumulation of friend counts. Eventually, users tend to explore outer communities to build new friendships and structural diversity will **increase again**. The rise-fall-rise pattern is illustrated in the analysis of both *static* and *dynamic* social networks. In addition, I show that traditional network models such as Watts-Strogatz [118] or Barabási- Albert [10] cannot explain this intriguing "friendship saturation" phenomenon. In lieu, I propose a novel "Open Closed Open" network generation model (OcOM) that captures these dynamics of structural diversity. The proposed model provides a potential mechanism behind the rise-fall-rise pattern.

To illustrate the potential utility of the new structural diversity metric, I analyze users' engagements in a leading online social platform, Snapchat. Snapchat, as a novel social media platform, presents a mixture of heterogeneous user behaviors within its app (see Figure 5.1). Particularly, users on Snapchat are involved in two different families of everyday interactions: narrowcasting (communicating with friends privately in direct messages) and broadcasting (posting/sharing to public online communities). I find that structural diversity plays heterogeneous roles in narrowcasting and broadcasting. In particular, users with high structural diversity (measured by Esd, i.e. embedding based average distance) are more likely to broadcast their generated contents and less likely to narrowcast them.

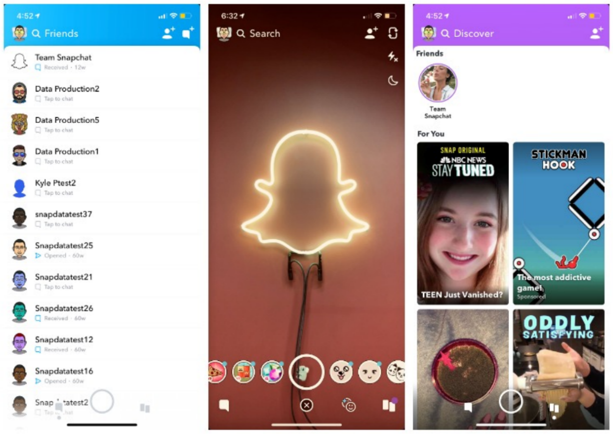


Figure 5.1: Snapchat in-app interface

The rest of the chapter contains several different parts. I summarize related work that motivates my study. I implement the new metric and analyze its effect in a special case. I describe the diversity pattern I found in the real large-scale social network and the model to simulate generation of this pattern. Finally, I explore the impact of structural diversity on user engagement and present the results.

5.2 Related Work

My work is motivated by previous research on diversity measurement and graph embedding. The various definitions of structural diversity and their implementations provide insights for me when I design the new measurement and explore its impact on the social network. The node embedding techniques developed in recent years pave the way to the proposed measurement of structural diversity and its effect on user engagement. A general discussion about defining diversity was included in Chapter II and Chapter III.

5.2.1 Effects of Structural Diversity

The study of structural diversity can be dated back to the initial study of network structure and its social influence. Weak tie theory [41] and structural theory [19] introduce the idea that impacts from diverse friends, who are not in people's local community, can be long-lasting and significant; in contrast, homophily [77] studies the impact of less diverse and more local communities.

The discussion of the mixed effects of different social influences extends to the research on online social media nowadays. Homophily and influence-based contagions [5] are what past research has focused on. In their findings, homophily explains more than 50% of perceived behavioral contagions. In [24], the study shows that users adopt online behaviors when receiving social reinforcement from a large number of friends.

On the other hand, much existing literature supports the effect of “diverse friendship.” In [108], researchers find the probability of contagion is tightly controlled by social structural diversity, instead of the number of friends. The number of friends even has a negative effect in the prediction of contagion. Researchers also find the positive social effect of structural diversity in various social settings, such as social contagion in exercise behaviors [6] and purchasing of social networking apps [99].

Promoting structural diversity has a heterogeneous social effect. It is proved to be potentially beneficial to break filter bubbles [96] and improve information novelty [7, 4]. But there is not enough evidence to support that it is beneficial to user online retention [105]. An extensive meta-analysis of the structural diversity of over one hundred social networks reveals the complexity of its social influence [32].

The mixture of positive and negative results of structural diversity motivates me to explore its effect in newly emerged social networks like Snapchat. I am particularly interested in the social influence brought by structural diversity on a platform with many potential interactions among users.

5.2.2 Structural Diversity Measurement in Social Networks

Just like the research on diversity in science of science, researchers have devoted efforts to measure structural diversity quantitatively in social networks since it can imply potential social implications as shown before. In addition, the metric itself is useful for tasks like predicting popularity of user-generated content [9] and users' reputations in platforms [131].

Most existing definitions are restricted to a *discrete space* (e.g., the topological structure of a graph). Some of them are as simple as the number of connected components (e.g., Ugander et al. [108]), while others are more complex and leverage more contextual information.

Dong et al. [32] define structural diversity as the number of connected components that comprise the common neighborhood. Zhang et al. [131] define weak structural diversity as the number of weakly connected components in the ego-network and strong structural diversity as the number of strongly connected components. Both of these metrics are close to the definitions of structural diversity in [108].

In recent research, researchers have begun to define the network metrics in online social media differently. Su et al. [105] define diversity as the average distance between neighbors, which take the binary incidence vectors in networks as users' representation, whereas in [116], a weighted average cosine distance to the center is used in a text embedding space to indicate whether a local community is general or specialized.

5.2.3 Graph Representation Learning

Recent developments in graph representation learning enable us to define and explore structural properties of a network in continuous embedding spaces, which brings in more capacity and flexibility to analyze network structures at the micro and macro levels. Given a connected graph, node embedding algorithms, such as deepwalk [85], LINE [106], and node2vec [42], transform nodes of graphs into a high-dimensional vector space and preserve their proximity in the topological space.

Networks in real life often contain millions of nodes, which makes some node embedding algorithms hard to scale up. To overcome the bottleneck of scalability, researchers have developed new large-scale node embedding technologies, such as ProNE [130], PytorchBiggraph [65], and Graphvite [132], to partition large networks and train graph embedding efficiently in a parallel way. In our study, I learn the embeddings for large-scale Snapchat networks using multiple node embedding technologies.

Graph Neural Network (GNN), such as Graph Convoluntional Network (GCN) [57] and Graph Attention Network (GAT) [113], utilize supervised information in the graph to learn node representation with rich contextual information. They perform well on multiple tasks like node classification, link prediction, and graph visualization. However, we are short on costly labeled information in very large-scale networks, and current graph neural networks do not scale up efficiently.

5.3 Embedding-based Structural Diversity

5.3.1 Diversity Measurement

As summarized in the related work, many structural diversity measurements are defined based on extracting patterns from a discrete space (i.e., the topological structure of networks). The alignment of nodes to communities is necessary to most measurements. However, community detection in real-world, large-scale networks is very challenging. Additionally, hard division of nodes into different groups will hide subtle relationships between communities and nodes. Indeed, nodes that belong to different groups may be very close to each other in some circumstances.

To tackle the complicated relationships between nodes within a large-scale social network, I leverage the recent developments in node embedding algorithms to project nodes into high-dimensional continuous spaces, in which one can easily define and compute the distance between two node vectors, distance that reflects the closeness of the two nodes in the original network structure.

Inspired by existing work that uses average distance to define diversity in discrete vector spaces ([105]), I design a new metric to characterize the structural diversity of a node, in the node embedding space, as the average cosine distance between the vector representations of its neighbors in the original network. This is the same metric that I designed in Chapter IV. I have proved its soundness within the axiomatic system and hope it can reveal some interesting patterns in the network embedding space.

