Tracking the Evolution of Community-based Research Topics in Dynamic Citation Networks

XIAO TAN, MSc. Applied Data Science, Utrecht University

Supervised by dr. Ioana Karnstedt-Hulpus, dr. A.A.A (Hakim) Qahtan and Vahid Shahrivari (daily supervisor)

Abstract

Understanding the evolution of research topics is essential to the development of any discipline. For funding agencies, academic institutions, individual researchers, and academic conference organizers, it helps them to understand the trends in their disciplines from a macro perspective and to make better decisions. In this paper, a method for understanding research topic evolution is proposed to answer the following questions: how can we use community detection approaches to locate research topics in citation networks? And how can we track the evolution of research topics in a dynamic citation network? This study used modularity-based algorithm for community detection, keyword word frequency for topic recognition using the tf-idf algorithm, and a clear definition of seven community events (Birth, Death, Growth, Contraction, Merging, Splitting and Continue). Based on this approach, research topics and disciplinary frontier developments can be better predicted and understood.

Keywords

Dynamic Networks, Network Communities, Topic evolution

Contents

1 Introduction

1.1 Motivation and context

Complex networks are widely utilized to explain and investigate interaction phenomena that occur in the real world theoretically and analytically. Numerous real-world systems, such as transportation networks, communication networks and social networks, can be represented in the digital realm as complex networks. Complex networks can be modelled as graph structures with nodes (individuals in the network) and edges (connections between nodes). The use of complex network analysis methods can help to better understand the structure of a domain.

Complex networks are also often applied in bibliometric studies, such as co-authorship network, citation network, etc. The volume of scientific publications is predicted to expand by 8-9% annually (Bornmann & Mutz, 2015). This extraordinary increase in research output indicates scientific advancement, but as a result, we must now cope with information overload. In recent years, many new research topics have emerged due to the increasingly interdisciplinarity of knowledge fields. It is important to discern patterns and trends in academic development at the general level because government organizations, academic institutions, and individual scholars can utilize these insights to develop more effective strategies for field development (Evans & Foster, 2011). However, it has become extremely difficult to detect prospective research topics and change research interests given the pace of overall academic output.

1.2 Research question

Generally, academic papers will cite other publications that are usually considered to be on the same research topic as itself. Therefore, communities in citation networks constructed by citation relations can be considered as a research topic. Community members are more connected internally, while community members are relatively loosely connected to external members. That is, members within each community in the citation network will cite each other more often, resulting in a research topic. In addition, the citation network itself gradually expands as new papers are cited, leading to changing communities (research topics).

Based on this current situation, I propose the following research questions:

- How can we use community detection approaches to locate research topics in citation networks?
- How can we track the evolution of research topics in a dynamic citation network?

2 Literature review

Previous research has examined the evolution of research topics in terms of communities in co-authorship and citation networks (Shibata, Kajikawa, Takeda, & Matsushima, 2008; Hopcroft, Khan, Kulis, & Selman, 2004). Communities are clusters of tightly connected nodes that are weakly connected to the rest of the network (Yang, Algesheimer, & Tessone, 2016). It is possible to detect the growth of enduring communities over time in dynamic networks. Moreover, these communities contain structural and temporal characteristics that may be utilized to predict their evolution, lifetime, and particular events such as merging and splitting (Goldberg, Magdon-Ismail, Nambirajan, & Thompson, 2011; Takaffoli, Rabbany, & Zaïane, 2014).

Communities in co-authorship networks consist of authors who cooperate closely with each other. Therefore, social events such as project partnerships and advisor-advisee relationship are characterized more by structural changes than by topic evolution. In the meanwhile, citation networks are used to describe the cross-referencing relationships between a series of papers. The communities in citation networks correspond to a set of publications assumed to be connected to an unidentified subject. Citation networks illustrate the current state of research information in a certain subject. Therefore, the network reflects groups revolving around particular research or a common research topic. Future research trends and the evolution of a field may be gleaned through an examination of how a community will develop in the future (Jung & Segev, 2014).

Previous investigations have categorized citation communities based on the most frequent words in papers. For example, Kusumastuti et al. (2016) used CitNetExplorer software to analyze the literature on successful aging and to summarize most frequency words in different citation networks to finally obtain research themes in the field. However, this approach does not help us understand the relationship of terms and concepts in the domain. Therefore, there are also studies that represent the research topic as a community of keywords in a dynamic co-occurrence network. Balili et al.

(2017) used co-occurrence word network analysis to summarize the interrelationships between concepts more precisely. But on the other hand, it does not show the evolution of the research topic well.

To address these issues, my work will consider communities as different research topics under the perspective of dynamic citation networks. I will summarize the behaviors of topic evolution through the evolution of these communities over time: birth, death, growth, continue, contraction, merging and splitting. In addition, I will use the TF-IDF method to analyze the abstracts of each publication to distinguish between different research topics. By doing this, I hope it can provide instructions to investigate the evolution of research topics in the research disciplines and help researchers to locate areas worthy of research.

3 Methods

3.1 Community detection

3.1.1 Network community and research topic

Since the dynamic nature of networks results in changing communities, a new field of study has evolved to examine dynamic communities. This area aims to develop methods for analyzing the collective behavior of networks and comprehending their development patterns.

