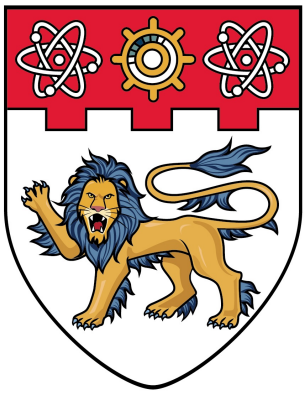


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

KNOWLEDGE EVOLUTION: FROM THE COMPLEX NETWORK PERSPECTIVE

LIU WENYUAN

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2019

KNOWLEDGE EVOLUTION: FROM THE COMPLEX NETWORK PERSPECTIVE

LIU WENYUAN

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[Input Date Here]

15/1/2019

Date

[Input Signature Here]

Liu Wenyuan

Liu Wenyuan

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[Input Date Here]

15/1/19

Date

[Input Supervisor Signature Here]

Cheong Siew Ann

Authorship Attribution Statement

This thesis contains material from 1 paper(s) published in the following peer-reviewed journal(s) where I was the first and/or corresponding author.

Chapter 2 is published as Liu, W., Nanetti, A., & Cheong, S. A. (2017). Knowledge evolution in physics research: An analysis of bibliographic coupling networks. PloS one, 12(9), e0184821. DOI: 10.1371/journal.pone.0184821.

The contributions of the co-authors are as follows:

- I designed research, performed research, analyzed data and wrote the paper.
- Assoc Prof Cheong Siew Ann designed research, wrote the paper.
- Assoc Prof Andrea Nanetti revised the manuscript.

[Input Date Here]

15/1/2019

Date

[Input Signature Here]

Liu Wenyuan

Liu Wenyuan

Abstract

As a rapidly developing research area, the science of science (SciSci) is devoted to quantifying, understanding, and predicting scientific research and its outputs. While much progress has been achieved on impact measuring and the collaboration network, the research on the dynamical evolution of whole research systems is much less studied. This is the key question we try to answer in this dissertation: how do we quantify, understand, and predict the evolution of scientific research, and more generally, the processes of innovation? Not only can the answer to this question help universities and research institutes recruit new scientists, and point governments and companies to the most fruitful research frontier to fund, it also opens up many new science questions: for example, are there any laws governing the enterprise of scientific research?

To answer this question, we first built a data-driven framework for studying knowledge evolution. Using the American Physical Society (APS) publications data sets, we constructed year-to-year bibliographic coupling networks, and identified validated communities — topical clusters (TCs) — that represent different research fields in them. We then visualized their evolutionary relationships in the form of alluvial diagrams, and showed how they remain intact through APS journal splits. Quantitatively, we saw that most fields

undergo weak mixing, and it is rare for a field to remain isolated or undergo strong mixing. The sizes of fields obey a simple linear growth with recombination. We can also reliably predict the merging between two fields, but not for the considerably more complex splitting. We reported a case study of two fields that underwent repeated merging and splitting around 1995, and how these Kuhnian events are correlated with breakthroughs on Bose-Einstein condensation (BEC), quantum teleportation, and slow light. This impact showed up quantitatively in the citations of the BEC field as a larger proportion of references from during and shortly after these events.

In addition to this empirical study of the APS data set, we also used the linguistic information available in their abstracts to study how scientific memes evolve during knowledge evolution. This can help us gain a more complete understanding of knowledge evolution beyond citations. We found that particular memes are associated with particular TCs, making memes good labels for the TCs' research contents, at the same time making the alluvial diagram more comprehensible. Like a TC, a meme also has a complex evolution process. We measured the co-occurrence probability for meme pairs, and found 'quantum' and 'optical' grew closer since 1981, which is consistent with the rise of quantum optics. The co-evolution between memes and TCs is also discussed.

Given the close relationship between evolution processes and scientific breakthroughs, it is important to be able to predict the future events. Having the predictive features describing a given TC and its known evolution in the next year, we can train a machine learning model to predict future changes of TCs, i.e., their continuing, dissolving, merging and splitting. We found the number of papers from certain journals, the degree, closeness, and betweenness to be the most predictive features. Additionally, betweenness of TCs revealed its significant increase for merging events. Our results represent a first step from a descriptive understanding of the SciSci, towards one that is ultimately prescriptive.

Keywords: knowledge evolution, complex network, community, meme, machine learning.

Acknowledgement

First and foremost I want to thank my advisor Assoc. Prof. Cheong Siew Ann. I would like to thank him for the insightful discussions, generous funding and responsible guidance. Without his help, the last two and half years research is impossible. I really appreciate his invaluable help with the writing. I thank my thesis advisory committee members: Assoc. Prof. Chew Lock Yue and Assoc. Prof. Xiao Gaoxi, for their discussions and suggestions. I thank my collaborators Assoc. Prof. Andrea Nanetti (Chapter 2), Asst. Prof. Tobias Kuhn (Chapter 3), Dr. Stanisław Saganowski (Chapter 4) and Prof. Przemysław Kazienko (Chapter 4) for their contributions in corresponding chapters.

I would like to thank my friends Darrell Tay, Teck Liang, Boon Kin, Woon Peng and Misha, for your companionship, support and encouragement. I thank Nanyang Technological University for providing a scholarship and a research environment that facilitate me to finish my PhD programme. I thank the staffs in SPMS for their help with all paperwork, which saved me a lot of time.

I owe my deepest gratitude to my father. His love and support encourage me to pursue my dreams. Thank you.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Modern science shapes the society in many important ways and the role of scientific knowledge continue to grow in human society. Because of its profound influence on our daily life, companies and governments invest a lot on scientific research to make our world better. In some sense this outlook resulted from the scientific revolution [Cohen, 1976], which made the practice of science increasingly professionalized and institutionalized, and scientific research more than just personal interest. These changes are closely associated related with industrial revolutions, which greatly boosted productivity, reinforcing people's positive attitude towards science and making scientific research an indispensable part of our society.

Given this importance in our society, understanding the process of scientific research becomes crucial. For example, universities and research institutes need to find qualified candidates; companies and governments need to develop funding policies; scientists also need to select their research directions and topics to have considerable outcomes to get promoted and obtain more research resources. These challenging issues have existed for

a very long time. In the past, people rely on intuition and experience to solve them. However, as science develops and expands, these traditional solutions become inadequate to handle the emerging problems and challenges. The effort made to meet this strong and urgent demand is a new research area called the *Science of Science* (SciSci) [Fortunato et al., 2018; Zeng et al., 2017].

As a rapidly developing field, SciSci aims to quantify, understand and model the practice of science, including not only the papers, but also the ideas, scholars, research organizations and so on. Two key conditions are crucial to the emergence of SciSci. The first is data availability. Nowadays scientific documents are digitized and can be accessed through the World Wide Web. Many data sources store and organize the massive scientific literature (arXiv, Scopus, PubMed, Web of Science, the U.S. Patent and Trademark Offices, and others). These datasets make the large-scale analysis of science and testing empirical data with model possible. The second is the rise of the field of network science. In the last two decades, the field of network science developed rapidly [Albert and Barabasi, 2002; Newman, 2010; Dorogovtsev et al., 2007]. As the name suggests, in the field of network science we focus on the study of the components of the system, such that the interactions between components are described by the network model and the whole system has non-trivial behavior, unlike their components. The methodology of network science have been found to be very useful in study of social network [Borgatti et al., 2009], the topological structure of the Internet [Doyle et al., 2005], transportation network [Banavar et al., 1999] and the human disease network [Goh et al., 2007], which are very hard to study using traditional methods. For the purpose of this thesis, we need to understand why network science is so helpful for the research of SciSci? The reason is that the practice of science is embedded within multiple networks. First, scientific ideas rely on previous achievements, and thus form a network of ideas. Second, scientists collaborate to solve difficult questions,

forming a social network. Finally, papers cite other papers, forming a citation network. All these facts suggest the science does not work alone, but through complex connections. To understand the structure and dynamics of science, we need to understand these connections. Network methods proved to be very useful in such a study. With the aid of complex networks, scientists are able to study the complex system of science, which includes ideas, researchers, research institutes, funding agencies and research articles. Their relations are indeed complex and intractable with traditional methods. It is also worthy to point out that such benefits are mutual — complex network theory help solve problems in SciSci, at the same time we ask new questions in SciSci stimulating the development of complex network theory.

In the last decade, significant progresses have been made in SciSci at the macroscopic and microscopic levels. At the microscopic level, that is individual paper or scientist, [Wang et al., 2013; Radicchi et al., 2008] have found the universality of citation distribution and predictability of long-term citation pattern. This universality suggests a general law that governs the citation dynamics in all disciplines, which involves preferential attachment [Jeong et al., 2001], attention decay [Parolo et al., 2015] and a fitness parameter that depends on the intrinsic quality of paper [Eom and Fortunato, 2011]. Interestingly, some papers deviate from this universal curve: they receive very little attention after publication but suddenly get a burst of citations some time later. This kind of papers are called *sleeping beauties* and are particularly important for SciSci because they are related to scientific breakthroughs [Ke et al., 2015]. As the producer of scientific knowledge, scientists' behavior also attracts our attention. Studies have shown that the success of a scientific career is indeed a complicated question, and is influenced by gender [Larivière et al., 2013], the scientist's mobility [Deville et al., 2014], reputation [Petersen et al., 2014], and career stage [Sinatra et al., 2016].

If we group papers or scientists, we move to the macroscopic level, to gain a big picture of science. As early as 1963, [de Solla Price, 1963] showed that the corpus of scientific documents grow exponentially, and this rapid growth continues today [Fortunato et al., 2018]. This process is a serious challenge to the entire scientific community, and its effect has been disscussed by [Pan et al., 2018]. Using bibliographic data, [Skupin, 2004] visualized a knowledge domain with cartographic means, and the result showed a clear landscape of research frontiers. The history and development of particular disciplines have also been studied in [Bettencourt and Kaur, 2011; Sinatra et al., 2015] for sustainability science and physics. This idea can even be extended beyond individual discipline. By studying the data set covering multiple disciplines, researchers create a map of the relations between disciplines [Boyack et al., 2005; Leydesdorff and Rafols, 2009]. On the other hand, the papers and scientists can also be group together based on nations or institutes to study scientific competition [Cimini et al., 2014; Cimini et al., 2016].

The research mentioned above provide valuable insight to SciSci, cover both small and large scales, include measurement, modelling and prediction. However, the mesoscopic picture—the level between microscopic scale and macroscopical scale is missed. In the language of complex network, the community level of science is not well studied. Community structure is the level between individual nodes and the whole component. The idea of community structure comes from the observation that there exist tightly knit groups inside many real-world networks, which include social networks [Girvan and Newman, 2002], transportation networks [Mossa et al., 2005], telephone networks [Blondel et al., 2008], neural networks [Hizanidis et al., 2016] and citation networks [Chen and Redner, 2010]. The nodes within such groups are well-connected while at the same time there are only looser connections between nodes in different groups. As a natural division within the network, community structure is very important to understand the large-scale structure of the

network, function of the network structure, the dynamical process within the network and also the principles for designing artificial networks. Unfortunately, the community is not a well-defined concept. Many different definitions of community have been proposed. One type is based on the difference between the internal and external edges, which is very easy to follow: since the community is the group of nodes have more insider connections than outside connections, such difference should be larger for "good" community division and small for "bad" community division. Based on this idea, the concept of modularity have been proposed [Newman and Girvan, 2004], and many algorithms have been devised to get community partition by trying to maximize the modularity. Unfortunately, finding an optimal graph partition which has maximum modularity is known in general to be an NP-hard problem, and therefore many efforts have been devoted to heuristic methods [Newman and Girvan, 2004; Blondel et al., 2008; Sobolevsky et al., 2014]. Over the course of these studies, people also realized that there are several problem with modularity. One is the modularity landscape has a larger number of distinct partitions, whose modularity are very close to the global maximum [Good et al., 2010], which may lead to meaningless partitions with high modularity. Another problem is about resolution limitation, that means modularity optimization may fail to extract communities smaller than a certain size, which depends on the total size of the network and on the degree of interconnectedness of the modules [Fortunato and Barthelemy, 2007].

Another class of algorithms utilize random walk to reveal the community structure. If nodes are well connected by intra-community links and loosely connected by external links, random walkers would be trapped inside each community for a long time, before finding a way out and moving to another community. Based on this idea, two popular methods **Walktrap** [Pons and Latapy, 2005] and **Infomap** [Rosvall and Bergstrom, 2008] are proposed . The Walktrap method calculate the similarity between nodes i and j by

the probability that a random walker moves from i to j in a fixed t steps. Such similarity should be large between nodes in the same community and small between nodes in different communities. The Infomap method try to yield the minimum description length of an infinitely long random walk. The code words that minimize the description length can then reveal the community structure.

A different way to tackle the problem of community detection is to use statistical inference. The stochastic block model (SBM) is widely used as a generative model of graphs with communities. Many different SBM models are proposed for community detection with the same goal: getting the parameters which maximize the likelihood of model to generate the network [Karrer and Newman, 2011; Peixoto, 2014]. Beside SBM, some models are proposed to explain the formation of community [Grönlund and Holme, 2004; Iñiguez et al., 2009]. A useful literature review about community detection can be found in [Fortunato, 2010; Fortunato and Hric, 2016].

In this dissertation, we would like to address the evolution of science at the mesoscopic scale—the community level. The work included in this dissertation is primarily motivated by the observation that the discipline and subdiscipline structure in science can be well described by the community structure. The study on this level may give rise to a new understanding of the trend of interdisciplinary research in recent decades. The history of science has shown that the interaction between different disciplines is very complex. On the one hand, different disciplines have their own questions, assumptions, and methods, which make them professional and different from other disciplines. One important driving force in history of modern science is differentiation, where starting from traditional philosophy, we now have mathematics, physics and chemistry. From the tradition of *Naturalis historia*, we have modern biology and geology. Such differentiation boost the research in both breadth and depth. On the other hand, communication between the different dis-

cipliness is also very important, perhaps even indispensable. Nowadays scientists face very complex questions, and need tools from different disciplines, a well known trend in interdisciplinary science and SciSci itself is a good example. Also, research in one field can inspire people in other fields, like topological insulators, fiber bundle, gauge theory, quantum computation, and quantum information, just to name a few. Despite the importance of this question, the evolution of science at the mesoscopic level is in its infancy. Researchers already noticed the community structure of the Physical Review citation network and tracked the important papers over 80 years [Chen and Redner, 2010]. The rapidly developing discipline Neuroscience be identified as the outcome of interaction between Psychology, Neurology and Molecular & Cell biology. [Rosvall and Bergstrom, 2010]. Research at the macroscopic level has already noted the importance of the mesoscopic level, like the subfield structure in physics research [Sinatra et al., 2015]. We will take a further step in this direction, to study the evolution of science at the mescoscopic level in a systematic way, propose a method to quantify this process and analyse the role of interactions between different communities, to shed light on SciSci at the mescoscopic level.

The dissertation is organized as follows. In Chapter 2 we define the knowledge evolution in terms of community structure evolution and study this evolution in the APS citation network. Chapter 2 is divided in two parts. In the first part, we argue that a group of papers published in the same time period sharing significant overlap of references can be defined as a research field, which we call a topical cluster (TC). TC is the elementary unit in knowledge evolution and the inheritance relation between TCs in consecutive years can be defined in terms of their reference's similarity, which we measure using intimacy indices. The alluvial diagram of knowledge evolution in the APS dataset is drawn based on TCs and their intimacy indices. This approach is general, and can also be used to track co-citation evolution with some modifications (see Chapter 4). In the second part, we analyse

the statistical distribution of merging and splitting, and the phenomenon of linear recombination in merging. We find that strong inter-community connection is predictive of an imminent merging event between two TCs, but a more complex correlation between intra-community structure and a splitting event arising in the TC. We report a case study of two fields that underwent repeated merging and splitting in the 1990s and how these events produce large impacts on the TCs' reference distributions. The results in this chapter were published in [Liu et al., 2017]

In Chapter 3, we study the knowledge evolution process from a different perspective—language. It is well known that different research fields use different terminologies and like the reference, terminology also evolves with time. Using the concept of scientific memes proposed in [Kuhn et al., 2014], which are words that successfully spread themselves through citations, we develop a method that can label TCs with highly-correlated memes, to significantly improve the readability of alluvial diagrams. Like TCs, memes also have rich interactions. We develop a method to measure the distance between memes and found 'optical' and 'quantum' to be among the strongest merging meme pairs in the last three decades, consistent with what we know from the history of physics. The coevolution between meme and TC is also discussed.

In Chapter 4 we move on to prediction, which has important implications for funding policies. It has been shown in Chapter 2 that the correlation between evolution events and network metrics is very complex. Therefore we introduce the machine learning techniques, which are good at handling complex relations between multiple variables, to predict the evolution event based on network structure. To do so, we modified the group evolution prediction (GEP) method used in social network analysis to classify the evolution events and predict future events based on a selected set of network metrics. The prediction performance is significantly higher than random guesses, tell us therefore that the knowledge

evolution process is intrinsically predictable. Furthermore, feature ranking suggests that betweenness is very informative for prediction, and this prompted us to do an in-depth analysis of the relationship between betweenness and the knowledge evolution events.

We conclude in Chapter 5, summarizing our findings and discussing the outlook enabled by this study.

CHAPTER 2

Evolution of community structure in the Physical Review citation network

According to Karl Popper, science progresses through repeated hypothesis testing [Popper, 2013]. Hypotheses contrary to empirical evidence must be rejected, while those consistent with data survive to be tested another day. In this picture of the scientific enterprise, our knowledge of the world around us is always tentative, but becomes more complete over time. On the other hand, Thomas Kuhn believes that the accepted knowledge of a given time is the result of consensus amongst scientists, based on evidences consistent with their theories [Kuhn, 1962]. However, when too many conflicting evidences are found, a new consensus can form around new theories in what he called a 'paradigm shift'. Kuhn gives special relativity and quantum theory as examples of paradigm shifts. Looking back, we realize these two theories have enormous impacts on how we understand the world today. But could there be paradigm shifts of various scales that have also contributed to reshaping our knowledge of physics?

Many historians of science have noted the strongly reductionistic flavor of scientific research in the last couple of centuries [Wootton, 2015]. Starting as natural philosophy, the

body of scientific knowledge became separated disciplines of astronomy, biology, chemistry and physics. Within physics itself, we also observe the emergence of high energy physics, condensed matter physics, biophysics, and photonics. These are the results of the splitting of science into more specialized fields. At the same time we also observe in parallel the merging of fields, such as the merging of astronomy and physics to give astrophysics, biology and chemistry to give biochemistry, and others "that arose by division and recombination of specialties already matured" [Kuhn, 1962]. These developments have been discussed extensively by philosophers and historians of science, but unlike our quantitative understanding of physics, our appreciation for the processes through which we acquired our knowledge of physics remains at a highly descriptive level. Some progress has been made in addressing this problem [Bollen et al., 2009; Kuhn et al., 2014; Jia et al., 2017]. In particular, the following three papers provide the inspiration for our study. Chen and Redner suggested that long-range connections can form between disparate fields because of the development of "a widely applicable theoretical technique, or cross fertilization between theory and experiment" [Chen and Redner, 2010]. Visualizing the cross citations between neuroscience journals, Rosvall and Bergstrom traced the growth and maturation of neuroscience as a discipline [Rosvall and Bergstrom, 2010]. Using embryology as a specific example, Chavalarias and Cointet created a phylomemetic network visualization for the evolution of science [Chavalarias and Cointet, 2013].

Before further discussion on the knowledge evolution, we would like to emphasize that in this thesis we focus on the quantitative study of knowledge evolution from the network perspective, instead of launching into a philosophical discussion. Having said this, we introduce the works of Karl Popper and Thomas Kuhn because our results suggest some connection or parallel between the empirical evolution of scientific knowledge and their philosophies of science. However a rigorous discussion of their philosophical opinions is

beyond the scope of our thesis, therefore we would like to give our own definition here to avoid potential confusion. The elementary unit of knowledge evolution in this thesis is a topical cluster, which is a idea-researcher-paper complex. As the first step, we consider the topical cluster at the paper level since evolution on multiple levels will make the analysis significantly harder (see Chapter 5 for more discussion). Inside this unit, ideas are closely connected with each other, researchers who have common interest collaborate with each other and the papers cite each other. Like the ship of Theseus, the idea-researcher-paper complex is also constantly changing. Considering the nature of these changes, it is very helpful to classify them into two qualitatively classes. Therefore, we introduce two operational definitions: Popperian processes and Kuhnian processes. A Popperian process is the **relatively smooth** change of topical cluster whereas a Kuhnian process is the **relatively severe** change in the topical cluster. These definitions are connected to the debate between Popper and Kuhn, although their meanings are not exactly as what Popper and Kuhn wrote in their books. An accurate and rigorous discussion about the connection between Popperian, Kuhnian process and repeated hypothesis testing, paradigm shift is beyond the scope of thesis. Therefore in the following chapters we only adopt our operational definitions.

2.1 Community structure of bibliographic coupling network

While these previous studies point to the evolution of scientific knowledge, they do not identify the entity that is recognizably 'knowledge', or they do not study the interactions between such objects. Therefore, it is instructive to discuss the definition of 'knowledge' first before studying its evolution. As the central topic of epistemology, 'knowledge' is a very subtle and complicated concept. We do not intend to engage with metaphysics in this thesis, but would instead we would like to give an operational definition of knowledge

which is useful from SciSci perspective. In this thesis, we define 'knowledge' as people's understanding of particular subjects. To motivate this definition, we have to understand that scientific research is a process of reaching consensus. Each paper, even after peer review, can only represent a single opinion. It is very common for people to hold different opinions on the same topic or study the same topic from different perspectives. They should all be considered as parts of the knowledge. Therefore we use "people" rather than "person" in our definition. At the same time, science is about finding answers to concrete questions, therefore people's understanding always concern certain subjects rather than something abstract. This definition is based on the features of scientific research, and is therefore suitable for the SciSci study. On the other hand, this definition is quite general: the method we take in following chapters is only one possible way to study knowledge evolution under this definition, and other methods are possible. While in this thesis we study knowledge evolution by tracing scientific papers, it is worth pointing out that knowledge evolution also happens at other levels. If one day the discussion and interaction between scientists during conference is available for research, knowledge evolution can also be studied using this data at the level of researchers. All of these are also consistent with our definition of 'knowledge'.

To clarify what constitutes knowledge, we start with the bibliographic coupling network (BCN) [Kessler, 1963], proposed by Kessler and used extensively in computer science [Yan and Ding, 2012; Huang et al., 2003]. In a BCN, nodes represent papers, and if two papers share w common references, we draw an edge with weight w between them (see Fig. 2.1). The BCN is suitable for our purpose for two reasons: (i) the BCN for a given year consists only of papers published that year and does not change after more papers are published later, so features in the BCN represent the state of knowledge in that year; and (ii) the appropriate collective unit of knowledge is a community in the BCN instead of a

few key papers or a journal. There is another way to construct the similarity network – co-citation network (CN) [Small, 1973]. As in the BCN, nodes represent papers in CN, and the edge between two nodes has weight w if these two papers are cited together by w other papers. The CN is very useful for citation analysis because it provide a forward-looking assessment on document similarity and thus we study the predictions on both BCN and CN in Chapter 4. However in this chapter we only focus on the BCN because compared to the CN, BCN is easier to interpret for knowledge evolution. As stated above, the BCN for a given year consists only of papers published that year and does not change after more papers are published later, so features in the BCN represent the state of knowledge in that year. By contrast, the CN can only reflect the papers (published before) which influence the knowledge in that year. In some sense, the BCN is the "first hand" network and the CN is the "second hand" network for knowledge evolution. Therefore in the following part of this chapter we only study the community evolution in the BCN.

Currently there are many large bibliographic data set available for research. The largest one we know is the Web of Science (WoS) database [web,]. According to its webpage, it covers publications in 256 disciplines in science, social science, arts, and humanities. It has more than 90 million records and each record includes citation indexing, author(s), title, abstract, publication year and so on. The width and depth of this database make it an ideal material for SciSci research, and many papers have been published based on the WoS database [Sinatra et al., 2015; Parolo et al., 2015; Uzzi et al., 2013]. However we do not have access to the WoS database, and therefore we do not use it in this thesis. Apart from the WoS data set, several other data sets are also good for SciSci research, including the APS data set, the NBER U.S. patent citations data set [usp,], the ACL anthology reference corpus [acl,], etc. Among all these data sets, we chose the APS data set as the main research subject for the following reasons. First, the APS data set is well known in citation

network research and many important results are already known about it. These are very useful for our own research, particularly the results in [Kuhn et al., 2014]. The second reason is that we are from a physics department, and therefore when compared to data sets from other disciplines, we have more domain knowledge on the APS dataset. This proved to be very important when we perform the meme analysis. In another data set, even we manage to find the representative memes, we will not understand the real meaning behind these memes. Although our method is applicable to all disciplines, domain knowledge is still very useful to gain particular insights when the research is in the primary stage.

The APS citation data is available from the American Physical Society for researchers who meet their criteria [APS,]. The data set includes two parts: one is for citing article pairs (if paper A cites paper B, there will be an entry consisting of the pair of DOIs for A and B), another is for metadata of papers (including DOI, journal, volume, issue, first page and last page, title, authors, affiliations, PACS codes, article type, etc.). Both of them are essential for our study, where we use the first part to construct the network, and the second part to mine more information about knowledge evolution beyond the citation level (PACS codes are used to test the community structure, titles and abstracts are used to extract scientific memes). The data set we used includes 541,448 papers from 1893 to 2013 and 6,040,030 citation pairs. Many important studies on SciSci have been done using the APS data set. One of the earliest study is from [Redner, 2004], where he studied the statistics of citations on APS papers from 1893 to 2003, and found it is consistent with linear preferential attachment. After this, more researchers studied this data set from different angles. Some researchers study the coauthorship network [Martin et al., 2013], which is very important to understand the trend of collaboration in modern physics. The problem of author name ambiguity has also been studied in the APS data set [Klosik et al., 2014; Deville et al., 2014]. Other studies focus on the citation distribution [Radicchi

and Castellano, 2011], citation prediction [Wang et al., 2013], or the measure of interdisciplinarity [Pan et al., 2012], just naming a few. In fact, these citation network research not only deepen our understanding on physics research, but the methods developed can also be used to study citation networks in other disciplines.

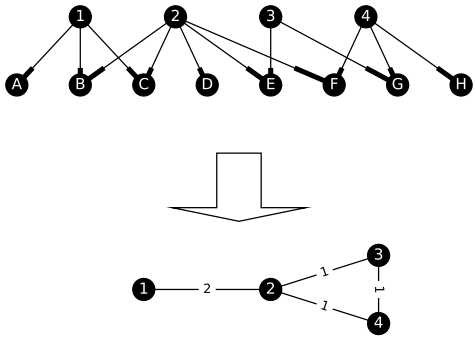


Figure 2.1: Building a BCN (lower) from a citation network (upper): circles with numbers are papers under consideration, circles with letters are their references, and numbers on edges are weights.

To determine the statistical significance of our empirical BCNs (Fig. 2.2(a)), we built a null model for comparison. In our null model, we fixed the out degrees and in degrees of all papers (citing and cited), so that we can directly compare the null model to the empirical BCN. To form the null model, we randomly rewired the edges of the citing and cited papers to get an ensemble of artificial bipartite citation networks. The ensemble of artificial BCNs is then obtained from these bipartite networks (Fig. 2.2(b)). If the empirical BCN is obtained purely by chance, its distribution of edge weights should be close to the distribution of edge weights of the ensemble of artificial BCNs. The results show that in the real BCN edge weights are far more heterogeneous (see Fig. 2.2(c)) than expected from our null model, suggesting that these weights are meaningful, and not the result of purely random connections. This heterogeneity can be explained by the presence of communi-

ties that we extracted using the Louvain method. Compared to the null model, the real BCN has more edges with high weight (in Fig. 2.2(c) all edges has weight more than 1 be considered high-weight edge). We suspect these are the most meaningful edges, arising from the paper's content. If two papers focus on the same topic, they will more likely have more than one reference in common. This effect also manifest itself in the degree distribution: the null model has a flatten degree distribution at small degrees because the edges are drawn by chance, whereas in the real BCN this coupling is based on content, meaning that papers will have edges mostly with papers that are trying to solve the same problems, so the real BCN will have more low-degree nodes, fewer high-degree nodes compared the null model. The most prominent feature of this content-sensitive citation is community structure: in the real BCN, papers focussed on the same topic share more common references with each other than papers focussed on different topics, so that the densities of edges within topics are much higher than between topics. Therefore the modularities of communities extracted by the Louvain method in the real network is much higher than in the null model, as shown in Fig. 2.2(d).

As mentioned, after building the BCN, we applied the Louvain method based on the maximization of modularity, to extract the community structure [Blondel et al., 2008]. We already mentioned in Chapter 1 that there are many different algorithms to extract community structure, and the Louvain method belongs to the family that tries to maximize the modularity. The main inspiration of the Louvain method is that there are several natural organization level—communities, sub-communities, or sub-sub-communities—inside most large real-world networks. To discover the community structure, the Louvain method is applied in two phases: we first identify small groups by optimizing the modularity locally on all nodes. This phase will finish when no further improvement on modularity can achieved. In the second phase, each small group is treated as a single node, and

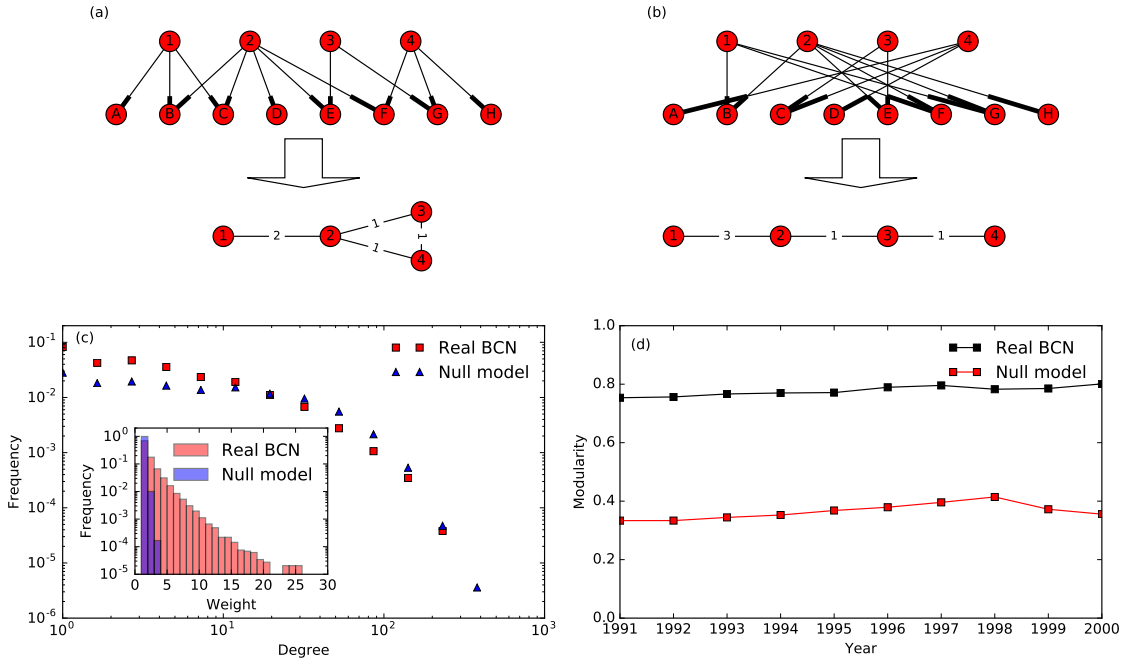


Figure 2.2: (a) Original citation network and its BCN. (b) A rewired citation network keeping in degrees and out degrees fixed and its BCN. (c) Comparison of the degree and weight distributions of papers published in 1991, between the real BCN and the null model. (d) Modularities of the best partitions extracted by the Louvain method for the real BCNs and the null model between 1991 and 2000. Results from null model are averaged over 10 different rewirings, and the error bars are much smaller than the marker size.

the local modularity optimization is repeated until the modularity cannot be increased by any further clustering. The hierarchical structure will be revealed naturally through the second phase. The Louvain method is widely used because it is efficient on large networks, and often chose as the benchmark for comparison between different algorithms. It is important to note that currently we still not have a 'silver bullet' to decide which methods give us the 'right' results in real-world network without prior information. Therefore the choice of algorithm depends on the network type and underlying mechanism. In our study, since we do not know much about the mechanism underlying the BCN, we try to

maximize the modularity as the first step to tackle this problem. Considering the fact that papers around the same topic should have more common reference that papers in different topics, maximizing modularity in the BCN to get the topical cluster is reasonable.

It is worth noting an important concern of community detection in any application, that is false group identification. More specifically, node 1 'in fact' belongs to group A, but the community detection give the result B. In the ideal case the detection algorithm should assign each node with the correct group label. In reality, the situation is much more complicated. First, because of the existence of interdisciplinary papers, there is no clear and hard boundary between topical clusters. Some papers are bridges between different topics, therefore it is very hard to decide which community we should put such papers in. In other words, some nodes are intrinsically harder to classify than other nodes. In our study, such nodes are more likely to be interdisciplinary papers. The result of community detection on these nodes then depends on the algorithm type or even initial condition. One possible way to solve this problem is to allow overlapping communities. However, overlapping communities will cause other problems with evolution quantification, a problem we will discuss in greater detail in Chapter 5. Second, and perhaps more importantly, is that in most cases we do not have information on the ground truth to test our detection results against. Although we realize the risk of putting node 1 into group B rather than into group A (the ground truth), we will not know beforehand that the ground truth is that node 1 belongs to group A (this is why we do the community detection). Because of these two reasons, we can not eliminate the false group identification easily. Notwithstanding this, we are confident that the effect of false group identification is negligible. The modularity of our community detection is very high (about 0.7). This suggests that the intra-community connection is much denser that inter-community connection, in other words, the boundary between communities is very clear, not fuzzy. The probability of false

group identification is low in this situation. We also use node attributes (the PACS number) to cross validate the community detection results, as given in the next paragraph. As far as PACS numbers are concerned, the communities are homogeneous (almost all papers in a community having the same PACS numbers), and this result is statistically significant.