I propose to define the embedding-based structural diversity (henceforth called

Esd) of a node x as:

Definition V.1 (Embedding-based Structural Diversity (Esd))**.**

$$\textsc{Esd}(x) = \left(\sum_{u,v \in N(x)} (1 - \cos(\vec{u}, \vec{v})) \right) \Big/ \binom{|N(x)|}{2}$$

where $N(x)$ denotes x's neighbors, $|\cdot|$ indicates cardinality, u, v are neighbors of x (w.l.o.g.), $\vec{u}, \vec{v}$ are their vector representations, and cos indicates cosine similarity.

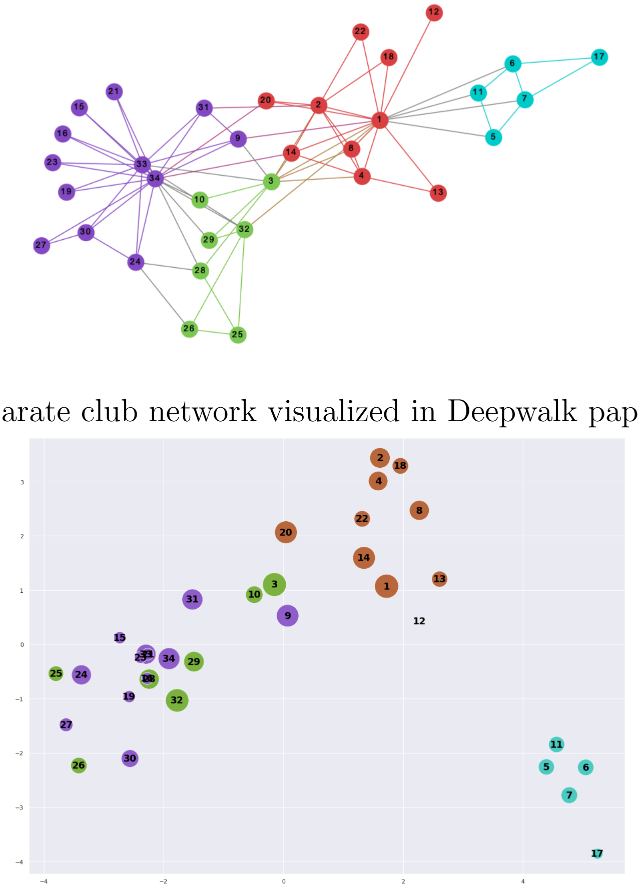
The metric is a special case of the average distance, as I mentioned in last chapter. I will use Esd as the alias of this metric in the rest of this chapter.

While the above definition is kept simple and intuitive, my study shows that this metric meets the expectation of structural diversity in real networks. Below, I demonstrate the properties of this metric using a classic case of social network analysis.

5.3.2 Case Study

Zachary's karate club [129] is a classic case and offers public data to demonstrate the separation of 34 club members in a dense social network. I embed the nodes in this network into a continuous space using DeepWalk just as it was done in [85] and visualize the nodes using the first two principal components of the PCA results. The labels (colors of nodes) come from the result of community detection using a modularity-based algorithm [85].

The graph in Figure 5.2(upper) is a replica of the graph in the DeepWalk paper, using different colors to demonstrate the community structure of the Karate Club network. Figure 5.2(bottom) presents the embedding space and the calculation of diversity scores. In particular, the coordinates of a node indicate its position in the embedding space, and the size of a node indicates the value of its structural diversity as computed in Definition V.1.



Karate club network visualized in Deepwalk paper

Structural diversity of nodes in an embedding space

Figure 5.2: Structural diversity score in embedding space of deepwalk for karate club network

We can observe that the two-dimensional embedding preserves the layout of nodes (or the network structure) well. Nodes within the same community tend to be embedded close to each other. People who play the role of "bridges" or span a "structural hole" between communities, such as 3 and 20, tend to have larger diversity scores than people who are far from the center of network, like 17. In addition, users with many friends are not necessarily diverse in the graph. The two centers in the graph, node 1 and node 34, are not the ones with the highest diversity scores: although they reside in the center of a community and have many friends, their friends are too alike (they are likely to also be from the same community).

To help readers better understand the unique property that the diversity score is measuring, I calculate a few other network metrics defined in the discrete topological space and compare the results with structural diversity. I find that structural diversity captures some characteristics which are not supplemented by other metrics. As shown in Table 5.1, structural diversity ranks nodes in a very different way from degree. Nodes with lots of edges, like node 34, do not yield large structural diversity.

A local clustering coefficient is used to measure the local density of a node's neighbors. An isolated node like node 17 has a large local clustering coefficient but definitely does not have a highly diverse neighborhood. One may wonder whether structural diversity is similar to the inverse of the local clustering coefficients, as a highly diverse neighborhood may also be "loose." Results show otherwise. We see that nodes with low local clustering coefficient (the second row in Table 5.1) usually are those with large degrees like 34 and 1, and part of nodes between communities like node 32. These nodes are not constantly aligned with high or low structural diversity. Different from the concept of centrality (defined in the fourth and fifth rows in Table 5.1), which identifies nodes at the "center" of a network, structural diversity can help find some nodes that are connected to heterogeneous neighbors but do not necessarily occupy the "center" position, such as node 31 in the network.

Metric	Top-10 nodes in ZKC
Degree	34, 1, 33, 3, 2, 32, 4, 24, 9, 14
Local Clustering Coefficient	17, 8, 16, 23, 27, 13, 22, 18, 15, 21
Inverse Local Clustering Coefficient	10, 34, 1, 28, 33, 32, 3, 20, 29, 26
Closeness Centrality	1, 3, 34, 32, 33, 9, 14, 20, 2, 4
Betweeness Centrality	1, 34, 33, 3, 32, 9, 2, 14, 20, 6
Number of Connected Components	1, 3, 34, 28, 32, 2, 10, 14, 20, 24
Average Distance in Discrete Space	32, 9, 28, 29, 20, 27, 31, 24, 14, 3
(Ours) **Embedding-based Structural Diversity**	1, 3, 32, 14, 20, 9, 34, 31, 29, 2

Table 5.1: Ranking of top-10 nodes in ZKC according to various potential node-level diversity metrics

I also implement the diversity metric defined in previous work based on the ZKC and the community label learned using a modularity-based community detection algorithm. The number of connected components used in [108] and average distance in discrete space in [105] are deployed and summarized in the sixth and seventh lines in Table 5.1. They will rank some weird nodes like node 28 and node 2 in the top ten results because the representations they relied on are not very accurate compared to the continuous embedding method.

5.3.3 Metric Embedding Consistency

Since the defined structural diversity metric relies on the embedding space of nodes, is it senstive to the node embedding algorithm? I implement four different node embedding techniques and use them to find the representations of the same network (the Snapchat static network mentioned in next section) and calculate the structural diversity of its nodes: deepwalk [85], LINE [106], ProNE [130], and PytorchBiggraph (pbg) [65]. Deepwalk and LINE are two of the very first methods proposed for large-scale node embedding. ProNE and PytorchBiggraph are two of the most recently developed, which are highly scalable.

These four node embedding methods used in this paper have different objectives.

