The University of Chicago

A Framework to Analyze the Evolution of Science and an Application for Broad Academic Fields

By

Rodrigo Valdés Ortiz

August 2018

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor:	*Preceptor:*
Johan Chu	Joshua Mausolf

Abstract

This thesis develops a framework to analyze the evolution of science. Using citation networks from 1985 to 2015 as primary data —886 M citations—, I create clusters of papers which belong to an academic field by year. To create those clusters, I utilize a two-step process. First, I use Infomap as a community detection algorithm, which output is a hierarchical tree. Second, I recursively prune that tree to create specific groups according to particular rules. Also, I utilize a recursive algorithm which uses t-SNE and k-means to order the groups while respecting the structure of Infomap's tree. I create a multi-slice network that connects yearly groups over time. Afterward, I elaborate on an application of this framework to analyze broad academic fields and their evolution. Medicine is the most persistent cluster in the last thirty years according to the results. Meanwhile, academic fields that last only a couple of years are characterized by their specificity or applied nature. The framework utilized for this thesis can be further applied to research on the creation, evolution, and decline of research areas, among other issues related to the evolution of science.

Contents

Chapter 1

Introduction

Science is not a heterogeneous phenomenon. Furthermore, science does not have the same meaning across all latitudes or disciplines. For instance, some scholars understand science as huge collaborations, others live science working alone in their cubicles. Some academic disciplines require an intense interdisciplinary dialogue because the progress of one discipline depends on another one. Other fields are more isolated, and seldom find cause to cite papers outside their own discipline. Moreover, patterns of interdisciplinary in science are also shaped by academic leaders, technological innovations, politics, sports' competitions, wars, and researchers' curiosity. These behaviors, traditions, necessities, and possibilities influence scientific discovery and we can track them through the fingerprints of scientific networks —citations.

This thesis elaborates on several methodological issues related to large-scale networks which evolve over time. However, analyzing citations networks inspires and guides the development of this study. In addition, this paper studies the structure of science over the last thirty years —the overall patterns on the continuity and discontinuity of subjects in broad academic fields.

I create a network of citations among scientific papers over the last thirty years, which reveals connections among disciplines and how these have changed over time. The most interesting part of this study is the creation of ties among yearly snapshots of science. Hence, it is possible to track changes in topics, size of the fields, most prevalent ideas and authors. However, in this paper, I focus only on the predominant subjects by each broad academic discipline, such as natural sciences or social sciences. An extension of this will be able to show isolated disciplines in the science space —the network—, new academic fields emerging from the fusion of heterogeneous disciplines, ideas of one field that become a discipline by themselves, and fields whose relevance drastically decrease.

Previous attempts to analyze this subject restricted their study to specific fields or journals. However, my analysis considers the networks within all papers, without restricting it to specific disciplines or journals. In comparison to previous studies, this analysis also examines patterns over time, which allows studying not only a screenshot but also the changes in scientific patterns.

This study contributes to creating a better understanding of the evolution of science. To this end, I consider three fundamental ideas. First, I utilize networks at the publication level —instead of a journal to journal level. Second, I identify academic disciplines using community detection algorithms, which allow me not to rely on third-party data for discipline classification. Third, I use multi-slice networks, which allows me not only to analyze a screenshot of science, but the path and evolution of academic fields.

The sequence of this paper is as follows. First, I present a literature review which introduces the concept of the science of science and talks about previous efforts to analyze citations networks. Second, I discuss the empirical methodology for the

analysis, including data, algorithms, and strategies to assess if the algorithms produce the expected results. Third, I talk about the empirical findings. Finally, I offer some concluding remarks and possible implications.

Chapter 2

Literature Review

2.1 Science of Science

This thesis is under the framework of the Science of Science, which aims to elucidate general questions about science and scientists. For instance, which is the optimal number of collaborators in a publication? What are the conditions to generate innovative science? Do funding strategies have an effect on scientific outcomes? Is the number of citations a good measure of scholar's performance? In recent years, the quality and quantity of information, as well as scientists' ability to use computation and large-scale data analysis, has made the study of these topics possible. This section introduces this field, especially the ideas related to citations, citation dynamics, and scientific communities. The information in this chapter is based on the summary of Science of Science made by Fortunato et al. (2018).

The number of scientific publications has increased exponentially but not the number of ideas. The conceptual territory in the scientific literature has increased linearly according to large-scale text analysis of articles' titles and abstracts. Then,

the increment in the number of articles and citations depict another trend in science, such as the increasing number of publications, journals, and virtual spaces (Milojevi, 2015, cited in Fortunato et al., 2018).

With respect to the distribution of citations, Radicchi and Fortunato (2008) argue that there is a universal distribution of citations, but the parameters of that distribution are unique for each field. For instance, if the number of citations of a paper is divided by the average number of citations collected by papers in the same discipline and year, the distribution is indistinguishable for all disciplines (Radicchi and Fortunato, 2008, cited in Fortunato et al. 2018). Then, if the distribution is the same for all disciplines, scholars can compare articles' impact considering the relative citations by discipline.

Regarding citation dynamics, high-impact papers accumulate citations according to a power law, which can be explained by preferential attachment, the process where publications that are already highly cited have a higher probability to be cited again (Golosovsky et al., 2012; Stegehuis et al., 2015; Thelwall, 2016; cited in Fortunato et al., 2018). However, a small fraction of publications is classified as sleeping beauties, which are papers that accumulate a few citations immediately after published, but they receive wide attention several years later (Van Raan, 2004; Ke et al., 2015; cited in Fortunato et al., 2018).

On scientific communities, scientists have analyzed communities of publications that frequently cite each other. Those groups, identified through network science, are clusters of authors that work in a similar topic or hold the same position on specific issues (Klavans & Boyack, 2016, cited in Fortunato et al., 2018). In the same vein, other authors have study diverse phenomena regarding academic communities. For instance, Shi, Forster, and Evans (2015) say the number of publications increased

in biomedical science, which reinforced communities in specific research domains. In addition, new publications "fall between things only one or two steps away from each other, implying that when scientists choose new topics, they prefer things directly related to their current expertise" (Shi et al., 2015; cited in Fortunato et al., 2018). The latter suggests that the current structure of science might be definitive to shape future topics for research. Along the same line, Bettencourt et al. (2009) show that successful fields undergo a process of unification that creates a giant component in a network of collaborators. Uzzi et al. (2013) find that publications with the highest impact combine well-established previous work with knowledge from a different field. Those publications double its probability to be high-cited.

2.2 Previous work on citations networks

Martin, Ball, Karrer, and Newman (2013) analyze networks of citations in the *Physical Review*. They examine collaboration patterns among citations and authors. Additionally, they investigate dynamics in the number of coauthors and speed of citations —a tendency to cite new papers—. Even though their study spans about a hundred years, they only analyze the field of physics.

Deville et al. (2014) use 420,000 papers to reconstruct academic trajectories. In specific, they study how changes in academic institution impact scientists' productivity. The authors find that switch from a high-rank institution to a low-rank institution correlates with a slight decrease in productivity. In contrast, moving from a low-rank institution to a high-rank institution does not increase productivity.

Sinatra, Deville, Szell, Wang, and Barabási (2015) scrutinize a century of citations in physics. They reveal subdisciplines and multidisciplinarity of this field. In order

to make their analysis, they define a physics paper in two different ways. First, those published in a physics journal. Second, those related to articles published in physics journals but not published in a physics journal. This second category comprises papers with references and citations to the core physics journal —papers published in physics journals— statistically higher than by chance and cited by articles in the core physics journal in the same fashion. After defining a physics paper, they study the productivity of scholars, average age of citations —mean life of articles that a paper cites—, average number of citations after ten years of publishing, fraction of citations by year after publication, and citations between core physics and interdisciplinary physics. Also, they identify subdisciplines by tagging each paper in one of the ten major Physics and Astronomy Classification Scheme (PACS), which is based on physics' subfields. However, only 5% of the physics papers not published in a physics journal were assigned to one category. Considering this discipline classification, they investigate self-references within the field —citing papers of the same subfield—, impact of papers by subdiscipline, and evolution of impact factors after publication.

The analysis of Sinatra et al. (2015) has one important limitation. They rely on the classification of subfields provided by a third-party source, i.e. PACS. Then, they are not able to classify in fields more than half of the papers relevant for the physics community. I propose to utilize algorithms of community detection, which do not depend on third-party data. Then, I am able to identify subfields not only for physics but also other scientific disciplines.

Chapter 3

Methodology and Data

Introduction

In order to analyze the evolution of science, I utilize data from citation networks. In these networks, nodes are papers and ties are citations among them —the direction of citations is not considered—. Citation network's information is used to run community detection algorithms that reveals patterns among papers.

In specific, the process to generate academic fields per year —groups of papers identified as one discipline— and ties among them —relationship across years— comprises three main steps. First, create a network where the nodes are papers and the links are citations. Second, run community detection algorithms to identify clusters of papers or academic fields, defined as groups of articles tightly connected. Although some of these clusters can be tagged with a name, such as "family economics" for instance, the scope of this project does not allow to name and identify each of the fields. This second step consists of two phases, (1) utilize the nodes and ties of the

citations network to feed a community detection algorithm —Infomap— in order to create a hierarchical graph, (2) design and run an algorithm that dynamically defines an academic field using the hierarchical graph. That is to say, this second phase is pruning the hierarchical graph and extract parts of it, where each part is a group of papers. Each group of papers represents a specific part of science. Third, connect these groups over time, that is to say, create a multi-slice network, where each slice is a screenshot of science in a defined time frame, and ties among groups in different slices are the overlapped papers between two groups. The multi-slice network, the outcome of this process, is the main tool to analyze changes in science because it not only provides information about one year of science but also the connection among academic fields in different years.

This section is organized as follows. First, a description of the data. Second, an explanation of Infomap, the community detection algorithm. Third, an explanation of the algorithms to create and order academic fields from the Infomap hierarchical graph. Fourth, an strategy to asses the groups' order. Finally, a description on how to create ties among groups in different years, and the application of it to track the continuity of subjects among broad academic fields.

3.1 Data

I use the data from the *Web of Science (WoS)* compiled by Thompson Reuters. At the time this thesis is being written, it is the most complete source of citation indexing

available for research. On the one hand, it provides citation data. That is to say, a list papers and all sources each paper cites, where each source has a unique id across time. Consequentially, researches can track papers that cite a specific article in a specific time frame. On the second hand, the *WoS* provides additional information about each article, such as a subject classification, keywords, name of the authors and their institutions, and detailed information about article's journals.

In this study, I utilize data from 1985 to 2014,[1] which consists on about 1 billion citations —tuple of (1) article which cites and (2) cited article—. The figure 3.1 summarize the main features of the data. Note that for 1985 the total number of citations is about 12 million. Meanwhile, for 2014 the total number of citations is above 60 million. Also, the average growth in the number of citations per year is 6%. However, in some extraordinary years, citations growth more than 10%. Only in 1986, the number of citations slightly decrease.

[1]The data for 2015 is not complete and available.