Communities in dynamic networks may develop or evolve over time. To give an instance, there are a number of communities in a citation network. One of these communities corresponds to publications on a research topic, and each community is considered as a research topic as mentioned before. The static community corresponds to all papers that have been cited at a given time. However, after a period of time, some new papers are cited and thus join the citation network, and the communities change accordingly. Based on this, we can see the evolution of the communities over time, in other words, the evolution of the research topics.

3.1.2 Modularity based community detection

In recent years, computer scientists have presented several community detection algorithms. There are two techniques to detect the potential structures inside a given dataset: the network topology-based method and the content-based method (Ding, 2011). The first technique is based on Graph Theory. Modularity Maximization is a commonly utilized community detection algorithm (Newman & Girvan, 2004). Modularity is used to access the robustness of network communities' divides. High modularity indicates strong links within communities but weak links across communities.

Initially conceived for unweighted and undirected networks, modularity has been expanded to weighted and directed networks. It has been shown that modularity cannot identify tiny communities. However, only the larger communities will be seen as research topics in this study because smaller communities mean that these studies are not cited by other studies and may be at the edge of the network. In this study, only communities with more than 100 members are considered as a research topic.

Clauset-Newman-Moore greedy modularity maximization is utilized to detect communities. It starts with each node in its own community and continually combines the pair of communities resulting in the greatest modularity until no further increase in modularity is achievable (a maximum) (Clauset, Newman & Moore, 2004).

3.2 Community labelling

Keywords of the publications are usually used to label the detected communities in citation networks. Most papers contain keywords corresponding to the papers' key perspective. With topic identification approaches treating labels as topics, a community may be recognized by a collection of keywords that occur frequently in its publications. This study will follow a similar methodology. Natural language processing will be utilized to detect a set of terms and keywords based on the abstract of publications. For each community C, NLP-executed text_c is generated from the abstracts of C's membership documents c.

Text_c is preprocessed by tokenizing each text using the space character as the separator and then remove punctuations, stop words and numbers, and finally lemmatize it. By using a sufficiently big corpus of documents, unrelated words will become more distinguishable and may be discarded with more certainty. To extract a set of typical keywords for each community, TF-IDF (Term Frequency/Inverse Document Frequency) will be employed.

TF/IDF is commonly used to extract a set of typical keywords from a corpus of documents and is well-known for its strong performance on large datasets. The IDF is determined as the ratio between the overall number of documents and the number of documents that include the keywords. TF/IDF is calculated by dividing each words' term frequency by its inverse document frequency, with each community's text as a document. By using this method, it is possible to obtain the keywords that distinguish each community from the others, thus helping to find differences between the research topics. In this study, TfidfVectorizer function was used to analyze the abstracts of each paper.

3.3 Community events

The durability throughout time of communities undergoing progressive changes is a crucial issue to address. As the paradox of the ship of Theseus demonstrated, determining whether an element consist of various entities at a given time is identical to another element composed of the some or even none of these entities at a later time is arbitrarily determined and cannot be answered unambiguously.

In the perspective of dynamic communities, the key to defining a dynamic network community is the presence and disappearance of nodes and edges. On a community scale, however, the process that determine community changes are more complicated and are referred as "events". Different events are specified in many studies, and they are all comparable and complementary (see Fig. 1). Palla er al. (2007) firstly established the systematic classification of the transformations that include communities by defining six of them (birth, death, growth, contraction, merge and split). Continue will also be added to these sometimes. Cazabet and Amblard (2014) suggested resurgence as an eighth events. A similar classification of community events was used in this study. Below is an explanation of the seven community events utilized in this study:

- Birth, when a new community arises at a given time. Community C_t is observed to appear at time t, but there is no corresponding community C_{t-1} at time t-1 (None of C_t contain more than 60% of the members of C_{t-1} and no any two of C_t contain more than 30% of members of C_{t-1} respectively).
- Death, when a community becomes extinct. Community C_t is observed at time t, but there is no corresponding community C_{t+1} at time t+1 (None of C_t contain more than 60% of the members of C_{t+1} and no any two of C_t contain more than 30% of members of C_{t+1} respectively).

- Growth, when a community gains more nodes. The growth of a community happens when the community C_t observed at time t is significantly larger than the community C_{t-1} observed at time t-1 (C_t should contain more than 60% of the members of C_{t-1} and a growth of 5% growth in the size of the community compared to the network).
- Contraction, when a community loses nodes. The contraction of a community happens when the community C_t observed at time t is significantly smaller than the community C_{t-1} observed at time t-1 (C_t should contain more than 60% of the members of C_{t-1} and a reduction of 5% growth in the size of the community compared to the network).
- Merging, when several communities merge into one community. The merging of a community happens when the communities (C_1, C_2) observed at time t merge into the same community C_{t+1} at time t+1 (C_{t+1} should contain more than 30% of the members of communities (C_1, C_2) respectively).
- Splitting, when a community splits into several communities. The splitting of a community happens when the community C_t observed at time t split into the communities (C_1, C_2) at time t+1 (communities (C_1, C_2) should contain more than 30% of the members of C_t respectively).
- Continue, when a community stays unchanged. The continue of a community happens when the community C_t observed at time t remains unchanged compared the community C_{t+1} at time t+1 (C_{t+1} should contain more than 60% of the members of C_t and the change in community size is going to be between -5% and 5% compare to the network).