As mentioned above, to verify that the communities extracted are really focussed on closely related questions, we check the Physics and Astronomy Classification Scheme (PACS) numbers of members of the communities. PACS number is a scheme for classifying physics and astronomy literature using a hierarchical set of codes. Each PACS code consists of six alphanumeric characters divided into three pairs, with the lowest-level term giving the most detailed information. For example, in PACS 2010 version, $03.XX.XX$ means **Quantum mechanics, field theories, and special relativity**. If we go two levels deeper to $03.65.XX$, then we are talking about **Quantum mechanics**. Finally, under **Quantum mechanics**, $03.65.Nk$ is about **Scattering theory**, $03.65.Ud$ is about **Entanglement and quantum nonlocality**. Such numbers are provided by the authors to indicate which subfields of physics their papers belong. In our case, we only use the first two digits of the PACS numbers, as a balance between accuracy and coverage since the TC is a mesoscopic entity. We list the first two digits of the PACS numbers in Appendix B.

To test whether the PACS numbers appearing in the communities could have occurred by chance, we choose one year t, build its BCN, extracting the community structure with sizes $\{s_1, s_2, ..., s_n\}$, and then randomly assign papers in year t into n pseudo-communities of the same sizes, to remove any potential size effects. For a community of size s, we then identify the largest subset of papers sharing the same PACS number. This PACS number can represent the subfield of the community to a certain extent, and the fraction of papers in the largest subset reflect the homogeneity of the community. The largest subset of pa-

pers sharing the same PACS number in a random collection of s papers is typically small. Dividing the sizes of the largest subsets in the empirical communities and in the random collections, we find ratios are larger than 1 for most cases (see Fig. 2.3). That is to say, for most communities, this is highly unlikely, so we conclude that the groupings of papers extracted are meaningful. The results show that the communities extracted are papers really focused on closely related topics, so we refer to these validated units of knowledge as *topical clusters* (TCs).

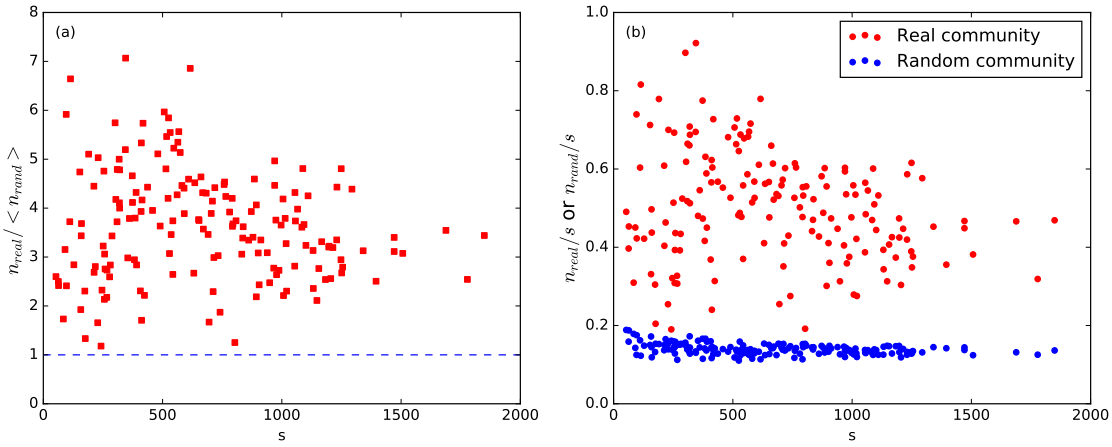


Figure 2.3: Comparison of PACS homogeneity between real BCN communities between 1991 and 2000 with more than 50 papers, and their corresponding random collections. (a) The red squares correspond to the sizes of the largest subsets of papers sharing at least one PACS number, n_{real}, in the empirical communities divided by the same quantity found in the corresponding random collections, n_{rand}, as a function of the community size s. (b) The fraction of the largest subset of papers sharing at least one PACS number as a function of s for real communities in the BCN and random collections. For clarity, the small error bars are not shown in the figures.

2.2 Evolution of community structure and alluvial diagram

To study how knowledge evolves, we investigate how TCs $\{\mathscr{C}^t\}$ in year t become $\{\mathscr{C}^{t+1}\}$ in year $t+1$. From now on we use C_m^t to denote a topical cluster m in time t and here $\mathscr{C}^t =$

$\{C_1^t,...,C_m^t,...,C_u^t\}$ and $\mathscr{C}^{t+1}=\{C_1^{t+1},...,C_n^{t+1},...,C_v^{t+1}\}$ are the collections of topical clusters in time t and $t+1$ respectively, namely all TCs in time t and $t+1$. The papers published in different years are distinct, but they do overlap in their references. Therefore we use this fact to define a *forward intimacy index* I_{mn}^f and a *backward intimacy index* I_{mn}^b:

$$\begin{aligned} I_{mn}^f &= \sum_i \frac{N\left(R_i,\mathscr{R}_n^{t+1}\right)}{N\left(R_i,\mathscr{R}^{t+1}\right)}\frac{N\left(R_i,\mathscr{R}_m^t\right)}{L\left(\mathscr{R}_m^t\right)}; \\ I_{mn}^b &= \sum_i \frac{N\left(R_i,\mathscr{R}_m^t\right)}{N\left(R_i,\mathscr{R}^t\right)}\frac{N\left(R_i,\mathscr{R}_n^{t+1}\right)}{L\left(\mathscr{R}_n^{t+1}\right)}, \end{aligned} \tag{2.1}$$

to quantify how close C_m^t is to C_n^{t+1}. Here we denote the references cited by papers in C_m^t and C_n^{t+1} as $\mathscr{R}_m^t=\mathscr{R}(C_m^t)=[R_{m1},...,R_{mp}]$ and $\mathscr{R}_n^{t+1}=\mathscr{R}(C_n^{t+1})=[R_{n1},...,R_{nq}]$; and $\mathscr{R}^t=\{\mathscr{R}_1^t,...,\mathscr{R}_m^t,...\}$ is the collection of all references cited in year t. $N(element,list)$ is the number of times $element$ occurs in $list$, and $L(list)$ is the length of $list$. Both forward and backward intimacy indices take on values between 0 and 1, see Appendix A for the proof. The larger the intimacy index, the clearer the inheritance relationship between two TCs. In this definition, we assume each citation instance in t will be uniformly distributed over all instances of the same citation in $t+1$, while each citation in $t+1$ receives equal contributions from all instances of the same citation in t. In general, this index is asymmetric, i.e. $I_{mn}^f \neq I_{mn}^b$, because the references are not cited the same number of times in the two years. Take Fig. 2.4 for example, to calculate the intimacy indices between ①② and ⑥⑦⑧, we observe that: $N(A,\mathscr{R}(⑥,⑦,⑧))=3$, $N(A,\mathscr{R}(①,②))=2$, $N(A,\mathscr{R}(⑥,⑦,⑧,⑨,⑩,⑪))=3$, $L(\mathscr{R}(①,②))=4, N(C,\mathscr{R}(⑥,⑦,⑧))=1, N(C,\mathscr{R}(①,②))=1, N(C,\mathscr{R}(⑥,⑦,⑧,⑨,⑩,⑪))=4$. Therefore, the forward intimacy index between ①② and ⑥⑦⑧ is $I^f=\frac{N(A,\mathscr{R}(⑥,⑦,⑧))}{N(A,\mathscr{R}(⑥,⑦,⑧,⑨,⑩,⑪))}\times\frac{N(A,\mathscr{R}(①,②))}{L(\mathscr{R}(①,②))}+\frac{N(C,\mathscr{R}(⑥,⑦,⑧))}{N(C,\mathscr{R}(⑥,⑦,⑧,⑨,⑩,⑪))}\times\frac{N(C,\mathscr{R}(①,②))}{L(\mathscr{R}(①,②))}=\frac{3}{3}\times\frac{2}{4}+\frac{1}{4}\times\frac{1}{4}=0.5625$. In a similar way, the backward intimacy index between ①② and ⑥⑦⑧ is $\frac{2}{3}\times\frac{3}{5}+\frac{1}{4}\times\frac{1}{5}=0.45$.

According to the above mentioned methodology we visualized the sequence of TCs

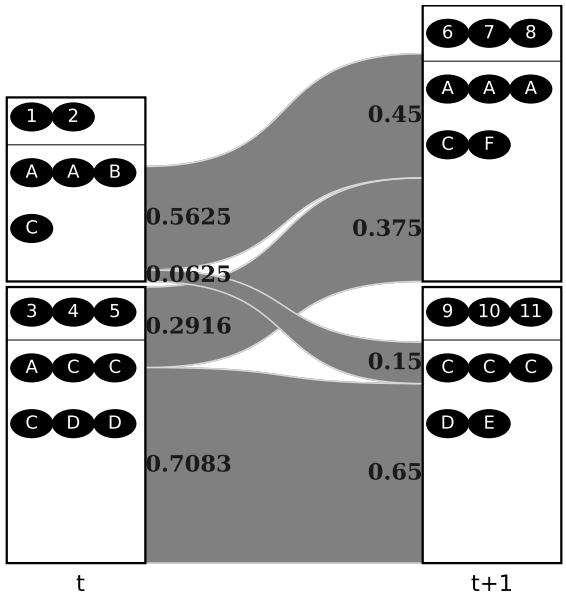


Figure 2.4: Topical clusters (papers 1, 2 citing reference A twice, reference B once, and reference C once; and papers 3, 4, 5 citing reference A once, reference C three times, and reference D twice) in year t (left) and (papers 6, 7, 8 citing reference A three times, reference C once, and reference F once; and papers 9, 10, 11 citing reference C three times, reference D once, and reference E once) in year $t+1$ (right), and their forward (right of the year-t TCs) and backward (left of the year-$t+1$ TCs) intimacy indices, shown as flows.

and their intimacy indices, the evolution of physics research they represent in the form of alluvial diagrams and the results are very clear. For example, in Fig. 2.5 we can clearly see the birth of PRA, PRB, PRC and PRD from PR in 1970. Each journal consist of several TCs, which existed even in the PR era. The editorial decision to split PR is consistent with the self-organized TCs even though it was done without classification analysis. We also plotted an alluvial diagram for 1991 to 2000, showing the splitting of PRA into PRA and PRE. As we can see from Fig. 2.6, before 1993, there were several PRA-dominated TCs. After the split in 1993, some PRA-dominated TCs remained PRA-dominated, whereas other PRA-dominated TCs became PRE-dominated. This means that even before 1993, papers in PRA were already divided into groups based on different topics, some of which are

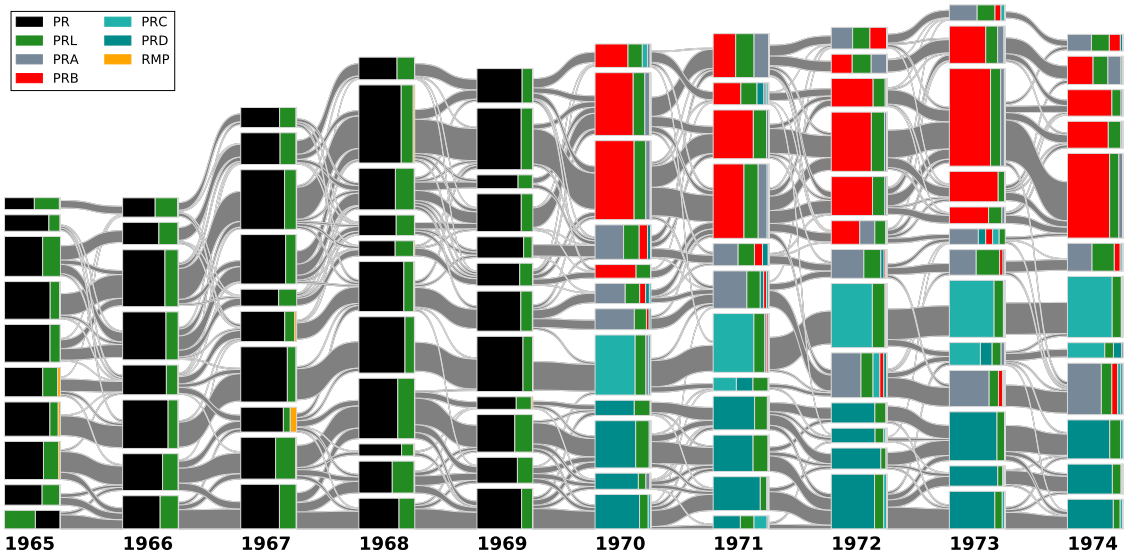


Figure 2.5: The alluvial diagram of APS papers from 1965 to 1974. Each block in a column represents a TC and the height of the block is proportional to the number of papers in the TC. Only communities comprising more than 100 papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the forward and backward intimacy indices. The different colors in a TC represent the relative contributions from different journals.

predecessors of the PRE TCs.

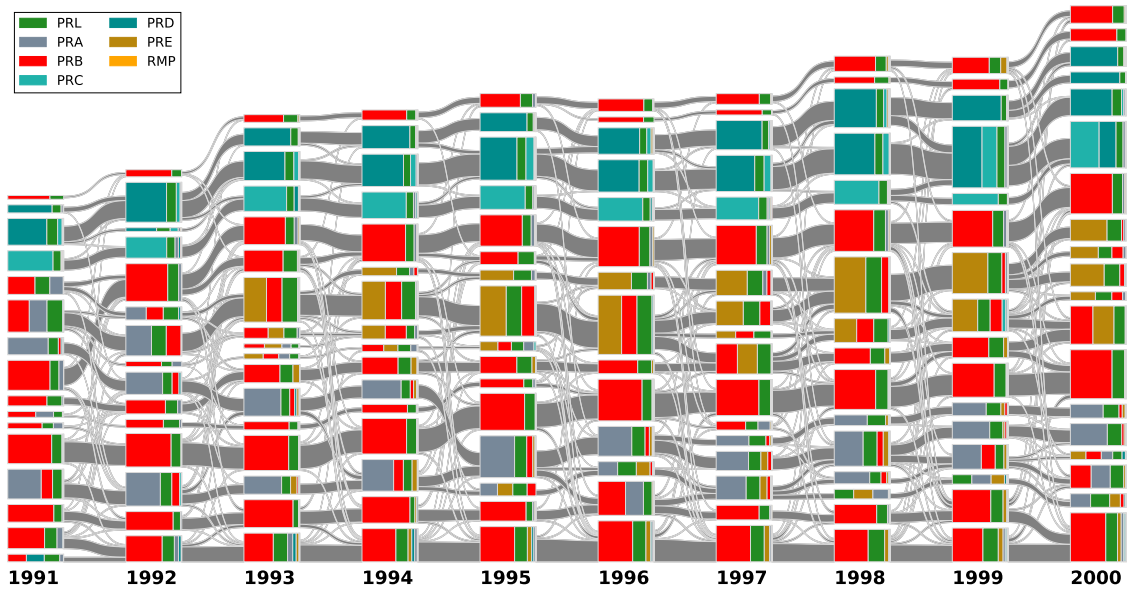


Figure 2.6: The alluvial diagram of APS papers from 1991 to 2000. Each block in a column represents a TC and the height of the block is proportional to the number of papers in the TC. Only communities comprising more than 100 papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the forward and backward intimacy indices. The different colors in a TC represent the relative contributions from different journals.

More importantly, from the alluvial diagram we can identity the key interactions between TCs that are correlated with important publications. Here we showcase one such episode between 1991 and 2000, involving interesting interactions between quantum optics (QO), quantum information (QI), and Bose-Einstein condensation (BEC). These three fields experienced breakthroughs in the 1990s. In Fig. 2.7 we highlighted the evolution of TCs that are related to these three topics and show the three most cited papers in these TCs in Appendix C. At the beginning of the decade, we see two PRA-dominated TCs. Based on the papers they contain, we can loosely associate one with quantum information (QI) and trapped atomic ions (BEC), and the other with quantum optics (QO). In 1993, the QI

+ BEC TC cited many QO papers, and in 1994, the QO TC cited many QI + BEC papers. Following this 'cross-fertilization', the two TCs merged in 1995, the same year Cornell *et al.* [Anderson et al., 1995] and Ketterle *et al.* [Davis et al., 1995] published their seminal papers demonstrating BEC in dilute atomic gases. In recognition of their works, Cornell, Wieman, and Ketterle were awarded the 2001 Nobel Prize in Physics. The PRA-dominated TC split after 1996 to give one that is exclusively BEC, and another that is still a combination of QI + QO. It was after Zeilinger demonstrated in 1997 experimental quantum teleportation [Bouwmeester et al., 1997] that the QI + QO TC split into a QI TC and a QO TC. After receiving more influence from another PRB-dominated TC, the QO cluster produced yet another breakthrough paper, in the form of ultraslow light in hot atomic gases [Kash et al., 1999]. Without the data visualization done here, few may suspect the existence of such connections between BEC, quantum teleportation and slow light.

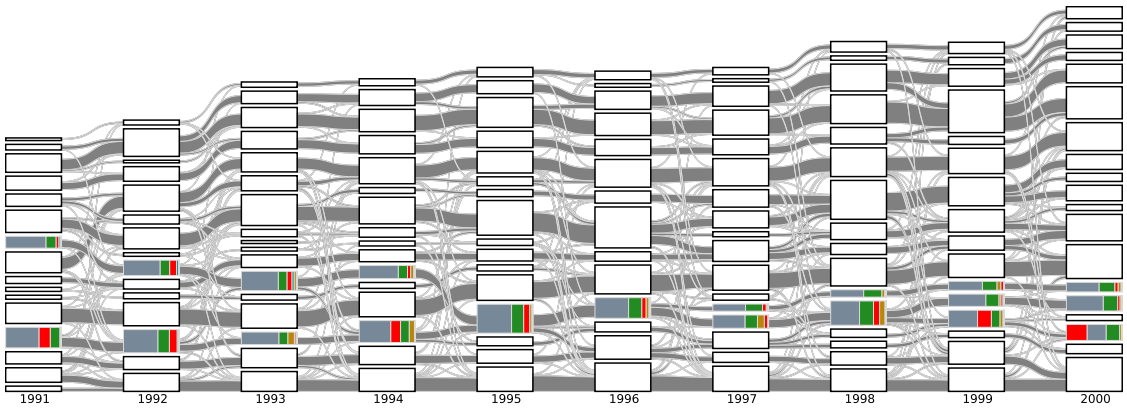


Figure 2.7: The same alluvial diagram of APS papers from 1991 to 2000 as Fig. 2.6, where we colored only TCs highly related to quantum optics, quantum information and Bose-Einstein condensation.

2.3 Analyses of the evolution

Some TCs have more references overlapping with those in the previous year, while other TCs have less. To quantify the evolution of references, we sum the forward and backward intimacy indices for each TC. These represent the percentage of a TC's references going to the next year, and the percentage of references the TC inherited from the previous year, which we think of as the 'outflow' and 'inflow' respectively. As shown in Fig. 2.8(a) and (b), most outflows and inflows are distributed within a narrow range, but there are exceptional cases as well: such as a single peak in Fig. 2.8 (b), whose references overlap significantly less than normal with the previous year. In the context of birth, death, growth, decay, split, and merge knowledge processes, we are inclined to call this event in 1993 the birth of a TC. Further analysis shows that most common PACS codes are: 03 (Quantum mechanics, field theories, and special relativity), 42 (Optics) and 63 (Lattice dynamics). Looking at the references of this TC, we find that most of these comes from 1990, 3 year before. This interesting phenomenon is therefore more appropriately identified as a *sleeping beauty* [Ke et al., 2015].

Every year, physicists absorb new references and drop old references as their fields progress. Although this 'metabolism' differ from TC to TC, the whole process is quite stable over all TCs, as shown in Fig. 2.8(c) and (d). This universal curve can be used as a benchmark for the test of scientific impact, as we have done in Fig. 2.16.

From Fig. 2.5 and Fig. 2.6 we see a diversity of inflows and outflows from one TC to another: some TCs are derived almost exclusively from one source, others receive strong contributions from a small number of sources, or weak contributions from a large number of sources. To quantify such diversity, we construct a forward mixing degree of community

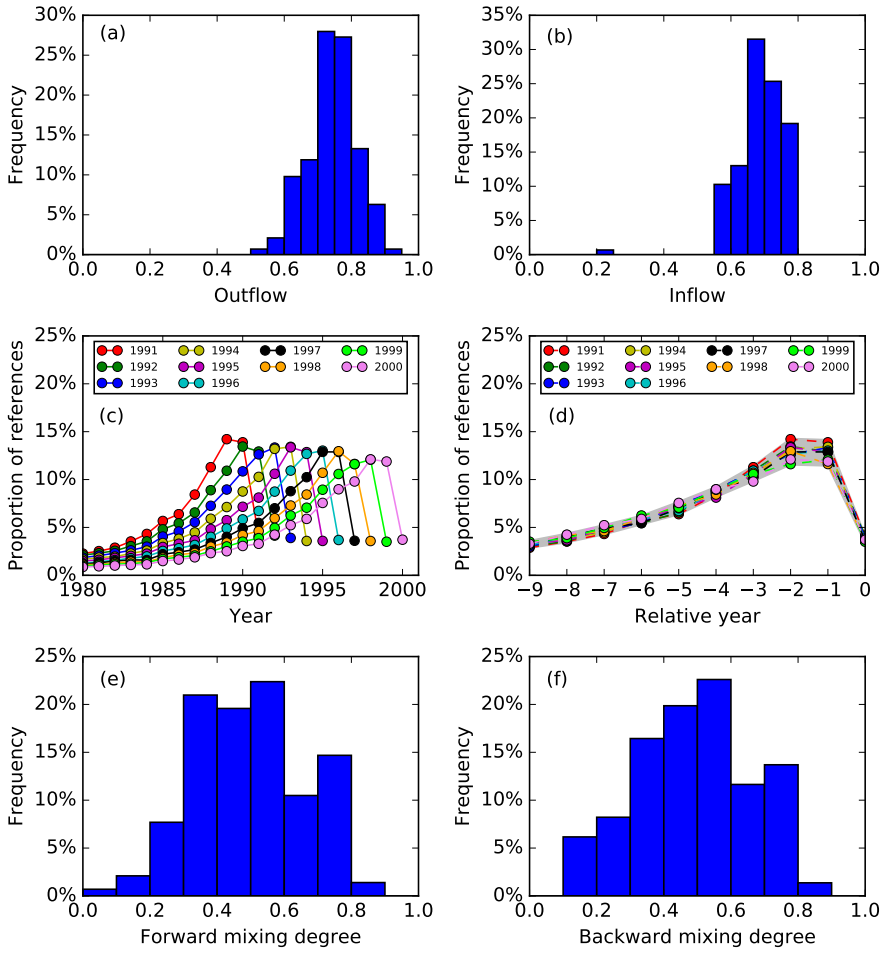


Figure 2.8: The metabolic analysis of APS papers in the 1990s. (a) The distribution of outflows of TCs. (b) The distribution of inflows of TCs. (c) Proportions of APS paper's references published in different years. (d) Proportions of APS paper's references published in different years, relative of the year 0 of publication. (e) The distribution of forward mixing degree of TCs. (d) The distribution of backward mixing degree of TCs.

C_m^t and backward mixing degree of C_n^{t+1} analogous to the Gini-Simpson index [Jost, 2006]:

$$\begin{aligned} M_m^f &= 1 - \sum_n \left(I_{mn}^f / \sum_{n'} I_{mn'}^f \right)^2, \\ M_n^b &= 1 - \sum_m \left(I_{mn}^b / \sum_{m'} I_{m'n}^b \right)^2, \end{aligned} \tag{2.2}$$

which measure the probabilities that two streams taken at random from the TC's outflow/inflow (with replacement) represent different streams. A TC with low forward/backward mixing degree has effectively one child/parent, whereas a TC with high forward/backward mixing degree undergoes/results from strong splitting/merging. As shown in Fig. 2.8 (e), (f), neither are frequent. It is more common to find weak mixing between TCs, however the mechanism behind this weak mixing is still not clear. It is wroth noting that the shapes in Fig. 2.8 (e), (f) are largely depend on our definition of intimacy indices and mixing degree and even the community detection algorithm. The mechanism behind not mixing – weak mixing – strong mixing is not well-understood, and this may be the key to understanding the evolution of scientific knowledge. Please refer to Chapter 5 for more discussion.

At this point, let us recall the Popperian and Kuhnian pictures of the evolution of knowledge, where we expect incremental growth punctuated by abrupt paradigm shifts. Certainly, at the aggregate level of PR series of premier physics journals, the number of articles published has grown over the years. When we partition these articles into TCs, we naively expect that some clusters will grow/shrink because of growing/declining interest in their topics. From the alluvial diagrams, we realize that the real picture is far more complex because of recombinations between TCs. Therefore, instead of measuring the growth rates of pure TCs, we need to measure the growth of recombined TCs. To do this, we assume that the contribution of C_m^t to the size of C_n^{t+1} is proportional to the size of C_m^t and also the normalized forward intimacy index $I_{mn}^f / \sum_n I_{mn}^f$, i.e.

$$L'(C_n^{t+1}) = \sum_m L(C_m^t)(I_{mn}^f / \sum_n I_{mn}^f). \tag{2.3}$$

When we plot the predicted sizes $L'(C_n^{t+1})$ against the observed size $L(C_n^{t+1})$ in Fig. 2.9, we find $(L'(C_n^{t+1}), L(C_n^{t+1}))$ scattered about a straight line with slope with 1.06, which is the

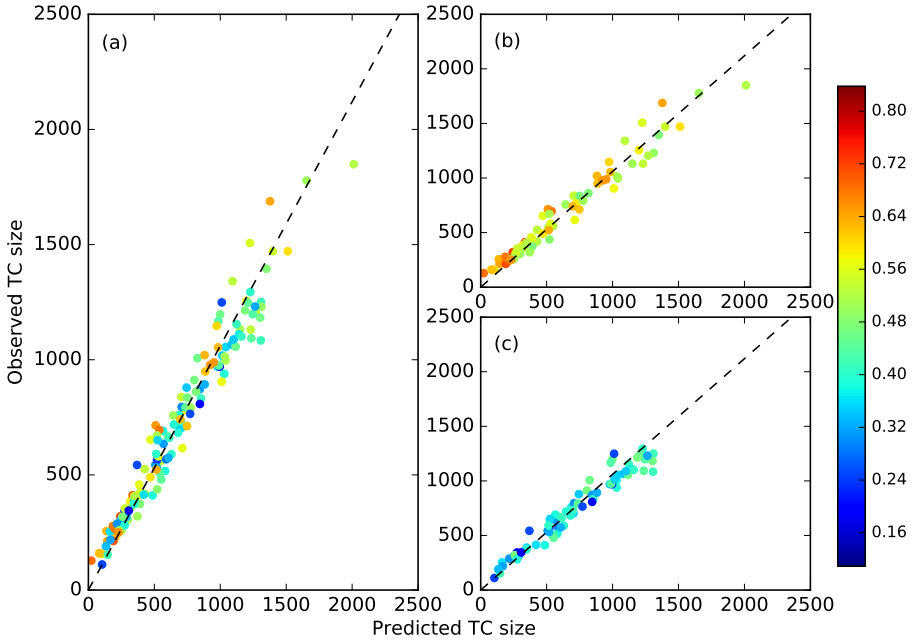


Figure 2.9: (a) Plot of observed (y-axis) against predicted (x-axis) sizes of recombined TCs, showing a linear growth with slope 1.06 (dashed line). This linear growth is the same for TCs with (b) high (red) or (c) low (blue) backward mixing degree.

annual growth rate of the number of papers in APS journals. This tells us that the growth of recombined TCs is also Popperian.

Next, we consider the Kuhnian processes of splitting and merging. For merging events, the similarity

$$S(C_m^t, C_{m'}^t) = \sum_n (I_{mn}^f / \sum_{n'} I_{mn'}^f)(I_{m'n}^f / \sum_{n''} I_{m'n''}^f) \tag{2.4}$$

measures the overlap between the offsprings of the TCs C_m^t and $C_{m'}^t$ in year t. If C_m^t and $C_{m'}^t$ merge perfectly into a single TC in year $t+1$, $S = 1$. On the other hand, if the offsprings of C_m^t and $C_{m'}^t$ are distinct, $S = 0$. In general, $0 \leq S \leq 1$. The value of S cannot be treated as a 'prediction', because we made use of information from years t and $t+1$ to compute it. As two TCs evolved from being distinct to merging into a single TC, we expect to find few,

low-weight edges between them in the BCN when they are distinct. This sum of weight of edges would gradually increase until the sum of weight between C_m^t and $C_{m'}^t$ is comparable to the sum of edges within C_m^t and $C_{m'}^t$. At this point, the two TCs merge. Therefore, to do the prediction, we define

$$T(C_m^t, C_{m'}^t) = W(C_m^t, C_{m'}^t)/(L(C_m^t)L(C_{m'}^t)), \tag{2.5}$$

where $W(C_m^t, C_{m'}^t)$ is the sum of weights of edges between papers in C_m^t and $C_{m'}^t$, normalized against the sizes of TCs involved. Fig. 2.10 shows that $S(C_m^t, C_{m'}^t)$ and $T(C_m^t, C_{m'}^T)$ are highly correlated. High $T(C_m^t, C_{m'}^t)$ leads with a large probability to a high $S(C_m^t, C_{m'}^t)$. Analyzing the APS papers in the 1990s, we found a Spearman's rank coefficient of 0.804 between $T(C_m^t, C_{m'}^t)$ and $S(C_m^t, C_{m'}^t)$ over all TCs (with at least 100 papers). However, because the average Pearson correlation coefficient is only 0.504, such a relation is not linear (see Fig. 2.11).

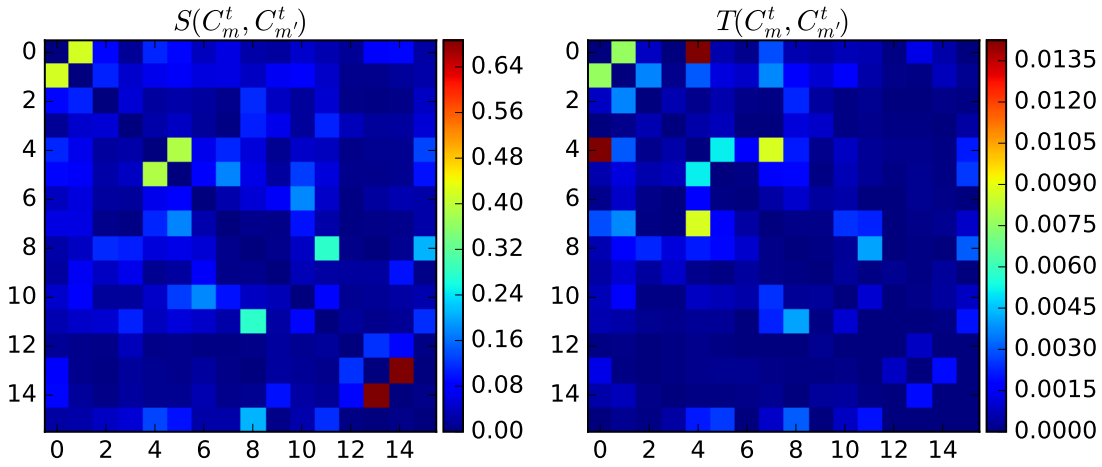


Figure 2.10: (left) $S(C_m^t, C_{m'}^t)$ of 16 TCs in 1991, computed using forward intimacy indices going from 1991 to 1992. (right) $T(C_m^t, C_{m'}^t)$ of the same 16 TCs, using information from 1991 only. We use the same ordering of TCs in both matrices.

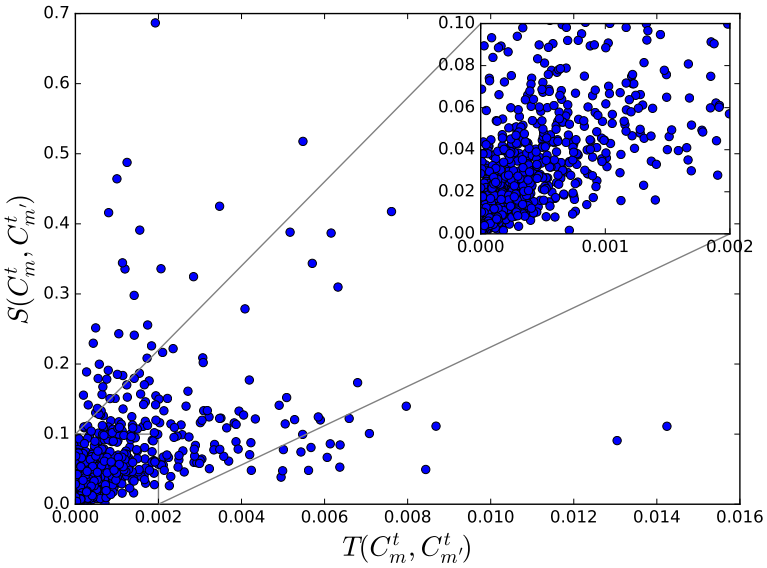


Figure 2.11: The scatter plot between $T(C^t_m, C^t_{m'})$ and $S(C^t_m, C^t_{m'})$ among all TCs (with at least 100 papers) in 1990s.