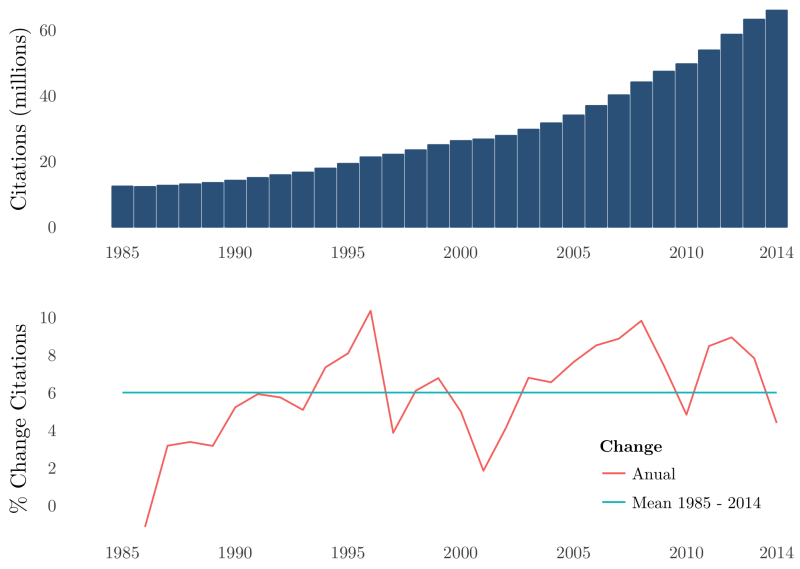


Figure 3.1: Number of citations per year

3.2 Infomap

In order to create communities of papers, I use an algorithm based on the map equation, Infomap (Edler & Rosvall, 2017). This algorithm minimizes the information needed to describe the process of information diffusion across the graph, where the proxy for information diffusion is a random walk.

With respect to information minimization, a useful analogy to understand this process is to think about a geographical graph. For example, different streets have the same name but are in different towns. For these streets, there are two options to uniquely identify each road. First, use two elements to identify each street: (1) the name of the town, (2) the name of the street. Second, assign a different name to each

road in the country. In the first option, the name of the town is stored only once. Meanwhile, in the second option, you save a specific different name for each road. Then, the first option requires less information to uniquely identify each road given that the name of the town is stored only once against the second option. That is to say, the number of bits used to describe all the streets in the first option is smaller than using a different name for each street.

Utilizing the analogy to explain the algorithm, Infomap creates clusters —towns— that allow minimizing the information needed to describe the graph. In the same vein, Infomap generates hierarchies —think about states, cities, neighborhoods, blocks, etcetera— to minimize the information needed to describe graphs' structures.

Infomap is superior against the most popular community detection algorithms. According to Lancichinetti and Fortunato (2009), Infomap has an excellent performance and low computational complexity, which makes it capable to perform large-scale data analysis with graphs containing millions of nodes and links. They tested the performance of several algorithms utilizing an extension of the Girvan and Newman (GN) benchmark, the Lancichinetti-Fortunato-Radicchi (LFR) benchmark. Regarding the GN benchmark, it assesses a model using a simple network model, a graph of 128 nodes, each with the expected degree of 16, and four groups. Lancichinetti et. al (2009) say that this benchmark utilizes a network with two assumptions that are not common among non-artificial networks. First, all the nodes have the same expected degree. Second, all groups have the same size. With respect to the second benchmark, LFR generalizes the GN benchmark introducing power law distributions of degree and community size. Because most of the community detection algorithms perform well

in the GN benchmark, the LFR benchmark allows testing algorithms in more challenging graphs, which help to identify algorithms' limitations. In Lancichinetti and Fortunato's test of the most common community detection algorithms, Infomap has the best performance in random graphs in both GN and LFR benchmarks.

Infomap is very efficient when graphs contain defined communities. Thus, when there are clear clusters in the data, the transitions of a random walker from one cluster to another should be rare. However, if the graph does not contain well-defined clusters or if the partition is not appropriate, the transitions between clusters should be frequent. Consequentially, there is a limited gain of using a two-level structure (town—street) to describe the graph (Fortunato & Hric, 2016).

Infomap is useful to identify diverse communities, such as those build from academic citations. According to Bohlin et. al. (2014), this algorithm can identify two-level, multi-level, and overlapping organization in weighted, directed, and multiplex networks. Then, "it naturally captures flow of ideas and citation flow, and is therefore well-suited for analysis of bibliometric networks" (3). Infomap eliminates the resolution limit —the algorithm' inability to detect small communities and aggregate them in bigger groups— for networks with nested multilevel modular structures when the hierarchical map equation is used. Then, Infomap is better to deal with the resolution limit against modularity (Kawamoto & Rosvall, 2015).

In spite of a couple of limitations, Infomap is currently the best algorithm for community detection in large networks. Lancichinetti and Fortunato (2009) say that Infomap is not able to analyze overlapping communities, which was a common problem

of the most popular algorithms when they did their analysis. Moreover, Lancichinetti et. al (2009) argue that more analysis was needed to assess algorithms using hierarchy graphs — with communities inside communities. However, after recent changes in the algorithm —including more than two levels—, the analysis of Bohlin et. al. (2014) indicates that both problems are not relevant anymore. In summary, Infomap is the best method available for this analysis.

On practical issues, the input of Infomap is a list of tuples of nodes. For instance, [(1,3), (1,4), ...]. In this case, the first two tuples mean that node 1 has connections with node 3 and 4. For this study, the direction of the connection is not considered. Additionally, the output of Infomap is a hierarchical tree, where the most relevant nodes are in the top branches and the less connected ones are in the last leaves. Then, the hierarchical tree is made of tuples, where each tuple contains (1) the name of a node, and (2) the position of that node in the tree. For example, (1, "1:1"), (3, "1:1:1"). In these cases, node 1 is in the position "1:1" and node 3 in the position "1:1:1." Consequentially, node 3 is inside the branch where node 1 is at the top.

3.3 Creation of scientific fields

In order to create groups of papers that represent an academic field, I dynamically prune the tree that is the output of Infomap. This output is a hierarchical graph where each node is a paper. Then, the most central, relevant, or cited papers are the top nodes of the hierarchical structure. In other words, the most important papers

are close to the trunk and the least relevant papers are the last leaves. Considering this tree structure, I create an algorithm to cut the tree, which is described in the next paragraphs.

To define and order specific academic fields, I execute a two-step process. First, I prune the tree —the output of Infomap— and define specific groups. Second, I reorder the groups with a recursive algorithm, which respects the structure of the Infomap's tree.

3.3.1 The algorithm to prune the tree

The input of this algorithm is a hierarchical tree and an integer η —desired number of elements per cluster–. In this paper, I use a $\eta = 1000$. Note that each branch contains several other branches, where each branch is potentially a group. The algorithm asses each of the branches according to their leaves,[2] recursively, according to the following rules:

1. All leaves with a number of members of at least η but below $2 * \eta$ are considered a group.

2. All leaves with a number of members below η are processed according to the following:

[2]Each leave is also a hierarchical tree, which can be one only one node or a tree of several levels.

(a) If the total number of members in leaves which members are below η is below $5 * \eta$ and above $\frac{1}{3} * \eta$, then, all those leaves are merged and assigned to only one group.

(b) If the total number of members in leaves which members are below η is above $5 * \eta$, then, the leaves with at least $\frac{1}{3} * \eta$ members are considered one group. However, the leaves with less than $\frac{1}{3} * \eta$ are dismissed.

(c) If the total number of members in leaves which members are below η is below $\frac{1}{3} * \eta$, then, those papers are not assigned into a group.

3. All leaves with a number of members above $2 * \eta$ are explored in the next level according to the following:

   (a) If in the next level all groups in the leaves are below η, then, the entire original leave —the one above $2 * \eta$ members— is saved as one group.

   (b) If in the next level there is at least one group above η members, the groups are created according to the rules 1 and 2.

4. The steps from 1 to 3 are repeated until all members of the tree are either assigned to a group or discarded according to the rules.

3.3.2 The algorithm to order groups

The process to prune the tree assigns publications —papers— to groups. However, it does not guarantee the best group ordering. That is to say, it does not ensure that the most similar groups are close to each other.

Groups have an implicit order because they come from pruning the hierarchical tree. For instance, Infomap assigns a place in the hierarchical tree for each paper, such as 1:3, then, that paper belongs to the branch 1 and it is the 3rd member of the branch. The output can have several levels, such as 1:1:1:1:1:3. Consequentially, the implicit order of a group which main nodes are 1:1, 1:2, and 1:3 is 1:1, 1:2, and 1:3. However, it is possible to increase similarity among adjacent groups if we reorder the groups. For example, one of the possible improved orders can be 1:1, *1:3*, and *1:2*. Note that this change preserves tree structure but moves leaves' order.

In order to increase the similarity among adjacent groups, I created another recursive algorithm to order groups. The steps are the following:

1. Create a matrix Σ, where each row is one group and each column is one discipline. This matrix has 274 columns, the total number of disciplines in the classification provided by the WoS. In the WoS, each paper is tagged to at least one specific discipline, for papers with available information. The element $\Sigma_{i,j}$ is the proportion of occurrences of discipline j in group i. For instance, if $\Sigma_{i,j} = 0.1$, it means that the occurrences of discipline j is 10% of the total number of tags related to that group. One example, if there are two papers in a group and the first paper is classified as "zoology" and "surgery," and the second is classified only in "zoology." Then, the proportion of "zoology" in this group is 2/3, and the proportion of surgery is 1/3.

2. Recursively, order top branches, intermediate branches, and finally leaves. Note that the matrix Σ is at the group level, however, each group has a unique code,

which comes from the structure of the hierarchical tree. Thus, to order the groups recursively, I order first the top branches. To create a new order for the top branches, I need to build a new aggregate matrix $\hat{\Sigma}$ where the columns are still the disciplines, but the rows are the branches. Therefore, the element $\hat{\Sigma}_{\hat{i},j}$ is the proportion of occurrences of the discipline j in the branch $\hat{i}$. The matrix $\hat{\Sigma}$ feed another algorithm —described in the next section— that returns a better order. The process of matrix generation and ordering is repeated at different levels of aggregation, recursively, ordering from the top branch to the last leave. Thus, this process generates a new order for all groups but it preserves the main structure of the original Infomap output.

3.3.2.1 Algorithm to order top branches, intermediate branches, and leaves

I tried three different approaches to order the groups per level: recursive t-SNE and KMeans, recursive Infomap, and the strings method. However, I only report the methodology and results for recursive t-SNE and KMeans, which was the method that showed significantly better results.

t-SNE and KMeans

I use a three-step process to create a new order for the groups. First, I reduce the dimensionality of the matrix Σ using t-SNE. Second, I utilize KMeans to assign each of the groups into a cluster. Third, I reorder the groups taking into account two

factors: (1) the cluster they were assigned by KMeans, and (2) the original ordering which corresponds to Infomap's output.

Considering the above mentioned, all groups assigned to the same cluster according to KMeans are next to each other. Yet the order inside those clusters is defined by the implicit order provided by Infomap's output. For instance, if all groups were "1:1", "1:2", and "1:3," the groups "1:1" and "1:3" were assigned to *cluster 1* by KMeans, and "1:2" to *cluster 2*. Then, the final order of the groups will be "1:1", "1:3", and "1:2." Note that "1:1" and "1:3" are first on the list because they belong to the first cluster according to KMeans —which does not have an implicit meaning—, and "1:1" is before "1:3" because the new order considers the original Infomap's order as a second reference to arrange the groups. Also, "1:2" is the last group because it belongs to *cluster 2* according to KMeans.

3.4 Assess groups' order

In the previous section, I describe how to reorder groups using the structure of Infomap's output. In the following paragraphs, I explain how to asses if a group order is better than another.