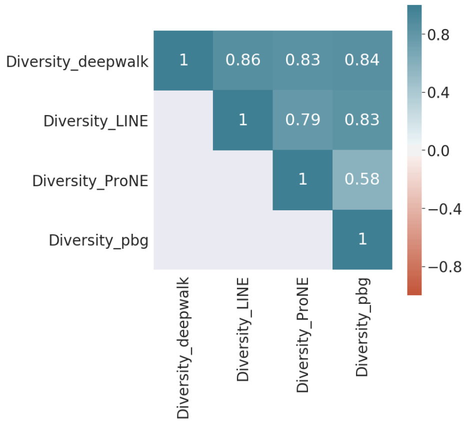


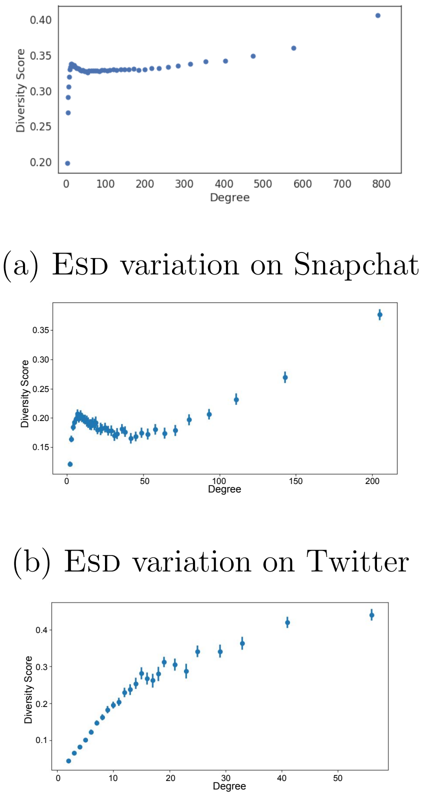
Figure 5.3: Correlation between structural diversity metrics

Deepwalk model graphs as random walk sequences and optimizes the occurrence of nodes in a random walk using the Skip-gram algorithm. LINE optimizes the first-order and second-order proximity directly for each individual node. ProNe frames node embedding as a matrix factorization task and optimizes the probability between each pair of nodes on edges. PytorchBigraph optimizes the rank of edges over syntheic edges created through negative sampling. The four methods have different objectives, but all of them, along with other node embedding methods not mentioned in this paper, keep local structures for networks in continuous embedding spaces enriched with global contextual information.

The results in Figure 5.3 illustrate a high consistency among the structural diversity scores calculated in different embedding spaces. This indicates that although the proposed diversity metric builds upon node embeddings, it is not sensitive to the choice of embedding algorithms.

5.4 Structural Diversity in Real-World Social Networks

Users in social networks establish their unique identities with various friending choices. By modeling and characterizing Esd over many users both statically and dynamically, we can better understand the intrinsic properties of social networks.



(a) Esd variation on Snapchat

(b) Esd variation on Twitter

(c) Esd variation on academic network

Figure 5.4: Open-closed-open (OcO) patterns in Snapchat (a), Twitter (b), and Academic (c) networks.

I now shift focus to a large, real-world social network: Snapchat. Snapchat is a leading mobile social platform where users can privately or publicly share images/videos with their friends, post to their stories and more. I select Snapchat users active in a given month and in a particular country to construct a friendship network. The network contains *1.2 million* nodes and *81.9 million* edges.

The study results in two related findings in both **static** and **dynamic** networks: patterns in the distribution of Esd in a static snapshot of the network, and temporal dynamics of users' Esd as they grow their own friend networks. Both findings demonstrate that users' Esd follows a rise-fall-rise pattern, which we call *"open closed open"*(OcO).

5.4.1 Open Closed Open in Static Networks

Based on ESD scores, I show the distribution of ESD for users with different numbers of friends (node degree). Figure 5.4(a) depicts this on the Snapchat social network and also illustrates an intriguing pattern: ESD presents an initial rapid rise when the node degree is small; it falls down when the degree reaches a certain threshold (near 50); and it presents a sustainable, slow growth thereafter. Such a pattern may represent a potential "friendship saturation" phenomenon: the ESD increases when users enter the network and begin to build friendships with multiple nodes. Naturally, a user tends to make new friends within their local community. The diversity among their friends increases rapidly at the beginning, and the increase will stop when the possible **novel** friendship resource is exhausted locally. Now, when the user keeps adding friends within their local community, they are essentially connecting to their "friend's friends," which results in a decrease of ESD. To further grow their friend networks, users eventually reach out to those people outside their local communities and the diversity of their neighborhood increases again correspondingly. Interestingly, the shape follows a similar pattern as the variation of community density described in [60], regarding the growth of communities. This sheds light on the similarity between the growth of users' local community and global social networks.

While my observation holds on the Snapchat social network, one may ask whether this pattern holds in other online social networks. It is generally challenging to find other large-scale, publicly available social network data, but I am able to evaluate on a public snapshot of the Twitter social network from [67], with 81K nodes and 1.7M edges. The Twitter dataset [76] contains 1,000 ego-networks and 4,869 circles. The ego-networks range in size from 10 to 4,964 nodes, which represent part of early users of Twitter. I find a similar relationship between degree and ESD of nodes, as shown in Figure 5.4(b). Variation of ESD in the Twitter network also reveals a clear OcO pattern, suggesting that this pattern may be general across social platforms. I also

degree range	(0,10)	(10,20)	(20,30)	(30,40)	(40,50)
correlation, p-value	0.45,<0.01	0.60,<0.01	0.63,<0.01	0.66,<0.01	0.68, <0.01
degree range	(50,60)	(60,70)	(70,80)	(80,90)	(90,100)
correlation, p-value	0.68,<0.01	0.68,<0.01	0.66,<0.01	0.71,<0.01	0.68,<0.01

Table 5.2: Pearson's correlation coefficient between ESD and the number of venues in which researchers published

extract a coauthor network that contains all the authors from 13 top-tier AI-related conference with 62K nodes and 184K edges from the same dataset as in Chapter IV. Figure 5.4(c) shows the pattern in the variation of ESD. Although the nature of academic collaboration is different from friendships in social apps, researchers also face the "saturation" of local collaborators. There also exists some rise-and-drop pattern of structural diversity in the academic network. However, since the friend number and relationship in coauthor network are not the same as those in common online social network, the "closed" stage is not as obvious as that in Snapchat and Twitter networks.

Furthermore, to verify the connection between ESD and some diversity-correlated metric, I test the relationships between ESD and the number of venues (natural labels of communities) in which researchers published in AI-related conferences. There is a fairly strong positive relationship (Pearson correlation coefficient >0.6) between ESD and the number of venues with the control of paper number (more detailed results are listed in Table 5.2). It reveals the close relationship between structural diversity in researchers' social networks and the research diversity of their publications in terms of publication venues. This result provides further evidence that the unsupervised ESD can still measure structural diversity effectively even if discrete labels of communities are not available.

5.4.2 Open Closed Open in Dynamic Networks

The above analysis illustrates the open-closed-open patterns in the snapshots of multiple social networks, but a static network prevents us from verifying whether the pattern is indeed driven by the dynamics of friending actions. To confirm our conjecture of "friendship saturation," I also investigate the variation in Esd over time, in dynamic networks.

I study the dynamics of Esd of 34,000 users who newly joined Snapchat one month before the data collection. By focusing on their behaviors, I can avoid attributing the OcO observed in the static network to old/existing nodes and links. Figure 5.5 demonstrates the overall trend across all the new users, showing an even more pronounced OcO pattern.