We also tried to predict the splitting events. The first factor we considered is TC's size. We divided all TCs in 1990s into two groups: one for TCs larger than median size, another for TCs smaller than median size. The medians and means for the forward mixing degree in these two groups are very close. Furthermore, the Pearson correlation coefficient between size and forward mixing degree is -0.031 (see Fig. 2.12). Therefore, we concluded that a TC does not become more likely to split as it grows larger. The second factor is the TC's internal structure. Here the situation is more complex: when we use the dendrogram extracted from the Louvain method to identify subcommunities, we found that different TCs have different internal structures (see Fig. 2.13), some have a few large subcommunities, while others have many small subcommunities. Naively, we expect the criterion for splitting is the opposite to merging, i.e. the easier it is to tell one subcommunity from

others, the higher the chances for a split. The *boundary index*

$$B = \frac{\sum_{i_1 \neq i_2} \sum_{\substack{j_1 \in C_{i_1} \\ j_2 \in C_{i_2}}} A(j_1, j_2) / \sum_{i_i \neq i_2} L(C_{i_1}) L(C_{i_2})}{\sum_i \sum_{j_1, j_2 \in C_i} A(j_1, j_2) / \sum_i L(C_i) L(C_i)}, \tag{2.6}$$

which is the ratio between inter-subcommunity edge density and intra-subcommunity edge density, measures how indistinct the subcommunities are in a TC. Here $A(j_1, j_2)$ is the weight of the edge between papers j_1 and j_2, and C_i is a subcommunity in the given TC. However the picture we find is not as simple as the merging case. When we plot forward mixing degree M^f against B, we find the expected decreasing trend, but at the same time, the large scatter makes it impossible to reliably predict a splitting event using B. To better understand the relationship between M^f and B, we use quantile regression [Sienkiewicz and Altmann, 2016] to find that the B has no 'prediction power' when M^f is small, but becomes 'predictive' when M^f is large. That is to say the relation between B and M^f depends on the decile, as shown in Fig. 2.14(a), (b). The slopes show that for the decile of most strongly splitting TCs, increasing the standardized B by one standard deviation will decrease M^f by about 0.05, whereas for the decile of the least strongly splitting TCs, there is no obvious trend.

We also define a *fragmentation index*

$$F = \sum_{i:j[i]} w_i S_{j[i]}^2 \tag{2.7}$$

where w_i is the size fraction of the top level subcommunity i, $s_{j[i]}$ is the relative size fraction of subsubcommunity j inside subcommunity i. The more fragmented a community is, i.e., more and smaller subcommunities, the closer F is to 0. Quantile regression between F and M^f gives very similar results as B and M^f, i.e., for the decile of most strongly splitting TCs, increasing the standardized F by one standard deviation will decrease M^f by about 0.06,

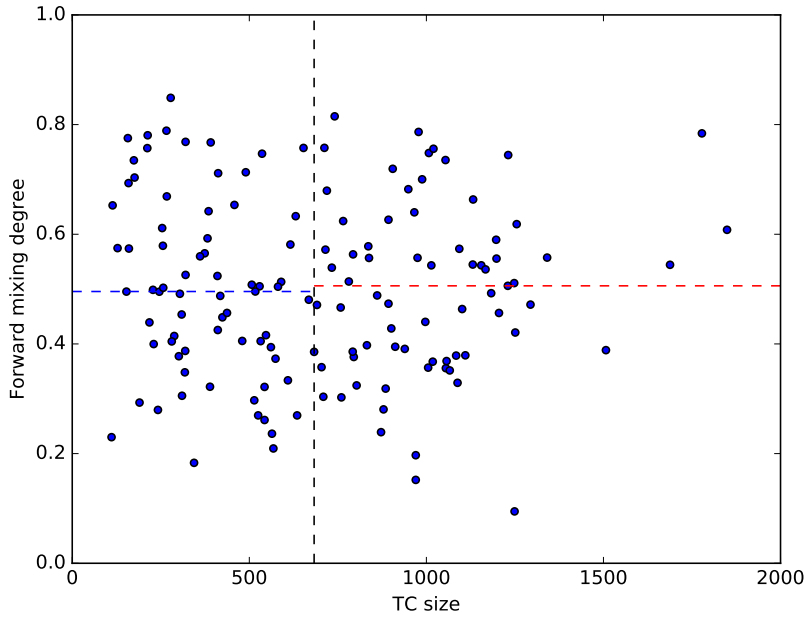


Figure 2.12: The scatter plot between size and forward mixing degree among all TCs (with at least 100 papers) in 1990s. The black dash line is median size, the blue dash line is the median of forward mixing degree for TCs are smaller than median size, the red dash line is the median of forward mixing degree for TCs are larger than median size.

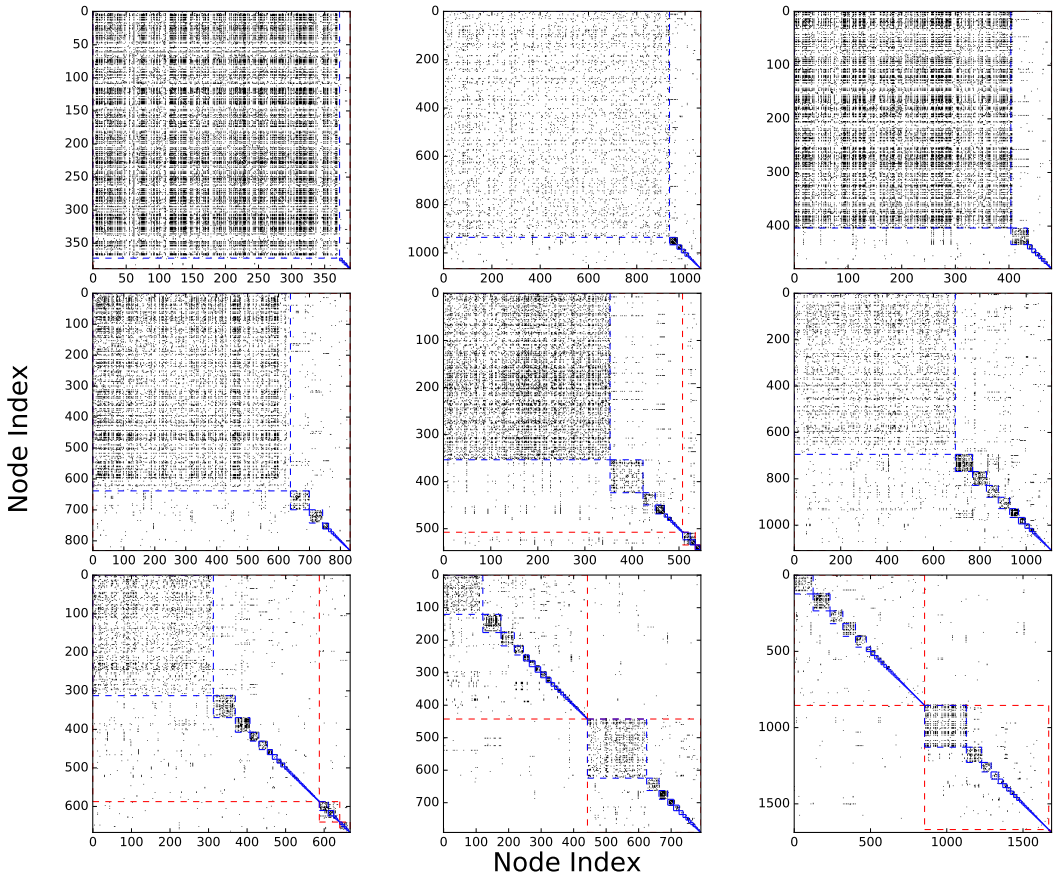


Figure 2.13: Adjacency matrices of TCs in the 1990s. The blue lines indicate the boundaries of subsubcommunities, the red lines indicate the boundaries of subcommunities. The red lines are absent from some plots because such TC have only one level when the Louvain algorithm terminated.

whereas for the decile of the least strongly splitting TCs, there is no obvious trend as β close to 0, as shown in Fig. 2.14(c), (d).

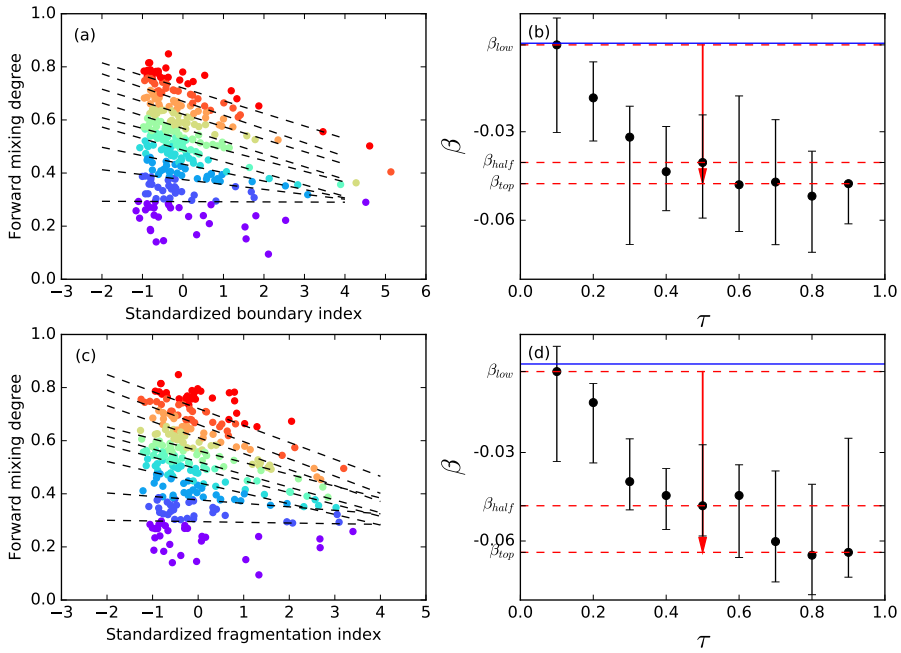


Figure 2.14: Relation between boundary index, fragmentation index and forward mixing degree of TCs in 1980s and 1990s. (a) Each dot corresponds to one TC, dash lines show QR results for quantiles $\tau = 0.1, 0.2, ..., 0.9$. (b) β coefficients (slopes of QR in the (a) as a function of τ. The red arrows show $\beta_{low} \equiv \beta(\tau = 0.1)$, $\beta_{half} \equiv \beta(\tau = 0.5)$ and $\beta_{top} \equiv \beta(\tau = 0.9)$, as, respectively, the nock, a circle on the shaft, and the head of the arrow, the blue solid line represents 0. (c) Each dot corresponds to one TC, dash lines show QR results for quantiles $\tau = 0.1, 0.2, ..., 0.9$. (d) β coefficients (slopes of QR in the (c) as a function of τ. The red arrows show $\beta_{low} \equiv \beta(\tau = 0.1)$, $\beta_{half} \equiv \beta(\tau = 0.5)$ and $\beta_{top} \equiv \beta(\tau = 0.9)$, as, respectively, the nock, a circle on the shaft, and the head of the arrow, the blue solid line represents 0. The color of a dot depends on its position in the quantile regression, i.e. red for $\tau = 0.9$, green for $\tau = 0.5$, and violet for $\tau = 0.1$.

Finally, we want to know the impacts of such merging and splitting events. To do this, our analysis should obey the principle of causality, that is any results must be due to something that happened in the past, and not something that happen in the future. This means we should use the backward intimacy index for correlation analysis, because

the backward intimacy index will provide information about the past, while the forward intimacy index is about what happens in the future. We first check for an increase in the number of publications after such events, but found an insignificant difference in paper numbers in strongly and weakly mixing TCs (see Fig. 2.9(b) and (c)). We suspected this is because our data set is confined to the APS publications, and a more careful check should include other physics journals to capture any 'influence spillover'. When we think of high-impact research, we also think of highly-cited papers. Therefore, to quantify the impact of strongly-splitting events in the alluvial diagrams, we counted the citations of TCs resulting from splittings. As shown in Fig. 2.15, we did this for number of citations 2 years after the events, and also 5 years after the events. There were no obvious trends. The results of backward mixing degree, i.e. merging, are similar.

Focusing on the highly productive chain of knowledge processes that led to experimental realizations of BEC, quantum teleportation and slow light, we checked the citation profiles between 1995 and 1998. While the 1995 BEC+QI+QO TC cited a slightly lower proportion of 1995 papers than the APS 0-year average, the 1996 BEC+QI+QO, the 1997 BEC TC, the 1998 BEC TCs all cited significantly more 0-year papers. The full effect of this BEC breakthrough can be seen in the large proportions of 1996 papers cited by the 1997 and 1998 TCs and the proportion of 1997 papers cited by the 1998 TC (see Fig. 2.16). Indeed, we have provided early evidence suggesting that strongly-mixing Kuhnian processes are associated with greater impact. It is worth noting that the relation we observed is only a correlation, not a causation. Our analysis above mainly focus on significant changes on the citation curves, but to really confirm the correlation between Kuhnian processes and scientific breakthroughs, a more systematic investigation is necessary.

To test the correlation between an abnormal reference distribution and the foreward/backward

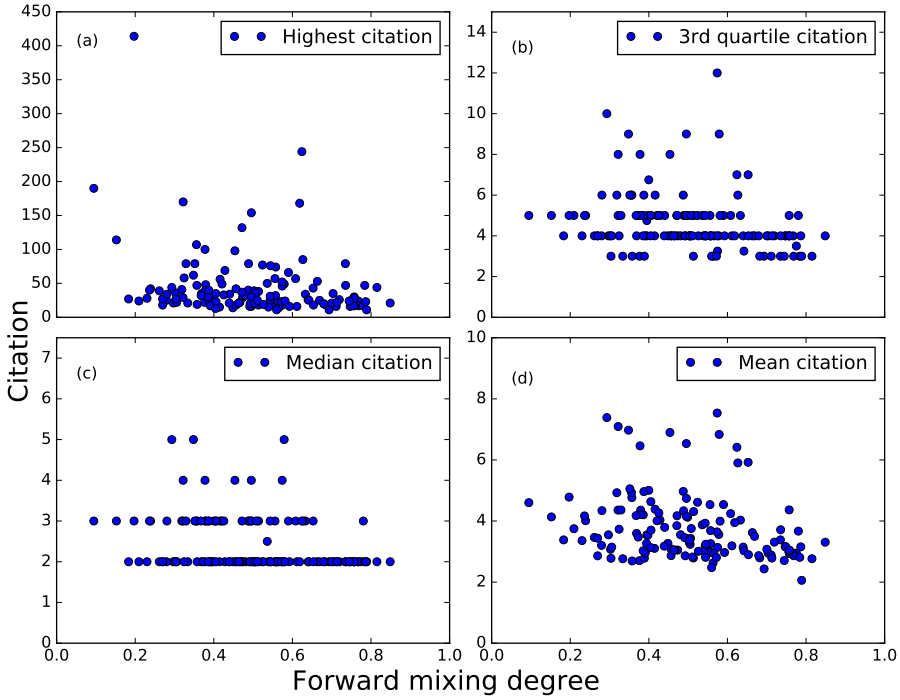


Figure 2.15: The scatter plot between different citations received during 2 years and forward mixing degree among all TCs (with at least 100 papers) in 1990s. (a) Highest citation, (b) Third quartile citation, (c) Median citation, (d) Mean citation.

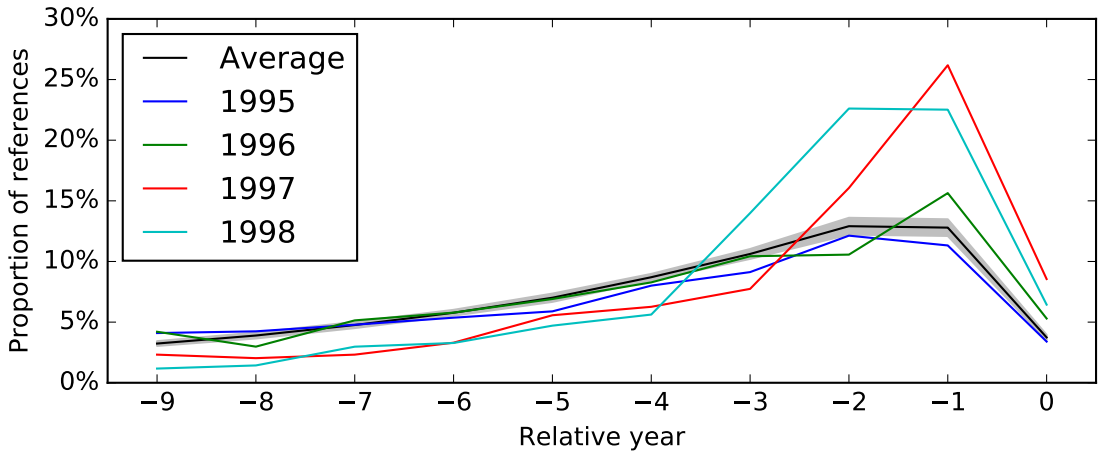


Figure 2.16: Proportions of a TC's references published in different years, relative to the year (0) of the TC. The black solid line is the proportions averaged over all TCs in the 1990s, while the area shaded gray is up to one standard deviation away from the mean. Other color lines represent the distribution of four different BEC related TCs.

mixing degree, we define the distance from the average distribution as:

$$D(TC) = \sum_i |TC(i) - M(i)|, \tag{2.8}$$

where i is the relative year, $TC(i)$ is the proportion of TC's reference in relative year i, $M(i)$ is the average proportion of all TC's reference in relative year i. If the TC's reference distribution is close to the average curve, the distance is close to 0, while the more it deviates from the average curve, the larger the distance. For example, if the average distribution is 20% in relative year 0, 40% in relative year -1, 30% in relative year -2, 10% in relative year -3, and the TC we want to study has a uniform distribution: 25% for relative year 0, -1, -2, -3, then the distance between the two is $|20\% - 25\%| + |40\% - 25\%| + |30\% - 25\%| + |10\% - 25\%| = 0.4$. From Fig. 2.17 we can see most TCs' reference curves are very close to the average curve: most distances are less 0.2, and only a few TCs deviate significantly

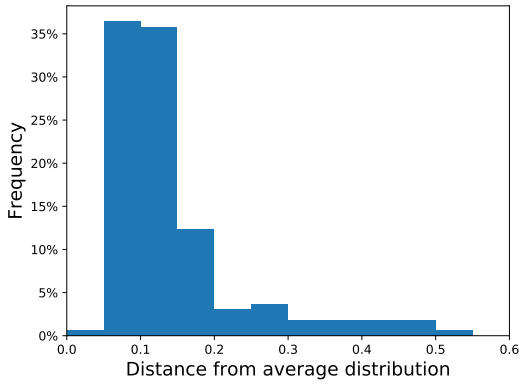


Figure 2.17: The distribution of distance from average reference distribution for all TCs in 1990s.

from the average curve, having distances larger than 0.2.

We then calculated the Pearson correlation coefficient between the forward/backward mixing degree and distance from average distribution. As we can see from Fig. 2.18, the coefficient is very close to zero. The results are similar for Spearman's coefficient. However it is too early to conclude that Kuhnian processes are not correlated with scientific breakthroughs. To illustrate this plausibility, we picked the 10 most deviant TCs based on their distance from the average distribution (see Fig. 2.19, Fig. 2.20), and found that two of them are from the BEC case study, six of them are closely related with birth, which should also be considered a Kuhnian process. However, limited by our measure of mixing degree (see Eq. 2.2), such birth-related events have very low mixing degrees in Fig. 2.18. This may be the source of the low Pearson and Spearman's coefficients. To study the correlation between the Kuhnian process and scientific breakthrough, an improved version of mixing degree is needed to give appropriate emphasis to birth processes. Please refer to Chapter 5 for a detailed discussion.

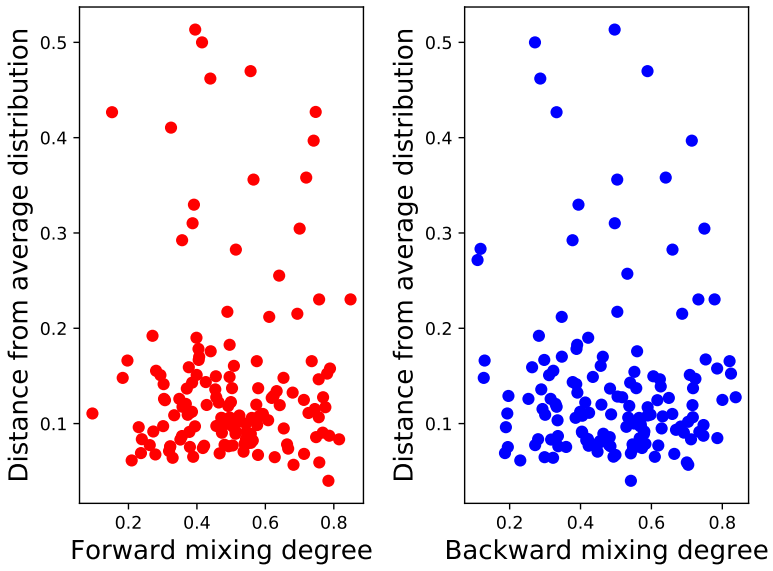


Figure 2.18: The scatter plots of forward/backward mixing degree and distance from average distribution for all TCs in 1990s.

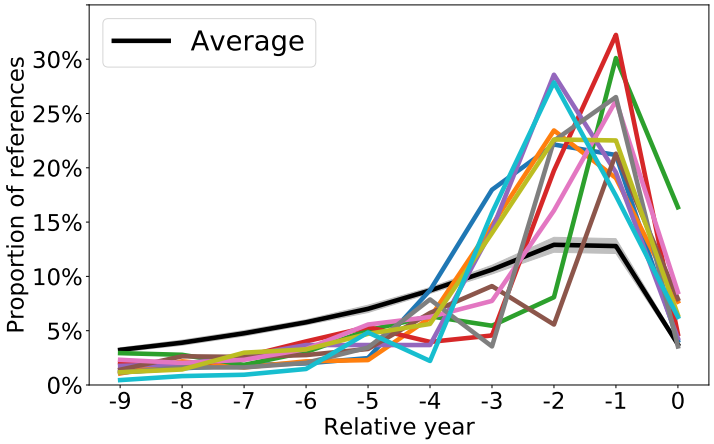


Figure 2.19: Proportions of a TC's references published in different years, relative to the year(0) of the TC. The black solid line is the proportions averaged over all TCs in the 1990s, while the area shaded gray is up to one standard deviation away from the mean. The other color lines represent the 10 most deviant TCs' distributions.

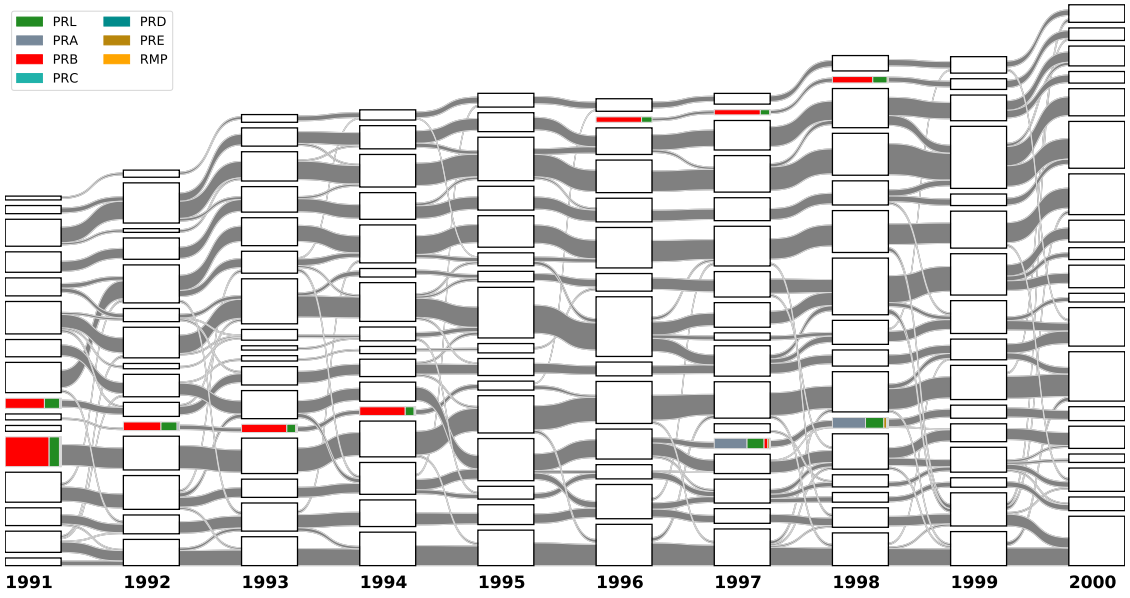


Figure 2.20: The same alluvial diagram of APS papers from 1991 to 2000 as Fig. 2.6, where we colored only the 10 most deviant TCs in Fig. 2.19

CHAPTER 3

Meme labelling of TCs and analyses

In Chapter 2 we developed a framework to study knowledge evolution in Physical Review journals. However, in this framework we include only citation information. One obvious shortcoming of this method is that even though we know the members of each TC, we do not know what research the TC represents until we read the papers' titles and abstracts. For example, to trace the evolution of quantum optics, quantum information and Bose-Einstein condensation (see Fig. 2.7), we checked the 10 most highly cited papers in each TC and judge whether it is related to quantum optics, quantum information or Bose-Einstein condensation research. This method is time-consuming, relies on the expert's subjective judgments and is therefore not scalable. To overcome such limitations, we would like to extract from each TC a set of 'keywords' that can help us roughly understand its research content, and how the knowledge represented by the TC evolves over time. This is our first motivation for doing meme labelling of TCs.

However, beyond its use as a label, meme can be the subject of a quantitative study of its own. In principle, knowledge evolution is a complex process and will show many

different telltale signs in research papers. Citation pattern is one, and other would be the language people used in papers, more precisely, the memes. When people make scientific discoveries, attention shift will occur since people's attention is limited. These changes will be reflected in the papers' language and memes. Although this sounds like an obvious statement, quantitative studies are limited. More importantly, this knowledge evolution occurs simultaneously in the TC level and meme level and the relation between them has never been discussed. In this chapter we will try to fill in this gap by investigating the evolution of scientific memes in the Physical Review journals and the co-evolution relation between memes and TCs.

3.1 Scientific memes in the Physical Review journals

The word 'meme' was first used by the evolutionary biologist Richard Dawkins in his book *The Selfish Gene* [Dawkins, 1976], to refer to the self-replicating unit in the diffusion of ideas and cultural phenomena — as the gene does in biological evolution. Although memes have mostly been used in the study of mass media and popular culture [Leskovec et al., 2009], like blog space and Twitter, this concept can also be used to study science, since science is also about the diffusion of ideas. Kuhn *et al.* propose a simple formulation that can extract memes from papers automatically and the results show that these memes are very close to the scientific concept representative of the papers [Kuhn et al., 2014]. Therefore we adopt their method and meme list to label the TCs and also to analyse meme evolution in the APS data set.

In a nutshell, a scientific meme is an n-gram that appears in the title and abstract of a paper, and those of a paper citing it (i.e., if paper A cites paper B and the titles, abstracts of A and B both include the n-gram M, then M can be considered a meme that was reproduced itself from B to A). This definition is simple and intuitive, but not all n-grams

that reproduce themselves through citations are equally interesting. For example, people use 'the' in almost every paper's title and abstract, so it is a meme according to the definition. However, this meme arise more from the English grammar than from real intellectual similarity, and is therefore less interesting than say 'gravitational wave'. To quantify this degree of *scientific interest*, Kuhn *et al.* introduced the propagation score P_m, which is high for a meme that appears frequently in papers that cite meme-carrying papers but rarely appears in papers that do not cite a paper that already contains the meme. Formally, they define the *propagation score* P_m and *meme score* M_m as

$$\begin{aligned} P_m &= \frac{d_{m\to m}}{d_{\to m}} \Big/ \frac{d_{m\to \not m}}{d_{\to \not m}}, \\ M_m &= f_m P_m, \end{aligned} \tag{3.1}$$

where $d_{m\to m}$ is the number of papers that contain meme m and also cite at least one paper containing meme m, while $d_{\to m}$ is the number of all papers that cite at least one paper that containing meme m, $d_{m\to \not m}$ is the number of meme-carrying papers that do not cite any paper that containing meme m, and $d_{\to \not m}$ is the number of all papers that do not cite publications containing meme m. Here f_m is simply the frequency of occurrence of m (i.e. the ratio of papers containing meme m). Some terms in Eq. 3.1 can be zero, especially for infrequent memes, therefore Kuhn *et al.* introduce a noise parameter δ to modify the propagation score:

$$P_m = \frac{d_{m\to m}}{d_{\to m} + \delta} \Big/ \frac{d_{m\to \not m} + \delta}{d_{\to \not m} + \delta}. \tag{3.2}$$

In this study we set $\delta = 3$ unless stated otherwise. The meme score M_m tells us whether a meme is important (f_m) and whether it is interesting (P_m), and can be used to extract ideas representative of the paper's content because n-grams associated with these ideas will have high meme scores. This feature can be seen very clearly from Tab. 3.1 and make memes

Table 3.1: Top 50 memes according to their meme scores from the APS data set. The symbol + indicates memes where the human annotators agree that this is an interesting and important physics concept, while the symbol ∗ indicates memes that are also found on the list of memes extracted from Wikipedia. Reproduced from [Kuhn et al., 2014].

1. Loop quantum cosmology+∗	14. Strange nonchaotic	27. Na_xCoO_2+	38. Inspiral+
2. Unparticle+∗	15. In $NbSe_3$	28. The unparticle+	39. Spin Hall effect+∗
3. Sonoluminescence+∗	16. Spin Hall+	29. Black	40. PAMELA
4. MgB_2+	17. Elliptic flow+∗	30. Electromagnetically induced	41. $BaFe_2As_2$+
5. Stochastic resonance+∗	18. Quantum Hall+∗	transparency+∗	42. Quantum dots+∗
6. Carbon nanotubes+∗	19. $CeCoIn_5$+	31. Light-induced drift+	43. Bose-Einstein condensates+
7. $NbSe_3$+	20. Inflation+	32. Proton-proton bremsstrahlung+	44. X(3872)∗
8. Black hole+∗	21. Exchange bias+∗	33. Antisymmetrized molecular	45. Relaxor+
9. Nanotubes+	22. Sr_2RuO_4+	dynamics+	46. Blue phases+
10. Lattice Boltzmann+∗	23. Traffic flow+∗	34. Radiative muon capture+	47. Black holes+∗
11. Dark energy+∗	24. TiOCl	35. Bose-Einstein+	48. $PrOs_4Sb_{12}$+
12. Rashba	25. Key distribution+	36. C_{60}+	49. The Schwinger multichannel method+
13. $CuGeO_3$+	26. Graphene+∗	37. Entanglement+	50. Higgsless+

very useful for labelling and content analysis. The details of scientific meme definition and analysis can be found in [Kuhn et al., 2014].

Through Tobias Kuhn who is now with the Vrije Universiteit Amsterdam we have obtained the full list of 1,578,079 memes in all 503825 APS journal papers published between 1964 to 2013. The average number of memes per paper is 9.26. A sample data is shown in Tab. 3.2.

3.2 Meme pair analyses

As a new scientific discovery is reported, interested researchers will follow this trend and use the same memes in their own papers. Over time, some memes will become more commonly used, while others will lose popularity and be forgotten. This phenomenon was discussed in [Kuhn et al., 2014] and shown in Fig. 3.1. A single meme can grow (become more popular), decay (become less popular), or even appear and disappear. In fact, most of the time scientific discoveries are not the *recitals* of individual memes, but the

Table 3.2: The sample data showing the list of memes for the paper "Random graphs with arbitrary degree distributions and their applications" [Newman et al., 2001]. In this table, memes are enclosed by quotation mark and separated by semi-colon.