To assess groups' order, I utilize modularity. I consider that a better order is when similar groups are close and dissimilar groups are distant. In order to evaluate similarity among neighbors, I merge adjacent groups to create *modularity groups*

—groups of groups—. Afterward, I calculate the modularity of the partition using *modularity groups*. According to this methodology, a partition with higher modularity is better than a partition with lower modularity. Please find a detailed explanation below.

The methodology to create *modularity groups* from groups is as follows. First, I define a level of analysis, which is the level to merge groups. For instance, in the group r-4-5 —root, four, five—, level 1 means the first level after root, the level of the 4. Note that the main branches belong to level 1. Second, I create rules to merge adjacent groups which are summarized below:

1. All adjacent groups which share the same group in the defined level of analysis should be merged. Example:
    - If the level of analysis is 1 and the groups r-4-5 and r-4-7 are next to each other, those groups will merge because they share the number 4.
2. Groups which do not share the same group in the defined level of analysis, but are between two clusters of groups that are merged due to rule 1 should be merged. Example:
    - If the level of analysis is 2, and the sequence of groups is r-4-5, r-4-7, r-3-7, r-8-9, r-5-1, r-5-5. Then, groups r-4-5 and r-4-7, as well as r-5-1 and r-5-5, are merged by rule 1. Therefore, r-3-7 and r-8-9 should be merged.
3. Groups which do not share the same group at the desired level of a analysis, but are at the beginning or ending of a groups sequence. Example:

- If the level of analysis is 1, and the sequence of groups is r-3-5, r-7-2-1-2, r-4-5, r-4-6. Then, groups r-4-5 and r-4-6 are already merged by rule 1. Consequentially, groups r-3-5 and r-7-2-1-2 should be merged.

Finally, to compare two different orderings, I create *modularity groups* for each of the orderings. Then, I calculate the modularity of the *modularity groups*. The ordering with a higher modularity is considered to be better.

3.5 Ties of academic fields over time

In the previous section, I describe the process of creating groups of articles per year —academic fields—. This section explains how to connect groups of adjacent years utilizing two criteria. First, the number of overlapped papers in groups θ_t and θ_{t+1} divided by the total number of papers of θ_t is the highest against any other group in $t+1$. Second, the number of overlapped papers in groups θ_t and θ_{t+1} divided by the total number of papers of θ_{t+1} is the highest against any other group in t. Then, connected groups are the best match from past to future and the best match from future to past.

Note that this process creates the multi-slice network. The groups per year are created with the algorithm to prune the tree —one tree per year. However, ties between groups in adjacent years generate connections to complete the multi-slice network, where each slice is one year, and relationships among groups are the ties

among slices.

3.6 Continuity of subjects

The ties across different years allow tracking changes in the prevalent topic in academic literature. This relative importance of topics can be analyzed at different levels. For example, it is possible to analyze very broad disciplines, such as medicine 1985-r-1 —year 1985, root, one—, or very specific fields, like cluster 1985-r-1-4-4 related to skin cancer. Given the scope of this thesis, I will only present results for the most aggregate academic fields. That is to say, for the main branches of the trees.

The methodology to identify and track the evolution of top subjects per branch is as follows.

1. I create branches —groups of groups—, which correspond to the top levels of Infomap's trees. Given that the analysis is only at the top of the braches, I extract from the group name the number related to the first level of the tree. For instance, in the group r-1-3-4, I extract 1, which indicates belonging to branch 1. Then, all groups with a 1 in level 1 are considered members of branch 1.

2. I get the top subjects by group of groups. For each branch, I extract the subject classification of its papers according to the WoS database. Then, I obtain the

top subjects related to each branch, which are the tags more frequently assigned to papers in each branch. Each tag is a specific field, such as "Nuclear physics" or "Film, radio and television." In this thesis, I utilize the top four subjects per branch.

3. I repeat the previous two steps for all groups and years.

4. Once I define branches per year, I create relationships among different years. I use data about ties of academic fields over time[3] —for example, that group 1990-r-1-2-4 is connected with group 1991-r-1-3-4— to define which top branch corresponds to another top branch in a different year. I consider that the evolution of top branch β_t is β_{t+1} if most of the groups in β_t evolve in groups contained in β_{t+1}. That is to say, if the number of group connections to β_{t+1} is higher than connections to any other branch in $t+1$.

5. In case the former step generates that one branch in $t+1$ connects to more than one branch in t, the algorithm will connect the branch in t with the biggest branch in $t+1$. That is to say, the branch in t will connect with the branch in $t+1$ with the lowest number in level 1. Note that this step prevents that one branch connects with more that one branch in the next adjacent year.

[3]The explanation about how to create this ties are in 3.5.

Chapter 4

Empirical Results

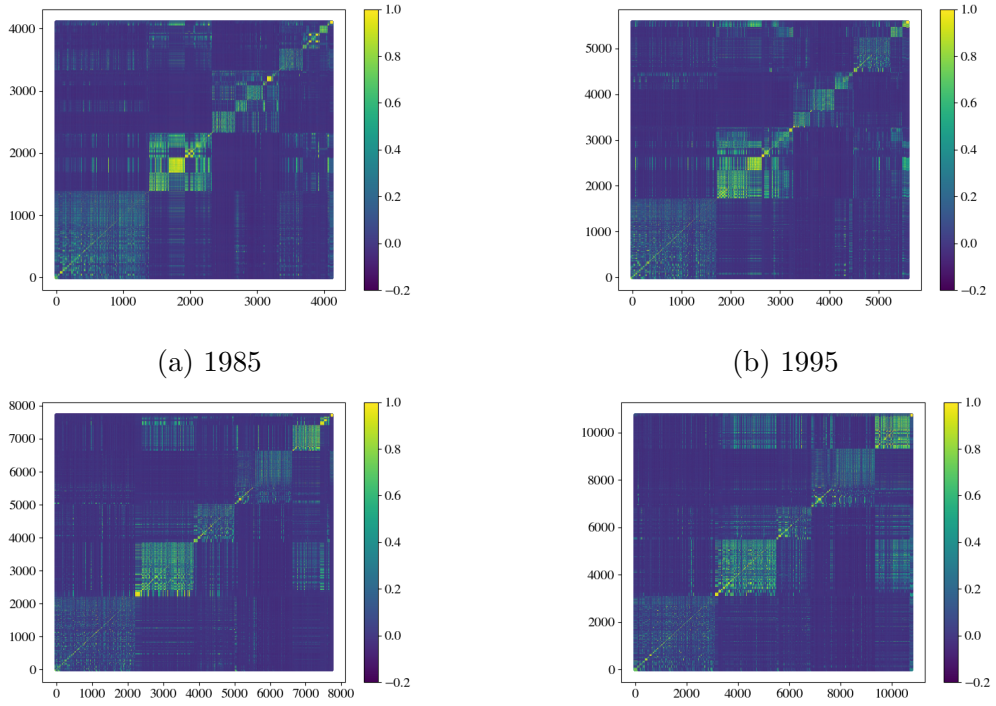
4.1 Assessment of scientific fields

In this section, I make an assessment of the group creation process described in Chapter 3. Ideally, we would like two features in our cluster of papers. First, papers in the same cluster relate to each other. Second, similar groups are close to each other and dissimilar groups far away. Consequentially, correlation plots among groups should depict similarity patterns across the line from the southwest corner to the northeast corner.[1]

The graphs 4.1 and 4.2 show correlation plots for 1985, 1995, 2005, and 2014,

[1]The process to create vectors which describe each group is in Chapter 3.

where the x and y axes are groups detected through the algorithm explained in 3.3. As expected, squares along the line from the northeast corner to the southwest corner indicate that the algorithm was successful to cluster groups of similar subjects. That is to say, groups of papers which deal with the same subjects are close to each other. Also, groups of papers of dissimilar subjects are not close to each other. This pattern is consistent for Pearson and Spearman correlations.



(a) 1985

(b) 1995

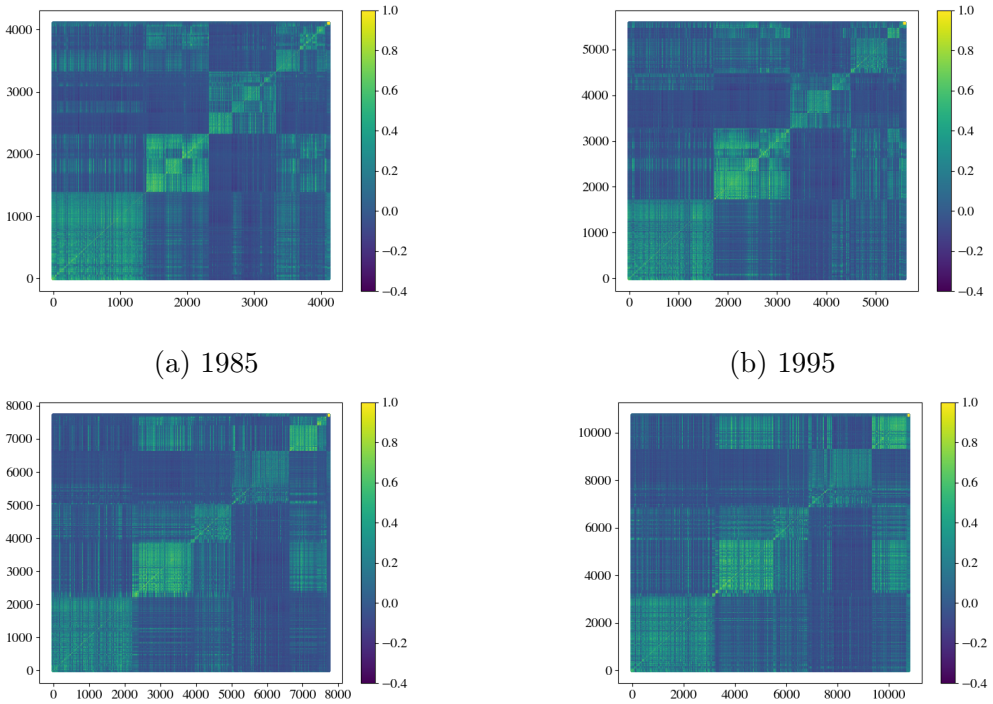
(c) 2005

(d) 2014

Figure 4.1: Pearson Correlation Plots

One interesting feature in 4.1 and 4.2 is that the pattern of colors in 2014 is not as contrasting as in the first years of the analysis. For instance, the contrast between deep blue and bright yellow is higher in 1985 than in 2014. Therefore, apparently, the range of topics in one group in 2014 have more in common with more groups than 30

years ago. There are two hypothesis about this. On one hand, Science has become more interdisciplinary, then, the frontiers across groups have become more diffuse. On the other hand, maybe the process to register subjects in the *WoS* database has changed over time. It is difficult to make a strong conclusion with the available data.



(a) 1985

(b) 1995

(c) 2005

(d) 2014

Figure 4.2: Spearman Correlation Plots

4.2 Assessment of groups' order

In this section, I show two pieces of evidence that help to assess which is the best ordering for the groups, the Infomap implicit order or the reordering with t-SNE and

KMeans, which methodology is described in 3.3.2. First, I describe correlation graphs as a visual introduction. Second, I explain the results of a quantitative test to assess this issue.