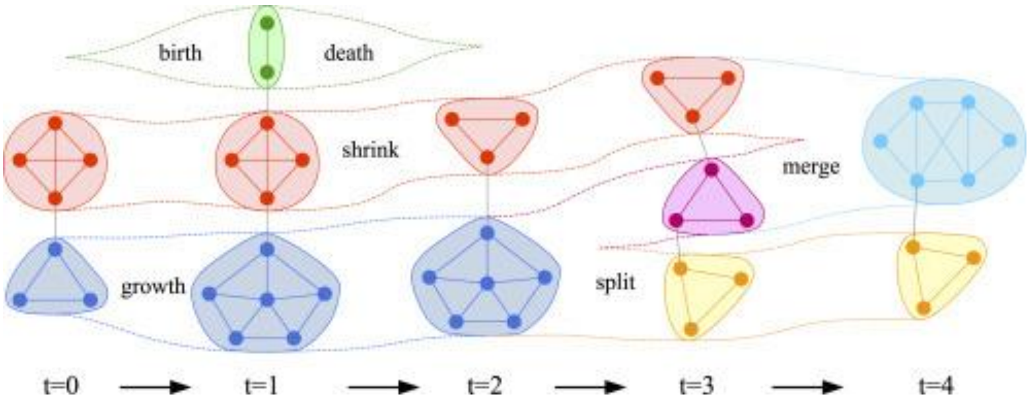

Figure 1 Community evolution in a dynamic network (Shang et al., 2016)

It is important to note that community events are not completely mutually exclusive by this standard. For example, a community may be shrinking and splitting into two different communities at the same time.

4 Data

4.1 Dataset

This investigation uses high-energy physics theory citation network dataset from SNAP (Stanford Network Analysis Project). The dataset contains 27770 papers from January 1993 to April 2003 (124 months). If a paper i cites paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any relevant information. Paper citation network, time of nodes and paper meta information are included in the dataset.

Table 1 Basic network attributes for the dataset

Nodes	27770
Edges	352807
Average clustering coefficient	0.3120
Number of triangles	1478735
Fraction of closed triangles	0.04331
Diameter (longest shortest path)	13
90-percentile effective diameter	5.3

The dataset is divided into 10 periods according to the publish year of papers. Papers published in 2003 were removed because the 2003 data contained only 4 months. Therefore, a total of 10 citation networks will be constructed and growing in size because each network will contain all papers from the previous year and newly cited publications for the year. According to Figure 2, the growth in the size of networks has remained generally stable, with an average of about 2,700 new publications per year.

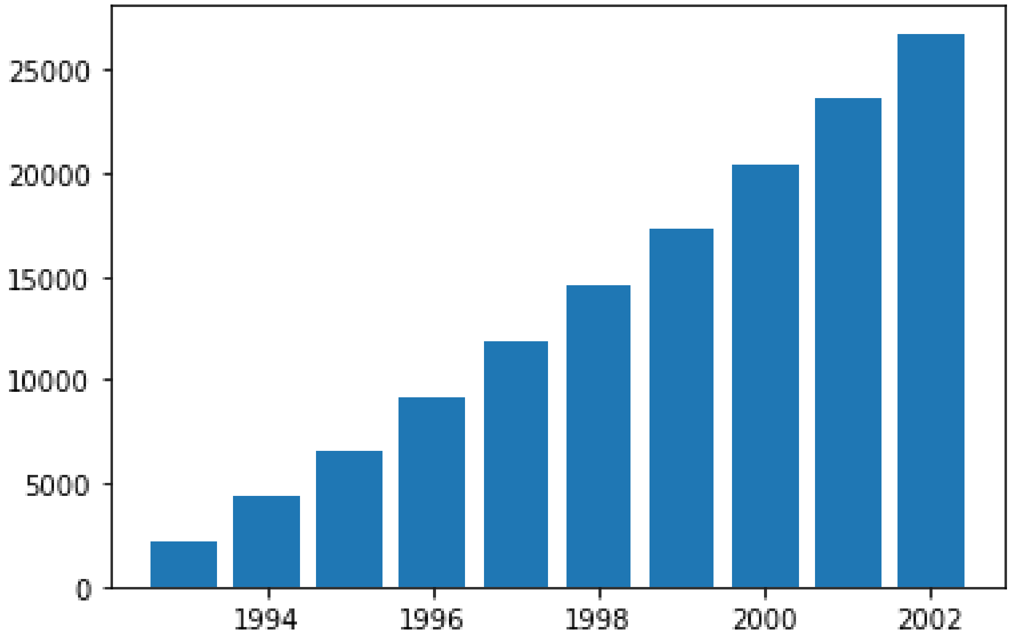

Figure 2 Number of nodes for each network

Greedy_modularity_communities function from networkx was utilized to detect communities for each network. As introduced before, it will distribute all publications to different communities without duplication based on modularity. Only the communities which are more than 100 members will be considered as research topic. According to Figure 3, each network has an average of 8.2 research topics.