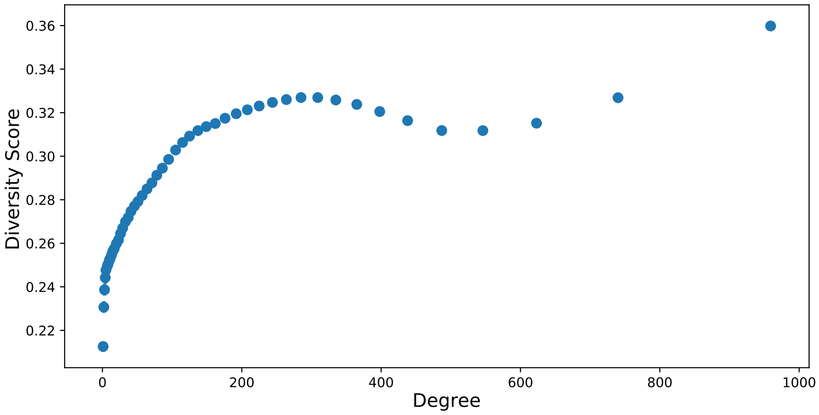


Figure 5.5: Open-closed-open (OcO) patterns across new Snapchat users and new friendships in dynamic networks.

For these new users, I additionally monitor the variation in diversity as they add new friends one-by-one (in temporal order). Figure 5.6 illustrates the diversity trajectories of several randomly sampled individual users with varying degrees; we can observe that although individual patterns are noisier than the aggregated pattern, all of the trajectories follow the OcO pattern to some extent. Interestingly, there are differences in the trajectories themselves: for example, the top-left user has a very steep rise among the first few friends (e.g., early friendships are across different social circles, like coworkers and childhood friends), whereas the bottom left user has a comparatively slow rise (e.g., early friendships are within the same circle).

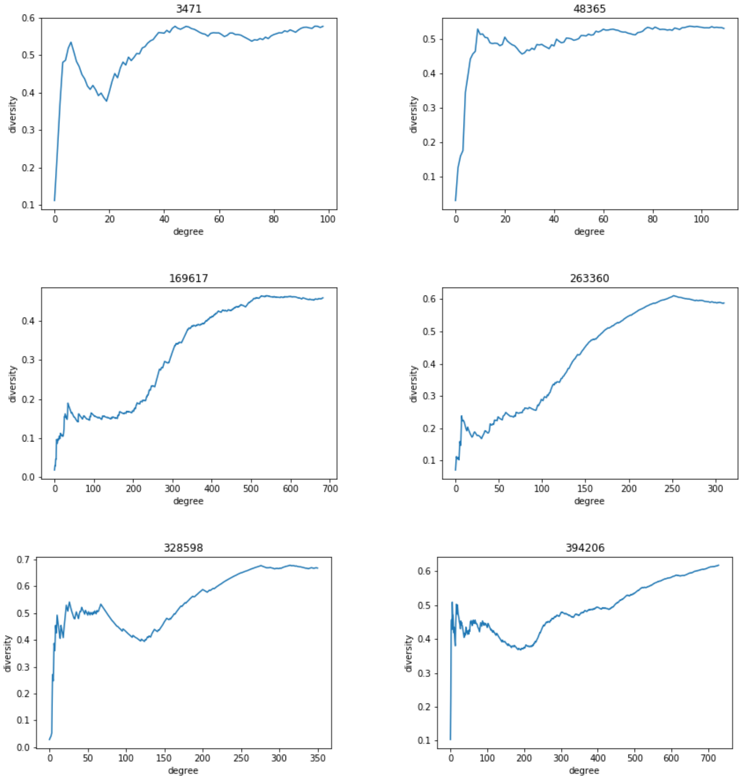


Figure 5.6: Sample OcO patterns in Esd trajectories of several new users on Snapchat, as they add friends one by one.

The results from static networks and dynamic networks illustrates the potential of explaining variation in structural diversity using the OcO mechanism. However, without actual behavior data, it is hard to claim this single mechanism is the only factor influence users' friending strategies. Thus, I propose a graph generation model in the next section to strengthen the possibility of the existence of this mechanism.

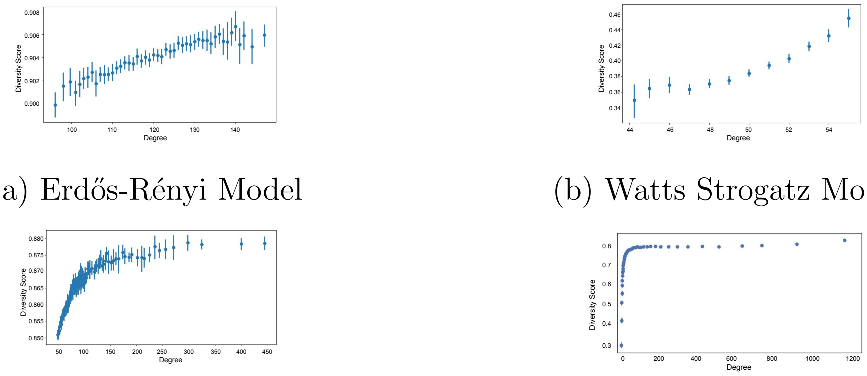
5.5 Network Generation Model

Encoded by the rise-fall-rise dynamics of structural diversity and explained by friendship saturation effects, we have seen that the growth of a user's egonet produces an interesting OcO pattern in diversity. Hypothetically, this diversity can be attributed to the process in which new users make friends with multiple people initially, stick to accumulating social capitals locally for a while, and step out of their comfort zone and start to make friends in outer communities eventually. Due to the limit of access to the users' behavior data, I verify this by building a random graph model to simulate the generation process of the network. A reasonable model of network growth should reflect such a process and capture the variation in structural diversity, while preserving other known properties of real-world networks.

5.5.1 Diversity in Classical Network Models

I first implement a few classical random graph generation models and track the variation pattern of structural diversity under these models. I include the Erdős-Rényi[35], Watts-Strogatz (small world) [118], Barabási-Albert (preferential attachment) [10], and Forest Fire models [66] to simulate the growth of a network and the dynamics of structural diversity. I generate networks using these four models and depict the variation in diversity scores. The parameters for these models are: Erdős-Rényi ($n = 2000$, $p = 0.06$), Watts-Strogatz ($n = 2000$, $p = 0.1$, $n_{neighbours} = 50$), Barabási-Albert ($n = 2000$, $n_{neighbours} = 50$), Forest Fire ($n = 5000$).

The simulation results in Figure 5.7 demonstrate that these models can at best capture partially, but not exactly, the "friendship saturation" phenomenon, or the rise-fall-rise dynamics of structural diversity in the real social network. The structural diversity in the Erdős-Rényi model is random since the links are added randomly without social consideration. The diversity scores fluctuate and reveal a positive relationship between degree and structural diversity. The curve in the Watts-Strogatz model appears to show a monotonic relationship between degree and structural diversity. The nodes connected by more rewired nodes simply obtain a higher degree and a higher structural diversity simultaneously. Both of these models have been shown to produce unrealistic degree distributions and thus are of limited applicability in modeling social growth. The Barabási- Albert model builds edges according to the preferential attachment mechanism. The structural diversity illustrates some concavity for low-degree nodes, but it reaches a plateau gradually later on. Finally, the Forest Fire model reflects the fast open stage at the very beginning but it fails to capture the decrease in structural diversity explicitly before the score quickly increases again after the saturation.



(a) Erdős-Rényi Model

(b) Watts Strogatz Model

(c) Barabási Albert Model

(d) Forest Fire Model

Figure 5.7: Simulated diversity variation with traditional random graph generation models.