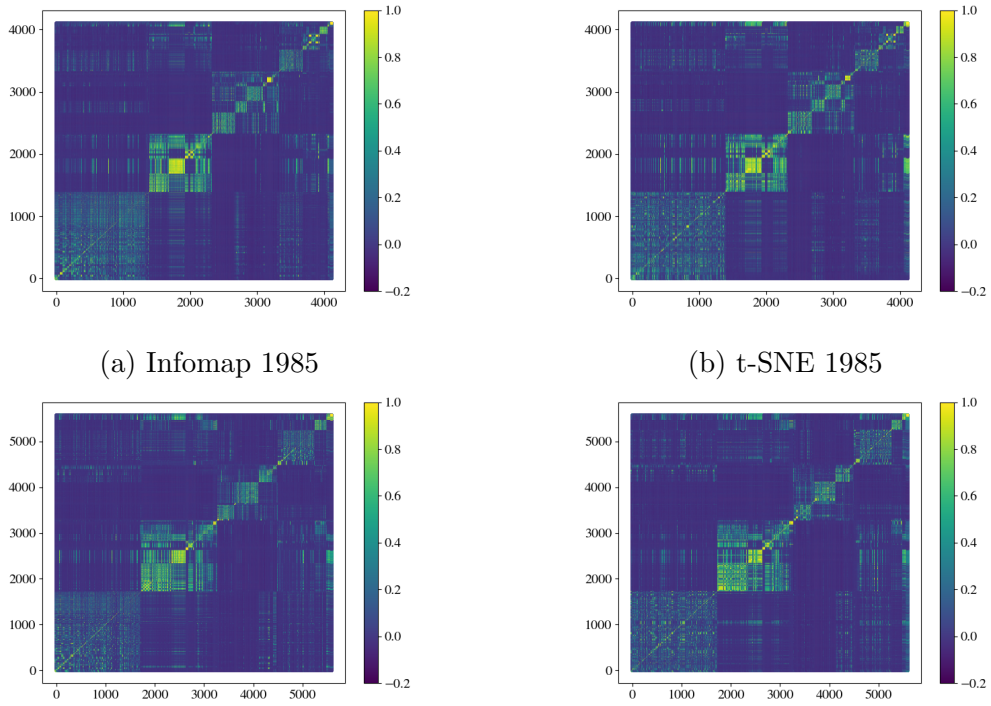
4.2.1 Correlation graphs

The graphs 4.3 and 4.4 shows slight differences on the Pearson correlation graph between Infomap and t-SNE + KMeans. The clustering among similar groups is not evidently better in any of both cases. In the same vein, for certain specific groups and years, the clustering appears to improve in the t-SNE case. However, in other cases, the trend appears to be the opposite. The Spearman correlation graphs in Appendix A.1 depict the same trend. Thus, a visual evaluation is not enough to assess which approach generates the best order.

4.2.2 Modularity of *modularity groups.*

The recursive process of group ordering described in 3.3.2, which utilize t-SNE and KMeans, increases the modularity of the modularity groups. Then, according to the methodology described in 3.4, the process to reorder groups is a better choice against only use the order of Infomap. The graph 4.5[2] shows that the ordering of t-SNE and KMeans increases modularity in 96% of cases against only Infomap. In 1989,

[2]The year 2003 is omitted due to inconsistency of the data.



(a) Infomap 1985

(b) t-SNE 1985

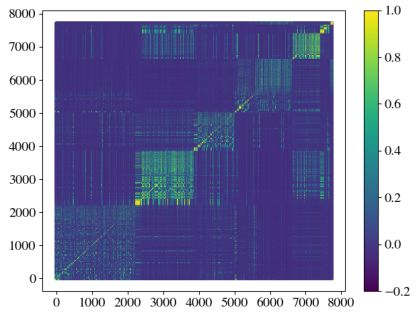
(c) Infomap 1995

(d) t-SNE 1995

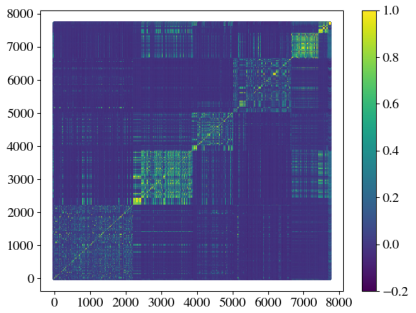
Figure 4.3: Pearson: Infomap vs t-SNE — 1

the modularity of only Infomap is higher than t-SNE and KMeans, however, the difference is minimal.

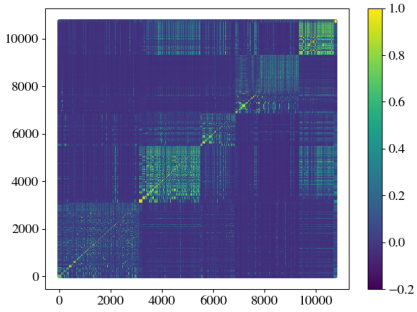
Tu sum up, the method to reorder groups preserving the structure of Infomap's tree is successful to put similar groups together and dissimilar groups far apart. Although in most of the cases this process provides better results against the baseline, the magnitude of the modularity improvement is modest. However, further research is needed to improve the methodology and get better results. For instance, some possible changes are creating a dynamic function to set the number of clusters in KMeans and the number of components in t-SNE, as well as experimenting with



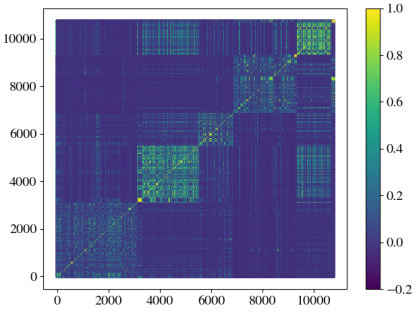
(a) Infomap 2005



(b) t-SNE 2005



(c) Infomap 2014



(d) t-SNE 2014

Figure 4.4: Pearson: Infomap vs t-SNE — 2

different combinations of parameters, such as early exaggeration and perplexity.

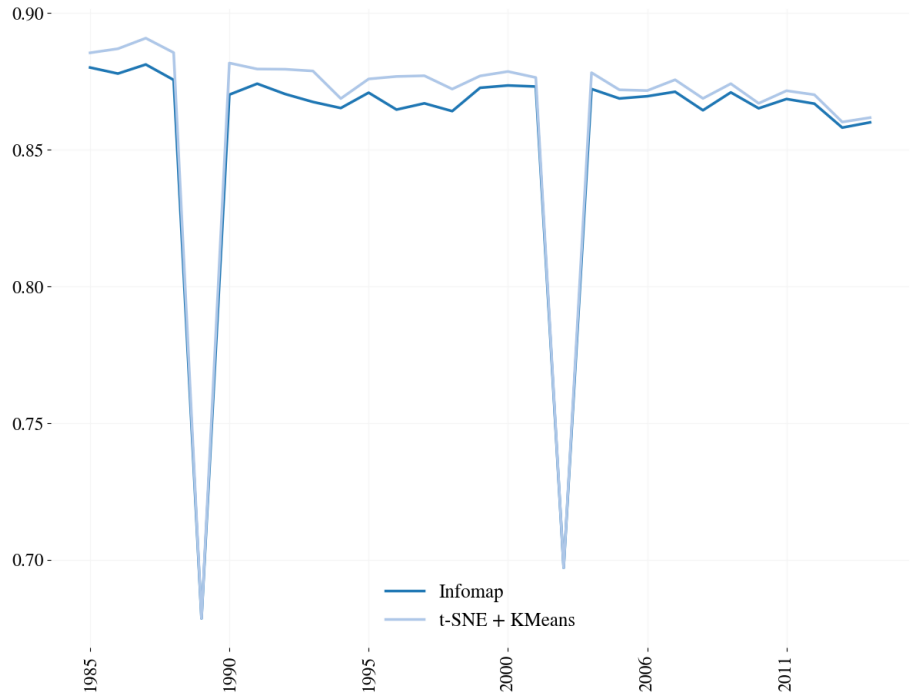


Figure 4.5: Modularity of the *modularity groups* partition.

4.3 Continuity of subjects

In this section, I show the most important subjects related to each of the intertemporal groups at level 1.[3] That is to say, the most relevant topics associated with the most aggregated clusters of science, those in the main branches of the trees. I name those branches as broad academic fields to facilitate the explanation. However, the data is not limited to a specific set of papers or disciplines.

On the continuity of topics, there are three main trends. First, some clusters have continuity across several years although its subjects evolve. Second, some groups are

[3]I describe the process to create intertemporal ties in 3.5.

preserved for a couple of years. Afterward, they do not appear in the main branches of the next year. Third, small groups of papers emerge one year, but they do not evolve into another branch next year.

Note that sometimes I analyze two clusters that appear to be the same broad discipline. However, I consider them as independent groups because the algorithm does not connect them across all years. This might be as a result of structural changes, but further research is needed to assess this possibility. One consequence of this approach is that graphs have empty spaces, however, this allows me to show the dates where these clusters appear and disappear. Also, I only consider the top four subjects per discipline in the graphs.

4.3.1 Continuity across several years

The cluster with the longest continuity is *Medicine*. Interestingly, this cluster is the only one preserved in all the thirty years of this study. It is relevant by its continuity, as well as by its size, the biggest cluster for 1985. As shown in 4.6, there are subjects that have been relevant in the last thirty years, such as biochemistry and molecular biology. In contrast, there are subject matters which relevance has decreased, such as general and internal medicine. However, other topics have increased its relative importance, such as surgery and cardiology. Finally, there are subjects which have intermittent patterns, such as pharmacology and pharmacy.

Natural Sciences 1 during the nineties is characterized by its stability, as depicted

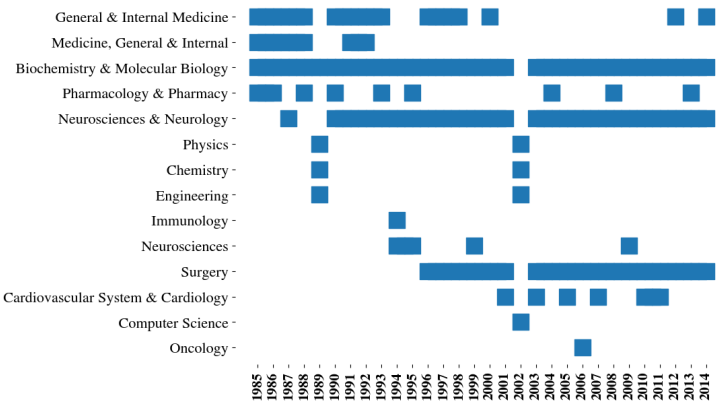


Figure 4.6: Medicine

in 4.7. The main disciplines are physics, chemistry, and material science. Also, note that during the second half of the nineties engineering substituted applied physics as one of the top four subjects.

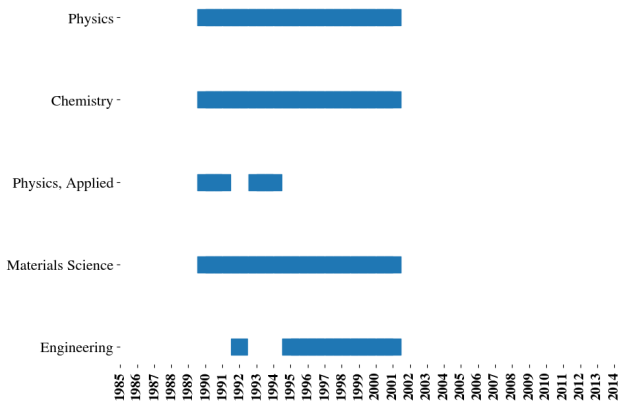


Figure 4.7: Natural Sciences 1

After 2002, there are two groups that look like the evolution of *Natural Sciences 1*. First, note that 4.8 shows a similar trend as *Natural Sciences 1*, however, engineering become one of the top disciplines in most of the years after 2002. Second, a new group emerges as shown in 4.9. An interesting feature of this second group is that the

top disciplines are mainly related to physics, such as astrophysics, particle physics, or interdisciplinary physics. Then, apparently, physics literature becomes sufficiently big, relevant, or isolated to create its own cluster.