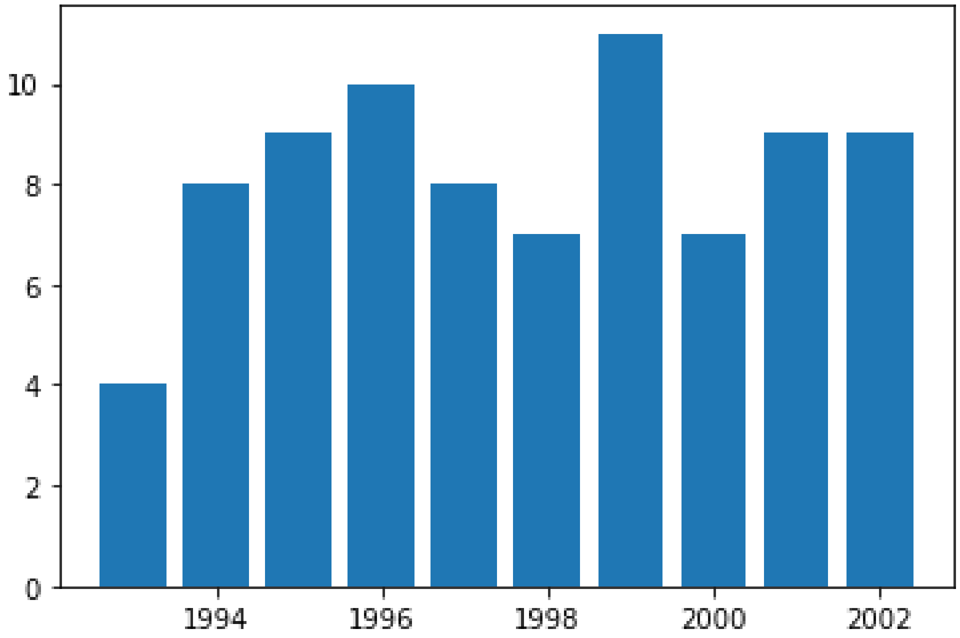

Figure 3 Number of research topics for each network

4.2 Community labelling

The abstracts of all the publications of each community were extracted and constitute a single document, so that there are 82 documents in total, corresponding to the research topics of each community. Next, the text is pre-processed using the data processing methods mentioned above, including tokenization, stop words removal, lemmatization etc.

Then, TfidfVectorizer function was utilized to get the key feature words for each document. After that, wordclouds are constructed for each research topic based on the words and tfidf values (Figure 4). Based on this, it is possible to summarize the research topics of each community.

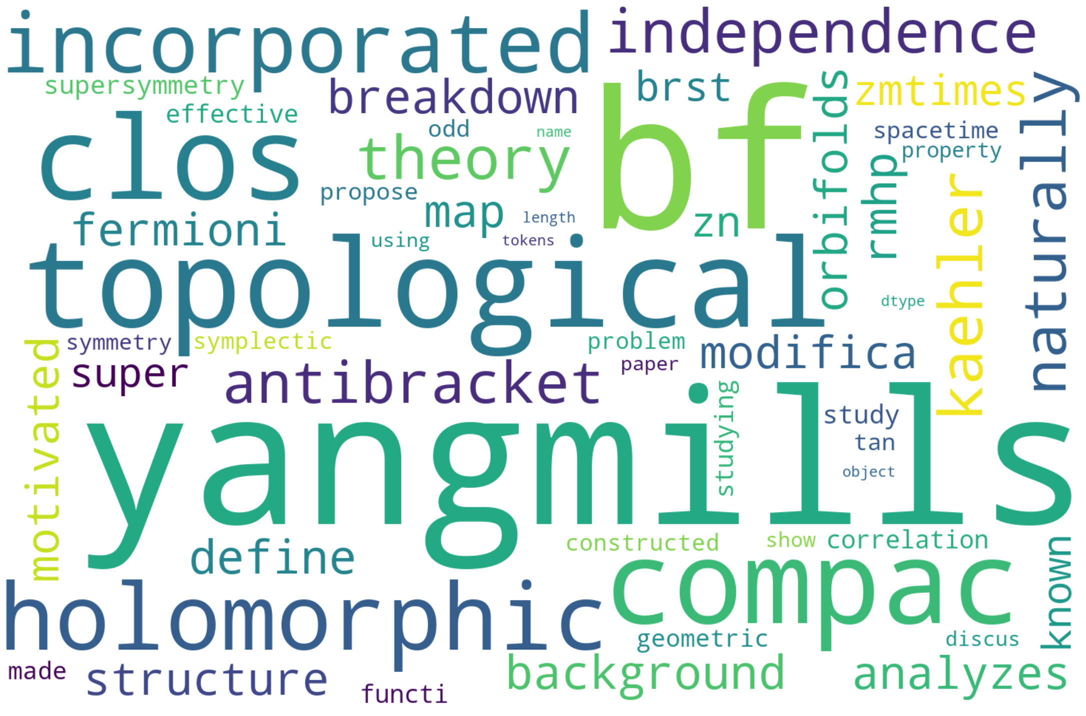

Figure 4 An example of research topic wordcloud

4.3 Tracking the evolution of communities

To determine the community events between C_t and C_{t+1}, firstly the overlap rate between them need to be calculated. Overlap rate means the number of publications present in both C_t and C_{t+1} as a percentage of number of papers in C_t. A matrix of overlap rates for all communities in adjacent years was calculated and used to determine the community events between them.