DOI	https://doi.org/10.1103/PhysRevE.64.026118
Title	Random graphs with arbitrary degree distributions and their applications
Abstract	Recent work on the structure of social networks and the internet has focused attention on graphs with distributions of vertex degree that are significantly different from the Poisson degree distributions that have been widely studied in the past. In this paper we develop in detail the theory of random graphs with arbitrary degree distributions. In addition to simple undirected, unipartite graphs, we examine the properties of directed and bipartite graphs. Among other results, we derive exact expressions for the position of the phase transition at which a giant component first forms, the mean component size, the size of the giant component if there is one, the mean number of vertices a certain distance away from a randomly chosen vertex, and the average vertex-vertex distance within a graph. We apply our theory to some real-world graphs, including the world-wide web and collaboration graphs of scientists and Fortune 1000 company directors. We demonstrate that in some cases random graphs with appropriate distributions of vertex degree predict with surprising accuracy the behavior of the real world, while in others there is a measurable discrepancy between theory and reality, perhaps indicating the presence of additional social structure in the network that is not captured by the random graph.
Meme list	', the'; 'been'; 'the network'; ', the mean'; 'simple'; 'that'; 'a giant component'; '. we'; 'the size of the giant component'; 'has'; 'results'; 'if'; 'between'; ', and'; 'which'; 'in'; 'properties of'; 'degree'; 'this'; 'size of'; 'and the'; ','; '.'; ', including'; 'size'; 'scientists'; 'on the'; 'exact'; 'in the'; 'of the'; 'some'; '. in'; 'other'; 'for'; ', including the'; 'networks'; 'our'; 'we'; 'network'; 'paper'; 'are'; 'and'; ', we'; 'of'; 'by'; 'have'; 'number of'; 'of scientists'; 'behavior'; 'on'; ', and the'; 'a'; 'studied'; 'transition at which'; 'one'; 'transition'; 'the'; 'with'; 'component'; 'the position of the'; 'collaboration'; 'giant component'; 'to'

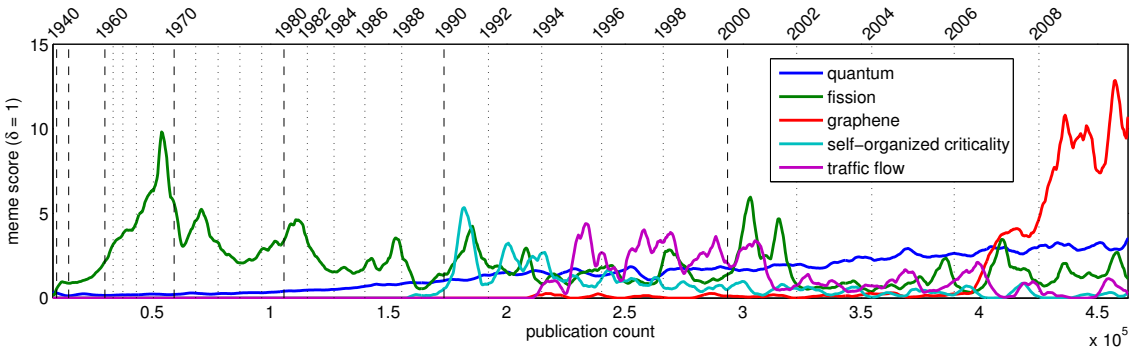


Figure 3.1: Meme scores of five exemplary memes from [Kuhn et al., 2014]. They exhibit very different histories: four of them show bursts at different points in time, while the fifth 'quantum' shows a very steady and almost linear path. The time axis is scaled by publication count. Reproduced from [Kuhn et al., 2014].

choruses of many memes. Science involve the complex interactions between many ideas, and therefore we expect it to show up as complex interactions between multiple memes at the linguistic level. This interaction of memes has been discussed in many papers [Weng et al., 2012; Gleeson et al., 2014], but they all focus on memes in social media. Research on scientific meme interactions is still lacking, and we will address this gap in the present section.

We start our study with the meme pair for two reasons: (i) the meme pair is the simplest model for interaction and what we learn from it will be useful for constructing more advanced models like triple interaction or general multiple interaction, (ii) this model is consistent with our understanding of the history of physics: physicists often propose the simplest theory first, before adding on higher-order corrections. At the meme level, we expect therefore the emergence of a core meme, before new memes are added one by one. This process can be well described by a meme pair model. For example, in the beginning we have *quantum mechanics*, then we add *relativistic correction* to *quantum mechanics* to obtain *relativistic quantum mechanics*. This process can be considered an interaction between the meme 'quantum mechanics' and meme 'relativistic' with the product 'relativistic quantum

mechanics'.

To figure out whether scientific memes can merge or split, we first counted all meme pairs in papers published between 1981 and 2010. To reduce the computational complexity and focus only on meaningful meme pairs, we cleaned the meme list before counting the pairs. The original meme list includes all n-grams that reproduce themselves through citation, and therefore include such n-gram as 'we', which has no scientific meaning. If we include 'we' in the meme-pair count, we will end up with many un-informative meme pairs like 'we'-'quantum', 'we'-'gravity' and so on. This will introduce noise into our analysis, but also produce a complete meme-pair list that is so large it cannot fit into our computer memory. Therefore we remove all memes in a blacklist, or contain a meme in the blacklist (see Tab. 3.3). Take Tab. 3.2 for example, 'a giant component' will be removed because it contains 'a' in the blacklist, while 'giant component' will be included in our meme-pair counting. After this, we also removed memes that only appeared once in one year, since most of them are rare long n-grams like 'via nonequilibrium molecular dynamics', which have complex meanings and therefore cannot be considered elementary memes. These two constraints can help us remove around 75% of the original memes. For example, in 1981 and 2010, the lengths of the original meme lists are 45,351 and 193,293, while the lengths after cleaning are 11,184 and 43,478. To focus on meme pairs that are activated between 1981 and 2010, we removed memes that appear less than or equal to 10 years during this time period because such memes are very rare in general and therefore are not well-accepted by the physics research community. Meme pairs containing such memes are therefore not very informative for our study. After this last filtering exercise 10,940 memes are left and we only consider the meme pairs between them.

We then counted the co-occurrences of meme pairs (two memes appearing together in at least three papers) between 1981 and 2010. The number of meme pairs under consider-

Table 3.3: The blacklist of memes that have no scientific meaning. All punctuations are enclosed by quotation mark.

Type	Memes
Article	the, a, an
Pronouns	that, those, this, these, which, we, it, all, both
Preposition	in, on, at, of, for, with, from, to, by, as, between, than, into, through, out, within
Conjunction	and, but, or, so, because, when, before, although, where, if
Verb	is, are, was, were, be, been, have, has, can
Determiner	our, such, its, their, other, no, not
Adverb	also, not, here, however, how, there
Punctuation	'.', ',', ';', ':'

ation is 1,067,422 and we noticed that many meme pairs appear only in a few isolated years because they have no significant relations. To remove such pairs that occur by chance, we removed all meme pairs that appeared fewer than 10 years between 1981 and 2010 and with this the number of meme pairs is reduced to 136,852. These meme pairs are the objects of our study in this section and we show a sample of this data in Fig. 3.2.

To identify meme pairs that grew closer with time, we ranked meme pairs by the ratio they grew between 1981 and 2010, that is $\frac{N(2010)}{N(1981)}$, in which $N(1981)$ is the number of times the meme pair appear together in 1981, and $N(2010)$ is the number of times the meme pair appear together in 2010. This formula can be improved by adding a control parameter δ as Kuhn *et al.* did in Eq. 3.2 to overcome the problem that some memes may not appear in 1981. Ultimately we calculated the *growth ratio* $\frac{N(2010)+\delta}{N(1981)+\delta}$ using $\delta = 5$ in this section unless stated otherwise. We list the top 10 growing meme pairs and top 10 decaying meme pairs in Tab. 3.4 and Tab. 3.5.

Although these lists of top 10 growing and decaying meme pairs are very useful, they do not tell us whether two memes are merging or splitting because growth ratio $\frac{N(2010)+\delta}{N(1981)+\delta}$ only consider the change of absolute number of the pair, which is high correlated with the populations of the individual memes. In the top 10 list of growing meme pairs, we

Index	MEME1	MEME2	1981	1982	1983	1984	1985	1986	1987	1988	
57132	model	results	290	319	322	341	368	409	400	519	56
67627	results	using	213	175	154	170	198	236	250	302	39
40012	field	magnetic	157	156	183	173	207	207	207	268	33
57350	model	using	176	138	153	140	155	209	208	268	31
57251	model	study	74	76	79	83	98	116	124	174	20
69692	state	states	148	156	152	152	214	198	188	219	25
62965	phase	transition	165	152	153	168	187	218	267	264	31
57330	model	two	157	161	143	159	184	207	183	244	30
65729	quantum	states	33	36	50	54	62	75	97	126	15
67519	results	study	66	77	89	95	107	133	160	189	21
65728	quantum	state	16	16	26	39	48	63	82	83	11
71533	two	using	74	76	63	58	92	101	92	126	15
34838	energy	using	184	145	156	145	152	187	220	221	26
67610	results	two	129	138	143	159	172	187	169	253	30
34385	energy	model	241	223	242	244	254	291	287	360	38
34635	energy	results	260	272	272	263	293	351	379	428	43
67437	results	show	72	75	91	77	95	112	134	154	21
53684	magnetic	magnetic field	99	98	129	111	149	155	152	208	25
40013	field	magnetic field	94	107	127	111	142	151	143	191	24
70353	study	using	48	45	41	34	57	58	81	77	11
38189	experimental	results	228	241	217	239	259	296	291	333	38
71181	theory	using	121	92	110	117	127	119	161	171	20
65762	quantum	two	21	25	28	40	45	53	56	81	10
67570	results	theory	228	215	247	234	305	303	285	315	35
65739	quantum	system	24	30	24	53	52	76	85	88	94
65767	quantum	using	20	19	21	17	26	43	47	46	59
68536	show	using	23	30	34	28	35	40	56	62	91
56996	model	phase	114	90	117	133	124	161	161	204	23
53878	magnetic	spin	62	63	74	68	78	85	84	103	11
69887	states	using	131	114	121	106	119	140	132	164	18

Figure 3.2: A screenshot of the DataFrame containing 139,248 meme pairs between 1981 and 2010. The first row tells us that the meme pair 'model' and 'results' occurred 290 times in 1981, 319 times in 1982 and so on.

Table 3.4: The top 10 growing meme pairs between 1981 and 2010 using growth ratio.

Rank	Meme pair	1981	2010
1	results, simulations	0	418
2	simulations, using	0	369
3	model, simulations	0	364
4	quantum, regime	0	322
5	coupled, quantum	0	302
6	model, standard model	0	293
7	different, dynamics	0	291
8	quantum, single	0	283
9	dynamics, state	0	277
10	simulations, study	0	273

Table 3.5: The top 10 decay meme pairs between 1981 and 2010 using growth ratio.

Rank	Meme pair	1981	2010
1	deduced, reactions	132	6
2	mev, reactions	293	19
3	angular, deduced	50	0
4	elastic scattering, mev	46	0
5	deduced, mev	119	8
6	mev, o16	39	0
7	mev, pion	38	0
8	distorted-wave, mev	36	0
9	mev, targets	35	0
10	c12, mev	35	0

notice that four meme pairs contain the meme 'simulations'. It is possible that these meme pairs are not really becoming closer to each other, but because the meme 'simulations' is becoming more popular, i.e. all other conditions remaining the same, but the numbers of meme A and meme B both increase about 10 times, then the number of A-B pair will also increase about 10 times.

To overcome this limitation and measure the real distance between memes, we proposed a simple probabilistic model: if there are N_1 meme1 and N_2 meme2, and the probability they appear together with each other (occurring in the same paper, but not necessarily consecutively) is p, then the probability there are N meme1-meme2 pairs is:

$$P(N) = \binom{N_1}{N}\binom{N_2}{N} p^N (1-p)^{\min\{N_1,N_2\}-N}. \tag{3.3}$$

Using the method of maximum likelihood estimation, we get the best estimation of p as:

$$\hat{p} = \frac{N}{\min\{N_1,N_2\}}. \tag{3.4}$$

We can then use $\hat{p}$ to check if two memes are become closer or further away by calculating

Table 3.6: The top 10 merging meme pairs between 1981 and 2010 based on the probabilistic growth ratio.

Rank	Meme pair	1981	2010
1	quantum, transitions	0	182
2	coupled, quantum	0	302
3	quantum, single	0	283
4	nuclear, quantum	0	87
5	optical, quantum	4	384
6	exchange, quantum	0	106
7	electronic, quantum	0	167
8	emission, quantum	0	113
9	cross section, mev	0	50
10	excitations, quantum	0	122

Table 3.7: The top 10 splitting meme pairs between 1981 and 2010 using probabilistic growth ratio.

Rank	Meme pair	1981	2010
1	quantum hall, shown	0	0
2	quantum hall, scattering	0	0
3	oscillations, quantum hall	0	0
4	confinement, dot	0	5
5	scanning tunneling microscopy, steps	0	0
6	mesoscopic, small	0	0
7	films, noise	4	0
8	proposed, scanning tunneling	0	4
9	dot, energies	0	7
10	band, wells	0	0

the ratio $\frac{\hat{p}(2010)}{\hat{p}(1981)}$. To overcome the problem that some term maybe 0, we introduce $\delta = 5$ to modify the Eq. 3.4 to

$$\hat{p} = \frac{N + \delta}{\min\{N_1, N_2\} + \delta}. \tag{3.5}$$

With this metric, the top 10 meme pairs that grew closer to each other and meme pairs that grew apart from each other are shown in Tab. 3.6 and Tab. 3.7. The top meme pairs are highly correlated with quantum optics, like 'quantum'-'transitions', 'optical'-'quantum', 'emission'-'quantum' and 'excitations'-'quantum'. Take 'quantum' and 'optical' for example: in 1981 we have 308 instances of 'optical' and 327 instances of 'quantum' with only 4 instances of 'quantum'-'optical' pair; in 2010 we have 1,527 instances of 'optical' and 3,635 instances of 'quantum' with 384 instances of 'quantum'-'optical' pair. Both 'quantum' and 'optical' became more popular going from 1981 to 2010, but the rise of their pair was even more significant than the rises of the memes themselves. Therefore these two memes really did get closer to each other. This is consistent with what we know about quantum optics, which has a long history — including an early part where most of the discussion is purely theoretical. With the advancement of laser techniques, physicists can finally test

Table 3.8: The top 10 growing/merging meme pairs between 1981 and 2010 and their ranks in the two ranking systems.

Rank in growth ratio	Meme pair	Rank in probabilistic growth ratio	Rank in probabilistic growth ratio	Meme pair	Rank in growth ratio
1	results, simulations	780	1	quantum, transitions	61
2	simulations, using	1392	2	coupled, quantum	5
3	model, simulations	1490	3	quantum, single	8
4	quantum, regime	312	4	nuclear, quantum	869
5	coupled, quantum	2	5	optical, quantum	29
6	model, standard model	75871	6	exchange, quantum	433
7	different, dynamics	241	7	electronic, quantum	76
8	quantum, single	3	8	emission, quantum	348
9	dynamics, state	334	9	cross section, mev	4989
10	simulations, study	4913	10	excitations, quantum	260

the theories in the lab and many papers are published after 1980s, and also stimulating other fields like quantum information and condensed matter physics.

Other meme pairs in this list are also informative: 'nuclear' and 'quantum' is highly correlated with the nuclear physics, 'exchange' and 'quantum' is highly correlated with calculation in quantum mechanism involving identical particles. However the same method does not work well with decaying pairs in Tab. 3.7. When we check the decaying pairs, we find that from 1981 to 2010, the populations of the high ranked pairs remain at very low levels, close to 0, but they both grew a lot separately. Because of the δ in Eq. 3.5, $\hat{p}$ is non-zero even the N is zero. Due the growth in $\min\{N_1, N_2\}$, $\hat{p}(2010)$ will be significantly smaller than $\hat{p}(1981)$. Therefore the pairs in Tab. 3.7 can not really represent splitting meme pairs, but may be artifacts because our equation amplify the change of probability in rare meme pairs.

Even with this shortcoming, the probabilistic method is still reliable in detecting rising

pairs, which is very useful for detecting emerging fields. Compared with simply counting the number of co-occurrences, the performance of this method is much better and we can see it from Tab. 3.8, in which top 10 rising pair in probabilistic growth ratio all have high rank in growth ratio, which means that their growth in absolute numbers are also high; on the other hand the top 10 rising pairs in growth ratio do not have high rank in probabilistic growth ratio in general, which means that their rise is more due to the rise in N_1 or N_2, and not really because they are growing closer. Examples like 'quantum' and 'optical', 'quantum' and 'transitions' show that emerging fields do change the linguistic habits of scientists and their co-occurrences can be used as a signal to detect this abstract trend. For splitting pairs, Tab. 3.5 provides better results. A more advanced method is needed to be able to simultaneously detect both merging and splitting pairs.

3.3 Meme labelling of TCs

Research thus far has already shown that memes are highly concentrated in their own medium-sized or small communities [Kuhn et al., 2014]. This characteristic inspired us to link memes with TCs because the correlations between memes and TCs can give us a much more comprehensive picture of the TCs' contents. In other words, we want to utilize this characteristic to label a TC with several key memes. If this labelling works well, we can roughly know the research content very quickly by reading the memes, as if they are 'keywords' provided by the authors, since keywords do not exist for Physical Review journals. This is much faster than reading the title and abstract, and the challenge of picking up relevant TCs (a problem we discussed in the beginning of this chapter) will be solved. For example, in Fig. 3.3 we intuitively realize that TC 'Group 1' is closely related to Meme 1, whereas Meme 2 appears in all Group 1, Group 2 and Group 3, and is therefore not a good label for any of these three TCs.

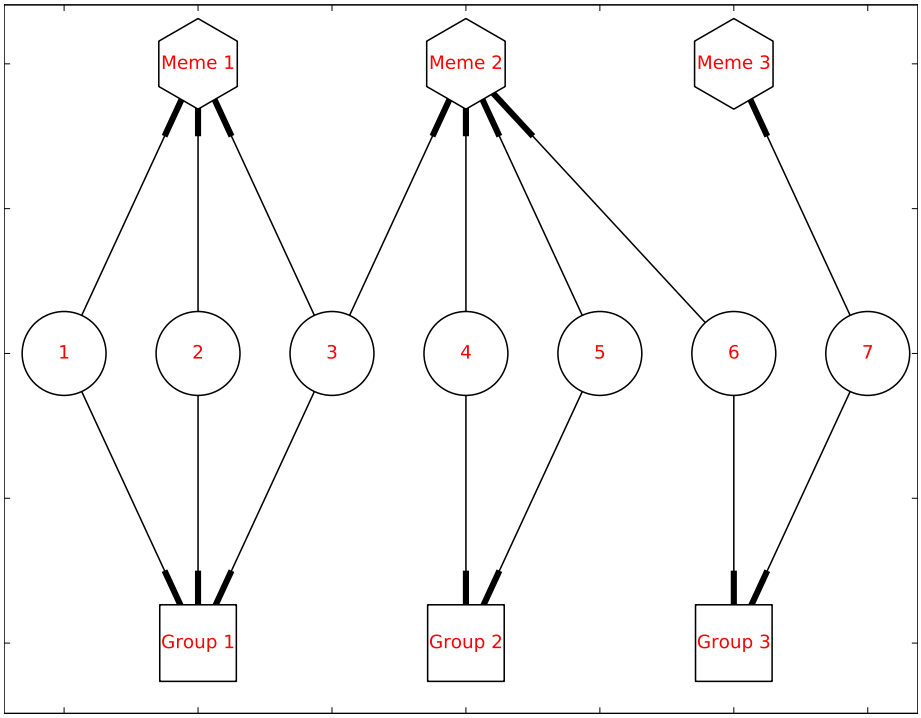


Figure 3.3: In this figure, circles represents papers, hexagons represents memes, and squares represents TCs. An edge from a paper to a meme means that the paper contains the meme, whereas an edge from a paper to a TC means that the paper belongs to the TC.

There are many different ways to measure the correlation between the topics and articles, like the *topic model* in machine learning and natural language processing [Blei, 2012]. In this section we will focus on two simple methods: *Jaccard index* and *mutual information* from the network science perspective. The Jaccard index is given by:

$$J(M_i, TC_j) = \frac{\left|P(M_i \cap TC_j)\right|}{\left|P(M_i \cup TC_j)\right|}, \tag{3.6}$$

in which $P(M_i \cap TC_j)$ represents the papers in TC_j containing meme M_i, $P(M_i \cup TC_j)$ represents the union of papers containing meme M_i and papers belonging to TC_j. The mutual information (MI) is given by:

$$I(M_i; TC_j) = H(M_i) - H(TC_j) - H(M_i, TC_j), \tag{3.7}$$

in which $H(M_i)$ is the marginal entropy of papers containing meme M_i, more precisely, it is the entropy of randomly picking up one paper containing the meme M_i; $H(TC_j)$ is marginal entropy of papers belonging to TC_j; $H(M_i, TC_j)$ is the joint entropy of papers belonging to TC_j and containing meme M_i. To remove the effect introduced by $H(M_i)$ and $H(TC_j)$, we can use the normalized mutual information (NMI):

$$NMI(M_i; TC_j) = \frac{I(M_i; TC_j)}{H(M_i) + H(TC_j)}. \tag{3.8}$$

We calculated the Jaccard index, MI and NMI between all memes and all TCs over the years 1981 to 2010. The results for 1981 are shown in Tab. 3.9.

Table 3.9: The top 5 memes for TCs in 1981 using MI, NMI, and Jacaard index (values enclosed in parentheses). The naming convention is such that 00 is the bottom block in Fig. 4.2, 01 is the block just above 00 and so on.

TC number	**Top 5 meme using MI**	**Top 5 meme using NMI**	**Top 5 meme using Jaccard index**
00	neutrino (0.0183), gauge (0.0147), leptons (0.0147), su(5) (0.0147), unified (0.0145)	neutrino (0.0553), leptons (0.0498), su(5) (0.0498), unified (0.0475), higgs (0.0407)	gauge (0,143), mass (0.138), neutrino (0.126), decay (0.124), weak (0.11)
01	quark (0.0211), quantum chromodynamics (0.0211), gauge (0.0182), production (0.0149), confinement (0.00921)	quantum chromodynamics (0.0507), quark (0.0493), gauge (0.0366), production (0.0292), confinement (0.0244)	gauge (0.14), production (0.131), quark (0.128), quantum (0.113), quantum chromodynamics (0.101)
02	reactions (0.0881), mev (0.071), reaction (0.0277), deduced (0.0267), measured (0.022)	reactions (0.119), mev (0.103), deduced (0.049), reaction (0.0484), nuclei (0.0386)	reactions (0.397), mev (0.35), measured (0.202), reaction (0.186), scattering (0.174)
Continued on next page			

Table 3.9 – continued from previous page

TC number	Top 5 meme using MI	Top 5 meme using NMI	Top 5 meme using Jaccard index
03	reactions (0.0469), mev (0.0256), fragments (0.0222), fusion (0.0196), + (0.0182)	fragments (0.0833), reactions(0.0806), fusion (0.0721), + (0.0628), fission (0.0559)	reactions (0.258), mev (0.197), angular (0.155), + (0.149), fragments (0.143)
04	laser (0.0191), coherent (0.0125), two-level (0.00797), bistability (0.00731), optical (0.00729)	laser (0.0506), coherent (0.0417), two-level (0.0295), bistability (0.0285), optical bistability(0.0236)	laser (0.171), optical (0.109), coherent (0.102), time (0.0954), field (0.0896)
05	cross (0.0138), collisions (0.0137), cross sections (0.0136), ions (0.0136), electron capture (0.0132)	electron caputre (0.0592), capture (0.0432), collisions (0.0376), ions (0.0356), projectile (0.0344)	collisions (0.152), ions (0.151), cross sections (0.145), cross (0.138), electron (0.132)
06	laser (0.00897), plasmas (0.00731), electron (0.00673), plasma (0.0065), laser field (0.00553)	plasmas (0.0356), laser field (0.0306), laser (0.0294), free-free (0.0267), plasma (0.0234)	laser (0.122), plasma (0.1), electron (0.0954), plasmas (0.0852), electrons (0.0754)

Continued on next page

Table 3.9 – continued from previous page

TC number	Top 5 meme using MI	Top 5 meme using NMI	Top 5 meme using Jaccard index
07	critical (0.0424), renormalization-group (0.0221), phase (0.0217), ising (0.0216), ising model (0.0185)	critical (0.0803), ising(0.0543), renormalization-group(0.054), ising model (0.0509), exponent (0.0424)	critical (0.262), phase (0.19), transition (0.151), two-dimensional (0.137), renormalization-group (0.136)
08	localization (0.0193), spin-glass (0.0153), random (0.0122), spin-glasses (0.0113), disordered (0.00983)	localization (0.0813), spin-glass (0.0669), spin-glasses (0.0538), random (0.0459), disordered (0.0407)	localization (0.157), random (0.133), spin-glass (0.128), conductivity (0.113), disordered (0.105)
09	auger (0.00908), graphite (0.009), spectra (0.00819), photoelectron (0.00773), 4f (0.00702)	graphite (0.0261), auger (0.0258), photoelectron (0.0229), 4f (0.0213), rydbery(0.0193)	spectra (0.118), states (0.105), ev (0.096), state (0.0918), atoms (0.0895)

Continued on next page

Table 3.9 – continued from previous page

TC number	Top 5 meme using MI	Top 5 meme using NMI	Top 5 meme using Jaccard index
10	self-consistent (0.029), band (0.0235), electronic (0.0215), reactions (0.0188), silicon (0.0173)	self-consistent (0.0558), band (0.0373), electronic (0.0345), silicon (0.0341), pseudopotential (0.0338)	band (0.168), electronic (0.158), surface (0.155), results (0.151), structure (0.145)
11	superconducting (0.0158), superconductivity (0.0118), tunneling (0.011), electron-phonon (0.00804), superconductors (0.00667)	superconducting (0.0579), superconductivity(0.0502), tunneling (0.0422), electron-phonon (0.0344), superconductors (0.0301)	superconducting (0.159), tunneling (0.121), superconductivity (0.113), k (0.0967), electron-phonon (0.087)
12	excitons (0.0102), electron-hole (0.00924), exciton (0.00767), luminescence (0.00577), electron-hole liquid (0.00557)	excitons (0.0664), electron-hole (0.0599), exciton (0.0483), electron-hole liquid (0.0432), luminescence (0.0377)	excitons (0.142), electron-hole (0.132), exciton (0.118), luminescence (0.094), gaas (0.0906)

Continued on next page

Table 3.9 – continued from previous page

TC number	Top 5 meme using MI	Top 5 meme using NMI	Top 5 meme using Jaccard index
13	solitons (0.0127), soliton (0.0127), polyacetylene (0.0115), one-dimensional (0.0112), chain (0.0107)	solitons (0.0624), soliton (0.06), polyacetylene (0.0583), nbse3 (0.0487), chain (0.0464)	one-dimensional (0.135), chain (0.119), soliton (0.118), solitons (0.109), polyacetylene (0.093)

In general, the top 5 memes are very informative for labelling. For example, the top 5 meme using MI for TC 00 is 'neutrino', 'gauge', 'leptons', 'su(5)' and 'unified'. Any researcher with background in physics can point out this field is about particle physics with emphasis on grand unified theory. With the help of these key memes, researchers in any field will be able to use the alluvial diagram method to trace the evolution in their field. Researchers no longer need to read the titles and abstracts to judge the research content of a TC: with the help of key memes, they can get a rough impression first, and read few paper to check, which is much faster.

From Tab. 3.9 we find that the performances of the three indictors are slightly different: in general the top 5 memes obtained using the Jaccard index are worse than those obtained using MI and NMI since the Jaccard index typically gives priority to very frequent memes like *decay*, *weak*, *time*, *field*, which are widely used by many fields, and therefore less informative to the contents of specific TCs. MI results are very close to NMI results in general, and in many cases they give the same top 5 memes, differing at most in ranks. In some cases the NMI result is slightly better than the MI result, like for TC 00, NMI picks the

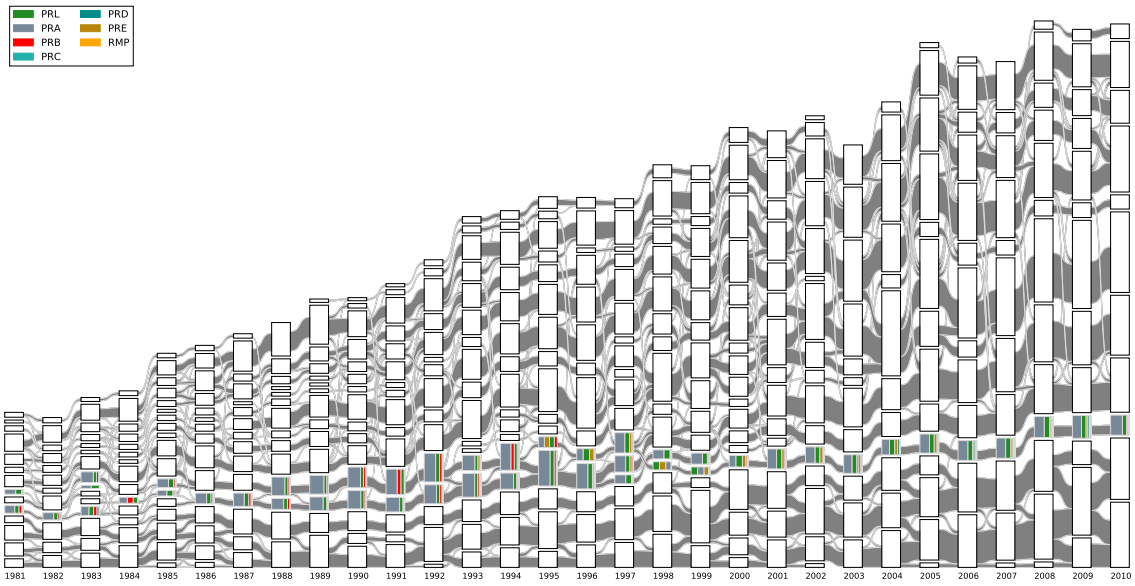


Figure 3.4: The alluvial diagram of APS papers from 1981 to 2010, where we colored only TCs which has NMI higher that 0.01 for the meme 'laser'.

higgs, for TC 02 NMI picks *nuclei*, which are very informative memes. Therefore we use the NMI for the later part of this chapter.

We can use this technique to automatically pick up streams that are high correlated with specific topics, like the case study in Chapter 2. In the BEC case study we had to read the 10 most cited papers' titles and abstracts to determine whether this TC is BEC-related or not. Using NMI, we can simply set a threshold and highlight the related streams. It's very fast and convenient compared to identifying the TCs manually. In Fig. 3.4 we highlighted the TCs highly correlated with the meme 'laser' (laser cooling is the key technique for Bose-Einstein condensation), and we notice that this figure is very similar to Fig. 2.7. This similarity is a testimony to the utility of our method.

In summary, in this section we tried to label the TCs using informative memes. To do this, the Jaccard index, MI, NMI can all be used as indictors to pick up the top represen-

tative memes. These top representative memes can then used as keywords in place of the research contents of TCs. A comparison between the three indictors show that the performance of NMI is better than MI and Jaccard index. Using NMI we can automatically pick the streams that are highly related to some special topics. This automatic technique will make our alluvial diagram method more efficient and friendly to researchers in any field who want to trace the evolution of their science.

3.4 Meme community structure

The previous results shown that different TCs have different representative memes. This suggest that we should also think about the community structure of memes since such community structure is very clear in the citation network, as shown in Chapter 2. The intuition that guide us here is that the language people used in one TC is not constructed randomly but is intelligently organized. Therefore particular memes will co-occur significantly more frequently than others. This is very much like a community structure, therefore we try to extract the community structure of memes from the APS data set.

We do so by first constructing a meme network, where nodes are the memes and an edges between two memes represent that the two meme appeared together in at least one paper, and the weight of the edge is the number of papers that two memes appeared together in. Thereafter we tried to use the Louvain method to detect the community structure in this meme network, but find unfortunately that the modularity of the best partition is very low. Take the meme network in 1981 for example, the network contain 15,575 nodes and 1,053,209 edges, and its average degree is 135.4. The modularity of best partition is only 0.194, and four largest communities contain about 27.1%, 25.9%, 23.9%, 17.7% of the nodes, while the remaining 6 communities only have 5.4% the nodes of the whole network. This modularity is much lower than the typical modularity of BCN (around 0.7), and the

Table 3.10: The top 10 frequently used memes in the 1st, 2nd, 3rd and 4th largest communities detected in 1981 meme network.

1st	2nd	3rd	4th
energy	model	observed	results
scattering	theory	temperature	states
measured	also	transition	using
data	two	measurements	found
reactions	field	magnetic	structure
cross	function	phase	obtained
state	one	behavior	calculations
mev	approximation	lattice	experimental
energies	system	range	calculated
potential	shown	dependence	effects

number of large communities is also much less than the typical number of TCs (between 10 and 20). We also checked the members of the four largest communities, and found that they are all widely used across many fields, instead of being closely related to specific fields (see Tab. 3.10). Based on the above reasons, we believe that such groups detected cannot be considered as meme communities in this section.

We also tried other methods to detect community structure, like Infomap [Rosvall and Bergstrom, 2008] and clique percolation method (CPM) [Palla et al., 2005]. Unlike Louvain method, which tries to optimize the modularity function, Infomap reveal community structure by compressing the description length of the probability flow in network and CPM tries to find the overlapping community structure in the form of cliques. However, neither of them can solve this problem well: Infomap extracted a gaint community, which is about 83.6% of whole meme network in 1991, and this is against our intuition of there being more than one meme communities. The CPM software CFinder [Palla et al., 2005] cannot finish the detection in a reasonable time because of the size of the network.

Given how well community detection works in the paper networks, and the demonstrated correlation between memes and specific TCs in the previous sections, why is it that

community detection cannot work for memes? We believe the real reason is the redundancy built into language, as well as the way we build the network. Because of redundancy in language, we will include many repeated memes, many of them sharing the same meanings. If any two memes appear in one paper, we will draw an edge (complete graph) even if they appear together just by chance. These two reasons together will introduce many 'noisy' edges into our graph and make the whole network much more homogeneous, making the real community structure very hard to detect no matter which method we used. To solve this problem, we can construct the network differently, including but not limited to merging similar memes into one meme or removing the memes that are widely used but not very informative about any specific field. Another possible way to overcome this problem is *backboning* [Serrano et al., 2009], which can remove insignificant edges to make the underlying community structure detectable.