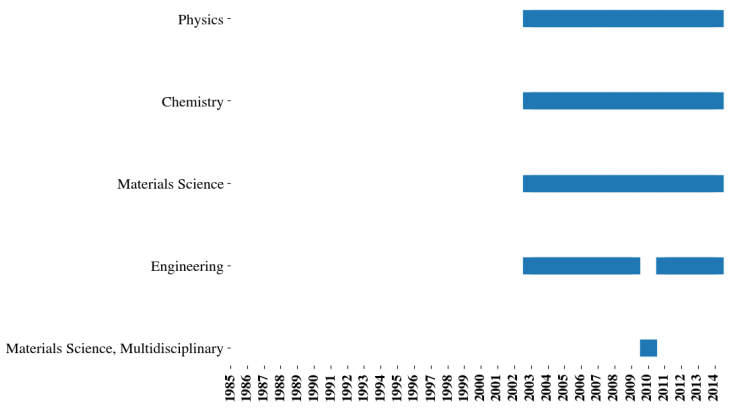


Figure 4.8: Natural Sciences 1: Evolution 1

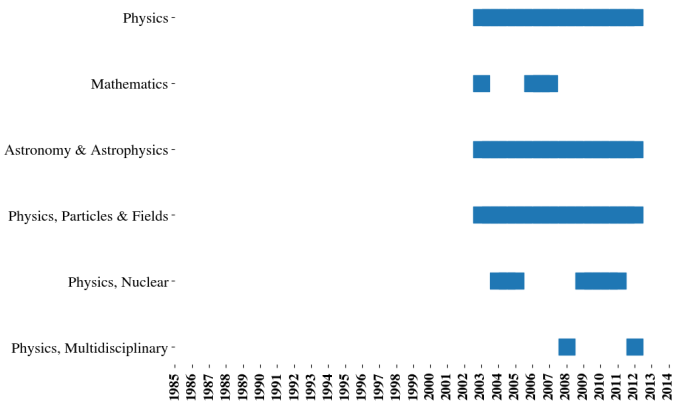


Figure 4.9: Natural Sciences 1: Evolution 2

Natural Sciences 2 during the nineties shows a clear predominance of environmental science and ecology, with geology also among the top subjects in most of the years, as shown in 4.10. Most of the other disciplines are intermittent, however, at the end of the period, engineering appears as one of the top disciplines. After 2002, there is a new group depicted in 4.11, which top subjects are very similar to those

in *Natural Sciences 2* during the nineties. However, agriculture is not consistently among the top disciplines, and engineering becomes one of the top disciplines after 2008.

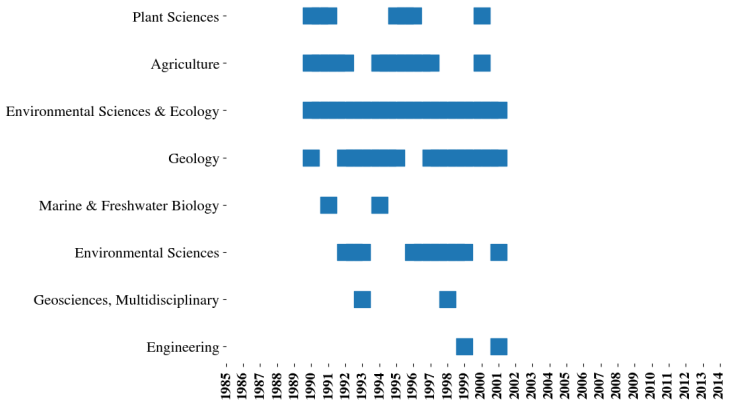


Figure 4.10: Natural Sciences 2

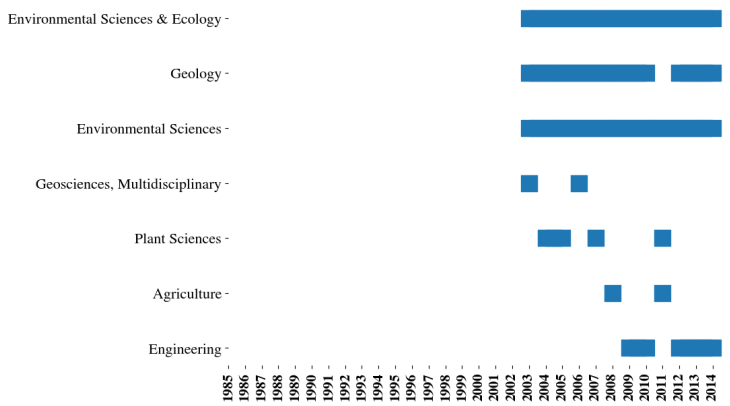


Figure 4.11: Natural Sciences 2: Evolution in the two-thousands

The *Computer Science et al. 1* cluster during the nineties shows stability in the top disciplines, such as mathematics, computer science, and engineering, as depicted in 4.12. However, interestingly, artificial inteligence becomes one of the top subjects at the beginning of the two-thousands. After 2002, *Computer Science et al. 2* contains

very similar topics, as illustrated in 4.13. Artificial intelligence continues in the top four disciplines until 2007, when the relative importance of mathematics increases.

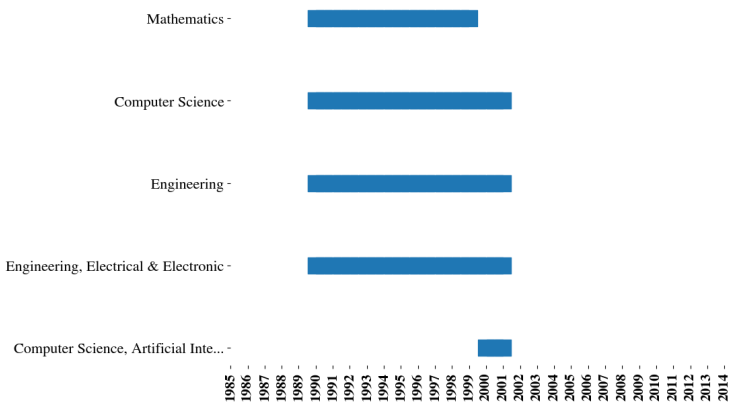


Figure 4.12: Computer Science et al. 1

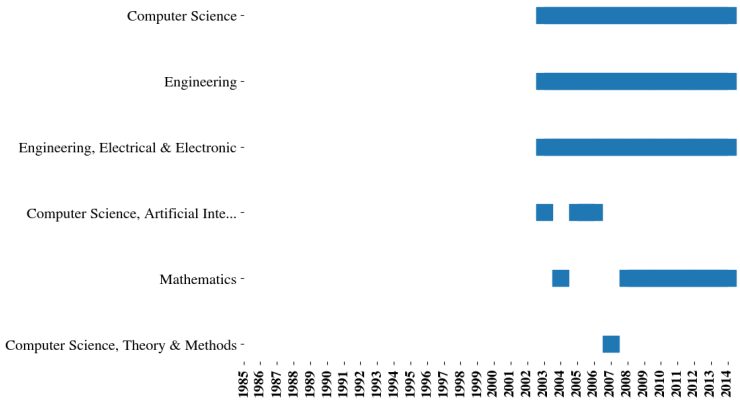


Figure 4.13: Computer Science et al. 1: Evolution

Social Sciences 1 is dominated by economics, business, and psychology during the nineties according to 4.14. At the beginning of the nineties, humanities papers were among the top four disciplines, which were substituted by history, psychiatry, and neurosciences at the end of the nineties. The cluster *Social Sciences 2*, depicted in 4.15, shows the evolution of social sciences after the nineties. As in the nineties, the most important disciplines are economics, business, and psychology. However, new

subjects, such as government and law, management, education, neurosciences, and computer science —as one main topic inside the social sciences!— emerge as relevant disciplines inside this cluster. Thus, this shows the dialogue of social sciences with applications and methodologies that used to be outside of the main topics for the social sciences.

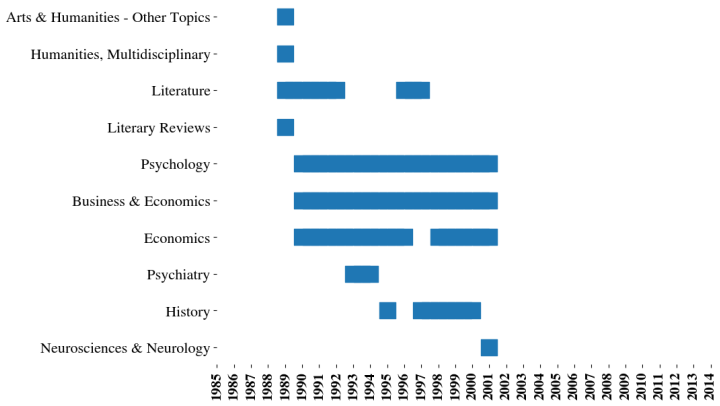


Figure 4.14: Social Sciences 1

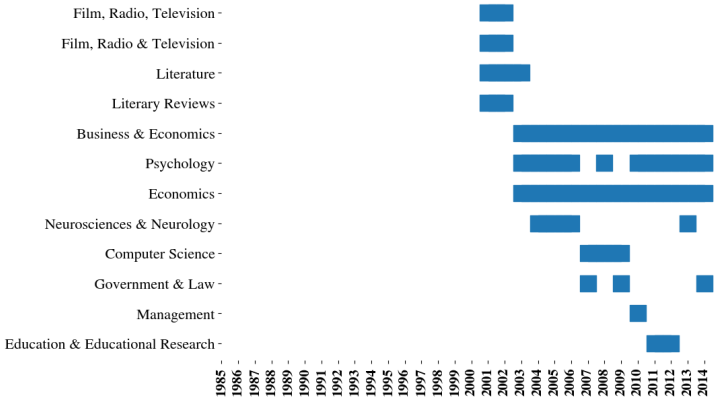


Figure 4.15: Social Sciences 2

4.3.2 Continuity for less than ten years

Several groups last less than ten years. Consequently, the total number of graphs to show each of them is huge. Then, for this section and the next, I show most of the graphs in the appendix B.2.

Some groups stay in the main branches for less than ten years but more than two years. Two main disciplines relate to these groups. On the one hand, some humanities' groups last less than five years as shown in B.1. Those branches relate to specific topics on literature, such as German, Dutch, and Scandinavian literature, or African, Australian, and Canadian Literature. In the same vein, others belong to the arts, such as film, radio, and television, or scenic arts, such as dance, theater, and music. Interestingly, there is a small cluster of history in the middle of the two-thousands. On the other hand, natural sciences' groups characterized by its applied nature. For instance, 4.16 shows in (a) and (b) subjects such as applied chemistry, pharmacology, and science and technology. Likewise, (d) contains topics like food science, veterinary sciences, agriculture, and animal science.