For example, as in Table 2, C_{1994_1} can be seen as a merging of C_{1993_2} and C_{1993_3}. C_{1994_2} is seen as a newly-born research topic.

Table 2 An example of overlap rate matrix between 1993 and 1994

	Overlap rate							
Community	**1994_1**	**1994_2**	**1994_3**	**1994_4**	**1994_5**	**1994_6**	**1994_7**	**1994_8**
1993_1	0	0.0169	0.5452	0.4237	0.0056	0	0.0028	0
1993_2	0.7492	0.0289	0.1576	0.0064	0.0032	0	0.0289	0.0032
1993_3	0.6944	0.2083	0	0	0.0046	0	0	0
1993_4	0.0625	0.5625	0.0156	0.1797	0	0	0	0

5 Result

5.1 Topic evolution analyzing

To analyze and visualize the overall topic evolution, overlap rate matrix and sankey diagram were adopted. Sankey diagram is a flowchart that represents the flow and transfer of energy, capital, etc. within a system. Therefore, the use of Sankey diagrams allows for a better representation of the flow of publications across research topics, and thus a better visualization of community events and topic evolution.

Figure 5 shows the evolution of the research topics from 1993 to 2002, in which the communities less than 100 members are ignored because they usually do not have a significant impact on the overall. Community events mentioned previously are shown in this diagram, including growth, shrink, merging, splitting etc. I will analyze representative community events in the following section.

In addition, the vast majority of keywords for research topics screened according to tfidf consisted of terms, concepts, theories, and names of researchers. Therefore, I will summarize the research topics through these keywords.

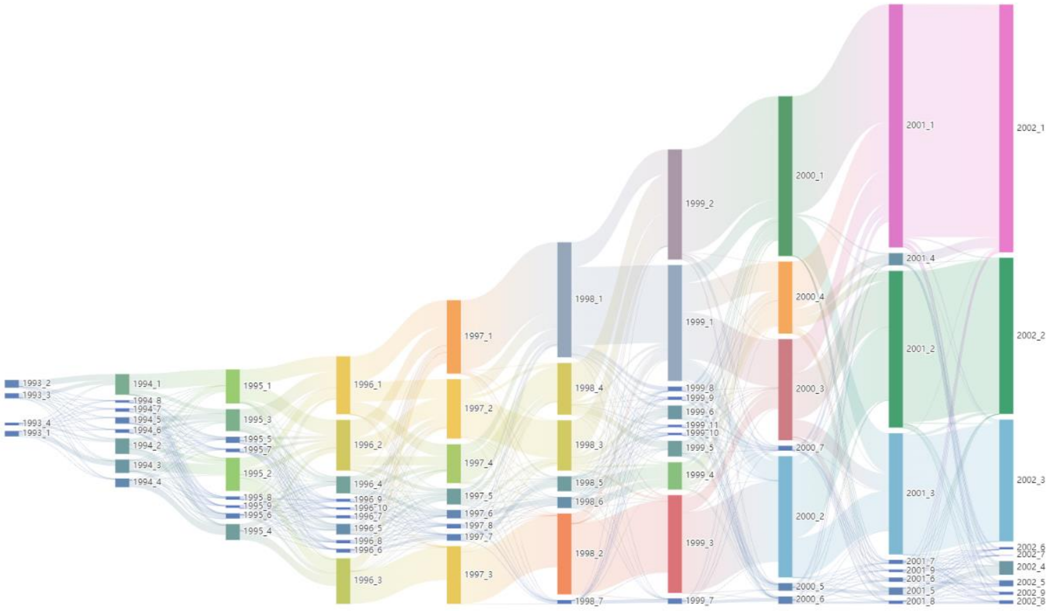

Figure 5 The overall research topics evolution

5.2 Community events

Figure 6 shows the distribution of community events for each year. Birth and merging were the most frequent community events, with 34 and 19 words respectively. This was followed by growth, splitting and death, which occurred 13, 11 and 10 times respectively. This also shows that the research topic is always in the process of evolving. The emergence and integration of new research topics is the mainstream of disciplinary development.