Even with these difficulties, we are eventually still be able to cluster the memes based on their distances along the TC dimensions. Instead of using co-occurrences in papers, we used the memes' distribution among the TCs. Specifically, we pick the 1000 most frequent memes, then count how many times they appear among different TCs to get their distribution vectors. For example, if meme A appears in TC 00 a_0 times, in TC 01 a_1 time and so on, the distribution vector of meme A is then $A = (a_0, a_1 ...)$. The distance between meme A and meme B can then be defined as the cosine similarity between vector $(a_0, a_1 ...)$ and $(b_0, b_1 ...)$:

$$D(A,B) = \frac{A \cdot B}{\|A\|\|B\|}. \tag{3.9}$$

Finally we performed hierarchical clustering using cosine similarity with the complete-linkage algorithm and the dendrogram of top 1000 memes is shown in Fig. 3.5. With a threshold very close to 1, like 0.99, we can easily cluster the 1000 memes into N groups, where N is the dimension of the distribution vectors, which is also the number of TCs for

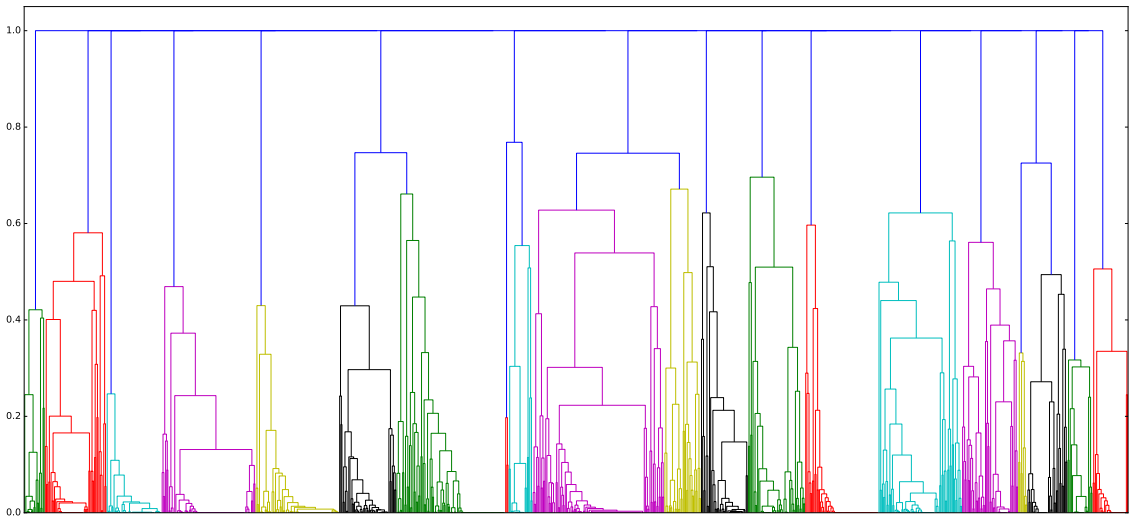


Figure 3.5: The dendrogram of top 1000 frequently used memes in 1991, each leaf represents one meme, and the y-axis is the cosine distance.

that year.

Although this method produces very clean clusters, we must emphasize that these do not form a 'real' meme community, because the distances between memes are based on their distributions among the TCs, and therefore are not their intrinsic distances. They are more like the label groups for the TCs instead of a meme community. However the memes that be clustered together are indeed closer to each other than they are to memes outside the clusters because they appear very often in the same TCs, so while the results are not really the meme community structure, we are one step closer to our goal. In other words, our problem of discovering how memes are associated with each other is only half solved and remains an open question. We will discuss potential solutions in Chapter 5.

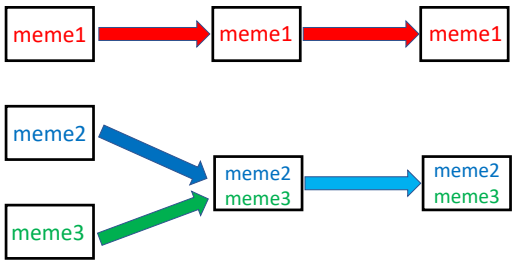


Figure 3.6: A example of trivial coevolution in form of alluvial diagram, each block represents a TC and the text inside is the meme it carries.

3.5 Coevolution between TCs and scientific memes

Inspired by our discovery that there are merging meme pairs like 'quantum' and 'optical', we want to know whether there is any correlation between meme interactions and TC interactions. We noticed the parallel between this two and believe they are two sides of the same coin: at the paper level, TCs can interact with each other to split or merge, at the same time the memes in the papers may have the same behavior: the merging of 'quantum' and 'optical' coinciding with the rise of the field of quantum optics. What is the correlation between TC interactions and meme interactions? This is the question we want to address in this section.

Indeed, this question is rather complex, and the reality is not quite as simple as that shown in Fig. 3.6, which is one of the simplest cases we can imagine. In this example, the top stream evolves independently and the meme it carries also does not change. In contrast, the TC carrying meme2 and the TC carrying meme3 merge together and the new TC contains both memes and continue evolve as a merged field '2+3'. In this case, the merging of TCs and merging of memes are perfectly synchronous. However, reality itself is much more complicated than what is suggested by this naive model, as memes may be distributed across multiple TCs and individual TCs may evolve in very complex fashions

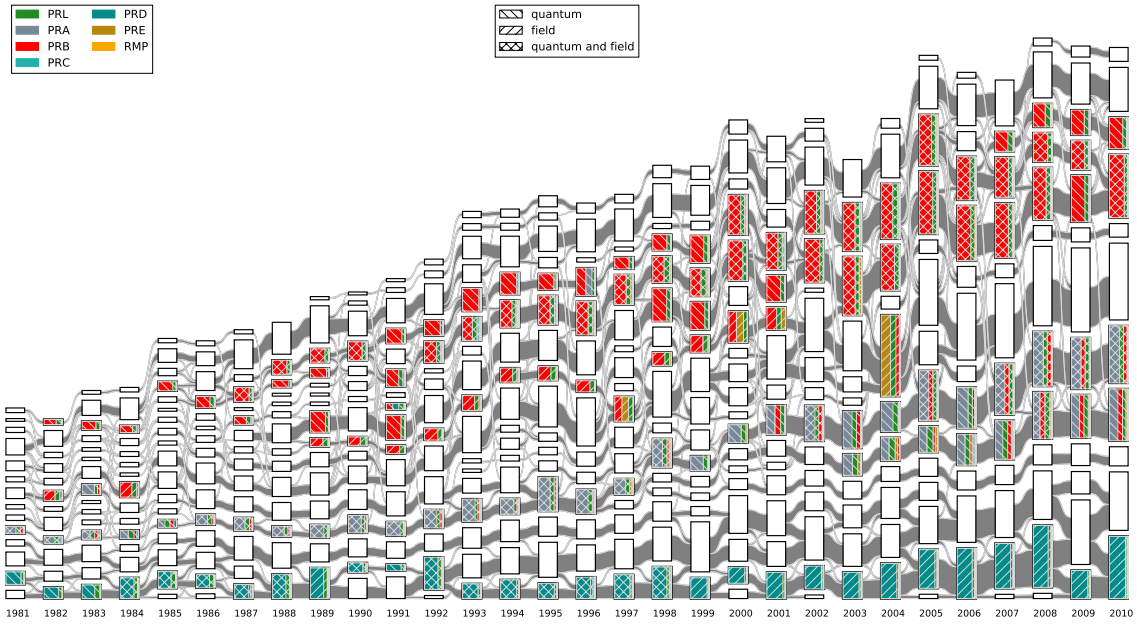


Figure 3.7: In this alluvial diagram, we only highlight the TCs which are highly correlated with "quantum" and "field". The threshold we used is Jaccard index = 0.07 for 'quantum' and Jaccard index = 0.05 for 'field'.

(for example, one TC may split and merge at the same time). Therefore it is very difficult to link TC merging and splitting events with meme merging and splitting events. Take Fig. 3.7 for example, the two memes 'quantum' and 'field' are popular, and therefore they are distributed across several TCs. In some streams, they developed independently, and in other streams they merged and evolved together. If we pick one evolution event (splitting, merging or continuing), then it is very easy to compare the meme populations before and after the event. However, each TC contains hundreds of memes, so how do we figure out which memes are correlated with the events? More importantly, as APS published more and more papers over time, many popular memes showed slow and steady growths. Thus we also need to distinguish the real signal from the background trend before we do the correlation analysis. Due to these factors, it is very challenging to quantify the correlation

of interacting memes.

To handle this challenge and develop a general theory of meme dynamics, we study this problem from another perspective: will splitting and merging affect the population of memes? In Fig. 2.9 we have shown that the size of a TC is governed by a simple linear recombination relation Eq. 2.3. Therefore we propose the same model for the meme population: the contribution of C_m^t to the meme population in C_n^{t+1} is proportional to the meme population in C_m^t and also the normalized forward intimacy index $I_{mn}^f / \sum_n I_{mn}^f$, i.e.

$$M'(C_n^{t+1}) = \sum_m M(C_m^t)(I_{mn}^f / \sum_n I_{mn}^f), \tag{3.10}$$

where $M'(C_n^{t+1})$ is the predicted population of meme M in TC C_n^{t+1} and $M(C_m^t)$ is the observed population of meme M in TC C_m^t. To test this model we first chose the top 1000 frequently used memes in 2010 and trace their populations in each TC between 1981 and 2010. For each meme in each TC between 1982 and 2010, we show in Eq. 3.10 the predicted meme population, compared to the real data. The regression results are similar to those shown in Fig. 2.9 but the fluctuations are larger. Then we divided all the TCs in two group based on their backward mixing degrees. The first group has backward mixing degree higher than median, the more merging group, and the second group has backward mixing degree lower than median, the less merging group. If a merging event can stimulate the spreading of memes in general, then the real value will tend to higher than the predicted value, or at least the more merging group will behave differently from the less merging group. However, we did not observe significant differences between (b) and (c) in Fig. 3.8.

This does not mean that we have to conclude that merging will not affect meme spreading, because of reasons stated below. In Fig. 3.8 we included all memes in a TC to test our hypothesis. It means that if a TC has high backward mixing degree, we include all data

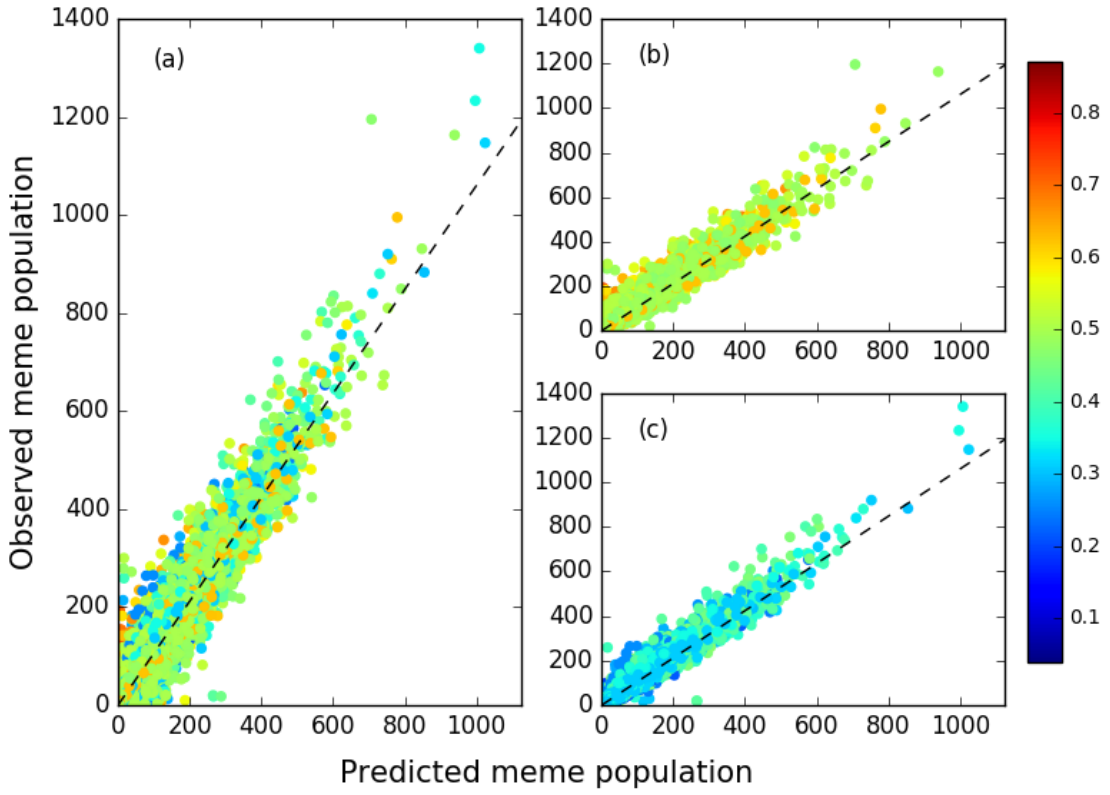


Figure 3.8: (a) Plot of observed (y-axis) against predicted (x-axis) meme population of recombined TCs, showing a linear growth with slope 1.06 (dashed line). This linear growth is the same for TCs with (b) high (red) or (c) low (blue) backward mixing degree.

points in (b) even when such memes are not relevant to new scientific discovery. Normally each TC will focus on one research direction and make contributions to that field. Therefore, if T_1 is a highly merging TC that focus on the field F, we would expect a boost only for memes related to the field F. However in Fig. 3.8 we have included all top 1000 memes to make our test general. Consequently even if there is really a boost effect, it will be masked by overwhelmingly many memes that do not benefit from the merging event. To overcome this problem, we need to pick memes are really correlated with the merging event. However, if we do that, we run the risk of circular reasoning.

To end this Chapter, let us remind ourselves that our initial goal was to quantify the correlations between meme interactions and TC interactions. However, the evolution on the two different levels are very complex indeed, and it is challenging to even study them within a unified framework. In fact, even after simplifying our question to test the boost effect of merging events, our simple linear recombination model Eq. 3.10 can fit the top 1000 meme populations between 1981 and 2010 well, but did not find any significant difference between memes in highly merging TCs and less merging TCs. We believe this is because of our poor choice of partition and we will discuss more about this in Chapter 5.

CHAPTER 4

Evolution prediction and betweenness analysis

In Chapter 2, we demonstrated the utility of visualizing and analysing scientific knowledge evolution for physics at the aggregated mesoscale through the use of alluvial diagrams [Liu et al., 2017]. In this picture, papers are clustered into groups (or communities) and these groups can grow or shrink, merge or split, new groups may arise while the others may dissolve. This shares a very strong parallel with what some researchers discovered in social group dynamics [Palla et al., 2007]. More importantly, many breakthroughs were made by scientists absorbing knowledge from other fields, often in a very short time. On the alluvial diagrams, these knowledge transformations manifest themselves as merging and splitting events. Clearly, funding agencies, universities and research institutes would want to promote growing research fields, and particularly those where breakthroughs are imminent. This is why it is important to be able to predict the future events. We attempted this in Chapter 2 by analysing the correlation between event types and several network metrics. Unfortunately, such predictions are very noisy. While merging events are highly correlated with interconnections between communities, the correlation between splitting

events and the internal structure of communities are much more complex; besides, the predictions of forming, dissolving, growing, shrinking were not considered at all.

Given the recent successes in the area of machine learning and artificial intelligence to a variety of prediction problems [Carrasquilla and Melko, 2017; Ahneman et al., 2018], as well as having developed and validated a general framework to predict social group evolution in [Saganowski et al., 2017], we decided to utilize machine learning techniques—more specifically, Group Evolution Prediction to fill the gap in predicting scientific knowledge events [Saganowski et al., 2015; İlhan and Öğüdücü, 2016; Pavlopoulou et al., 2017]. The overall idea behind the Group Evolution Prediction (GEP) method is to build a classification model trained with historical observations in order to predict the future group changes based on their current characteristics, such as size, density, average degree of nodes, etc. A single historical observation consists of a set of features describing the group at a given point in time, and an event type that this group just experienced. The profile of the group may reflect its structure (e.g. density), dynamics (e.g. average age of its member articles) or context (e.g. the journals which the articles—group members—come from). The GEP method is a general framework rather than a fixed algorithm. Therefore it is very convenient to apply it to different type of networks and test the effect of different factors. This gives us a lot of freedom to test different classifiers. In total, we used over 100 features, some of which were already known to the literature, whereas the others focusing on the dynamics and context are the new, unique features proposed in this paper. Indeed, when we rank the most valuable features contributing to successful prediction of knowledge evolution events, the new features are among the best ones. In order to be able to perform prediction of future group changes, we have to track and learn the model on the historical cases. For that purpose, the group changes from the past (historical evolution) need to be defined and discovered using the methods successfully applied to the

Social Network Analysis field, e.g. the GED method [Bródka et al., 2013], Tajeuna *et al.* method [Tajeuna et al., 2015] or other [Bródka et al., 2014]. Most of the methods consider the similarity between the groups in the consecutive time windows as a major factor to match similar groups and further to identify the evolution event type between them. In this chapter, we apply the GED method, which varies both the group quantity (the number of common members) and the group quality (the importance of common members), in order to match related groups. By comparing the inclusion measure between groups with control parameters α and β, GED can assign groups with event labels accordingly. By adjusting the parameters α and β, GED can be used in any type of evolving networks without changing any other parts. This feature gives the GED method high flexibility. This allows us to enrich the co-citation evolution network with information about member relations, which is depicted in the Social Position measure [Brodka et al., 2009].

The entire analytical process consists of several steps that are primary defined by the Group Evolution Prediction (GEP) framework. The GEP method is the first generic approach for the prediction of the evolution of groups [Saganowski et al., 2017], in our case groups correspond to TCs. The GEP process consists of six main steps: (i) time window definition, (ii) temporal network creation, (iii) group detection, (iv) group evolution tracking, (v) evolution chain identification and feature calculation, and (vi) classification using machine learning techniques. Thanks to its adaptable character, we were able to apply it to the BCN and CN differently. First, the bibliographic coupling network (BCN) and co-citation network (CN) are extracted from the references placed in the papers from a given time window, see Fig. 4.1, and this is carried out separately for each period. As a result, we get a time series of BCNs/CNs. Next, paper groups called topical clusters (TCs) are extracted using the Louvain clustering methods, independently for each BCN/CN in the time series. Having TCs for consecutive periods, we were able to identify changes in

TC evolution using the Group Evolution Discovery (GED) method that appropriately labels the TC changes. Each group is described by the set of predictive features. Finally, we applied the Auto-WEKA tool to find the best predictive model and its parameters from the wide range of all possible solutions. The commonly known average F-measure was used as a prediction performance measure. Independently, the features ranking and its validation were performed to find the most valuable TC measures. Based on this ranking, a structural measure node betweenness was selected for the more in-depth studies as the early signal for splitting or merging.

In this chapter, we extract groups—topical clusters (TCs)—from the bibliographic coupling networks (BCNs) and independently from the co-citation networks (CNs) for the period 1981-2010. Next, the GED method is utilized to label four types of evolution events (changes of TCs): continuing, dissolving, merging and splitting. Then, we use an auto-adaptive mechanism to find the most predictive machine learning model together with its parameters for each network. Additionally, two scenarios were considered for each network: when the number of events of each kind is imbalanced (the original case) and balanced by equally sampling. In general, the prediction quality was satisfactory good for all event types, with F-measures substantially exceeding 0.5. Such values are significantly greater than the baseline F-measures as of 0.14–0.21 for both networks. The feature ranking tells us that the most informative features are context-based like the number of PRE, PRB, and RMP papers belonging to the group, and the structural features like the degree, closeness, and betweenness. While looking more carefully at the betweenness of papers from two *merging* TCs, we find the significantly higher betweenness for papers that are linked across these two TCs than those connected inside the TCs. No such enhancement in betweenness was found for *continuing* TCs, while a significant decrease in average betweenness was found for *splitting* TCs. In summary, our findings suggest that evolution-

ary events in the landscape of physics research can be predicted accurately using various machine learning models, and understanding this predictive power in terms of important features is a worthwhile future research direction.

4.1 Training data for evolution prediction

Physics research evolution for 1981-2010

We begin with studying how scientific knowledge evolved in terms of communities of research papers, and how these communities changed over time. There are several studies on evolution of knowledge within the set of whole journals [Rosvall and Bergstrom, 2010], which is considered as the analysis on the macroscopic level. Also some research has been carried out for the collection of papers, usually involving some subjective criterion provided by the authors, e.g. only papers cited at least 100 times [Chen and Redner, 2010]. As a result, they focus only on a small subset—the most prominent, frequently cited papers, which do not represent the whole diverse domain knowledge. This kind of analysis is considered as microscopic. In our approach, we assume that the most informative way is to analyse neither the entire journal, nor the most cited papers, but whole communities of closely related papers. These communities emerge naturally since they share the same citation patterns. The analysis at such level provides better balance between high and low granularity. We call this kind of analysis as mesoscopic, because it is in-between the macroscopic scale of journals and the microscopic scale of individual papers. However, if we perform community detection directly on the citation network, we might end up with communities consisting of both old and recent papers simultaneously. In such case, it is difficult to interpret how scientific knowledge has evolved from the past to the present. We should be able to explain that such and such communities represent scientific

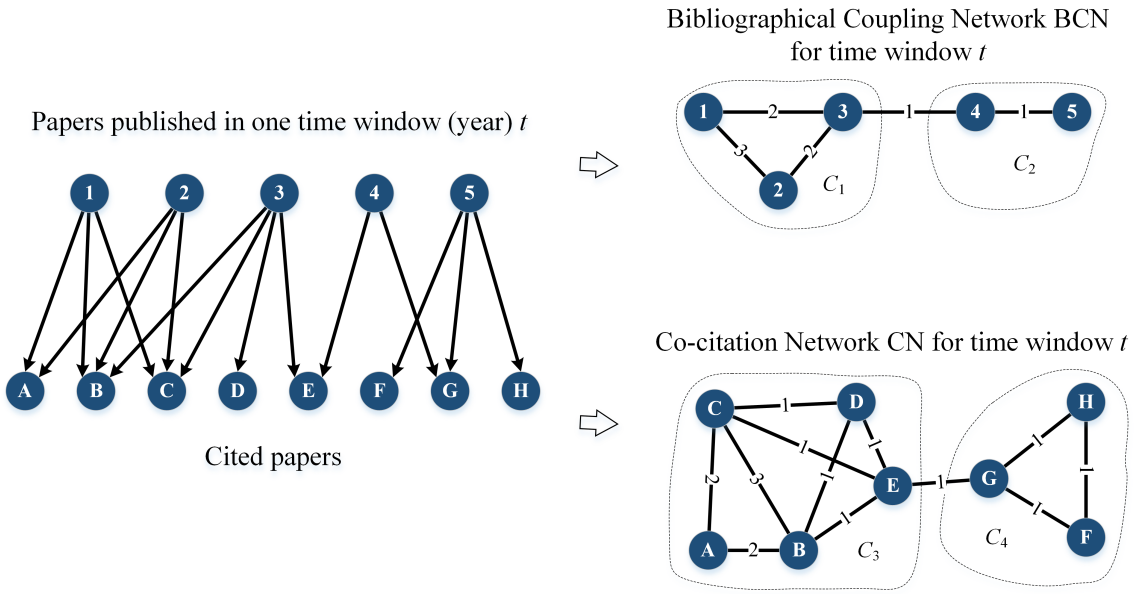


Figure 4.1: The process of building a Bibliographical Coupling Network (BCN) and Co-citation Network (CN) from the citation bipartite network for a given period—year t. Both BCN and CN are undirected and weighted; the weights denote the number of shared citations (BCN) or co-citing papers (CN). Separate topical clusters are extracted for BCN (C_1, C_2) and CN (C_3, C_4). Nodes with numbers are papers from a given period being considered and nodes with letters are their references.

knowledge from an earlier year, whereas the other communities correspond to scientific knowledge from another consecutive year. This enable us to compare them and to distill a picture of how scientific knowledge has evolved from past to present. It requires, however, to construct the networks from research papers that are published in a given year (bibliographic coupling), or papers that are cited in a given year (co-citation). The bibliographic coupling network (BCN) reflects the relation between present publications while the co-citation network (CN) represent the relation between papers which have strong influence on recent publications. In this way, we can detect communities over the years, and study how they evolve year by year.

In the BCN and CN, nodes represent papers and undirected but weighted edges denote

the bibliographic coupling strengths and co-citation strengths, respectively. That is, if two papers share w common references, the BCN edge between them would have a weight of w. For example, papers 1 and 2 in Fig. 4.1 share three citations: A, B, and C, whereas papers 3 and 4 commonly cite only one paper—E. On the other hand, if two papers are cited together by w' papers, the edge between them in the CN receives weight w'. Papers A and B are cited together by two other papers: 1 and 2, but papers B and C by three, i.e. additionally by paper 3. Both BCN and CN are temporal networks, in which the nodes are all papers published within a specific time window (BCN) or papers cited within a given time window (CN). We assume that the reasonable time window for bibliographical data is one year to facilitate the analysis of changes in scientific knowledge, i.e. changes in topical clusters year by year. For the BCN, only the giant component, which in most cases occupies 99% of the whole BCN, will be considered for the TC detection and evolution analysis. For the CN, we do not use all papers cited in the given time window because most of them are cited only a small number of times, and thus they have little influence on the broader knowledge evolution. Therefore, we rank all available N papers $p_1, p_2, \ldots, p_N$ in the descending order by the number of times they are cited in this time window (year): $f_1, f_2, \ldots, f_N, f_1 \geq f_2 \geq ... \geq f_N$. Next, we choose the top n papers $p_1, p_2, \ldots, p_n$, that totally gathered $\frac{1}{4}$ of all citations, i.e. such that $n < N$ is the smallest integer to satisfy $\sum_{i=1}^{n} f_i \geq \frac{1}{4}\sum_{j=1}^{N} f_j$.

After building BCN and CN, the Louvain method was used to extract the community structures. By checking the Physics and Astronomy Classification Scheme (PACS) numbers of the papers in these communities, we have shown that the BCN communities are meaningful and reflect the real structure of the scientific communities in Chapter 2. For the CN communities, this validation is tricky because of two problems: (i) the old Physics Review papers have no PACS numbers, and (ii) PACS was revised several times, so the

same numbers in different versions can potentially refer to different topics, or the same topics are referred to by different numbers in different versions. Nevertheless, systematic validation seems to be impossible although a quick check on some CN communities after 2010 suggests that CN community structure also reliably reflects the actual scientific community. We refer to these validated units of knowledge evolution as topical clusters (TCs) in BCN and CN.

In Fig. 4.2, we provide the alluvial diagram that depicts the evolution of TCs within the BCNs for the period between 1981 to 2010. The equivalent alluvial diagram for the CNs is shown in Fig. 4.3. In both alluvial diagrams, we visualized the sequences of TCs, their inheritance relations, which can be intimacy indices (for the BCN communities), fraction of common members or inclusion measures (for the CN communities), and the evolution processes they undergo. The events (changes) that we can discern from the alluvial diagram (shown in Fig. 4.2 and Fig. 4.3) are analogous to those recognized in social group evolution [Palla et al., 2007]. They represent forming, dissolving, growing, shrinking, merging and splitting. We found in Chapter 2 that the prediction of such events is hard, since the correlation between them is nonlinear and complex. This challenge is addressed in the following section by tapping into the power of machine learning.

Event labelling

The Group Evolution Discovery (GED) method [Bródka et al., 2013] was used for tracking group evolution for historical cases——to learn the classifier and for testing cases to validate classification results. The GED method makes use of the similarity between groups in the following years as well as their sizes to label one of six event types: continuing, dissolving, merging, splitting, growing, shrinking. However, we have adapted the GED method to label only four types of events: continuing, dissolving, merging, splitting, as these are the

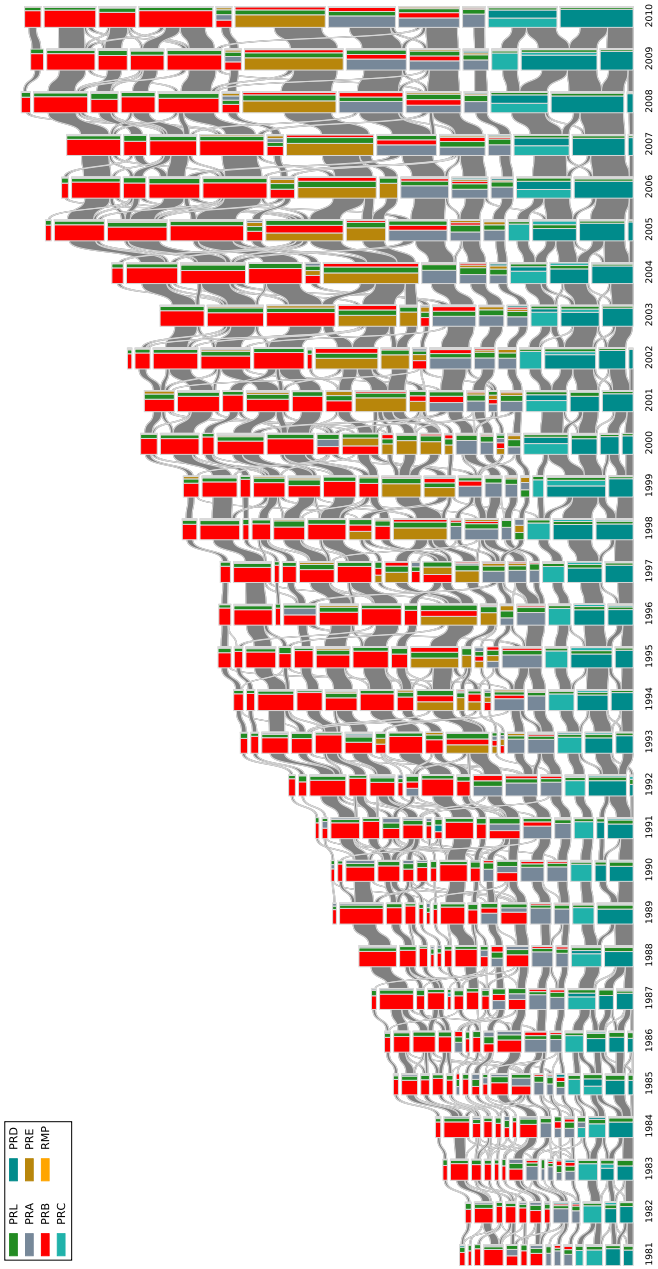


Figure 4.2: The alluvial diagram of APS papers from 1981 to 2010 for the BCN. Each block in a column represents a TC and the height of the block is proportional to the number of papers in the TC. For clarity reason only TCs comprising more than 100 papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the forward and backward intimacy indices. The colours inside a TC represent the relative contributions from different journals.