In addition, *Public Health and Medicine* contains subjects related to applications of medicine, as depicted in 4.17. For instance, health care sciences, education research, and public, environmental and occupational health. To sum up, groups which stay as independent branches for a few years tend to study applied subjects. For example, specific literature topics in humanities, technology in natural sciences, and public health in medicine. Further research is needed to understand the process that generates this patters, and if those groups merge with the main branches or gradually

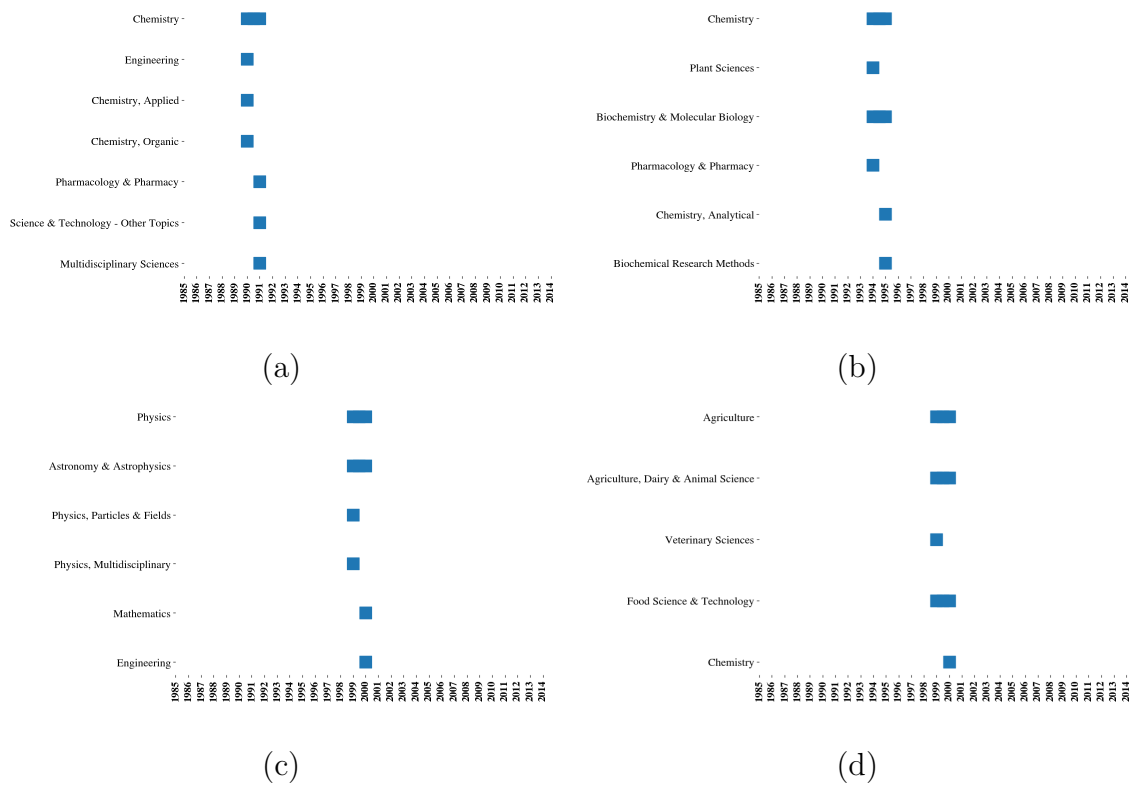


Figure 4.16: Exotic Natural Sciences

disappear.

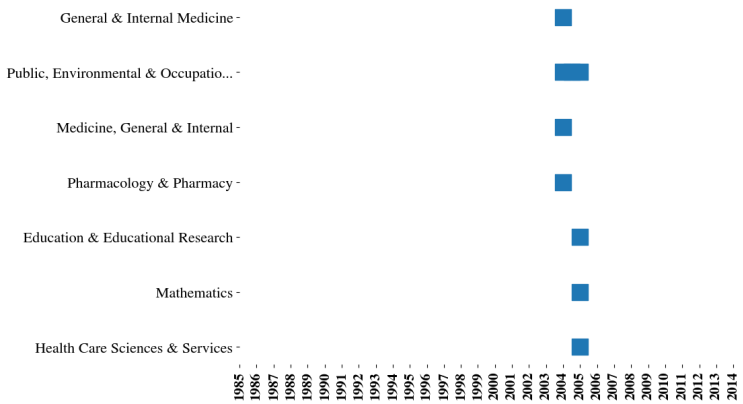


Figure 4.17: Public Health and Medicine

4.3.3 Continuity for one year

Interdisciplinary medical clusters last only one year, such as B.2. Overall, the size of the main medicine cluster each year is the biggest among those in the main branches. Then, medicine groups outside the main medicine branch are surprising. These groups are specific and interdisciplinary. For example, the group (a) comprises occupational health, psychology, and internal medicine; cardiology and physics; and (c) material science, education, and pharmacy.

Among groups which last only one year, the most repeated discipline is chemistry. In most of the cases, these groups are applied science, such as food science and agriculture depicted in 4.18 and B.3. Interestingly, these subjects repeatedly appear among groups which last only one year. Then, maybe this pattern reflects that discoveries in some academic fields —and the submission of papers— come in waves. However, several groups linked with chemistry are not evidently applied but they deal with more specific knowledge against groups which last longer, see B.4. For example,

some subjects in these groups are molecular biology, computational chemistry, or spectroscopy.

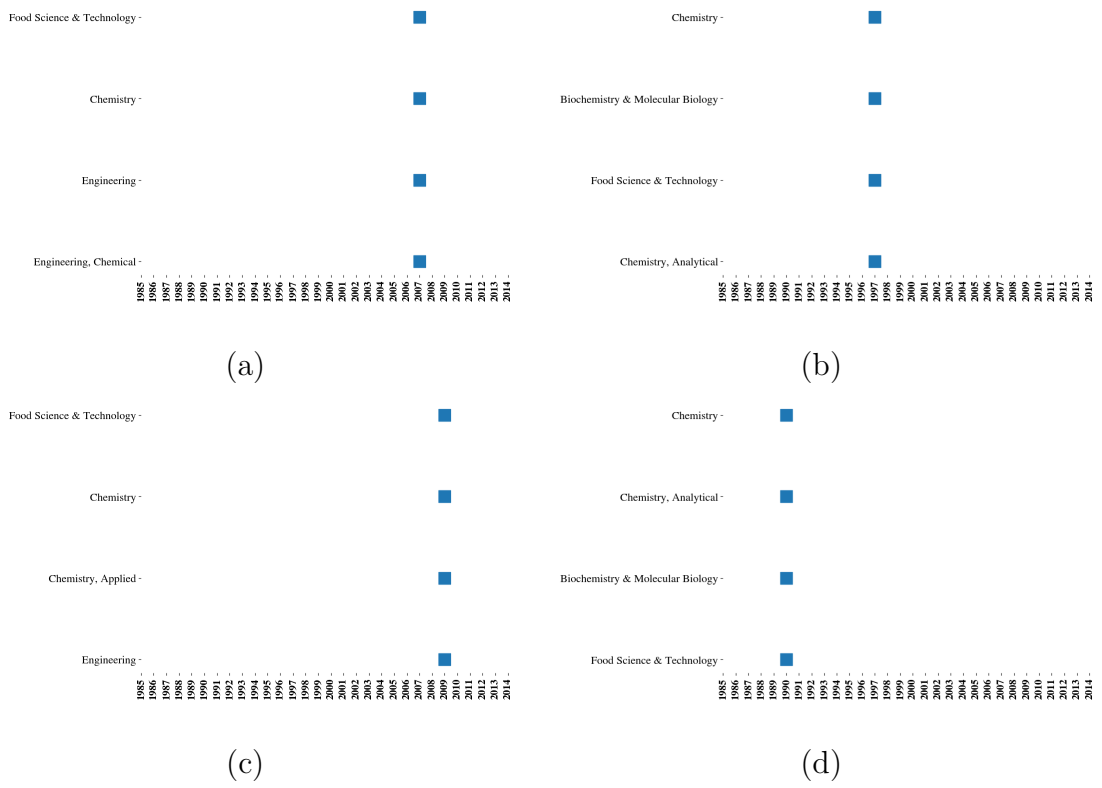
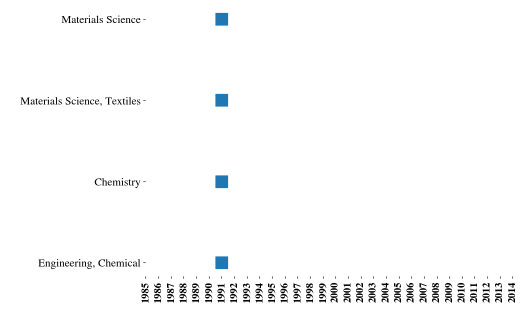


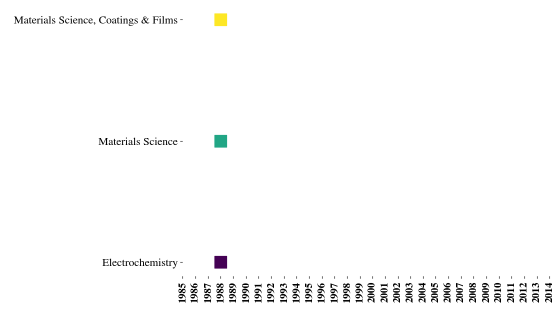
Figure 4.18: Food Science

Engineering stands out as another discipline with several groups that last one year. The figures 4.19 and B.5 show two subfields of engineering. On the one hand, material science, which specific topics change by year, coatings and films in 1988, textiles in 1991, and biomaterials in 2003. On the second hand, mechanical engineering with slightly differences in three years, 1999, 2007, and 2009.

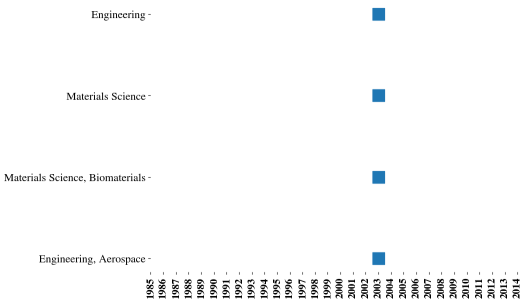
Exotic natural sciences only last one year. For instance, geology and science and technology in 1985, applied mathematics for chemistry in 1988, and computational



(a)



(b)



(c)

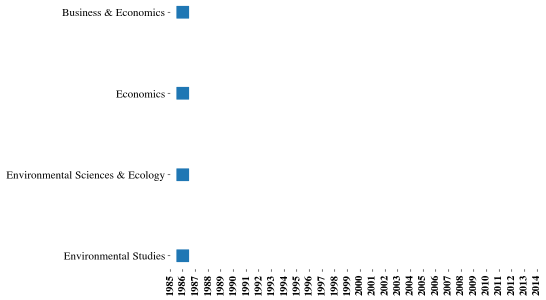
Material Science groups with continuity of only one year.

Figure 4.19: Material Science

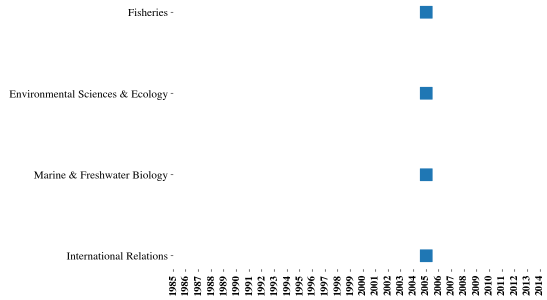
biology in 1997, depicted in B.6. One hypothesis to explain this pattern is that these groups merged with the main groups of science in the following years. Alternatively, those are big and atypical observations of groups that are smaller in other years, then, those are not detected by the algorithm in other years.