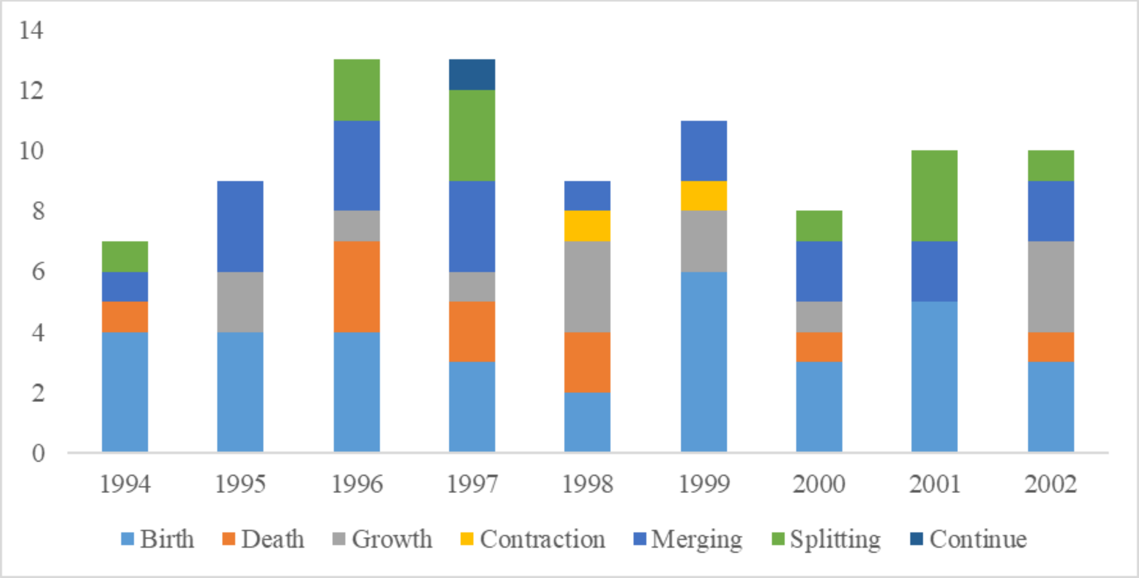

Figure 6 Distribution of community events

5.3 Case study

At the beginning of this section, I will first focus on the formation of the typical community events and give the corresponding examples, shown in Figure 6. Based on this, we can see the evolution between communities and how the keywords for each research topic evolved. For example, we can clearly see how C_{1993_2} and C_{1993_3} merged into C1994_1 and have studied how the subject keywords evolved.

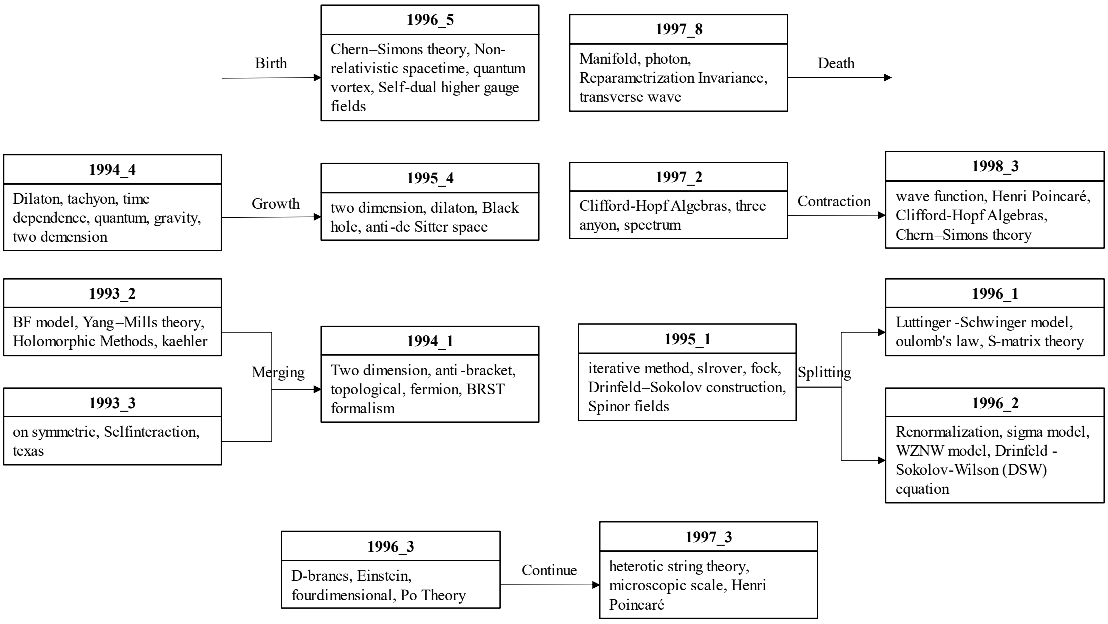
Birth
1996_5
Chern–Simons theory, Non-relativistic spacetime, quantum vortex, Self-dual higher gauge fields
1997_8
Manifold, photon, Reparametrization Invariance, transverse wave
Death
1994_4
Dilaton, tachyon, time dependence, quantum, gravity, two demension
Growth
1995_4
two dimension, dilaton, Black hole, anti-de Sitter space
1997_2
Clifford-Hopf Algebras, three anyon, spectrum
Contraction
1998_3
wave function, Henri Poincaré, Clifford-Hopf Algebras, Chern–Simons theory
1993_2
BF model, Yang–Mills theory, Holomorphic Methods, kaehler
Merging
1994_1
Two dimension, anti-bracket, topological, fermion, BRST formalism
1993_3
on symmetric, Selfinteraction, texas
1995_1
iterative method, slrover, fock, Drinfeld–Sokolov construction, Spinor fields
Splitting
1996_1
Luttinger -Schwinger model, oulomb's law, S-matrix theory
1996_2
Renormalization, sigma model, WZNW model, Drinfeld - Sokolov-Wilson (DSW) equation
1996_3
D-branes, Einstein, fourdimensional, Po Theory
Continue
1997_3
heterotic string theory, microscopic scale, Henri Poincaré