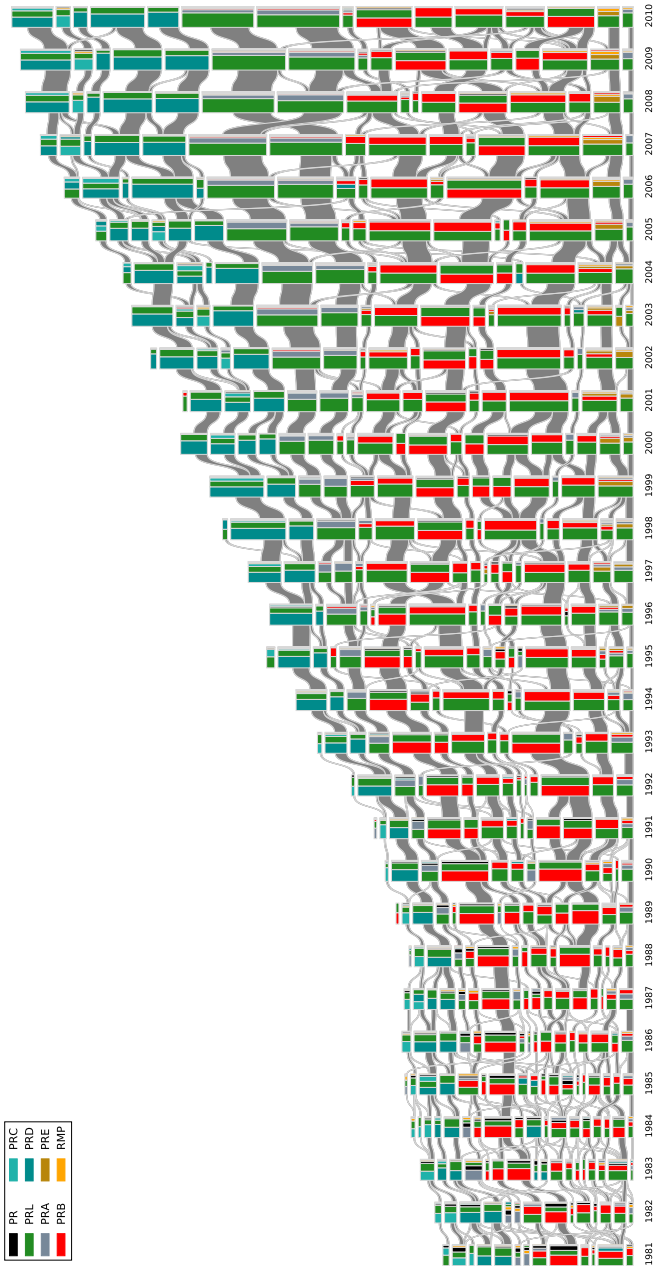


Figure 4.3: The alluvial diagram of APS papers' references from 1981 to 2010. Each block in a column represents a TC extracted from the CN. The height of the block is proportional to the number of papers in the TC. For clarity, only TCs comprising more than 1% of all papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the relative overlap percentage. The colours inside a TC represent the relative contributions from different journals.

most important to us. The other two (growing and shrinking) are covered by continuing. In general, the GED method allows us to use various metrics as a similarity measure between groups. Therefore, the intimacy indices defined in Eq. 2.1 were used for the BCN to match similar groups in the consecutive time windows. However, the original GED inclusion measures were used for the CN. It means that the similarity between two groups from two successive time windows is reflected by the inclusion measure, which is calculated for two scenarios: inclusion $I(C_n^t, C_m^{t+1})$ of a group C_n^t from time window t in another group C_m^{t+1} from time window $t+1$ (forward, Eq. 4.1), and inclusion $I(C_m^{t+1}, C_n^t)$ of this second group C_m^{t+1} from $t+1$ in the first group C_n^t from t (backward, Eq. 4.2). The inclusion measure makes use of the Social Position $SP(p)$, which is a kind of weighted PageRank. It denotes an importance of paper p being cited among all other papers [Brodka et al., 2009]. The inclusions for CN are defined as follows:

$$I(C_n^t, C_m^{t+1}) = \overbrace{\frac{\|C_n^t \cap C_m^{t+1}\|}{\|C_n^t\|}}^{\text{group quantity}} \cdot \underbrace{\frac{\sum_{p \in (C_n^t \cap C_m^{t+1})} SP(p)}{\sum_{p \in (C_n^t)} SP(p)}}_{\text{group quality}} \cdot 100\%, \tag{4.1}$$

$$I(C_m^{t+1}, C_n^t) = \overbrace{\frac{\|C_m^{t+1} \cap C_n^t\|}{\|C_m^{t+1}\|}}^{\text{group quantity}} \cdot \underbrace{\frac{\sum_{p \in (C_m^{t+1} \cap C_n^t)} SP(p)}{\sum_{p \in (C_m^{t+1})} SP(p)}}_{\text{group quality}} \cdot 100\%. \tag{4.2}$$

The GED method has two main parameters (alpha and beta), which are the levels of inclusion that groups in the consecutive years have to cross in order to be considered as matching groups. The theoretical range of values for alpha and beta is between 0% and 100%. However, the most common values are selected from the range from 30% to 70%, depending on the density of the network and node's fluctuation year by year. In

general, the selection of parameters should reflect the needs of researchers. For example, one may choose very high value (e.g. 80%) in order to preserve only very similar groups. In another case, it might be necessary to set very low value, e.g. 10% if the network is sparse or the fluctuation is high. In our study, we ran the GED method with alpha and beta parameters varying from 5% to 100%, to see how the number of events varies. Our goal was to have at least one event assigned to each TC. As the splitting and merging events involve several groups, we aimed to have on average slightly more than one event per TC. With this assumption, we selected 30% for both alpha and beta parameters in case of BCN, and 10% for alpha and beta parameters in case of CN. This values produced in total 479 events per 430 groups for BCN, and 492 events per 457 groups for CN. In both networks, the events distribution was imbalanced with the continuing event dominating over all other types, see Fig. 4.5A1 and Fig. 4.5B1. If both inclusions (CN) or both intimacy indices (BCN) are greater than the percentage thresholds alpha and beta, the method labels the event continuing. If at least one inclusion or one intimacy index exceeds one of the thresholds, the splitting and merging events is considered, the proper event is assigned depending on the number of similar groups in t and $t+1$. If both inclusions or both intimacy indexes are below the thresholds, i.e. the group has no corresponding group in the next time window, the dissolving event is assigned.

The details of four events considered in this chapter are:

- *continuing*—a research field is said to be continuing when the problems identified and solutions obtained from one year to another are of an incremental nature. It corresponds to the repeated hypothesis testing picture of the progress of science proposed by Karl Popper [Popper, 2013]. Therefore, in the CN, this would appear as a group of papers that are repeatedly together cited year by year. In the BCN, this shows up as groups of articles from successive years sharing more or less the same

reference list.

- *dissolving*—a research field is thought to disappear in the following year if the problems are solved or abandoned, and no new significant work is done after this. For the CN, we will find a group of papers that are cited up to a given year, but receives very few new citations afterwards. In the BCN, no new relevant papers are published in the field, hence, the reference chain terminates.

- *splitting*-—a research field splits in the following year, when the community of scientists who used to work on the same problems, start to form two or more subcommunities, which are more and more distant from one another. In terms of the CN, we will find a group of papers that are almost always cited together up till a given year, breaking up into smaller and disjoint groups of papers that are cited together in the next year. In the BCN, we will find the transition between new papers citing a group of older papers to new papers citing only a part of this reference group.

- *merging*-—multiple research fields are considered to have merged in the following year when the previously disjoint communities of scientists found mutual interest in each other's field so that they solve the problems in their own domain using methods from another domain. In the CN, we find previously distinct groups of papers that are cited together by papers published after a given year. In the BCN, newly published papers will form a group commonly citing several previously disjoint groups of older papers.

Correlation between overlap and inclusion measure

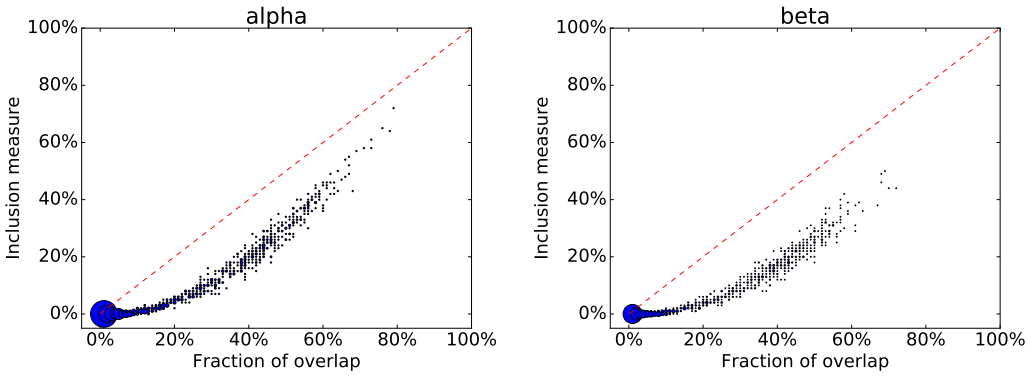
For BCN, we use the forward and backward intimacy indices to measure the closeness between TCs in consecutive time windows (years). For CN, we considered two types of

measure: (i) a simple overlap measure of two groups (the relative fraction of common members), and (ii) an overlap of two groups enriched with the information about the importance of the common members. The latter is suggested by the GED method authors, who named their similarity measure the inclusion measure. One way to evaluate the importance of TC members is to use node centrality measures to rank them within the group. In this chapter, we are using the Social Position measure [Brodka et al., 2009] (as suggested in the GED method), an idea based on the PageRank algorithm [Page et al., 1999]. [Saganowski et al., 2012] found that using a richer similarity measure allows us to track group evolution more reliably. To better understand the difference between the simple overlap measure and the inclusion measure we compared values obtained with both measures in Fig. 4.4. It turned out that the inclusion measure is on average 20% lower than the simple overlap measure, and the corresponding values, i.e. 30% for the simple overlap and 10% for the inclusion measure, produce roughly the same number of the evolution events. However, the more complex version of the similarity measure (i.e. the inclusion measure), provided slightly better initial prediction results. Therefore, we finally utilized the inclusion measure in our calculations for CN.

4.2 Prediction and feature ranking

Future events prediction

The machine learning approach to prediction requires dividing the data into two parts: the training data set and test data set. The training data is used to learn classifier, which can then label events in the test data. The labelled values are compared with the event labels and the prediction performance is calculated. More than 450 observations were used to train the classifiers. Each observation contained 77 features (preselected from the initial



(A) (B)

Figure 4.4: The scatter plots for simple overlap measure and inclusion measure for CNs between 1981 to 2010. The left panel (A) is for the alpha parameter, i.e. how the groups in t are close to groups in $t+1$. The right panel (B) is for beta parameter, i.e. how the groups in $t+1$ are close to groups in t. The sizes of circles are proportional to the number of instances. The red dash lines are $y = x$ for reference only.

100) divided into three categories: microscopic features (related to nodes in the group, e.g. node degree), mesoscopic features (related to the entire group, e.g. the group size), and macroscopic features (related to the whole network, e.g. network density). Mesoscopic features calculated for individual nodes are commonly aggregated for all nodes from the group, e.g. average node degree or betweenness in the group. See the Appendix D for the complete list of features used. To automatically select the best classification algorithm (model) as well as its hyper-parameter settings to maximize the prediction performance, the Auto-WEKA software package [Kotthoff et al., 2016] was utilized. The Auto-WEKA package is based on WEKA, a unified workbench that allow researchers easy access to state-of-the-art algorithms in machine learning [Hall et al., 2009]. However, for researchers outside of machine learning but would like to adopt the power of this technique, it is very hard to choose the appropriate algorithm and set its hyper-parameters [Thornton et al., 2013]. By Bayesian optimization, Auto-WEKA is able to automatically and simul-

taneously select a learning algorithm and set its hyper-parameters to optimize empirical performance. For each network, we ran the Auto-WEKA for 48 hours, which allowed us to validate nearly 20,000 configurations per network. The metric being maximized was the F-measure, commonly used for multi-class classification. The overall classification quality was calculated as the average F-measure for all event types, treating them as equally important.

The predicted output variable (event labels) had an imbalanced distribution. Commonly, classifiers tend to focus on the dominant event type (class), which is very well predicted, but at the expense of the minority event types. For the imbalanced BCN data set, the best performance was achieved with the Attribute Selected Classifier (with the SMO as base classifier), which performs feature selection [Platt, 1999]. The percentage of the correctly classified instances was 80.6%, while the average F-measure was only 0.50 due to classifier focusing on continuing, which was the most frequently occurring event type, see Fig. 4.5A. For this event, the F-measure value was equal to 0.89, and only 7 events out of 352 were incorrectly classified. The worst classified was the splitting event, whose F-measure was only as of 0.11. Most of the splitting events were incorrectly classified as continuing (31 out of 33 events). The second worst was merging, with F-measure 0.35. Again, the majority of the merging events were wrongly classified as continuing events: 38 out of 56. Interestingly, the splitting and merging events were never cross-classified mistakenly. For the imbalanced CN data set, the best performance was achieved with a lazy classifier, which uses locally weighted learning [Christopher et al., 1997]. The percentage of the correctly classified instances was 73.3%, while the average F-measure was only 0.53, again due to the classifier concentrating on the dominating continuing event type, see Fig. 4.5B. The F-measure value for the continuing event was only 0.83, however, as many as 50 continuing events (out of 337) were wrongly classified as dissolving. Alike to BCN,

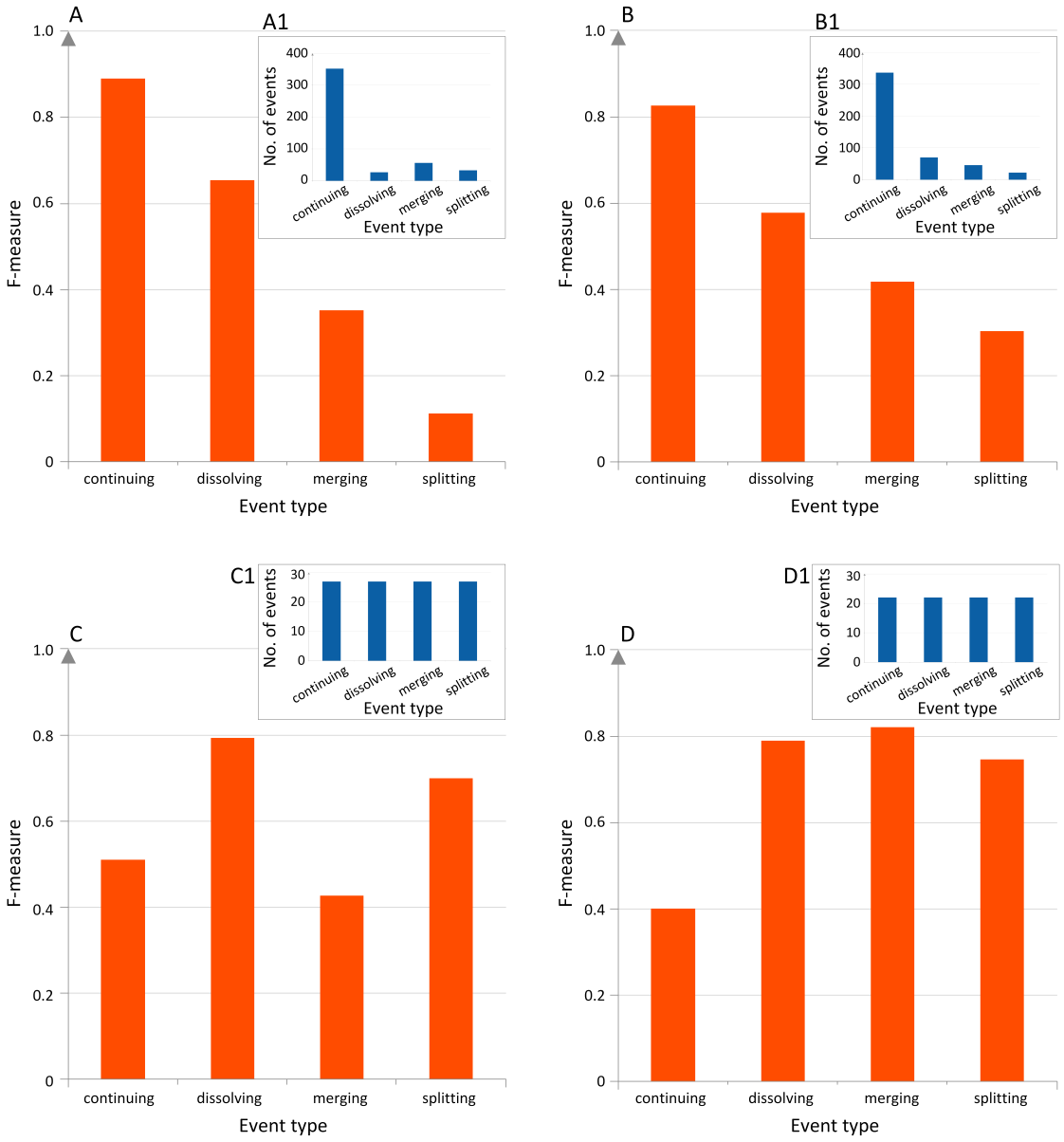


Figure 4.5: The prediction quality of classification results. The F-measure values for the imbalanced BCN (A) and CN (B) data sets, as well as the balanced BCN (C) and CN (D) data sets. The distribution of classes in the training sets are provided for each data set: A1, B1, C1, D1, respectively. For the imbalanced data sets, the classifier focused on the dominating continuing event. Balancing the data sets increased the overall prediction quality by over 20%.

many splitting and merging events were incorrectly classified as continuing: 17 out of 22 events, and 24 out of 46 events, with F-measure equal to 0.30 and 0.42, respectively.

By balancing the imbalanced training data sets (i.e. by equally sampling them), we force the classifiers to pay more attention to the features rather than to the number of occurrences of the particular majority event type. As a result of balancing data sets, the previously minor event types (dissolving, merging, and splitting) were predicted much better, but with a significant drop in performance of the continuing event classification. More importantly, by balancing the data sets we increased the overall prediction quality by over 20%. For the balanced BCN data set, the best performance was achieved by means of the boosting-based classifier AdaBoost with the Bayes Net as the base model. The percentage of the correctly classified instances was 62.0% and the average F-measure was 0.61. The biggest sources of errors were merging events, which were wrongly classified as continuing and dissolving, as well as continuing wrongly classified as splitting. The best classified event was dissolving (only 4 mistakes in 27 classifications, the overall score 0.79) followed by the splitting event (6 mistakes in 27 classifications, overall F-measure 0.70). For the balanced CN data set, the Attribute Selected Classifier (with the PART as base classifier) provided the best results—the percentage of the correctly classified instances was 69.32%, while the average F-measure was 0.69 [Frank and Witten, 1998]. The dissolving, merging, and splitting events were classified very well with the F-measure values equal to 0.79, 0.82, and 0.75 respectively. Most of the continuing events were wrongly classified as splitting (13 out of 22), which resulted in lower F-measure value 0.40.

What is interesting for us to note is that the prediction results for the CN being slightly better than for the BCN. A possible explanation is that for the CN we used a richer similarity measure containing users importance information. Thus the event tracking and therefore the ground truth could be more accurate. Overall, the prediction quality expressed

by the average F-measure was very good for the imbalanced as well as for the balanced data sets, as the baseline results obtained with the ZeroR classifier were much worse: F-measure 0.21 for both, BCN and CN, imbalanced data sets, 0.18 for the balanced BCN and 0.14 for the balanced CN. For each data set different classifier turned out to be the best, however most models were wrapped with the boosting or meta classifiers.

Predictive feature ranking

The feature selection technique is used in machine learning to find the most informative features, to avoid classifier overfitting, to eliminate (or at least to reduce) the noise in the data as well as to provide some explanations about phenomena. Rankings of the most prominent features was obtained by repeating the feature selection 1000 times using a basic evolutionary algorithm [Yang and Honavar, 1998], as proposed in [Saganowski et al., 2017]. By repeating the feature selection 1000 times, we obtain 1000 sets of selected features. Next, we calculate how many times each feature has been selected, thus, receiving the ranking of the most often selected features. For the BCN, the context-based features dominated the ranking. It referred, especially the number of papers from the Physical Review E, Physical Review B, and Physical Review A, see Fig. 4.6A. Beside the context, the network features based on degree, betweenness, size and closeness measures were most informative, which tells us that the structural properties are as important as context awareness. The context-based feature, i.e. the number of papers published in Review of Modern Physics, was the most often selected for the CN data set. It is followed by closeness- and degree-based features in the ranking, see Fig. 4.6B. For both networks macroscopic features were ranked rather low, which suggests that the general network profile is not very important, perhaps because of the smooth changes in the entire network. Surprisingly, the dynamic features, e.g. related to the average age of references (for BCN) and age of arti-

cles (for CN) did not show informative value and were ranked very low for both networks. The rankings were validated in the additional two years of data available (2010-2012). The prediction was performed twice: (i) using all features, and (ii) using the top 10 ranked features only. Selecting only the top 10 features, boosted the quality of the prediction by 11% for the CN, and by 2% for the BCN, which underlines the necessity of the feature selection process.

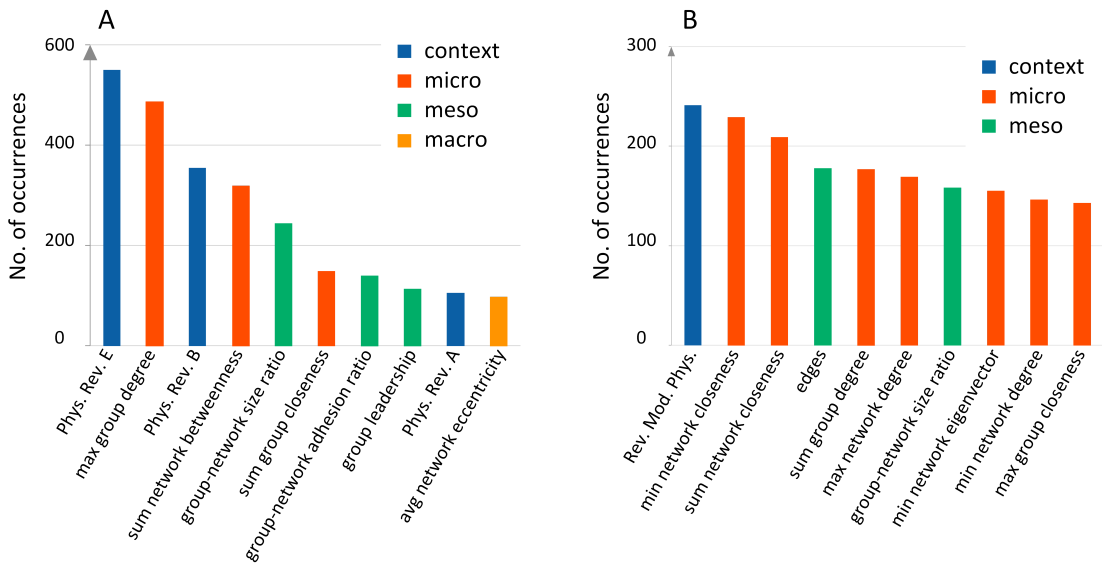
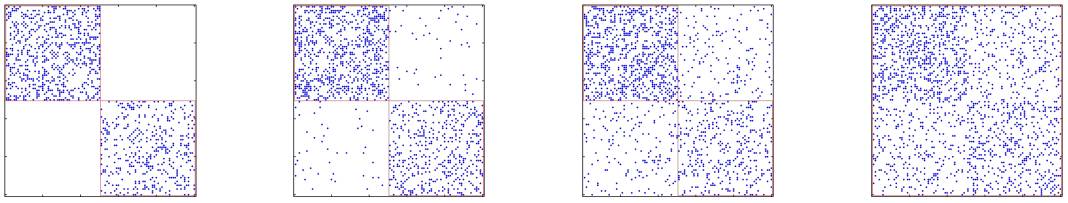


Figure 4.6: Feature ranking. The most frequently selected features in 1000 iterations for the BCN (A) and CN (B) data sets. The context-based features (number of papers published in a given journal) turned out to be the most informative, followed by the microscopic structural measures, especially closeness, degree and betweenness.

4.3 Changes to the Betweenness Distributions Associated with Merging and Splitting Events in BCN

Having the list of best predictive features, Fig. 4.6, we can analyse some of them more in-depth to look for early warning signals. Basically, we believe that scientific knowledge

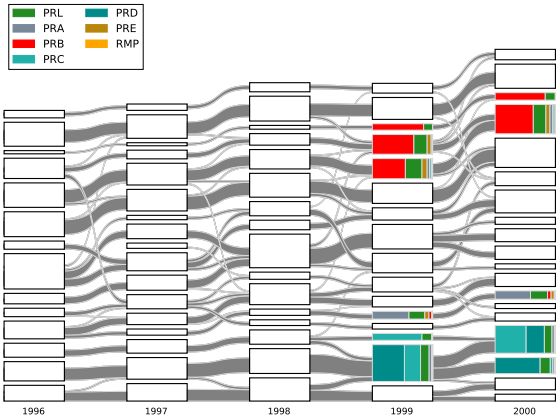
evolves slowly, and this slow evolution drives the evolution of citation patterns. Therefore, there must be specific changes in citation patterns that precede merging and splitting events. Besides the number of PRE papers in a TC, the sum_network_betweenness is also a strongly predictive feature, see Fig. 4.6A. This suggests that we should look at the betweenness of papers in the BCN more carefully. The betweenness of the node denotes what percentage of shortest paths between all pairs of nodes in the network passes a given node. Values of nodes' betweenness can be aggregated (sum, average, max, min) for all nodes in the TC, as what we list in Tab. D.1. However, in this section we only focus on the distribution of original node betweenness. Naively, when we consider the part of the BCN adjacency matrix corresponding to two TCs that ultimately merged, we expect to find few links between TCs at first. But as the number of links between TCs increase over time, the modularity-maximizing Louvain method will eventually merge the two TCs into a single TC. This is shown schematically in Fig. 4.7, where in general betweenness will increase on average with time as the two TCs merge. In reality, there are always links between



(A) (B) (C) (D)

Figure 4.7: Part of the BCN adjacency matrix for two TCs (red boxes) that ultimately merged. (A) No links between the two TCs at first. (B) Few links between the two TCs. (C) More links between the two TCs. (D) Many links between the two TCs, leading to their identification as a single merged TC (big red box) by the Louvain method.

TCs, and the numbers and strengths of these links fluctuate over time. To develop a more quantitative description of the merging events outlined in Fig. 4.2, as well as splitting and continuing events, we focus on five events going from 1999 to 2000, shown in Tab. 4.1.

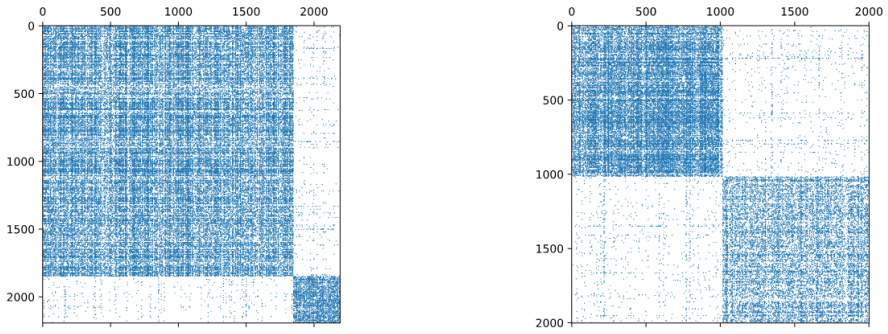


TC in 1999	event	TC in 2000
1999.01	split	2000.02, 2000.03
1999.01, 1999.02	merge	2000.03
1999.04	continue	2000.06
1999.11, 1999.12	merge	2000.15
1999.13	continue	2000.16

Table 4.1: The five evolution events from 1999 to 2000 in the BCN alluvial diagram Fig. 4.2 that we will study quantitatively. The naming convention for TC is that four digits before '.' is the year of TC, two digits after '.' is the position of the TC in the diagram, starting with 00 for the bottom TC, the one just above bottom is 01 and so on. In the left panel, we highlight the related TCs.

1999.01 + 1999.02 → 2000.03

Let us consider the part of the BCN associated with the TCs. For example, for 1999.01 and 1999.02, we can see from Fig. 4.8(A) that connections within 1999.01 and 1999.02 are very dense, but there are also some links between the two TCs. In fact, we find 164 out of 1849 papers in 1999.01 with non-zero bibliographic coupling to 144 papers in 1999.02 (344 papers). The natural question we then ask is: are the betweenness of the 164 papers in 1999.01 that are coupled to 1999.02 larger, equal, or smaller than the betweeness of the rest 1685 papers in 1999.01 not coupled to 1999.02? Alternatively, if we think of the 164 papers as randomly sampled from the 1849 papers in 1999.01, are we sampling the 164 betweenness in an unbiased fashion? To distinguish the different parts of the TC, we call all papers in 1999.01 which have coupling with papers in 1999.02 as 1999.01a, and the rest of papers as 1999.01b. For more detail analysis, we will divide 1999.01a and 1999.01b into 1999.01$a\alpha$, 1999.01$a\beta$, 1999.01$b\alpha$, 1999.01$b\beta$. 1999.01$a\alpha$ consist of 17 papers in 1999.01a



(A) (B)

Figure 4.8: (A)The adjacency matrix of the BCN associated with the TCs 1999.01 (top dense block) and 1999.02 (bottom dense block). (B)The adjacency matrix of the BCN associated with the TCs 1999.11 (top dense block) and 1999.12 (bottom dense block).

that do not have references in common with papers in 1999.01b, 1999.01$a\beta$ consist of 147 papers in 1999.01a that have references in common with papers in 1999.01b, 1999.01$b\alpha$ are 907 papers in 1999.01b that have references in common with papers in 1999.01a and 1999.01$b\beta$ represents 778 papers in 1999.01b that do not have references in common with papers in 1999.01a.

In Tab. 4.2, we show the 25th, 50th and 75th percentiles of the papers in these smaller groups, compared to those of 1849 papers in 1999.01 and 344 papers in 1999.02. As we can see, the 25th, 50th, 75th percentile betweenness in the connecting parts (1999.01a and 1999.02a) are all higher than the 25th, 50th, 75th percentile betweenness in the non-connecting parts (1999.01b and 1999.02b). More importantly, these percentile betweenness are higher than the 25th, 50th, 75th percentile betweenness of the TCs 1999.01 and 1999.02 themselves. To test how significant these quartiles are in 1999.01a, we randomly sampled 164 betweenness values from 1999.01 10^6 times, and measured the quartiles of these samples. When we draw random samples from a TC, the 25th percentile, the 50th percentile, and the 75th percentile, depends on the size of the TC. There is more variabil-

	percentile		
	25	50	75
1999.01	8.06×10^{-6}	5.73×10^{-5}	2.05×10^{-4}
1999.01a	5.90×10^{-5}	1.58×10^{-4}	4.67×10^{-4}
1999.01$a\alpha$	7.77×10^{-6}	1.95×10^{-5}	2.44×10^{-4}
1999.01$a\beta$	5.29×10^{-6}	4.96×10^{-5}	2.48×10^{-4}
1999.01b	6.22×10^{-6}	5.04×10^{-5}	1.88×10^{-4}
1999.01$b\alpha$	8.59×10^{-6}	6.00×10^{-5}	2.14×10^{-4}
1999.01$b\beta$	7.97×10^{-6}	5.32×10^{-5}	1.83×10^{-4}
1999.02	2.47×10^{-6}	5.54×10^{-5}	2.13×10^{-4}
1999.02a	3.08×10^{-5}	1.13×10^{-4}	3.17×10^{-4}
1999.02b	2.14×10^{-7}	1.44×10^{-5}	1.60×10^{-4}
1999.11	1.73×10^{-5}	9.04×10^{-5}	2.81×10^{-4}
1999.11a	6.38×10^{-5}	1.98×10^{-4}	4.61×10^{-4}
1999.11b	9.91×10^{-6}	6.17×10^{-5}	2.17×10^{-4}
1999.12	6.56×10^{-6}	4.54×10^{-5}	1.62×10^{-4}
1999.12a	2.74×10^{-5}	9.08×10^{-5}	2.33×10^{-4}
1999.12b	2.52×10^{-6}	2.69×10^{-5}	1.20×10^{-4}

Table 4.2: The 25th, 50th and 75th percentiles of the betweenness of 1849 papers in 1999.01, the 164 papers in 1999.01a, the 17 papers in 1999.01$a\alpha$, the 147 papers in 1999.01$a\beta$, the 1685 papers in 1999.01b, the 907 papers in 1999.01$b\alpha$, the 778 papers in 1999.01$b\beta$; the 344 papers in 1999.02, the 144 papers in 1999.02a, and the 200 papers in 1999.02b; the 1014 papers in 1999.11, the 299 papers in 1999.11a, the 715 papers in 1999.11b and the 988 papers in 1999.12, the 347 papers in 1999.12a, the 641 papers in 1999.12b.

ity in these quartiles in smaller samples than they are in larger samples. Therefore, in the test for statistical significance, the observed quartile has to be tested against different null model quartiles for samples of different sizes. To do this, we draw samples with a range of sizes from the same set of betweenness, and for a given quartile (25%, 50%, or 75%), fit the minimum quartile value against sample size to a cubic spline, and the maximum quartile value against sample size to a different cubic spline. With these two cubic splines, we can then check whether the observed quartile value for a sample of size n is more than or less than the null model minimum or maximum using cubic spline interpolation. From the histograms shown in Fig. 4.9(A), we see that the betweenness quartiles of $1999.01a$ are statistically larger than random samples of the same size from 1999.01, at the level of $p < 10^{-6}$, which means the papers in $1999.01a$ have significantly larger betweenness than other papers in 1999.01.

We also checked the statistical significance of the larger betweenness values of $1999.02a$, against 10^6 random samples of the same length (144) from 1999.02. From Fig. 4.9(B), we can derive that the quartiles of $1999.02a$ are only a little larger than the tails of the quartile histograms of the random samples, but their statistical significance is still at the level of $p < 10^{-6}$.