Diverse environmental disciplines last one year. These groups, depicted in 4.20, combine environmental sciences with another discipline. For instance, one of the groups uses chemistry and law in 1985, economics in 1986, and fishery and international relations in 2005. Then, apparently, certain topics on environmental sciences and ecology become important in certain years, and those are captured by the algo-

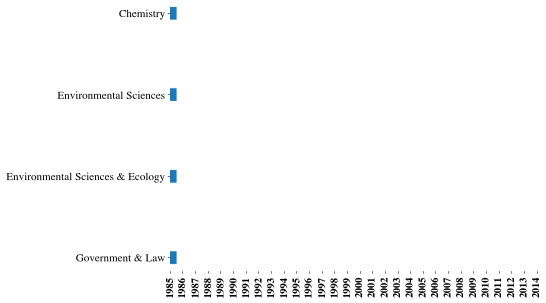
rithm.



(a)



(b)



(c)

Environment groups with continuity of only one year.

Figure 4.20: Disciplines related to the Environment

Finally, peculiar humanities groups last one year, see B.7. For instance, there are groups in 1997 and 2003 which deal with African, Australian, and Canadian Literature and multidisciplinary humanities. In the same vein, another group in 1991 studies multidisciplinary humanities, microbiology, and pharmacology. Then, intuitively, some multidisciplinary humanities are isolated and they do not belong to the main humanities cluster. However, they create its own cluster certain years.

Chapter 5

Conclusions

This thesis develops a framework to analyze the evolution of science with multi-slice networks, which allows us to move beyond analyzing snapshots at specific points in time to making hypotheses and analyses comprising several years. My work explores some of the possibilities that this framework provides.

The correlation graphs show that the two-stage process to create the science clusters is successful. Groups which are similar to each other are close to each other and those that are dissimilar to each other are far apart. This trend is consistent for both Pearson and Spearman correlations.

The reordering of groups which utilize t-SNE and KMeans increases the similarity among groups close to each other. This ordering respects the structure of Infomap's

tree. That is to say, it changes the order of the branches, but it does not allow to move elements from one branch to another branch in any level of the tree. However, more research is needed to optimize this process. In particular, creating a dynamic process to set up parameters.

Among the groups which last several years, the most stable cluster is *Medicine*. An analysis of the evolution of the relative importance of topics in this group, and how those relate to diseases or funding strategies is a promising area of research. *Social Sciences 1* and *Social Sciences 2* show an interesting pattern on changes of the most relevant subjects in the field. For instance, at the beginning of the nineties, humanities were among the top four disciplines, which were substituted by history, psychiatry, and neurosciences by the beginning of the millennium. In the same vein, one unexpected result is that computer science was among the top subjects in the social sciences for four years during the two-thousands, which might indicate structural changes on how social sciences research is done.

Most of the groups that last less than ten years belong to humanities, natural sciences, and medicine. In all cases, these clusters are applied subjects or related to very specific topics, such as German, Dutch, and Scandinavian literature, pharmacology, science and technology, food science, or occupational health.

The groups that last only one year are relatively exotic. For instance, medicine groups are unusual because they investigate in the intersection of cardiology and physics, or material science, education, and pharmacy. Further, other groups study specific material science topics, such as coatings and films, textiles or biomaterial.

Moreover, natural sciences groups research computational biology or geology and technology. Finally, humanities groups that examine microbiology and pharmacology.

Among the groups that last only one year, some clusters research on applied science, such as food science and agriculture. In the same vein, other groups study environmental sciences along with other fields, like economics, marine biology, international relations, or government and law.

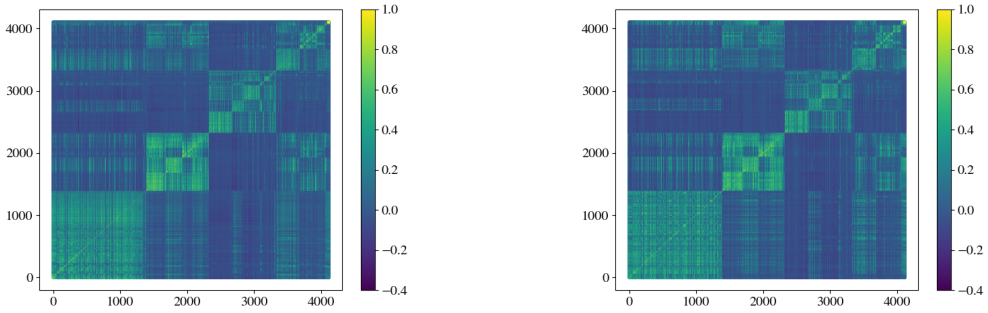
Overall, the groups with last less than ten years tend to be more specific, unusual, or interdisciplinary. A future area for research is analyzing the origin of those groups. Those can result from the evolution of some part of the main branches or can appear spontaneously. Also, the end of those groups is another interesting topic to investigate. It is an open question if those clusters integrate into other disciplines in the big branches, or if they gradually disappear.

Several topics can be explored with the methodology developed in this thesis. For instance, slight modifications in the algorithms allow analyzing the emergence of interdisciplinarity, identify authors that publish in several unassociated clusters, and track patterns behind the creation of academic fields from the ideas of diverse or unconnected disciplines. This will enable us to track the evolution of how an idea becomes an academic field on its own.

Appendix A

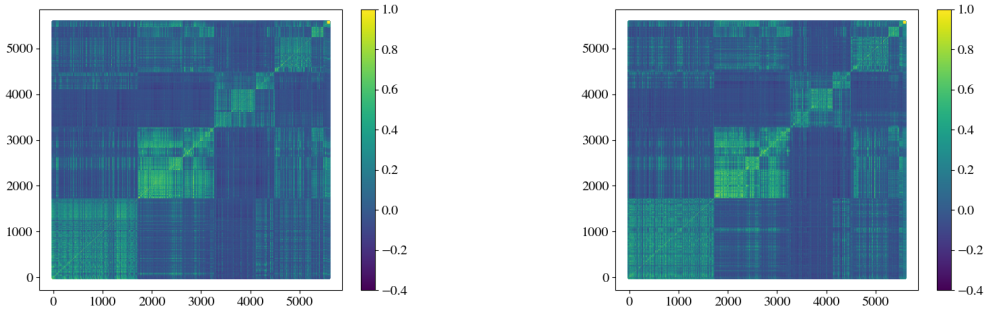
Complementary graphs

A.1 Spearman Correlation: Infomap vs t-SNE



(a) Infomap 1985

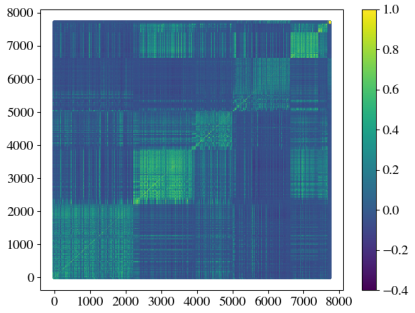
(b) t-SNE 1985



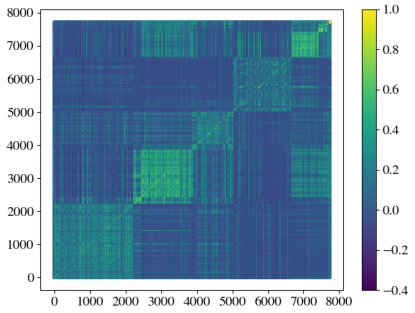
(c) Infomap 1995

(d) t-SNE 1995

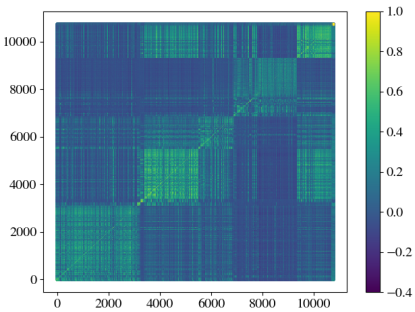
Figure A.1: Spearman: Infomap vs t-SNE — 1