5.5.2 The OcO Model

Can a random graph model capture friend saturation and the real dynamics of structural diversity at all? Motivated by this question, I propose a new random graph generation model called the OcOM (OcO Model), which simulates the growth of the network and the variation in ESD. The model is simple and intuitive, which better reflects diversity variation than the above-mentioned models while preserving the other classical properties of real-world networks such as degree distribution, small-world, and shrinking diameters.

The full OcOM model is described in Algorithm 1. The model begins from an initial graph G with many isolated nodes V. New nodes are added into the graph one by one, and they will initially attach to a node x, randomly selected from

Algorithm 1 Open Closed Open (OcO) Model

```
Input: initial Graph $G = (V, E)$ with disconnected nodes, parameter $\alpha$, threshold $T$ ($T << |V|$), and number of new nodes $N$.
for $i \in \{1, \cdots, N\}$ do
    Add node to Graph $G$
    Select a random node $x \in V$ , add undirected edge $e(i, x)$
    $c(i) \leftarrow \{x\}$
    $L \leftarrow 0$
    while $L \leq T$ do
        $p \leftarrow \frac{1}{(1+e^{-\alpha L})} - \frac{1}{2}$
        $\forall p_0 \in \mathcal{U}(0, 1)$
        if $p_0 < p$ then
            Select $y$ from a different component
            Add undirected edge $e(i, y)$
            $c(i) \leftarrow \{y\}$
        else
            cand $\leftarrow \emptyset$
            for $j$ in $c(i)$ do
                left $\leftarrow$ Neighbors($j$) $- c(i)$
                cand $\leftarrow$ cand $\cup$ left
            end for
            Sample $y \in$ cand
            Add undirected edge $e(i, y)$
            $c(i) \leftarrow c(i) \cup \{y\}$
        end if
        $L \leftarrow L + 1$
    end while
end for
```

the graph. The new node i then undertakes a procedure to make new connections. Similar to the design of return and exploration steps in [48], at every step, the node i randomly chooses from **two connection strategies**: (1) *local connection:* connect to a neighbor's neighbor. This process is similar to a Polya's Urn [52], so that if a node y appears in the neighborhood of multiple existing neighbors of i, it will be more likely to be connected to i; (2) *jump connection:* connect i to a random node y in a component which does not overlap with the current node's 2-hop neighbors. A random variable controls the trade-off between these two strategies, which depend on the number of connections of i. Naturally, a node is more likely to connect locally first when it has fewer friends, and tends to move to other communities later when it has more and more friends, thereby simulating the process of friendship saturation.

Note that the local connection process naturally ensures a high clustering coefficient. It also encodes preferential attachment, as a node with a high degree is more likely to be connected to by a new node. This "rich gets richer" process leads to a highly skewed degree distribution. The jump connection process creates long-range

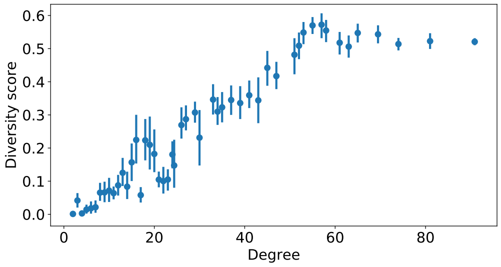


Figure 5.8: Diversity variation of OcOM. The model simulates the rise-fall-rise pattern in structural diversity

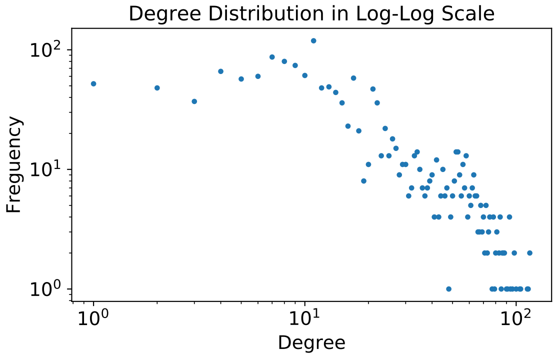


Figure 5.9: Degree distribution in log-log form for OcOM.

links and thus shortens the paths. I expect the OcOM model to capture the OcO pattern while also preserving other classical properties of real-world networks such as the scale-free and the small-world properties.

I implement OcOM and depict the properties of the generated network. The parameters for this simulation are $T = 50$, $\alpha = 0.002$, $N = 1600$, which reflect the number of nodes to connect to for a new node, a scalar controlling jump probability threshold, and the number of nodes to grow the graph. I set 200 separate nodes as the initial graph and keep adding new nodes into the graph. Figure 5.8 shows the variation in Esd of the graph generated by OcOM. I can observe the rapid increase in structural diversity in the very beginning along with an explicit sharp descent later on; the structural diversity score boosts after the "saturation" stage and increases to a plateau in the later stage. Figure 5.9 depicts the degree distribution for the generated model. It preserves a clear heavy-tail pattern that is close to many real

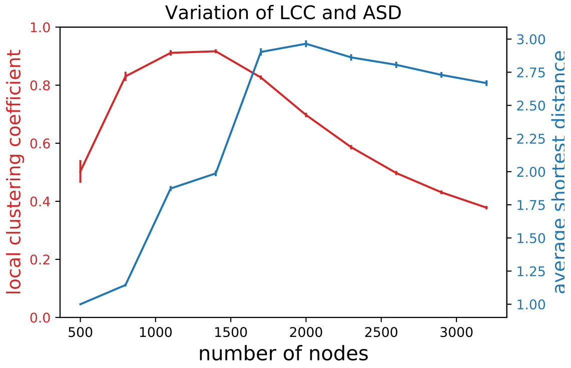


Figure 5.10: Variation of average local clustering coefficient (lcc, red) and average shortest distance (ASD, blue) for OcOM.

degree distributions of social networks. The Figure 5.10 shows the variation in average local clustering coefficient and average shortest distance in the generated graph. The local clustering coefficient remains at a high level (>0.5) most of the time and the average shortest distance remains short (<3), which illustrates the generation of a "small world". The average shortest distance grows in the very beginning when some initial connections between nodes appear and it drops slowly later on, which caters to the diameter shrinkage phenomenon found in [66], i.e., the diameter of a network decreases during network growth.

The OcOM model successfully simulate the OcO pattern existing in online social networks. It illustrates that the mechanisms of exhausting low friendship resources and expanding later can generate the rise-drop-rise pattern we observed in real world.

5.6 Structural Diversity and User Engagement

After modeling the structural diversity using the new proposed metric, I pay attention to the potential social implications, leveraging the structural diversity. One of the most important implications is to find the effect of structural diversity on users' engagement.

I collect engagement data for the Snapchat social network (as described in Section

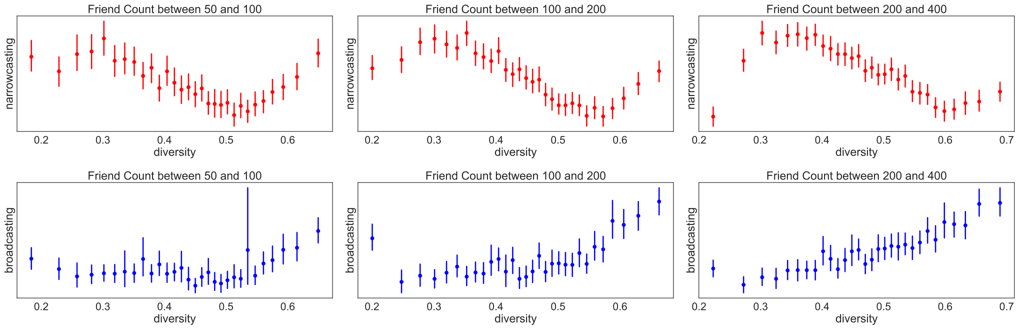


Figure 5.11: Variation of broadcasting and narrowcasting metrics

5.4), which contains ten engagement metrics that reflect users' behavior in Snapchat. To explore the association between ESD and several Snapchat engagement metrics, I divide users into ten groups according to their structural diversity scores. I run one-way ANOVA tests to compare the differences in engagement metrics among these groups. The tests reveal significant differences (p-value <0.01) in all the engagement metrics. The p-values also pass multiple testing correction using the Benjamini-Hochberg procedure. Users with different levels of structural diversity do behave differently in online communities.