Figure 7 An example of community events

In order to discover how the research topic evolves, Community C_{2002_1} was taken as an example. C_{2002_1} is a research topic mainly focusing on gravitational singularity, super fields, and quasi-exact solvability based on the tfidf. Based on the community events formulated above, I can briefly track back the evolution of C_{2002_1} as shown in the Figure 7. This research topic dates back as far as 1996 and has undergone several community evolutionary events, including growth and merging. In 1996, the community was first working on BF theory and Noncommutative Geometry. As the research evolved, more theories and concepts were included, such as Batalin-Vilkovisky Formulism, Neveu-Schwarz-Ramond Superstring Model, Finite theory, Landau–Lifshitz–Gilbert equation, and so on.

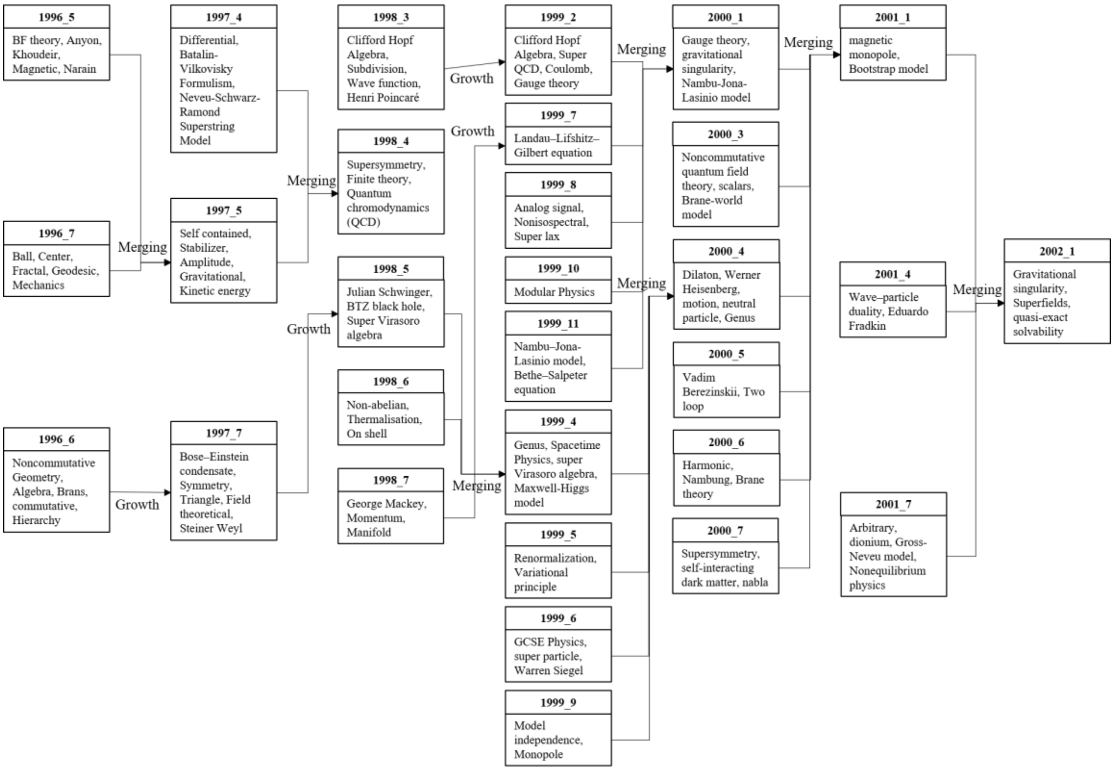
1996_5
BF theory, Anyon, Khoudeir, Magnetic, Narain
1996_7
Ball, Center, Fractal, Geodesic, Mechanics
1996_6
Noncommutative Geometry, Algebra, Brans, commutative, Hierarchy
Merging
Growth
1997_4
Differential, Batalin-Vilkovisky Formulism, Neveu-Schwarz-Ramond Superstring Model
1997_5
Self contained, Stabilizer, Amplitude, Gravitational, Kinetic energy
1997_7
Bose–Einstein condensate, Symmetry, Triangle, Field theoretical, Steiner Weyl
Merging
Growth
1998_3
Clifford Hopf Algebra, Subdivision, Wave function, Henri Poincaré
1998_4
Supersymmetry, Finite theory, Quantum chromodynamics (QCD)
1998_5
Julian Schwinger, BTZ black hole, Super Virasoro algebra
1998_6
Non-abelian, Thermalisation, On shell
1998_7
George Mackey, Momentum, Manifold
Growth
Growth
Merging
1999_2
Clifford Hopf Algebra, Super QCD, Coulomb, Gauge theory
1999_7
Landau–Lifshitz–Gilbert equation
1999_8
Analog signal, Nonisospectral, Super lax
1999_10
Modular Physics
1999_11
Nambu–Jona-Lasinio model, Bethe–Salpeter equation
1999_4
Genus, Spacetime Physics, super Virasoro algebra, Maxwell-Higgs model
1999_5
Renormalization, Variational principle
1999_6
GCSE Physics, super particle, Warren Siegel
1999_9
Model independence, Monopole
Merging
Merging
2000_1
Gauge theory, gravitational singularity, Nambu-Jona-Lasinio model
2000_3
Noncommutative quantum field theory, scalars, Brane-world model
2000_4
Dilaton, Werner Heisenberg, motion, neutral particle, Genus
2000_5
Vadim Berezinskii, Two loop
2000_6
Harmonic, Nambung, Brane theory
2000_7
Supersymmetry, self-interacting dark matter, nabla
Merging
2001_1
magnetic monopole, Bootstrap model
2001_4
Wave–particle duality, Eduardo Fradkin
2001_7
Arbitrary, dionium, Gross-Neveu model, Nonequilibrium physics
Merging
2002_1
Gravitational singularity, Superfields, quasi-exact solvability