$\mathbf{1999.01 \rightarrow 2000.02 + 2000.03}$

When a TC splits into two in the next year, we expect the links between two parts a and b in the TC to have thinned out to the point that the modularity Q of the whole is lower than the modularities Q_a and Q_b of the two parts. However, in general, we would not know how to separate the TC into the two parts a and b. Fortunately, for the $1999.01 \rightarrow 2000.02 + 2000.03$ splitting event, we also know the part $1999.01a$, which merged with $1999.02a$, became 2000.03. Therefore, we might naively expect $1999.01b$ to be the part that split from 1999.01

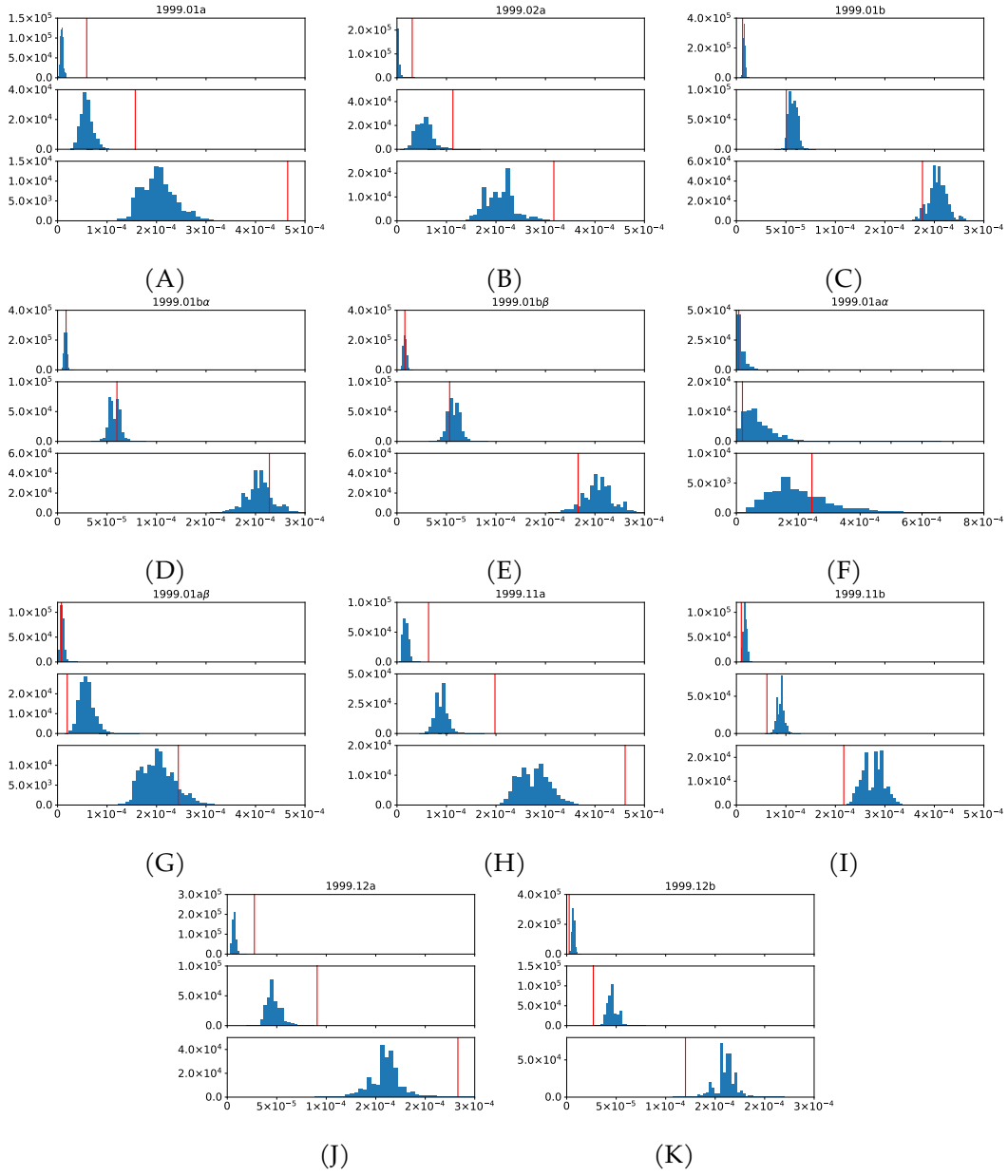


Figure 4.9: The lower quartile (top), median (middle), and top quartile (bottom) of the betweennesses in (A) 1999.01a, (B) 1999.02a, (C) 1999.01b, (D) 1999.01$b\alpha$, (E) 1999.01$b\beta$, (F) 1999.01$a\alpha$, (G) 1999.01$a\beta$, (H) 1999.11a, (I) 1999.11b, (J) 1999.12a, (K) 1999.12b shown as red vertical lines, and 10^6 random samples of the same number of betweennesses from 1999.01 (A, C, D, E, F, G) or 1999.02 (B) or 1999.11 (H, I) or 1999.12 (J, K) shown as blue histograms. All x-axes are quartile value, all y-axes are null model density.

to become 2000.02. If we test the quartiles of 1999.01b, against random samples of the same size from 1999.01, we find the histograms shown in Fig. 4.9(C). As we can see, the betweenness quartiles of 1999.01b are quite a bit lower than the typical values in 1999.01, but this difference is statistically not as significant as the quartiles of 1999.01a. Thinking about this problem more deeply, we realized that while papers in 1999.01b have no references in common with 1999.02, some of them do share common references with 1999.01a. Let us call these sets of papers 1999.01$a\alpha$ (papers do not have references in common with papers in 1999.01b), 1999.01$a\beta$ (papers have references in common with papers in 1999.01b), 1999.01$b\alpha$(papers have references in common with papers in 1999.01a), and 1999.01$b\beta$ (papers that do not have references in common with papers in 1999.01a). In Fig. 4.9(D), we learn from the histograms that the betweenness quartiles of 1999.01$b\alpha$ are indistinguishable with random samples of the same size from 1999.01. On the other hand, from the histograms in Fig. 4.9(E), we find out that while the lower betweenness quartile of 1999.01$b\beta$ is indistinguishable with the random samples of the same size from 1999.01, its median and upper quartile are both on the low sides of the random sample distributions. This suggests a split of 1999.01 to (1999.01a + 1999.01$b\alpha$) + 1999.01$b\beta$.

Just to be safe, we also checked the betweenness quartiles of 1999.01$a\alpha$ and 1999.01$a\beta$, against random samples of the same sizes from 1999.01. As we can see from Fig. 4.9(F) and (G), the lower quartiles and medians are lower than those obtained from random samples, but the upper quartiles are decidedly higher. However, the difference between 1999.01$a\alpha$ and 1999.01$a\beta$ is not as obvious as difference between 1999.01$b\alpha$ and 1999.01$b\beta$, one possible reason is the smaller sample size (17, 147 vs. 907, 778). Again, these results are consistent with the picture that the rise in betweenness in 1999.01a is driving the merging with 1999.02a, while the fall in betweenness in 1999.01$b\beta$ is driving a splitting inside 1999.01.

1999.11 + 1999.12 → 2000.15

Although a small part split off from each of 1999.11 and 1999.12, the main event associated with the two TCs was a symmetric merging. Looking again into the relevant parts of the BCN, we found 299 out of 1014 papers in 1999.11 coupled to 347 out of 988 papers in 1999.12, and we call them 1999.11*a* and 1999.12*a*, respectively. As we can see from the histograms in Fig. 4.9(H) and (J), the betweenness quartiles in 1999.11*a* and 1999.12*a* are significantly higher than one would expect from random samples of 1999.11 and 1999.12. Simultaneously, the betweenness quartiles in 1999.11*b* and 1999.12*b* are significantly lower than in random samples of 1999.11 and 1999.12 (see Fig. 4.9(I) and (K)). Therefore, what we are seeing here might be the early warning signals of merging, as well as that of the asymmetric splitting.

1999.04 → 2000.06 and 1999.13 → 2000.16

So far we have learnt that a decrease in betweenness within a TC signals a possible split, whereas an increase in betweenness of the part of the TC coupled to another TC signals a merger between the two TCs. For this story to be consistent, we must not see these signals in the continuing events 1999.04 → 2000.06 and 1999.13 → 2000.16. However, if we go through the full BCN, we find that 370 out of 389 papers in 1999.04 and 308 out of 319 papers in 1999.13 are coupled to papers outside of these TCs, which suggests the possibility of merging or splitting. However, as we can conclude from Tab. 4.3, while the lower betweenness quartiles of the coupling parts of 1999.04 and 1999.13 with other TCs may be significantly larger than those of random samples of the two TCs, the highest betweenness quartiles are never significantly larger. Therefore, at the same level of confidence that we have set for the precursors of merging between 1999.01 and 1999.02, as well as between 1999.11 and 1999.12, we have to say that there is no significant precursors for 1999.04 and

	1999.04				1999.13			
	size	percentile			size	percentile		
		25	50	75		25	50	75
1999.00	12	9.0×10^{-5}	1.1×10^{-3}	2.3×10^{-3}	1	-	-	1.8×10^{-3}
1999.01	56	1.6×10^{-4}	4.2×10^{-4}	1.0×10^{-3}	6	2.0×10^{-4}	4.9×10^{-4}	6.5×10^{-4}
1999.02	6	3.0×10^{-4}	5.1×10^{-4}	7.4×10^{-4}	2	6.0×10^{-4}	-	2.6×10^{-4}
1999.03	25	1.6×10^{-5}	4.3×10^{-4}	8.1×10^{-4}	0	-	-	-
1999.04	-	-	-	-	8	1.5×10^{-4}	4.8×10^{-4}	8.0×10^{-4}
1999.05	179	4.9×10^{-5}	1.7×10^{-4}	4.5×10^{-4}	4	2.2×10^{-4}	4.3×10^{-4}	6.5×10^{-4}
1999.06	110	8.7×10^{-5}	2.0×10^{-4}	6.2×10^{-4}	40	5.9×10^{-5}	1.6×10^{-4}	4.5×10^{-4}
1999.07	29	1.7×10^{-4}	5.6×10^{-4}	1.2×10^{-3}	44	1.4×10^{-4}	3.1×10^{-4}	5.5×10^{-4}
1999.08	63	1.1×10^{-4}	3.2×10^{-4}	8.6×10^{-4}	17	2.2×10^{-4}	5.2×10^{-4}	8.5×10^{-4}
1999.09	49	7.8×10^{-5}	2.6×10^{-4}	8.0×10^{-4}	99	8.0×10^{-5}	2.5×10^{-4}	4.8×10^{-4}
1999.10	53	1.2×10^{-4}	3.8×10^{-4}	8.2×10^{-4}	254	3.6×10^{-5}	8.8×10^{-5}	2.7×10^{-4}
1999.11	89	1.0×10^{-4}	3.2×10^{-4}	9.2×10^{-4}	71	1.4×10^{-4}	3.4×10^{-4}	5.7×10^{-4}
1999.12	53	8.7×10^{-5}	2.9×10^{-4}	9.3×10^{-4}	39	1.3×10^{-4}	2.7×10^{-4}	4.6×10^{-4}
1999.13	9	1.3×10^{-4}	4.2×10^{-4}	1.1×10^{-3}	-	-	-	-
1999.14	62	1.4×10^{-4}	4.8×10^{-4}	1.0×10^{-3}	210	4.2×10^{-5}	1.0×10^{-4}	2.7×10^{-4}
1999.15	17	1.8×10^{-4}	3.6×10^{-4}	9.7×10^{-4}	176	5.1×10^{-5}	1.3×10^{-4}	3.1×10^{-4}
b	88	2.1×10^{-6}	2.2×10^{-5}	5.8×10^{-5}	27	9.1×10^{-11}	4.3×10^{-6}	1.8×10^{-5}

Table 4.3: The distributions of betweennesses of papers in 1999.04 and 1999.13 that share common references with the other TCs in 1999 (1999.00 to 1999.15). Four columns below '1999.04' and '1999.13' denote: the first column shows how many papers have common references with the other TCs, while the second, third, and fourth column show the lower, median, and upper quartile values of betweennesses of these papers, respectively. For example, there are 25 papers in 1999.04 that share common references with papers in 1999.03, and the betweennesses of these papers have a lower quartile value of 1.6×10^{-5}, a median value of 4.3×10^{-4}, and an upper quartile value of 8.1×10^{-4}. Similarly, there are 254 papers in 1999.13 that share common references with papers in 1999.10, and the betweennesses of these papers have a lower quartile value of 3.6×10^{-5}, a median value of 8.8×10^{-5}, and an upper quartile value of 2.7×10^{-4}. The bottom row 'b' represent 1999.04b and 1999.13b respectively, which are papers in 1999.04 and 1999.13 have no references in common with papers in other TCs. A betweenness value in red means that it is larger than the maximum of the corresponding quartile distribution of 10^6 random samples, and a betweenness value in blue denotes it is smaller than the minimum of the corresponding 10^6 random samples.

1999.13 to merge with other TCs.

What about splitting then? A TC is likely to split into two if at least one of two parts has reduced betweenness. We see in Tab. 4.3 that betweenness in the coupling parts of 1999.04 and 1999.13 are not significantly lower than those of random samples. Therefore, we look at the non-coupling part, i.e. papers in 1999.04 and 1999.13 which have no references in common with papers in other TCs, but they may have common references with papers in the same TCs. We call these non-coupling parts 1999.04*b* and 1999.13*b*, respectively (the bottom row in Tab. 4.3). Only the top betweenness quartile of 1999.04*b* falls below that of random samples from 1999.04 in Tab. 4.3. Therefore, the early warning for a splitting event in the next year is not strong enough. For 1999.13*b*, on the other hand, all three betweenness quartiles fall below that of random samples from 1999.13, even after we have accounted for the small size of 1999.13*b*. This suggests that the probabiiity of a splitting event next year is high, but 1999.13 continued on to 2000.16, which thereafter continued to 2001 without merging or splitting. This might be because additional conditions, like the size of TC being large, must be satisfied before a splitting can occur.

CHAPTER 5

Conclusions

5.1 Summary

In this dissertation we set out to study the knowledge evolution at the community level quantitatively using empirical citation network data set. As the importance of scientific research continues to grow in our society, understanding the mechanisms and rules of scientific progress become more and more crucial. In addition to its enormous implication on funding agencies, industrial R&D, scientists' performance evaluations and identification of promising scientific frontiers and so on, such research can also shed light on other areas of complexity science since science itself is a complex system of researchers, ideas and papers. The discussion of the essence of science dates back to Plato and Aristotle, however for a long time such discussion is treated as a sub-field of philosophy. Recently, with the help of digitized scientific documents and complexity science theory, the research on science became quantitative and form a new field *science of science* (SciSci). Many studies have been done to quantify, understand and model the development of science on both

the microscopic level (paper) and the macroscopic level (journal or discipline). However, research on the mesoscopic level is still limited.

On mesoscopic level, the basic unit is a group of papers, which share features related to our pursuit of scientific knowledge. This group can represent the human knowledge on a specific field both in accuracy (comparing with journal or discipline level) and coverage (comparing with single paper), and can therefore provide an informative picture of knowledge evolution that is missed by microscopic and macroscopic approaches. Inspired by this idea, we built a framework for evolution at the mesoscopic level, in terms of topical clusters (TCs) and intimacy indices. Using these information we can visualize the evolution in the form of an alluvial diagram. Different types of events (birth, death, growth, decay, merging and splitting) are observed in the alluvial diagram, and these are typical events others have found in social group evolution. To make our framework more informative and easy to use by researchers from any background, we label TCs using scientific meme and also try to model the meme evolution. Some case studies revealed that merging and splitting events are closely related to scientific breakthroughs, therefore we want to predict such events to better understand knowledge evolution. Using machine learning techniques, we can achieve good prediction performance according to the F-measure while feature ranking showed that betweenness is very informative for prediction. A detail analysis on betweenness suggested that the distribution of betweenness will change a lot before merging and splitting. This finding shed light on we can model community splitting and merging. In the following sections we will summarize our contributions to this topic and also the numerous insights derived from our work.

5.2 Contributions

Our **first contribution** was to identify the elementary unit of knowledge evolution: topical cluster and propose a framework to quantify and visualize the evolution process. Because of the practice of modern science, scientists will list the references that they think are related with their own work, either providing background information or to support their conclusions from the perspectives of others. We used this feature to construct the bibliography coupling network (BCN), and found that papers focusing on the same topic have denser connections with each other than with papers on other topics. This paper group is a unit of knowledge evolution since it can represent the status of human's knowledge on a specific field, so we call it a topical cluster (TC). As a unit at the mesoscopic level, a TC contains more complete information than a single paper (microscopic level) and more specific and accurate information for the field than all papers in one journal or one discipline (macroscopic level). Therefore a TC at the mesoscopic level provides a more informative picture for knowledge evolution. More importantly, showing this evolution as an alluvial diagram also makes clear the rich interactions between TCs that are significantly different from the behavior of a single paper or a whole journal or discipline. The evolution of multiple fields under one discipline is like a circulatory system: on one hand, each stream has its own focus and function; on the other hand, different streams also interact with each other with different intensities, with some interactions being crucial for their functions. Comparing with our mesoscopic picture, the macroscopic picture is oversimplified since it treat the whole journal or discipline as a homogeneous system, whereas a microscopic picture is too localized and neglect the high similarity between papers in the same field. Furthermore, we validated our partition of TCs against a null model of PACS number, and show that individual TCs are significantly more homogeneous than the null model with respect to the PACS numbers. By measuring the similarity between references of TCs, we

proposed the forward intimacy index and backward intimacy index, which can be used to quantify the inheritance relation between TCs in successive years. TCs and intimacy indices are the essential components of our framework of knowledge evolution and this framework can be visualized in the form of an alluvial diagram. In our alluvial diagram of Physical Review journals, many event types were found: birth, death, growth, decay, split, and merge, which is analogous to events in social group evolution. Using our alluvial diagram, two big events in the history of Physical Review journals: split of Physical Review into Physical Review A, B, C and D in 1970, split of Physical Review A into Physical Review A and Physical Review E were captured. We then identified the streams related to Bose-Einstein condensation and showed the correlation between scientific breakthroughs and merging, splitting events.

Apart from informative visualization, many quantitative analyses were also carried out in our framework. We proposed a linear recombination model for the sizes of TCs and found the empirical data fit quite well with our model. The correlation between breakthroughs and merging, splitting events in the Bose-Einstein condensation related branches suggests that these events may be the signal of scientific breakthroughs, therefore making the prediction of merging and splitting events very useful. The Spearman's correlation rank coefficient between merging events and inter-TC connections is very high, while the splitting event does not show such clear correlations, which means that merging is easy to predict while the splitting is much harder. Besides the question of the predictability of scientific breakthroughs, we are also interested in the question of how impactful these breakthroughs are. To answer this question, we checked the reference distribution of the TC before and after the breakthrough and the results show that before the breakthrough, the reference distribution is very close to the general distribution curve, whereas after the breakthrough the distribution will become significantly different from the general distri-

bution curve. Such localized deviation is the impact of the scientific breakthrough and can be used as indicator for the breakthrough measurement.

The analyses on Bose-Einstein condensation are enlightening, but to pick the related TCs we have to read a number of titles and abstracts manually to judge whether the TC is related or not. This manual selection requires us to know something about the background of field and the whole process is very slow and subjective. These limitations represent an important shortcoming of our framework: although we know the papers in each TC, we do not know the research direction unless we read the titles and abstracts. To overcome these limitations and make our framework more comprehensible, we introduce scientific memes into our framework and analyse the correlations between memes and TCs. This is our **second contribution**. Scientific memes are n-grams that can reproduce themselves through citations, and can be extracted automatically [Kuhn et al., 2014]. Because each TC has a well-defined topic, naturally they are strongly correlated with some memes and weakly correlated with other memes. We can exploit this feature by calculating the Jaccard index, mutual information and normalized mutual information between memes and TCs. Our results show that the top 5 memes tell us a lot about the TC's scientific content, and can therefore be used as labels for TCs. By choosing appropriate memes and thresholds, we can pick relevant TCs automatically in much shorter time than doing it manually. For example, using the meme 'laser' and a threshold of 0.01 for NMI, we picked up TCs from 1981 to 2010 (see Fig. 3.4), which are almost the same as Bose-Einstein condensation related TCs we picked up manually (see Fig. 2.7). Beside labelling, memes are worth studying on their own because they can provide information of knowledge evolution at the linguistic level. As we mentioned in Chapter 1, science is a network of ideas, researchers and papers. Our framework is mainly about the network of papers, but if we introduce the memes into our analyses, the evolution picture will be more complete. As a first step to-

wards studying the interaction of scientific ideas, we analyzed meme pairs that occur more frequently than by chance. We proposed a probabilistic model to measure the correlation between two memes and found that several pairs that are increasing more rapidly in the field of quantum optics, which is consistent with our knowledge of the history of physics. However, unlike for papers, we did not find clear community structures in the meme network because of the redundancy of language. If we define the distance between memes in term of their distributions among TCs, hierarchical clustering will give a nice partition of memes, which is not really a community structure but will be useful for further analysis. To quantify the correlation between meme evolution and TC evolution, we proposed a linear recombination model for the meme population. Our results thus far suggest that merging does not have a significant boost effect for top 1000 frequently used memes.

The Bose-Einstein condensation case study, on the other hand, suggests the strong correlation between scientific breakthroughs and merging/splitting events. This relation inspired us to do correlation analyses about the merging/splitting events and several network structure indicators in Chapter 2. The results showed the merging event is highly correlated with inter-community connections while the splitting event is more complex. Although these analyses gave us some predictive power, the results have certain limitations. First, these results are not 'real' predictions, but correlation analyses. We only know that inter-community connections are correlated with merging events, but we cannot tell if two TCs will merge in the next year given the information we have on this year. Second, our analysis is restricted to merging and splitting events, while other event types like birth, death, growing, shrinking were not discussed. Our **third contribution** is therefore the training of a machine learning model to predict future changes of TCs. By using GED methods, we are able to label the evolution events automatically and these events, continuing, dissolving, splitting and merging, constitutes the event labels for the machine learning

model. The Auto-WEKA software package was used to perform the prediction task and 56 features were considered in our model. The prediction performances were evaluated using the F-measure. For both imbalanced and balanced training sets, the F-measures are significantly higher than random guesses. We then used the feature selection technique to find the most informative features for prediction. Features based on the degree, betweenness, size and closeness measure are most informative among the structure-based features. Following this, we conducted a detailed analysis on the betweenness distribution and found that higher betweenness is associated with merging TCs, no such enhancement is found in continuing TCs, while a significant decrease in average betweenness was found in splitting TCs.

5.3 Discussion and Outlook

The study done in this dissertation opens up numerous productive perspectives. The citation network we studied in this dissertation is restricted to APS journals, which are prestigious journals and likely reflect the frontiers of physics research, but are incomplete in two ways. First, many important physics papers are published outside of the APS journals, like Nature, Science and Applied Physics Letters, etc. These papers are clearly important components of physics research. The second limitation is that even for the APS papers, we only have citations pointing back to APS papers. This means that the BCN between APS papers is not complete. For example, if two APS papers cite the same Nature paper, but no APS paper in common, this information will not be captured in the data set. The distortion due to the network being incomplete is hard to estimate, because we are dealing not only with missing nodes, but also with missing edges. One possible way to overcome this limitation is using a more complete data set which include all papers of interest to physicists, as suggested in [Sinatra et al., 2015]. The third limitation is the timescale, an important

factor for understanding evolving networks [Darst et al., 2016]. In this thesis we always use one year as one time step, but other timescales are also possible. If we use half a year as one time step, we will have 20 steps for one decade, meanwhile if we use two years as one time step, we will have only 5 steps for one decade. In general, if we set one time step too long, we will lose a lot of detail; if we set one time step too short, the fluctuation may overwhelm the real signal. Recent research suggested that the timescale may also affect community detection [Medo et al., 2018]. To balance these effects, we choose one year as one time step, while the optimal choice of timescale remains an open question.

In our framework, TCs are disjoint groups of papers, which means that each paper can only belong to one TC. It is a nice model to begin with because of its simplicity, but it does not mean that this model is the optimal one. As physics research becomes more and more interdisciplinary, different fields might overlap with one another, and one paper may belong to multiple fields. When we use the Infomap method to detect community structure in the BCN, our results show that the overlapping communities have a shorter description length compared to non-overlapping communities, which suggests that overlapping communities can give a more accurate picture of knowledge distribution. However, if we allow a paper to belong to multiple TCs, we also have to redefine all evolutionary concepts in our framework. Interaction between TCs can also become ambiguous: if TC A contains some members of TC B, what does it mean for A to interact with B? Because of this difficulty, we adopted the disjoint TC in description in this dissertation. If we can incorporate the overlapping communities into our framework, it may reveal even richer information of knowledge evolution than our current work.

In Chapter 2 we have observed the continuous distribution of mixing degree (including forward and backward), it seems that almost-no-mixing and strong-mixing are quite rare, and the weak mixing is the dominant type. This conclusion maybe appealing at first

glance, however we must consider it with caution. The distribution of mixing strength is closely related with the evolution mechanism, and a misinterpretation of Fig. 2.8 (e), (f) will prevent us from understanding the evolution mechanism properly. There is at least one factor known that can introduce fictional mixing: false group identification. If node 1 in fact belongs to group α but was wrongly put into group β, and group α becomes group a, group β becomes group b next year. Such a false group identification will increase the intimacy indices between β and a and decrease the intimacy indices between α and a, thereby making groups α, β and a more mixing in terms of the mixing degree. To understand the phenomenon of weak mixing, such effects should be addressed first. If the weak mixing persist after such effects are addressed, it will be very useful for understanding splitting/merging. There are at least two potential sources of weak mixing, the first coming from interdisciplinary papers. The references of such papers show high diversity in the context of TCs. No matter where we put such papers, they will introduce some mixing. Another source is some papers citing papers in other fields to positions themselves in the big picture. If the effects of these factors can be understood, we will be able to determine the threshold between weak mixing and significant mixing. It will not only help with prediction of splitting/merging, but also further our understanding of the evolution mechanism.

Even though we are very excited to find Kuhnian processes correlated with changes in the citation curve in the BEC case study in Chapter 2, it is too early to conclude that Kuhnian processes will lead to scientific breakthroughs. The main contribution of the BEC case study is that it demonstrate local effects are more useful than global effects to trace the impact of Kuhnian processes. More systematic investigation is necessary to understand the connection between Kuhnian processes and scientific breakthroughs, but when we tried to generalize the method used in the BEC case study, we faced many difficulties. The most

crucial one is the scope of 'local effect'. In BEC case study, we limit ourselves to only the BEC-related branches, i.e., the branches highlighted in Fig. 2.7. However, for a general and systematic analysis, we cannot do the same thing. If we include the whole alluvial diagram into the analysis, we will end up with no correlation as shown in Fig. 2.15, which is not true if we analyse at the local scale. However, there is no clear boundary between local scale and global scale. This is the reason we only have the case study, and not a systematic analysis between Kuhnian processes and scientific breakthroughs. We hope future mesoscopic studies will help us define a proper index that can measure the local influence of Kuhnian processes.

In Chapter 3 we mentioned that our meme population model includes all top 1000 memes whether or not they are really correlated with the TCs' research interests. This introduces heavy noise making the real signal hard to detect. One possible way to solve this problem is to use our meme labelling technique to pick the top memes automatically and only using the key memes for analysis. However, when we use such a technique, we need to check if we have done some 'overfitting' to get false positives. If we filter our data set to test the model, we are in danger of circular reasoning. These effects need therefore to be considered carefully when we try to study the correlation between meme evolution and TC evolution.

To be able to identify changes in TCs in Chapter 4, we needed to define time windows used for network creation and community detection. The natural choice for bibliographical data was the usage of single years, since the publishing process may last many months. Obviously, another granularity may be considered like multiple years, e.g. 2 or 5 years. In our approach, i.e. both for BCN and CN, every citation has the same importance. However, there are some other concepts like fractional counting of citations [Leydesdorff and Opthof, 2010]. It assumes that the impact of each citation is proportionate to the number

of references in the citing document. Additionally, it can be differentiated depending on e.g. the quality of the journal. We decided to analyse more in-depth only on one feature describing structural profile of TCs, namely node betweenness. It was primary caused by the limited amount of resources and complexity of analyses. The entire process required much human assistance and could not have been easily automated. In our experiments, we utilized the raw, imbalanced or artificially flattened—balanced data sets. However, depending on the study purpose, we can bias some classes we are more interested in e.g. split. It can be achieved either by means of appropriate balancing—sampling for the learning set, or reformulating the problem into the binary question—is split expected (true) or not (false). As of now, the betweenness analysis is still limited to several case studies, in future a more rigorous framework will be desired. The idea of analysing science by discovery of knowledge changes is general and can be applied to all bibliographical data containing citations. We focus solely on APS journals, however, also papers indexed by PubMed, Web of Science or Google Scholar may be studied. In Chapter 4 we try to do short-term prediction using information only from the previous year. However, longer-memory effect may exist in the knowledge evolution, that is, the event that happens in year $t+1$ not only depends on what happened in year t, but also on the events in years $t-1$, $t-2$ and so on. For future prediction studies, these factors should be tested and included.

Finally, from a more theoretical perspective, although we have a framework to quantify knowledge evolution, we do not have any model. More precisely, even though we observe a plethora of merging, splitting, birth, death, growth and decay events, the mechanisms underlying these events are still unknown. For example, when multiple TCs merge, how do their boundaries disappear? Conversely, when a TC splits, how do boundaries appear inside the TC until they divide the old TC into several new ones? One possible way to find out is using sliding windows to trace the progress of an event. Although this question

sounds very theoretical, understanding it has huge implications on knowledge evolution. Such models can help us identify the driving forces behind the evolution event and would allow us to do predictions beyond the machine learning model.

Bibliography

[acl,] ACL Anthology Network (All about NLP), see clair.eecs.umich.edu/aan/index.php.

[APS,] APS Data Sets for Research, see http://journals.aps.org/datasets.

[usp,] The NEBR U.S. Patent Citations Data File: Lessons, Insights, and Methodological Tools, see www.nber.org/patents/.

[web,] Web of Science, see www.webofknowledge.com.

[Ahneman et al., 2018] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D., and Doyle, A. G. (2018). Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190.

[Albert and Barabasi, 2002] Albert, R. and Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

[Anderson et al., 1995] Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E., and Cornell, E. A. (1995). Observation of Bose-Einstein Condensation in a Dilute Atomic Vapor. *Science*, 269(5221):198–201.

[Banavar et al., 1999] Banavar, J. R., Maritan, A., and Rinaldo, A. (1999). Size and form in efficient transportation networks. *Nature*, 399(6732):130–132.

[Bettencourt and Kaur, 2011] Bettencourt, L. M. A. and Kaur, J. (2011). Evolution and structure of sustainability science. *Proceedings of the National Academy of Sciences*, 108(49):19540–19545.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Bollen et al., 2009] Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., and Balakireva, L. (2009). Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE*, 4(3):e4803.

[Bonacich, 1972] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120.

[Borgatti et al., 2009] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, 323(5916):892–895.

[Bouwmeester et al., 1997] Bouwmeester, D., Pan, J.-W., Mattle, K., Eibl, M., Weinfurter, H., and Zeilinger, A. (1997). Experimental quantum teleportation. *Nature*, 390:575–579.

[Boyack et al., 2005] Boyack, K. W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3):351–374.

[Brodka et al., 2009] Brodka, P., Musial, K., and Kazienko, P. (2009). A Performance of Centrality Calculation in Social Networks. In *2009 International Conference on Computational Aspects of Social Networks*, pages 24–31. IEEE.

[Bródka et al., 2013] Bródka, P., Saganowski, S., and Kazienko, P. (2013). GED: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14.

[Bródka et al., 2014] Bródka, P., Saganowski, S., and Kazienko, P. (2014). Community Evolution. In *Encyclopedia of Social Network Analysis and Mining*, pages 220–232. Springer New York, New York, NY.

[Carrasquilla and Melko, 2017] Carrasquilla, J. and Melko, R. G. (2017). Machine learning phases of matter. *Nature Physics*, 13(5):431–434.

[Chavalarias and Cointet, 2013] Chavalarias, D. and Cointet, J. P. (2013). Phylomemetic Patterns in Science Evolution-The Rise and Fall of Scientific Fields. *PLoS ONE*, 8(2):e54847.

[Chen and Redner, 2010] Chen, P. and Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, 4(3):278–290.

[Christopher et al., 1997] Christopher, A., Andrew, M., and Stefan, S. (1997). Locally Weighted Learning. *Artif Intell Rev*, 11:11–73.

[Cimini et al., 2014] Cimini, G., Gabrielli, A., and Sylos Labini, F. (2014). The Scientific Competitiveness of Nations. *PLoS ONE*, 9(12):e113470.

[Cimini et al., 2016] Cimini, G., Zaccaria, A., and Gabrielli, A. (2016). Investigating the interplay between fundamentals of national research systems: Performance, investments and international collaborations. *Journal of Informetrics*, 10(1):200–211.

[Cohen, 1976] Cohen, I. B. (1976). The Eighteenth-Century Origins of the Concept of Scientific Revolution. *Journal of the History of Ideas*, 37(2):257.

[Darst et al., 2016] Darst, R. K., Granell, C., Arenas, A., Gómez, S., Saramäki, J., and Fortunato, S. (2016). Detection of timescales in evolving complex systems. *Nature Publishing Group*, (November):1–8.

[Davis et al., 1995] Davis, K. B., Mewes, M. O., Andrews, M. R., van Druten, N. J., Durfee, D. S., Kurn, D. M., and Ketterle, W. (1995). Bose-Einstein Condensation in a Gas of Sodium Atoms. *Physical Review Letters*, 75(22):3969–3973.

[Dawkins, 1976] Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

[de Solla Price, 1963] de Solla Price, D. J. (1963). *Little science, big science*. Columbia University Press New York.

[Deville et al., 2014] Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., and Barabási, A. L. (2014). Career on the move: Geography, stratification, and scientific impact. *Scientific Reports*, 4:1–7.

[Dorogovtsev et al., 2007] Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2007). Critical phenomena in complex networks. 80(December).

[Doyle et al., 2005] Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R., and Willinger, W. (2005). The "robust yet fragile" nature of the Internet. *Proceedings of the National Academy of Sciences*, 102(41):14497–14502.

[Eom and Fortunato, 2011] Eom, Y. H. and Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE*, 6(9):1–7.

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[Fortunato and Barthelemy, 2007] Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

[Fortunato et al., 2018] Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L. (2018). Science of science. *Science*, 359(6379):eaao0185.

[Fortunato and Hric, 2016] Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

[Frank and Witten, 1998] Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 144—-151. Morgan Kaufmann Publishers Inc.

[Freeman, 1978] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[Gleeson et al., 2014] Gleeson, J. P., Ward, J. A., O'Sullivan, K. P., and Lee, W. T. (2014). Competition-induced criticality in a model of meme popularity. *Physical Review Letters*, 112(4):1–5.

[Goh et al., 2007] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

[Good et al., 2010] Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106.

[Grönlund and Holme, 2004] Grönlund, A. and Holme, P. (2004). Networking the seceder model: Group formation in social and economic systems. *Physical Review E*, 70(3):036108.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[Harary, 1969] Harary, F. (1969). *Graph theory*. Addison-Wesley, Reading, MA.

[Hizanidis et al., 2016] Hizanidis, J., Kouvaris, N. E., Gorka, Z. L., Díaz-Guilera, A., and Antonopoulos, C. G. (2016). Chimera-like States in Modular Neural Networks. *Scientific Reports*, 6(January):1–11.

[Huang et al., 2003] Huang, M. H., Chiang, L. Y., and Chen, D. Z. (2003). Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies. *Scientometrics*, 58(3):489–506.

[Iñiguez et al., 2009] Iñiguez, G., Kertész, J., Kaski, K. K., and Barrio, R. A. (2009). Opinion and community formation in coevolving networks. *Physical Review E*, 80(6):066119.