Snapchat (and several other social platforms) encourage users to generate content and enable them to share directly/privately with friends, or post/exhibit content to their Stories (see Figure 5.1). I refer to this first type of behavior as *narrowcasting*, and the latter as *broadcasting*. Prior work [44, 74, 107] shows that these phenomena have different motivations and user considerations. I show variation in these metrics for users with different diversity scores in Figure 5.11. The top row represents variation in a narrowcasting metric and the bottom row indicates variation of a broadcasting metrics. The y-axis in Figure 5.11 indicates normalized scores for these metrics. Due to privacy concerns, I will not disclose any specific metric names or scales of metrics in this figure. Furthermore, I divide users into a spectrum of bins according to their friend count (degree) to control for the previously observed degree effects. I focus on one narrowcasting behavior and one broadcasting behavior, which users choose between when deciding where to share recorded content and I observe clear differences.

For example, the top row in narrowcasting shows concavity much more clearly than that in broadcasting. This shows that for users with moderate Esd (0.2-0.3) have a higher propensity to narrowcast, whereas the same cannot be said about their propensity to broadcast (the trend is fairly noisy). Conversely, we can see that for users who have many friends (the third column), those who tend to broadcast the most have markedly higher diversity scores (0.6-0.7, "right-leaning") than those who tend to narrowcast the most (0.25-0.35, "left-leaning" in the concave). Intuitively, these results make sense: users with few friends and low diversity are less incentivized to broadcast their content (perhaps due to a more closed-off/less socially diverse nature), whereas users with high diversity are more likely to do so (possibly due to extroverted tendencies and a large following). In short, I find that the more structurally diverse a user is, the less likely that they will choose close friends to narrowcast with, and the more likely that they will treat Snapchat as a broadcasting channel.

5.7 Discussion

In this work, I utilize the metric derived from Chapter IV for measuring structural diversity defined using continuous embedding spaces, in lieu of alternative discrete topological metrics. The metric shows promise in capturing subtle aspects of diversity which other metrics cannot, and at the same time shows interesting surface associations with user engagement. I discuss a few opportunities for extensions.

Firstly, this chapter tackles a novel approach in defining a network measure in embedding space rather than in discrete space, which is the traditional approach. Continuous, embedding-based analogs for other node-level metrics are interesting to study (especially for those that are difficult to compute in a discrete space), equipped with the prevalence of new embedding methods in modern representation learning.

The benefits brought by continuous embedding pave the way to better understanding the relationships between objects in complicated settings.

Secondly, the current metric only considers embeddings which utilize only network connectivity. Other information like node and edge attributes can be incorporated using recent advances in GNNs; comparing structural diversity in these embedding spaces is insightful, and could open the door to inductive diversity computation.

Thirdly, I propose a potential explanation for the friendship saturation phenomenon in this chapter and create a graph generation model to simulate this process. However, the variation in structural diversity is influenced by the size of networks and there may be other confounding factors to change users' behaviors. For example, the rise-drop-rise pattern may be a result of self selection from users or the recommender system can shape users' friending strategies implicitly. Some future studies, especially well-designed controlled experiments, may be helpful to model users' friending behaviours in the complex context of online social media.

Finally, these discoveries suggest further explorations and applications of structural diversity in practice. The method proposed in this chapter is general, and easy to incorporate as a node-level descriptor to associate with various other user properties and metrics (i.e., personality traits, opinion polarization, spamminess, etc.), and to use as features in recommendation and inference tasks. For example, the OcO dynamics of nodes suggest that users grow their networks in two different phases: exploiting local connections and exploring external connections; a smart recommender system should be able to leverage this pattern and make different recommendations accordingly.

5.8 Summary

In this chapter, I propose a new method to measure structural diversity for users in large social networks in continuous node embedding spaces. By calculating the structural diversity of users in a large-scale Snapchat network, we observe an intriguing novel pattern of "friendship saturation" that reflects a user's variation between two friending strategies. I introduce a new random graph generation model which successfully simulates the observed pattern of structural diversity variation. The analysis based on the proposed metric indicates the complex relationship between structural diversity and heterogeneous types of user behaviors with either narrowcasting or broadcasting natures.

CHAPTER VI

Conclusion

The studies within this dissertation reflect the importance and complexity of data-driven research on unsupervised diversity measurement and its applications based on large-scale data. The size of a dataset forces people to figure out how to measure diversity without supervision and validate the effect of new measurements in real applications. The dissertation proposes several conclusions regarding these challenges. The metric design, axiomatic analysis, and empirical studies on real applications result in some findings for a broad audience.

Metrics in unsupervised representation: Several representation learning techniques, especially topic modeling, text embedding, and graph embedding, are adopted to convert data objects into representations in both discrete space and continuous space. I design several metrics within these spaces and successfully measure scholars' research diversity in science of science and users' structural diversity in online social networks.

Axiomatic analysis of metric choices: An axiomatic analysis method is introduced in this dissertation to dive into the properties of existing and proposed diversity metrics. It is found that we cannot choose an optimal diversity metric in discrete space to meet all of the requirements for a good diversity metric. It is important to decision makers to understand what perspective on diversity is critical before they choose diversity metrics. However, a simple and intuitive diversity metric, average distance, performs well in axiomatic analysis in continuous space. It can be applied easily to different social applications. Additionally, there are concerns involved in real applications and the choice between discrete and continuous representations.

Social applications of diversity metrics: Through my initial exploration in two real-world applications, I find a significant effect of research diversity in science of science and a variation in user structural diversity in online social media. I reveal a strong relationship between research diversity and research impact and disclose their dynamics through a trajectory analysis. In the research on online social networks, I capture the “friendship saturation” phenomenon and propose a graph generation model to simulate it accurately. The findings regarding the influence of structural diversity on user engagement shed light on the potential to model user experience in online social media.

6.1 Implications

The dissertation has explored the potential of implementations of diversity in the real-world applications. It offers implications for both academia and our society.

The axiomatic analysis covered in this dissertation provides a solid testbed to judge whether a metric is sound. It plays as a lower bound to check whether heuristics-based metrics can meet the standard of a good metric in theory. Researchers in social science can leverage the results of axiomatic analysis to choose a theoretically sound metric when diversity is a key factor to be studied in their research. For researchers in the data science community, the analysis and the axiomatic analysis method can be a source to refer to when selecting objectives to optimize in machine learning frameworks.

My studies of real-world data reveal the positive influences of research diversity and structural diversity. They suggest that with a carefully designed metric and well-learned data representation, decision makers can pick appropriate metrics for diversity and try to utilize them in a reasonable and efficient way in real social applications, such as scientific policy making and user friendship recommendations.