Figure 8 Case study: an evolutionary community

6 Conclusion and discussion

The most important issue for research topic detection and evolution is how to define a research topic and the progression of its evolution. In general, publications mostly cite literature related to their own research topics. That is, the citation network consisting of citations largely reflects the distribution of research topics. A community is a dense group of members in a network. The connections between the nodes within each community are relatively strong, and the connections between individual communities are relatively sparse. Therefore, I can determine the research topic based on the communities in the citation network.

Then, how to determine the specific research content of each topic? To simply this problem, I used the tf-idf approach to analyze the abstracts of all publications within each community. The abstract is a highly summarized part of the paper, according to which the research topic of the paper can be easily summarized. The tf-idf algorithm can help extract keywords that distinguish documents from other documents, which can help better identify research differences between communities.

At the same time, however, communities and research topics are not static but evolve dynamically. Therefore, based on previous studies, I manually designed seven community events and defined criteria for their definition, including birth, death, growth, contraction, continue, merging and splitting. Based on this, the evolution of community research topics can be better studied and guidance can be provided to researchers in the field on the direction of research and shifts in research interests.

The case study conducted in this research has demonstrated the evolution of the research topics in 7 forms. In all 7 of these forms, merging and splitting lead to the emergence of new research topics. Often, there are precursors that precede the emergence of a new topic in a research field. Exploring the precursors and causes will help to predict the changing dynamics of topics and research frontiers. In addition, growth means that a research topic is expanding, while contraction means that a research topic is gradually decaying. These evolutions are also significant and instructive in research.

In this study, I used a method to identify research topics based on word frequency. However, there is also an approach to dynamic co-word networks based on the relationships between keywords (Wang, Cheng, & Lu, 2014). This method could also be adopted in future studies to increase the accuracy of generalization. In addition, this study artificially defines community evolutionary events. A more complex community evolution verification algorithm can be used in the future research and this method can be applied in different research areas. Suitable similarity thresholds are found in different research areas and matching community evolution algorithms are selected for different disciplines.

List of references

Balili, C., Segev, A., & Lee, U. (2017, December). Tracking and predicting the evolution of research topics in scientific literature. In 2017 IEEE international conference on big data (big data) (pp. 1694-1697). IEEE.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11), 2215-2222.

Cazabet, R., & Amblard, F. (2011, August). Simulate to detect: a multi-agent system for community detection. In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (Vol. 2, pp. 402-408). IEEE.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.

Ding, Y. (2011). Community detection: Topological vs. topical. Journal of Informetrics, 5(4), 498-514.

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. Science, 331(6018), 721-725.

Goldberg, M., Magdon-Ismail, M., Nambirajan, S., & Thompson, J. (2011, October). Tracking and predicting evolution of social communities. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 780-783). IEEE.

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. Proceedings of the National Academy of Sciences, 101(suppl_1), 5249-5253.

Jung, S., & Segev, A. (2014). Analyzing future communities in growing citation networks. Knowledge-Based Systems, 69, 34-44.

Kusumastuti, S., Derks, M. G., Tellier, S., Di Nucci, E., Lund, R., Mortensen, E. L., & Westendorp, R. G. (2016). Successful ageing: A study of the literature using citation network analysis. Maturitas, 93, 4-12.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. Nature, 446(7136), 664-667.

Shang, J., Liu, L., Li, X., Xie, F., & Wu, C. (2016). Targeted revision: A learning-based approach for incremental community detection in dynamic networks. Physica A: Statistical Mechanics and its Applications, 443, 70-85.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 28(11), 758-775.

Takaffoli, M., Rabbany, R., & Zaïane, O. R. (2014, August). Community evolution prediction in dynamic social networks. In 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (pp. 9-16). IEEE.

Wang, X., Cheng, Q., & Lu, W. (2014). Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics, 101(2), 1253-1271.

Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. Scientific reports, 6(1), 1-18.