[Jeong et al., 2001] Jeong, H., Neda, Z., and Barabasi, A. L. (2001). Measuring preferential attachment for evolving networks. *Europhysics Letters*, 76(5):753–759.

[Jia et al., 2017] Jia, T., Wang, D., and Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1(march):0078.

[Jost, 2006] Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

[Karrer and Newman, 2011] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(1):1–10.

[Kash et al., 1999] Kash, M. M., Sautenkov, V. A., Zibrov, A. S., Hollberg, L., Welch, G. R., Lukin, M. D., Rostovtsev, Y., Fry, E. S., and Scully, M. O. (1999). Ultraslow Group Velocity and Enhanced Nonlinear Optical Effects in a Coherently Driven Hot Atomic Gas. *Physical Review Letters*, 82(26):5229–5232.

[Ke et al., 2015] Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, 2015(35):40.

[Kessler, 1963] Kessler, M. (1963). An experimental study of bibliographic coupling between technical papers (Corresp.). *IEEE Transactions on Information Theory*, 9(1):49–51.

[Klosik et al., 2014] Klosik, D. F., Bornholdt, S., and Hütt, M.-T. (2014). Motif-based success scores in coauthorship networks are highly sensitive to author name disambiguation. *Physical Review E*, 90(3):032811.

[Kotthoff et al., 2016] Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2016). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 17:1–5.

[Kuhn et al., 2014] Kuhn, T., Perc, M., and Helbing, D. (2014). Inheritance patterns in citation networks reveal scientific memes. *Physical Review X*, 4(4):1–9.

[Kuhn, 1962] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*, volume 31.

[Larivière et al., 2013] Larivière, V., Ni, C., Gingras, Y., Cronin, B., and Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213.

[Leskovec et al., 2009] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 497, New York, New York, USA. ACM Press.

[Leydesdorff and Opthof, 2010] Leydesdorff, L. and Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, 61(11):2365–2369.

[Leydesdorff and Rafols, 2009] Leydesdorff, L. and Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2):348–362.

[İlhan and Öğüdücü, 2016] İlhan, N. and Öğüdücü, u. G. (2016). Feature identification for predicting community evolution in dynamic social networks. *Engineering Applications of Artificial Intelligence*, 55:202–218.

[Liu et al., 2017] Liu, W., Nanetti, A., and Cheong, S. A. (2017). Knowledge evolution in physics research: An analysis of bibliographic coupling networks. *PLoS ONE*, 12(9):1–19.

[Martin et al., 2013] Martin, T., Ball, B., Karrer, B., and Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E*, 88(1):012814.

[Medo et al., 2018] Medo, M., Zeng, A., Zhang, Y.-c., and Mariani, M. S. (2018). Optimal timescale of community detection in growing networks. pages 1–20.

[Mossa et al., 2005] Mossa, S., Turtschi, A., Amaral, L. a. N., Guimera, R., Guimerà, R., Mossa, S., Turtschi, A., Amaral, L. a. N., Guimera, R., Guimerà, R., Mossa, S., Turtschi, A., Amaral, L. a. N., Guimera, R., Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. a. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–7799.

[Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

[Newman and Girvan, 2004] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69(2 2):1–15.

[Newman et al., 2001] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical report.

[Palla et al., 2007] Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

[Palla et al., 2005] Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

[Pan et al., 2018] Pan, R. K., Petersen, A. M., Pammolli, F., and Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3):656–678.

[Pan et al., 2012] Pan, R. K., Sinha, S., Kaski, K., and Saramäki, J. (2012). The evolution of interdisciplinarity in physics research. *Scientific Reports*, 2:1–8.

[Parolo et al., 2015] Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., and Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, 9(4):734–745.

[Pavlopoulou et al., 2017] Pavlopoulou, M. E. G., Tzortzis, G., Vogiatzis, D., and Paliouras, G. (2017). Predicting the evolution of communities in social networks using structural and temporal features. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 40–45. IEEE.

[Peixoto, 2014] Peixoto, T. P. (2014). Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047.

[Petersen et al., 2014] Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E., and Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, 111(43):15316–15321.

[Platt, 1999] Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185—-208. IEEE.

[Pons and Latapy, 2005] Pons, P. and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. volume 10, pages 284–293.

[Popper, 2013] Popper, K. R. (2013). *All life is problem solving*. Routledge.

[Radicchi and Castellano, 2011] Radicchi, F. and Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E*, 83(4):046116.

[Radicchi et al., 2008] Radicchi, F., Fortunato, S., and Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272.

[Redner, 2004] Redner, S. (2004). Citation Statistics From More Than a Century of Physical Review. pages 1–12.

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

[Rosvall and Bergstrom, 2010] Rosvall, M. and Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE*, 5(1).

[Saganowski et al., 2012] Saganowski, S., Bródka, P., and Kazienko, P. (2012). Influence of the User Importance Measure on the Group Evolution Discovery. *Foundations of Computing and Decision Sciences*, 37(4).

[Saganowski et al., 2017] Saganowski, S., Bródka, P., Koziarski, M., and Kazienko, P. (2017). Analysis of group evolution prediction in complex networks. *arXiv preprint*.

[Saganowski et al., 2015] Saganowski, S., Gliwa, B., Bródka, P., Zygmunt, A., Kazienko, P., and Koźlak, J. (2015). Predicting Community Evolution in Social Networks. *Entropy*, 17(5):3053–3096.

[Serrano et al., 2009] Serrano, M. A., Boguna, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.

[Sienkiewicz and Altmann, 2016] Sienkiewicz, J. and Altmann, E. G. (2016). Impact of lexical and sentiment factors on the popularity of scientific papers. *Royal Society Open Science*, 3(6):160140.

[Sinatra et al., 2015] Sinatra, R., Deville, P., Szell, M., Wang, D., and Barabási, A.-L. (2015). A century of physics. *Nature Physics*, 11(10):791–796.

[Sinatra et al., 2016] Sinatra, R., Wang, D., Deville, P., Song, C., and Barabasi, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239–aaf5239.

[Skupin, 2004] Skupin, A. (2004). The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5274–5278.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269.

[Sobolevsky et al., 2014] Sobolevsky, S., Campari, R., Belyi, A., and Ratti, C. (2014). General optimization technique for high-quality community detection in complex networks. *Physical Review E*, 90(1):012811.

[Tajeuna et al., 2015] Tajeuna, E. G., Bouguessa, M., and Wang, S. (2015). Tracking the evolution of community structures in time-evolving social networks. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

[Thornton et al., 2013] Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855.

[Uzzi et al., 2013] Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, 342(6157):468–472.

[Wang et al., 2013] Wang, D., Song, C., and Barabasi, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

[Weng et al., 2012] Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, 2:1–9.

[White and Harary, 2001] White, D. R. and Harary, F. (2001). The Cohesiveness of Blocks In Social Networks: Node Connectivity and Conditional Density. *Sociological Methodology*, 31(1):305–359.

[Wootton, 2015] Wootton, D. (2015). *The invention of science: a new history of the scientific revolution*. Penguin UK.

[Yan and Ding, 2012] Yan, E. and Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, 63(7):1313–1326.

[Yang and Honavar, 1998] Yang, J. and Honavar, V. (1998). Feature Subset Selection Using a Genetic Algorithm. In *Feature Extraction, Construction and Selection*, pages 117–136. Springer US, Boston, MA.

[Zeng et al., 2017] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., and Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714-715:1–73.

APPENDIX A

Proof of I^f_{mn} and I^b_{mn} are in range $[0,1]$

We define the *forward intimacy index* I^f_{mn} and *backward intimacy index* I^b_{mn} using Eq. 2.1 in Chapter 2 as following:

$$I^f_{mn} = \sum_i \frac{N\left(R_i, \mathscr{R}^{t+1}_n\right)}{N\left(R_i, \mathscr{R}^{t+1}\right)} \frac{N\left(R_i, \mathscr{R}^t_m\right)}{L\left(\mathscr{R}^t_m\right)};$$
$$I^b_{mn} = \sum_i \frac{N\left(R_i, \mathscr{R}^t_m\right)}{N\left(R_i, \mathscr{R}^t\right)} \frac{N\left(R_i, \mathscr{R}^{t+1}_n\right)}{L\left(\mathscr{R}^{t+1}_n\right)}.$$

And we denote the references cited by papers in C^t_m and C^{t+1}_n as $\mathscr{R}^t_m = \mathscr{R}(C^t_m) = [R_{m1}, ..., R_{mp}]$ and $\mathscr{R}^{t+1}_n = \mathscr{R}(C^{t+1}_n) = [R_{n1}, ..., R_{nq}]$; and $\mathscr{R}^t = \{\mathscr{R}^t_1, ..., \mathscr{R}^t_m, ...\}$ is the collection of all references cited in year t. $N(element, list)$ is the number of times $element$ occurs in $list$, and $L(list)$ is the length of $list$.

According to our notation $\mathscr{R}^t = \{\mathscr{R}^t_1, ..., \mathscr{R}^t_m, ...\}$, this is obvious that

$$\begin{aligned} N\left(R_i, \mathscr{R}^t\right) &= \sum_m N\left(R_i, \mathscr{R}^t_m\right), \quad \forall R_i \in \mathscr{R}^t; \\ N\left(R_i, \mathscr{R}^{t+1}\right) &= \sum_n N\left(R_i, \mathscr{R}^{t+1}_n\right), \quad \forall R_i \in \mathscr{R}^{t+1}. \end{aligned} \tag{A.1}$$

Additionally, $N\left(R_i, \mathscr{R}^t_m\right)$ and $N\left(R_i, \mathscr{R}^{t+1}_n\right)$ are nonnegative integers. Therefore it is easy to get

$$\begin{aligned} 0 <= \frac{N\left(R_i, \mathscr{R}^t_m\right)}{N\left(R_i, \mathscr{R}^t\right)} <= 1, \quad \forall R_i \in \mathscr{R}^t; \\ 0 <= \frac{N\left(R_i, \mathscr{R}^{t+1}_n\right)}{N\left(R_i, \mathscr{R}^{t+1}\right)} <= 1, \quad \forall R_i \in \mathscr{R}^{t+1}. \end{aligned} \tag{A.2}$$

Furthermore,

$$\begin{aligned} L\left(\mathscr{R}_m^t\right) &= \sum_i N\left(R_i, \mathscr{R}_m^t\right), \\ L\left(\mathscr{R}_n^{t+1}\right) &= \sum_i N\left(R_i, \mathscr{R}_n^{t+1}\right). \end{aligned} \tag{A.3}$$

Combining Eq. A.2 and Eq. A.3, we can obtain

$$\begin{aligned} 0 = \sum_i 0 \frac{N\left(R_i, \mathscr{R}_m^t\right)}{L\left(\mathscr{R}_m^t\right)} <= I_{mn}^f = \sum_i \frac{N\left(R_i, \mathscr{R}_n^{t+1}\right)}{N\left(R_i, \mathscr{R}^{t+1}\right)} \frac{N\left(R_i, \mathscr{R}_m^t\right)}{L\left(\mathscr{R}_m^t\right)} <= \sum_i 1 \frac{N\left(R_i, \mathscr{R}_m^t\right)}{L\left(\mathscr{R}_m^t\right)} = \frac{\sum_i N\left(R_i, \mathscr{R}_m^t\right)}{L\left(\mathscr{R}_m^t\right)} = 1, \\ 0 = \sum_i 0 \frac{N\left(R_i, \mathscr{R}_n^{t+1}\right)}{L\left(\mathscr{R}_n^{t+1}\right)} <= I_{mn}^b = \sum_i \frac{N\left(R_i, \mathscr{R}_m^t\right)}{N\left(R_i, \mathscr{R}^t\right)} \frac{N\left(R_i, \mathscr{R}_n^{t+1}\right)}{L\left(\mathscr{R}_n^{t+1}\right)} <= \sum_i 1 \frac{N\left(R_i, \mathscr{R}_n^{t+1}\right)}{L\left(\mathscr{R}_n^{t+1}\right)} = \frac{\sum_i N\left(R_i, \mathscr{R}_n^{t+1}\right)}{L\left(\mathscr{R}_n^{t+1}\right)} = 1. \end{aligned} \tag{A.4}$$

Thus, we get $I_{mn}^f, I_{mn}^b \in [0,1]$. ■

APPENDIX B

Top two codes from PACS 2010

Table B.1: First two digits of PACS 2010 and their meaning.

PACS code	Meaning
00	**General**
01	Communication, education, history, and philosophy
02	Mathematical methods in physics
03	Quantum mechanics, field theories, and special relativity
04	General relativity and gravitation
05	Statistical physics, thermodynamics, and nonlinear dynamical systems
06	Metrology, measurements, and laboratory procedures
07	Instruments, apparatus, and components common to several branches of physics and astronomy
10	**The Physics of Elementary Particles and Fields**
11	General theory of fields and particles
12	Specific theories and interaction models; particle systematics
13	Specific reactions and phenomenology
14	Properties of specific particles
20	**Nuclear Physics**
21	Nuclear structure
23	Radioactive decay and in-beam spectroscopy
24	Nuclear reactions: general

Continued on next page

Table B.1 – continued from previous page

PACS code	Meaning
25	Nuclear reactions: specific reactions
26	Nuclear astrophysics
27	Properties of specific nuclei listed by mass ranges
28	Nuclear engineering and nuclear power studies
29	Experimental methods and instrumentation for elementary-particle and nuclear physics
30	**Atomic and Molecular Physics**
31	Electronic structure of atoms and molecules: theory
32	Atomic properties and interactions with photons
33	Molecular properties and interactions with photons
34	Atomic and molecular collision processes and interactions
36	Exotic atoms and molecules; macromolecules; clusters
37	Mechanical control of atoms, molecules, and ions
40	**Electromagnetism, Optics, Acoustics, Heat Transfer, Classical Mechanics, and Fluid Dynamics**
41	Electromagnetism; electron and ion optics
42	Optics
43	Acoustics
44	Heat transfer
45	Classical mechanics of discrete systems
46	Continuum mechanics of solids
47	Fluid dynamics
50	**Physics of Gases, Plasmas, and Electric Discharges**
51	Physics of gases
52	Physics of plasmas and electric discharges
60	**Condensed Matter: Structural, Mechanical and Thermal Properties**
61	Structure of solids and liquids; crystallography
62	Mechanical and acoustical properties of condensed matter
63	Lattice dynamics
64	Equations of state, phase equilibria, and phase transitions
65	Thermal properties of condensed matter
66	Nonelectronic transport properties of condensed matter
67	Quantum fluids and solids
68	Surfaces and interfaces; thin films and nanosystems (structure and nonelectronic properties)
70	**Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties**

Continued on next page

Table B.1 – continued from previous page

PACS code	Meaning
71	Electronic structure of bulk materials
72	Electronic transport in condensed matter
73	Electronic structure and electrical properties of surfaces, interfaces, thin films, and low-dimensional structures
74	Superconductivity
75	Magnetic properties and materials
76	Magnetic resonances and relaxations in condensed matter, Mössbauer effect
77	Dielectrics, piezoelectrics, and ferroelectrics and their properties
78	Optical properties, condensed-matter spectroscopy and other interactions of radiation and particles with condensed matter
79	Electron and ion emission by liquids and solids; impact phenomena
80	**Interdisciplinary Physics and Related Areas of Science and Technology**
81	Materials science
82	Physical chemistry and chemical physics
83	Rheology
84	Electronics; radiowave and microwave technology; direct energy conversion and storage
85	Electronic and magnetic devices; microelectronics
87	Biological and medical physics
88	Renewable energy resources and applications
89	Other areas of applied and interdisciplinary physics
90	**Geophysics, Astronomy, and Astrophysics**
91	Solid Earth physics
92	Hydrospheric and atmospheric geophysics
93	Geophysical observations, instrumentation, and techniques
94	Physics of the ionosphere and magnetosphere
95	Fundamental astronomy and astrophysics; instrumentation, techniques, and astronomical observations
96	Solar system; planetology
97	Stars
98	Stellar systems; interstellar medium; galactic and extragalactic objects and systems; the Universe

APPENDIX C

Top 3 most cited papers in the case study Fig. 2.7

To illustrate the utility our knowledge evolution framework can offer, we use as a case study the interesting interactions between quantum optics (QO), quantum information (QI), and Bose-Einstein Condensation (BEC). These three fields experienced breakthroughs in the 1990s. Tab. C.1 shows the three most cited papers in these TCs, which are highlighted Fig. 2.7. Key merging and splitting events are reported in Chapter 2, as are important publications these events are correlated with.

Table C.1: The three most cited papers in quantum optics, quantum information theory, quantum computation and Bose-Einstein condensation related TCs.

Year	TC	DOI	Title
1991	Upper	10.1103/PhysRevLett.67.661	Quantum cryptography based on Bells theorem
		10.1103/PhysRevLett.66.2593	Observation of electromagnetically induced transparency
		10.1103/PhysRevLett.67.1855	Enhancement of the index of refraction via quantum coherence
	Lower	10.1103/PhysRevA.44.5674	Above-surface neutralization of highly charged ions: The classical over-the-barrier model
		10.1103/PhysRevB.43.13401	Strong magnetic x-ray dichroism in 2p absorption spectra of 3d transition-metal ions
		10.1103/PhysRevLett.66.2601	Dynamic stabilization of hydrogen in an intense, high-frequency, pulsed laser field
1992	Upper	10.1103/PhysRevLett.69.2881	Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states
		10.1103/PhysRevLett.69.3314	Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity
		10.1103/PhysRevLett.68.580	Wave-function approach to dissipative processes in quantum optics
	Lower	10.1103/PhysRevLett.68.1943	X-ray circular dichroism as a probe of orbital magnetization
		10.1103/PhysRevLett.68.3535	High-order harmonic generation from atoms and ions in the high intensity regime
		10.1103/PhysRevLett.69.1383	Absorption of ultra-intense laser pulses
1993	Upper	10.1103/PhysRevLett.70.1895	Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels
		10.1103/PhysRevA.47.4114	Threshold and resonance phenomena in ultracold ground-state collisions
		10.1103/PhysRevLett.70.1244	Measurement of the Wigner distribution and the density matrix of a light mode using optical homodyne tomography: Application to squeezed states and the vacuum
	Lower	10.1103/PhysRevLett.71.1994	Plasma perspective on strong field multiphoton ionization
		10.1103/PhysRevLett.70.1599	Above threshold ionization beyond the high harmonic cutoff
		10.1103/PhysRevLett.70.774	High-order harmonic generation in rare gases with a 1-ps 1053-nm laser
1994	Upper	10.1103/PhysRevA.50.67	Squeezed atomic states and projection noise in spectroscopy
		10.1103/PhysRevLett.72.3439	Statistical distance and the geometry of quantum states
		10.1103/PhysRevLett.73.58	Experimental realization of any discrete unitary operator
	Lower	10.1103/PhysRevA.49.2117	Theory of high-harmonic generation by low-frequency laser fields
		10.1103/PhysRevLett.73.1227	Precision Measurement of Strong Field Double Ionization of Helium
		10.1103/PhysRevA.50.1540	Modeling harmonic generation by a zero-range potential
1995		10.1103/PhysRevLett.75.3969	Bose-Einstein Condensation in a Gas of Sodium Atoms
		10.1103/PhysRevLett.74.4091	Quantum Computations with Cold Trapped Ions
		10.1103/PhysRevA.52.R2493	Scheme for reducing decoherence in quantum computer memory
1996		10.1103/PhysRevA.54.3824	Mixed-state entanglement and quantum error correction
		10.1103/PhysRevLett.77.1413	Separability Criterion for Density Matrices
		10.1103/PhysRevLett.77.2360	Collective Excitations of a Trapped Bose-Condensed Gas
1997	Upper	10.1103/PhysRevLett.78.985	Bose-Einstein Condensation of Lithium: Observation of Limited Condensate Number
		10.1103/PhysRevLett.78.586	Production of Two Overlapping Bose-Einstein Condensates by Sympathetic Cooling
		10.1103/PhysRevLett.78.5	Demonstration of the Casimir Force in the 0.6 to 6*mum* Range
	Lower	10.1103/PhysRevLett.78.5022	Entanglement of a Pair of Quantum Bits
		10.1103/PhysRevLett.78.3221	Quantum State Transfer and Entanglement Distribution among Distant Nodes in a Quantum Network
		10.1103/PhysRevLett.79.3306	Noiseless Quantum Codes
1998	Upper	10.1103/PhysRevLett.81.3108	Cold Bosonic Atoms in Optical Lattices
		10.1103/PhysRevLett.81.938	Atomic Scattering in the Presence of an External Confinement and a Gas of Impenetrable Bosons
		10.1103/PhysRevLett.81.742	Spinor Bose Condensates in Optical Traps
	Lower	10.1103/PhysRevA.57.120	Quantum computation with quantum dots
		10.1103/PhysRevLett.80.2245	Entanglement of Formation of an Arbitrary State of Two Qubits
		10.1103/PhysRevLett.81.5932	Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication
1999	Upper	10.1103/PhysRevLett.83.2498	Vortices in a Bose-Einstein Condensate
		10.1103/PhysRevLett.83.5198	Dark Solitons in Bose-Einstein Condensates
		10.1103/PhysRevLett.82.1975	Entanglement of Atoms via Cold Controlled Collisions
	Middle	10.1103/PhysRevLett.83.4204	Quantum Information Processing Using Quantum Dot Spins and Cavity QED
		10.1103/PhysRevB.59.2070	Coupled quantum dots as quantum gates
		10.1103/PhysRevLett.82.2417	Dynamical Decoupling of Open Quantum Systems
	Lower	10.1103/PhysRevLett.82.5229	Ultraslow Group Velocity and Enhanced Nonlinear Optical Effects in a Coherently Driven Hot Atomic Gas
		10.1103/PhysRevLett.83.2845	Transmission Resonances on Metallic Gratings with Very Narrow Slits
		10.1103/PhysRevLett.83.967	Liquid-Crystal Photonic-Band-Gap Materials: The Tunable Electromagnetic Vacuum
2000	Upper	10.1103/PhysRevLett.84.806	Vortex Formation in a Stirred Bose-Einstein Condensate
		10.1103/PhysRevLett.85.1795	Stable 85Rb Bose-Einstein Condensates with Widely Tunable Interactions
		10.1103/PhysRevLett.85.3745	Regimes of Quantum Degeneracy in Trapped 1D Gases
	Middle	10.1103/PhysRevA.62.062314	Three qubits can be entangled in two inequivalent ways
		10.1103/PhysRevLett.84.2722	Inseparability Criterion for Continuous Variable Systems
		10.1103/PhysRevA.62.012306	Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures
	Lower	10.1103/PhysRevLett.85.5214	Double Resonant Raman Scattering in Graphite
		10.1103/PhysRevLett.85.154	Electronic Structure of Deformed Carbon Nanotubes
		10.1103/PhysRevB.62.13104	Carbon nanotubes, buckyballs, ropes, and a universal graphitic potential

APPENDIX D

The full list of predictive features

As we mentioned in Chapter 4, each observation contained 77 features (preselected from the initial 100). The full list of 100 features are showed in Tab. D.1. Many features in this list are proposed for directed social network, therefore are inappropriate for our undirected BCN and CN. The symbol $+$ indicates this feature was used in BCN prediction, while the symbol $*$ indicates this feature was used in CN prediction.

Table D.1: List of all features used in the study. Features proposed in this study are shown in bold.

Features group	Feature name	Feature description
Members/microscopic	sum_group_degree_in	The sum of indegree [Freeman, 1978] of nodes belonging to the community calculated within the community. Indegree is a node measure defining the number of connections directed to the node
	avg_group_degree_in	The average value of indegree of nodes belonging to the community calculated within the community
	min_group_degree_in	The minimum value of indegree of nodes belonging to the community calculated within the community
	max_group_degree_in	The maximum value of indegree of nodes belonging to the community calculated within the community
	sum_group_degree_out	The sum of outdegree [Freeman, 1978] of nodes belonging to the community calculated within the community. Outdegree is a node measure determining the number of connections outgoing from the node
	avg_group_degree_out	The average value of outdegree of nodes belonging to the community calculated within the community
	min_group_degree_out	The minimum value of outdegree of nodes belonging to the community calculated within the community
	max_group_degree_out	The maximum value of outdegree of nodes belonging to the community calculated within the community
	sum_group_degree_total+∗	The sum of total degree of nodes belonging to the community calculated within the community. Total degree is the sum of indegree and outdegree
	avg_group_degree_total+∗	The average value of total degree of nodes belonging to the community calculated within the community
	min_group_degree_total+∗	The minimum value of total degree of nodes belonging to the community calculated within the community
	max_group_degree_total+∗	The maximum value of total degree of nodes belonging to the community calculated within the community
	sum_group_betweenness+∗	The sum of betweenness [Freeman, 1978] of nodes belonging to the community calculated within the community. Betweenness is a node measure describing the number of the shortest paths from all nodes to all others that pass through that node
	avg_group_betweenness+∗	The average value of betweenness of nodes belonging to the community calculated within the community
	min_group_betweenness+∗	The minimum value of betweenness of nodes belonging to the community calculated within the community
	max_group_betweenness+∗	The maximum value of betweenness of nodes belonging to the community calculated within the community

Continued on next page

Table D.1 – continued from previous page

Features group	Feature name	Feature description
	sum_group_closeness+∗	The sum of closeness [Freeman, 1978] of nodes belonging to the community calculated within the community. Closeness is a node measure defined as the inverse of the farness, which in turn, is the sum of distances to all other nodes
	avg_group_closeness+∗	The average value of closeness of nodes belonging to the community calculated within the community
	min_group_closeness+∗	The minimum value of c of nodes belonging to the community calculated within the community
	max_group_closeness+∗	The maximum value of closeness of nodes belonging to the community calculated within the community
	sum_group_eigenvector+∗	The sum of eigenvector [Bonacich, 1972] of nodes belonging to the community calculated within the community. Eigenvector is a node measure indicating the influence of a node in the network
	avg_group_eigenvector+∗	The average value of eigenvector of nodes belonging to the community calculated within the community
	min_group_eigenvector+∗	The minimum value of eigenvector of nodes belonging to the community calculated within the community
	max_group_eigenvector+∗	The maximum value of eigenvector of nodes belonging to the community calculated within the community
	avg_group_eccentricity+∗	The average value of eccentricity [Harary, 1969] of nodes belonging to the community calculated within the community. Eccentricity of a node is its shortest path distance from the farthest other node in the graph
	min_group_eccentricity+∗	The minimum value of eccentricity of nodes belonging to the community calculated within the community
	max_group_eccentricity+∗	The maximum value of eccentricity of nodes belonging to the community calculated within the community
	sum_network_degree_in	The sum of indegree of nodes belonging to the community calculated within the network
	avg_network_degree_in	The average value of indegree of nodes belonging to the community calculated within the network
	min_network_degree_in	The minimum value of indegree of nodes belonging to the community calculated within the network
	max_network_degree_in	The maximum value of indegree of nodes belonging to the community calculated within the network
	sum_network_degree_out	The sum of outdegree of nodes belonging to the community calculated within the network
	avg_network_degree_out	The average value of outdegree of nodes belonging to the community calculated within the network
	min_network_degree_out	The minimum value of outdegree of nodes belonging to the community calculated within the network
	max_network_degree_out	The maximum value of outdegree of nodes belonging to the community calculated within the network

Continued on next page

Table D.1 – continued from previous page

Features group	Feature name	Feature description
	sum_network_degree_total+∗	The sum of total degree of nodes belonging to the community calculated within the network
	avg_network_degree_total+∗	The average value of total degree of nodes belonging to the community calculated within the network
	min_network_degree_total+∗	The minimum value of total degree of nodes belonging to the community calculated within the network
	max_network_degree_total+∗	The maximum value of total degree of nodes belonging to the community calculated within the network
	sum_network_betweenness +∗	The sum of betweenness of nodes belonging to the community calculated within the network
	avg_network_betweenness+∗	The average value of betweenness of nodes belonging to the community calculated within the network
	min_network_betweenness+∗	The minimum value of betweenness of nodes belonging to the community calculated within the network
	max_network_betweenness+∗	The maximum value of betweenness of nodes belonging to the community calculated within the network
	sum_network_closeness+∗	The sum of closeness of nodes belonging to the community calculated within the network
	avg_network_closeness+∗	The average value of closeness of nodes belonging to the community calculated within the network
	min_network_closeness+∗	The minimum value of closeness of nodes belonging to the community calculated within the network
	max_network_closeness+∗	The maximum value of closeness of nodes belonging to the community calculated within the network
	sum_network_eigenvector+∗	The sum of eigenvector of nodes belonging to the community calculated within the network
	avg_network_eigenvector+∗	The average value of eigenvector of nodes belonging to the community calculated within the network
	min_network_eigenvector+∗	The minimum value of eigenvector of nodes belonging to the community calculated within the network
	max_network_eigenvector+∗	The maximum value of eigenvector of nodes belonging to the community calculated within the network
	avg_group_coefficient [Wasserman and Faust, 1994]+∗	The average of the local clustering coefficients of all the nodes in the community
	avg_network_coefficient [Wasserman and Faust, 1994]+∗	The average of the local clustering coefficients of all the nodes in the network
Group/mesoscopic	group_size+∗	The number of nodes in the group
	group_density [Wasserman and Faust, 1994]+∗	The number of connections between nodes in the group in relation to all possible connections between them
	group_cohesion [White and Harary, 2001]+∗	The vertex connectivity of the community
	group_coefficient_global [Wasserman and Faust, 1994]+∗	The ratio of the triangles and the connected triples in the community
	group_reciprocity [Newman, 2010]	A fraction of edges that are reciprocated within the community

Continued on next page

Table D.1 – continued from previous page

Features group	Feature name	Feature description
	group_leadership [Freeman, 1978]+∗	A measure describing centralization in the community (the largest value is for a star network)
	neighborhood_out	The number of nodes outside the community that have incoming connection from the nodes inside the community divided by the number of nodes in the community
	neighborhood_in	The number of nodes outside the community that have outgoing connection to the nodes inside the community divided by the number of nodes in the community
	neighborhood_all+∗	The number of nodes outside the community that are connected to the nodes inside the community divided by the number of nodes in the community
	group_adhesion [White and Harary, 2001]+∗	The minimum number of edges needed to be removed to obtain a community which is not strongly connected
	alpha [Bródka et al., 2013]	The GED inclusion measure of group G_i from time window T_n in group G_j from T_{n+1}
	beta [Bródka et al., 2013]	The GED inclusion measure of group G_j from time window T_{n+1} in group G_i from T_n
	network_ratio_size+∗	The ratio of *group_size* to *network_size*
	network_ratio_density+∗	The ratio of *group_density* to *network_density*
	network_ratio_cohesion+∗	The ratio of *group_cohesion* to *network_cohesion*
	network_ratio_coefficient_global+∗	The ratio of *group_coefficient_global* to *network_coefficient_global*
	network_ratio_coefficient_average+∗	The ratio of *group_clustering_coefficient* to *network_clustering_coefficient*
	network_ratio_reciprocity	The ratio of *group_reciprocity* to *network_reciprocity*
	network_ratio_leadership+∗	The ratio of *group_leadership* to *network_leadership*
	network_ratio_eccentricity+∗	The ratio of *avg_group_eccentricity* to *network_avg_eccentricity*
	network_ratio_adhesion+∗	The ratio of *group_adhesion* to *network_adhesion*
	phys_rev∗	**The number of articles belonging to the group that were published in the Physical Review journal**
	phys_rev_a+∗	**The number of articles belonging to the group that were published in the Physical Review A journal**
	phys_rev_b+∗	**The number of articles belonging to the group that were published in the Physical Review B journal**
	phys_rev_c+∗	**The number of articles belonging to the group that were published in the Physical Review C journal**
	phys_rev_d+∗	**The number of articles belonging to the group that were published in the Physical Review D journal**
	phys_rev_e+∗	**The number of articles belonging to the group that were published in the Physical Review E journal**
	phys_rev_lett+∗	**The number of articles belonging to the group that were published in the Physical Review Letters journal**

Continued on next page

Table D.1 – continued from previous page

Features group	Feature name	Feature description
	phys_rev_stab+∗	**The number of articles belonging to the group that were published in the Physical Review STAB journal**
	phys_rev_stper+	**The number of articles belonging to the group that were published in the Physical Review STPER journal**
	physics∗	**The number of articles belonging to the group that were published in the Physics journal**
	rev_mod_phys+∗	**The number of articles belonging to the group that were published in the Review of Modern Physics journal**
	sum_group_age+∗	**The sum of age of articles belonging to the group. In the co-reference network the age of an article is the average age of the articles it references to. In the co-citation network the age of an article is the age of the articles being cited.**
	avg_group_age+∗	**The average age of articles belonging to the group**
	min_group_age+∗	**The minimum age of articles belonging to the group**
	max_group_age+∗	**The maximum age of articles belonging to the group**
	network_ratio_avg_group_age+∗	**The ratio of avg_group_age to the average age of all articles in the network**
	time_window+∗	The number of time window from which the community instance was obtained
Network/macroscopic	network_size+∗	The number of nodes in the network
	network_density+∗	The number of connections between nodes in the network in relation to all possible connections between them
	network_cohesion+∗	The vertex connectivity of the network
	network_coefficient_global+∗	The ratio of the triangles and the connected triples in the network
	network_coefficient_average+∗	The average of the local clustering coefficients of all the nodes in the network
	network_reciprocity	A fraction of edges that are reciprocated within the network
	network_leadership+∗	A measure describing centralization in the network (the largest value is for a star network)
	network_avg_eccentricity+∗	The average value of eccentricity of nodes within the network.
	network_adhesion+∗	The minimum number of edges needed to be removed to obtain a graph which is not strongly connected