On the other hand, although the theoretical analysis suggests some good metrics to use, based on the empirical studies on research diversity and structural diversity, it is hard to claim that diversity is the dominating decisive power in social implications. There is no easy rule-of-thumb to follow when we considering whether to add diversity into an existing system and what the best metrics are for real applications. It is better to run a considerable number of field studies before accepting all the results from theoretical research. Additionally, it is more important to figure out what is unit of analysis and what is good data representation than it is to implement any diversity metrics when facing real challenges.

6.2 Future Work

My research on diversity measurement and applications occupies one small niche within a large complicated research domain. I hope the findings and the potential outcomes from the proposed studies can motivate deeper and broader research on large-scale unsupervised diversity research in the future. In particular, there are two general research directions that deserve more attention.

Design heuristics-based metrics with modern representation learning: The developing representation learning techniques pave the way to better understanding of datasets, especially those depicting objects in high-dimensional space in an accurate manner. These techniques have been incorporated in all of the studies in this dissertation. The effects of diversity, which are revealed in Chapter IV and Chapter V, also partially come from these techniques. Diversity is one example of merging traditional metric design methods with the recent complicated representation learning methods. There are many other factors which stakeholders care about that could be redesigned and incorporated using the methodologies introduced in this paper. For example, researchers discussed readability, freshness, fluency, and perplexity in text generation. These are important criteria for judging whether a generated text is in good shape. How to measure them when facing representations learned using complicated language models is very challenging and intriguing. The axiomatic analysis and application design in this dissertation may be useful for researchers who want to build a reliable system based on existing heuristics-based metrics.

Optimizing diversity in a machine learning framework: In Chapter IV and Chapter V, I have shown the complicated relationship between research diversity and research impact, and the influence of structural diversity on users' behavior. Diversity plays intriguing and positive roles in the social implications. A follow-up question that naturally arises is whether we are able to optimize diversity along with other objectives. Average distance is a simply formed yet hard to derive definition. This metric or other diversity metrics can be incorporated into machine learning frameworks. Researchers can add these metrics as regularizers or constraints for complicated optimization goals and tune the parameters. Thus, diversity can not only be a measure to understand social dynamics but also a goal to pursue computationally. It can be a good factor to use in nudging both human and machine learning development.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. Adams, L. Jackson, and S. Marshall. Bibliometric analysis of interdisciplinary research. *Report to Higher Education Funding Council for England*, 2007.

[2] J. Alcalde-Unzu and M. Vorsatz. Measuring the cohesiveness of preferences: an axiomatic analysis. *Social Choice and Welfare*, 41(4):965–988, 2013.

[3] E. Amigó, D. Spina, and J. Carrillo-de Albornoz. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 625–634, 2018.

[4] S. Aral and P. Dhillon. Unpacking novelty: The anatomy of vision advantages. *Available at SSRN 2388254*, 2016.

[5] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[6] S. Aral and C. Nicolaides. Exercise contagion in a global social network. *Nature communications*, 8:14753, 2017.

[7] S. Aral and M. Van Alstyne. The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171, 2011.

[8] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[9] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng. Popularity prediction in microblogging network: a case study on sina weibo. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 177–178. ACM, 2013.

[10] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[11] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, 2019.

[12] M. Benhenda. Chemgan challenge for drug discovery: can ai reproduce natural chemical diversity? *arXiv preprint arXiv:1708.08227*, 2017.

[13] S. Bervoets and N. Gravel. Appraising diversity with an ordinal notion of similarity: an axiomatic approach. *Mathematical Social Sciences*, 53(3):259–273, 2007.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[16] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6), 2009.

[17] W. Bossert, P. K. Pattanaik, and Y. Xu. Similarity of options and the measurement of diversity. *Journal of Theoretical Politics*, 15(4):405–421, 2003.

[18] K. W. Boyack and R. Klavans. Measuring multidisciplinarity using the circle of science. In *From WRK1: Tracking and Evaluating Interdisciplinary Research, Workshop at ISSI*, volume 87122, 2009.

[19] R. S. Burt. Structural holes: The social structure of competition. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1992.

[20] K. Cao and S. Clark. Latent variable dialogue models and their diversity. *arXiv preprint arXiv:1702.05962*, 2017.

[21] N. Carayol and T. U. N. Thi. Why do academic scientists engage in interdisciplinary research? *Research evaluation*, 14(1):70–79, 2005.

[22] V. Caselles, J.-M. Morel, and C. Sbert. An axiomatic approach to image interpolation. *IEEE Transactions on image processing*, 7(3):376–386, 1998.

[23] L. Cassi, W. Mescheba, and E. De Turckheim. How to evaluate the degree of interdisciplinarity of an institution? *Scientometrics*, 101(3):1871–1895, 2014.

[24] D. Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[25] S. Chen, C. Arsenault, and V. Larivière. Are top-cited papers more interdisciplinary? *Journal of Informetrics*, 9(4):1034–1046, 2015.

[26] S. Chen, Y. Gingras, C. Arsenault, and V. Larivière. Interdisciplinarity patterns of highly-cited papers: A cross-disciplinary analysis. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–4, 2014.

[27] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[28] M. Cochran-Smith. *Walking the road: Race, diversity, and social justice in teacher education.* Teachers College Press, 2004.

[29] H. Conley, F. Colgan, C. Creegan, A. McKearney, and T. Wright. Equality and diversity policies and practices at work: lesbian, gay and bisexual workers. *Equal Opportunities International*, 2007.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[31] Y. Dong, R. A. Johnson, and N. V. Chawla. Will this paper increase your h-index? scientific impact prediction. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 149–158, 2015.

[32] Y. Dong, R. A. Johnson, J. Xu, and N. V. Chawla. Structural diversity and homophily: A study across more than one hundred big networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 807–816. ACM, 2017.

[33] C. Dowden. A new axiomatic approach to diversity. *arXiv preprint arXiv:1101.5305*, 2011.

[34] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.

[35] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[36] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2004.

[37] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487, 2005.

[38] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[39] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. Torr, and P. K. Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8513–8521, 2018.

[40] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, pages 381–390, 2009.

[41] M. S. Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[42] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[43] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450, 2018.

[44] H. Habib, N. Shah, and R. Vaish. Impact of contextual factors on snapchat public sharing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[45] B. He, Y. Ding, J. Tang, V. Reguramalingam, and J. Bollen. Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective. *Journal of Informetrics*, 7(1):117–128, 2013.

[46] I. N. Herstein and J. Milnor. An axiomatic approach to measurable utility. *Econometrica, Journal of the Econometric Society*, pages 291–297, 1953.

[47] T. Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in neural information processing systems*, pages 914–920, 2000.

[48] T. Hu, Y. Xia, and J. Luo. To return or to explore: Modelling human mobility and dynamics in cyberspace. In *The World Wide Web Conference*, pages 705–716, 2019.

[49] U. Jain, Z. Zhang, and A. G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6485–6494, 2017.

[50] P. Jensen and K. Lutkouskaya. The many dimensions of laboratories' interdisciplinarity. *Scientometrics*, 98(1):619–631, 2014.

[51] Y. Jo, J. E. Hopcroft, and C. Lagoze. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web*, pages 257–266, 2011.

[52] N. L. Johnson and S. Kotz. Urn models and their application; an approach to modern discrete probability theory. 1977.

[53] M. Kalkbrener. An axiomatic approach to capital allocation. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 15(3):425–437, 2005.

[54] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, and V. Ramavajjala. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 955–964, New York, NY, USA, 2016. Association for Computing Machinery.

[55] M. Karlovčec and D. Mladenić. Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, 102(1):433–454, 2015.

[56] J. Kaur, F. Radicchi, and F. Menczer. Universality of scholarly impact metrics. *Journal of Informetrics*, 7(4):924–932, 2013.

[57] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[58] J. M. Kleinberg. An impossibility theorem for clustering. In *Advances in neural information processing systems*, pages 463–470, 2003.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] R. Kumar, J. Novak, and A. Tomkins. *Structure and Evolution of Online Social Networks*, pages 337–357. Springer New York, New York, NY, 2010.

[61] V. Larivière and Y. Gingras. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1):126–131, 2010.

[62] V. Larivière, S. Haustein, and K. Börner. Long-distance interdisciplinarity leads to higher scientific impact. *Plos one*, 10(3), 2015.

[63] R. Laxton. The measure of diversity. *Journal of theoretical biology*, 70(1):51–67, 1978.

[64] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[65] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA, 2019.

[66] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[67] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[68] J. Levitt and M. Thelwall. The most highly cited library and information science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, 78(1):45–67, 2009.

[69] J. M. Levitt and M. Thelwall. Is multidisciplinary research more highly cited? a macrolevel study. *Journal of the American Society for Information Science and Technology*, 59(12):1973–1984, 2008.

[70] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1565–1568, 2010.

[71] D. Li, Q. Huang, X. He, L. Zhang, and M.-T. Sun. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018.

[72] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[73] W. H. Lin, K.-T. Chen, H. Y. Chiang, and W. Hsu. Netizen-style commenting on fashion photos: dataset and diversity measures. In *Companion Proceedings of the The Web Conference 2018*, pages 395–402, 2018.

[74] Y. Liu, X. Shi, L. Pierce, and X. Ren. Characterizing and forecasting user engagement with in-app action graph: A case study of snapchat. *arXiv preprint arXiv:1906.00355*, 2019.

[75] A. E. Magurran. *Measuring biological diversity*. John Wiley & Sons, 2013.

[76] J. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *NIPS*, volume 2012, pages 548–56. Citeseer, 2012.

[77] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[78] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[79] P. R. Morales, R. Lamarche-Perrin, R. Fournier-S'Niehotta, R. Poulain, L. Tabourier, and F. Tarissan. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115, 2021.

[80] L. G. Nichols. A topic model approach to measuring interdisciplinarity at the national science foundation. *Scientometrics*, 100(3):741–754, 2014.

[81] H. Nilforoshan and N. Shah. Silcendice: Mining suspicious multi-attribute entity groups with multi-view graphs. *arXiv preprint arXiv:1908.07087*, 2019.

[82] J. Paparrizos and L. Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870, 2015.

[83] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato. On the predictability of future impact in science. *Scientific reports*, 3(1):1–8, 2013.

[84] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[85] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[86] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[87] I. V. Ponomarev, B. K. Lawton, D. E. Williams, and J. D. Schnell. Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, 100(3):755–765, 2014.

[88] A. Porter, A. Cohen, J. David Roessner, and M. Perreault. Measuring researcher interdisciplinarity. *Scientometrics*, 72(1):117–147, 2007.

[89] A. Prasad, S. Jegelka, and D. Batra. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *Advances in Neural Information Processing Systems*, pages 2645–2653, 2014.

[90] K. Prewitt. Racial classification in america: where do we go from here? *Daedalus*, 134(1):5–17, 2005.

[91] I. Rafols, L. Leydesdorff, A. O'Hare, P. Nightingale, and A. Stirling. How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research policy*, 41(7):1262–1282, 2012.

[92] I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010.

[93] D. Roessner, A. L. Porter, N. J. Nersessian, and S. Carley. Validating indicators of interdisciplinarity: linking bibliometric measures to studies of engineering research labs. *Scientometrics*, 94(2):439–468, 2013.

[94] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[95] J. Ruscio, F. Seaman, C. D'Oriano, E. Stremlo, and K. Mahalchik. Measuring scholarly impact using modern citation-based indices. *Measurement: Interdisciplinary Research and Perspectives*, 10(3):123–146, 2012.

[96] J. Sanz-Cruzado and P. Castells. Enhancing structural diversity in social networks by recommending weak ties. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 233–241. ACM, 2018.

[97] M. Sedighi. Interdisciplinary relations in some high-priority fields of science and technology. *Library Review*, 2013.

[98] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017.

[99] Y. Shmargad. Structural diversity and tie strength in the purchase of a social networking app. *Journal of the Association for Information Science and Technology*, 69(5):660–674, 2018.

[100] F. N. Silva, F. A. Rodrigues, O. N. Oliveira Jr, and L. d. F. Costa. Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, 7(2):469–477, 2013.

[101] E. H. Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

[102] T. W. Steele and J. C. Stier. The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science*, 51(5):476–484, 2000.

[103] C. Sternitzke and I. Bergmann. Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, 78(1):113–130, 2009.

[104] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.

[105] J. Su, K. Kamath, A. Sharma, J. Ugander, and S. Goel. An experimental study of structural diversity in social networks. *arXiv preprint arXiv:1909.03543*, 2019.

[106] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[107] X. Tang, Y. Liu, N. Shah, X. Shi, S. Wang, and P. Mitra. Knowing your fate: Friendship, action and temporal explanations for user engagement prediction on social apps. In *SIGKDD*, 2020.

[108] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

[109] D. Valcarce, J. Parapar, and Á. Barreiro. Axiomatic analysis of language modelling of recommender systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(Suppl. 2):113–127, 2017.

[110] F. J. Van Rijnsoever and L. K. Hessels. Factors associated with disciplinary and interdisciplinary research collaboration. *Research policy*, 40(3):463–472, 2011.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[112] T. Velden and C. Lagoze. The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, 64(12):2405–2427, 2013.

[113] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster.

[114] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[115] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner. Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of informetrics*, 5(1):14–26, 2011.

[116] I. Waller and A. Anderson. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, pages 1954–1964. ACM, 2019.

[117] J. Wang, B. Thijs, and W. Glänzel. Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PloS one*, 10(5), 2015.

[118] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[119] W. Weaver. Recent contributions to the mathematical theory of communication. *ETC: a review of general semantics*, pages 261–281, 1953.

[120] J. W. Weibull. An axiomatic approach to the measurement of accessibility. *Regional science and urban economics*, 6(4):357–379, 1976.

[121] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131, 2005.

[122] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522, 2013.

[123] G. J. Woeginger. An axiomatic analysis of egghe's g-index. *Journal of Informetrics*, 2(4):364–368, 2008.

[124] J. Xu, X. Ren, J. Lin, and X. Sun. Dp-gan: Diversity-promoting generative adversarial network for generating informative and diversified text. *arXiv preprint arXiv:1802.01345*, 2018.

[125] E. Yan. Finding knowledge paths among scientific disciplines. *Journal of the Association for Information Science and Technology*, 65(11):2331–2347, 2014.

[126] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1247–1252, 2011.

[127] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[128] A. Yegros-Yegros, I. Rafols, and P. D'Este. Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PloS one*, 10(8), 2015.

[129] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[130] J. Zhang, Y. Dong, Y. Wang, J. Tang, and M. Ding. Prone: Fast and scalable network representation learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4278–4284. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[131] Y. Zhang, L. Wang, J. J. Zhu, X. Wang, and A. Pentland. The strength of structural diversity in online social networks. *arXiv preprint arXiv:1906.00756*, 2019.

[132] Z. Zhu, S. Xu, M. Qu, and J. Tang. Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference*, pages 2494–2504. ACM, 2019.