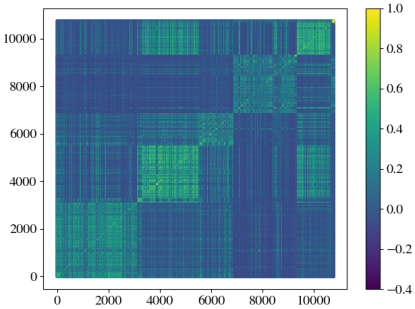
(a) Infomap 2005



(b) t-SNE 2005



(c) Infomap 2014



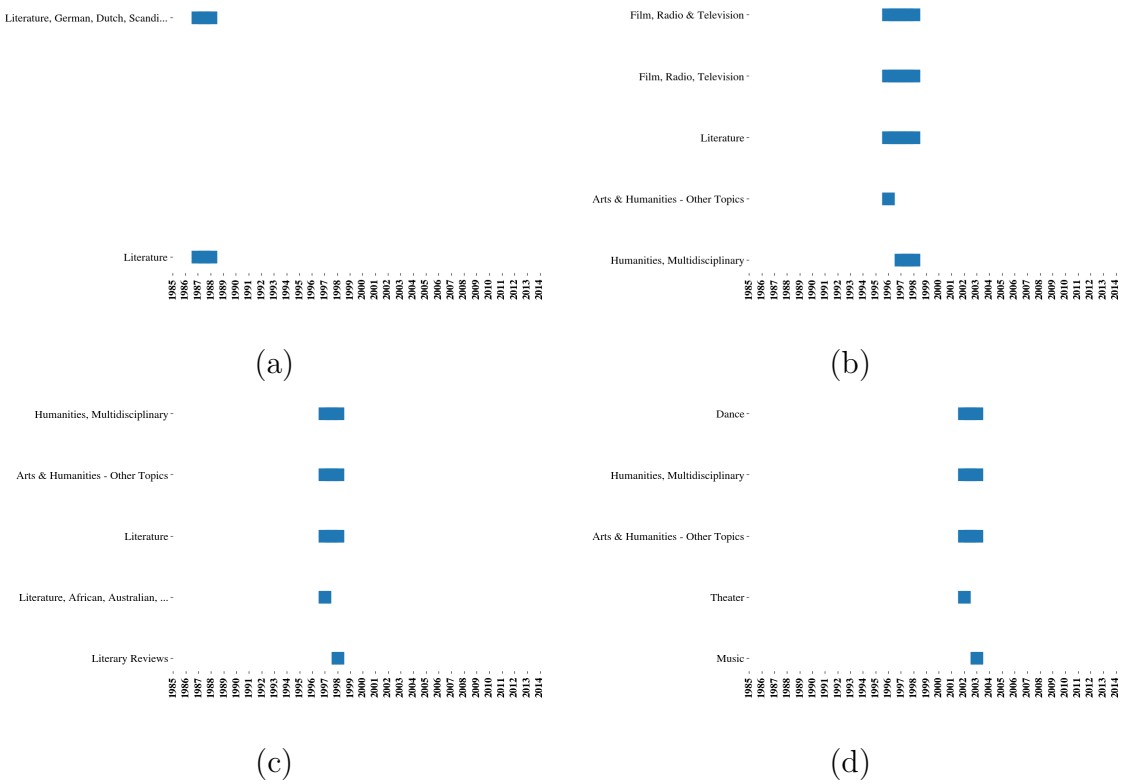
(d) t-SNE 2014

Figure A.2: Spearman: Infomap vs t-SNE — 2

Appendix B

Continuity graphs

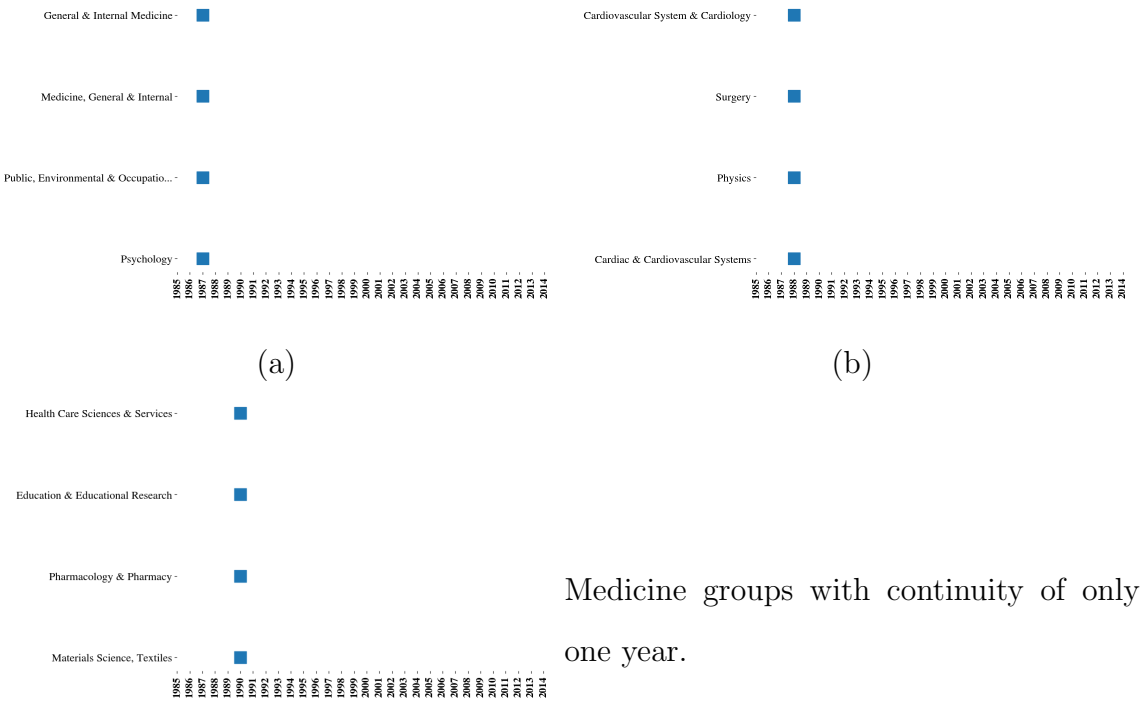
B.1 Continuity graphs of groups which last less than ten years



Humanities groups which last less than ten years.

Figure B.1: Humanities groups

B.2 Continuity graphs of groups which last one year



Medicine groups with continuity of only one year.

Figure B.2: Isolated Medicine Groups

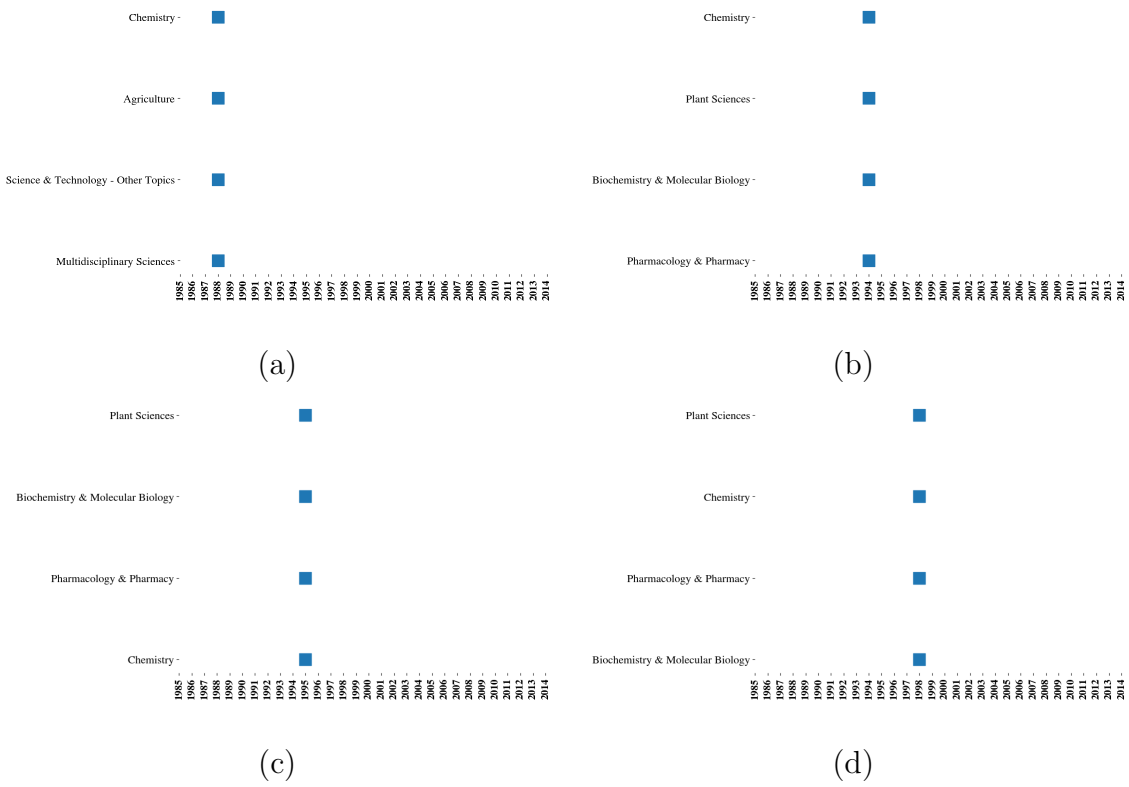
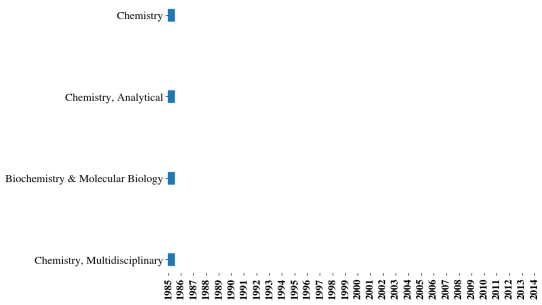
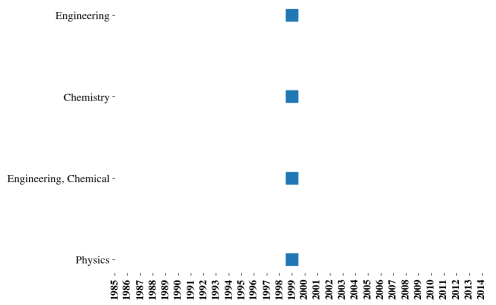


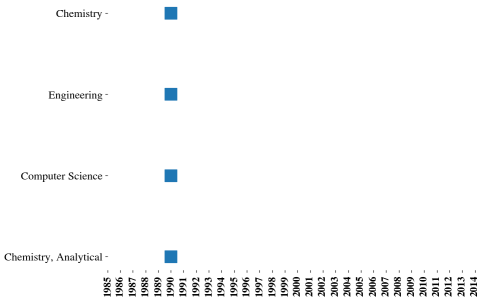
Figure B.3: Plant Sciences



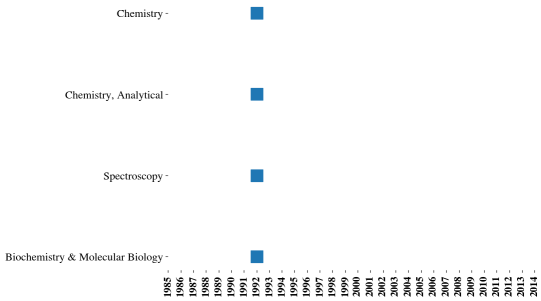
(a)



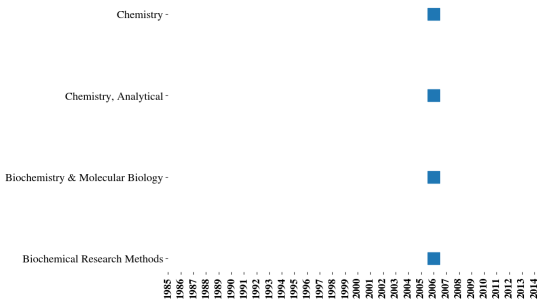
(b)



(c)



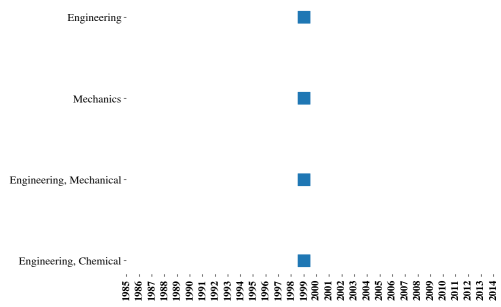
(d)



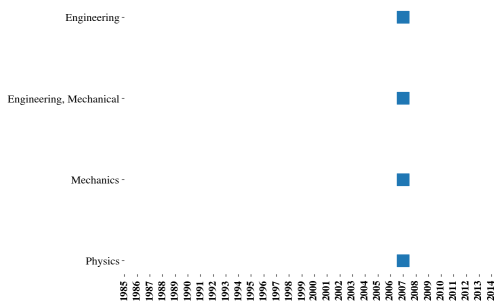
(e)

Groups related to Chemistry with continuity of only one year.

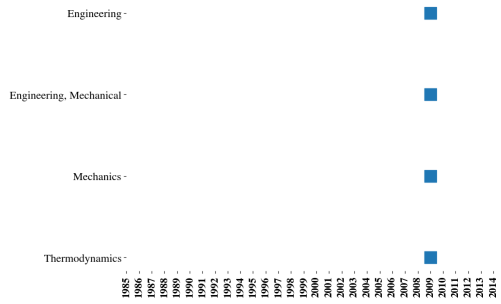
Figure B.4: Exotic Chemistry



(a)



(b)

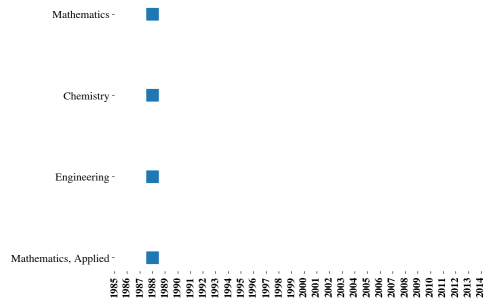


(c)

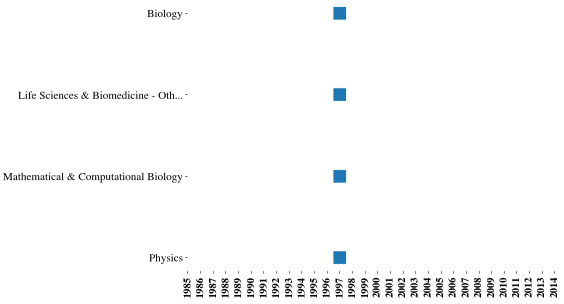
Mechanical Engineering groups with continuity of only one year.

Figure B.5: Mechanical Engineering

(a)



(b)



(c)

Natural Science groups with continuity of only one year.

Figure B.6: Isolated Natural Sciences

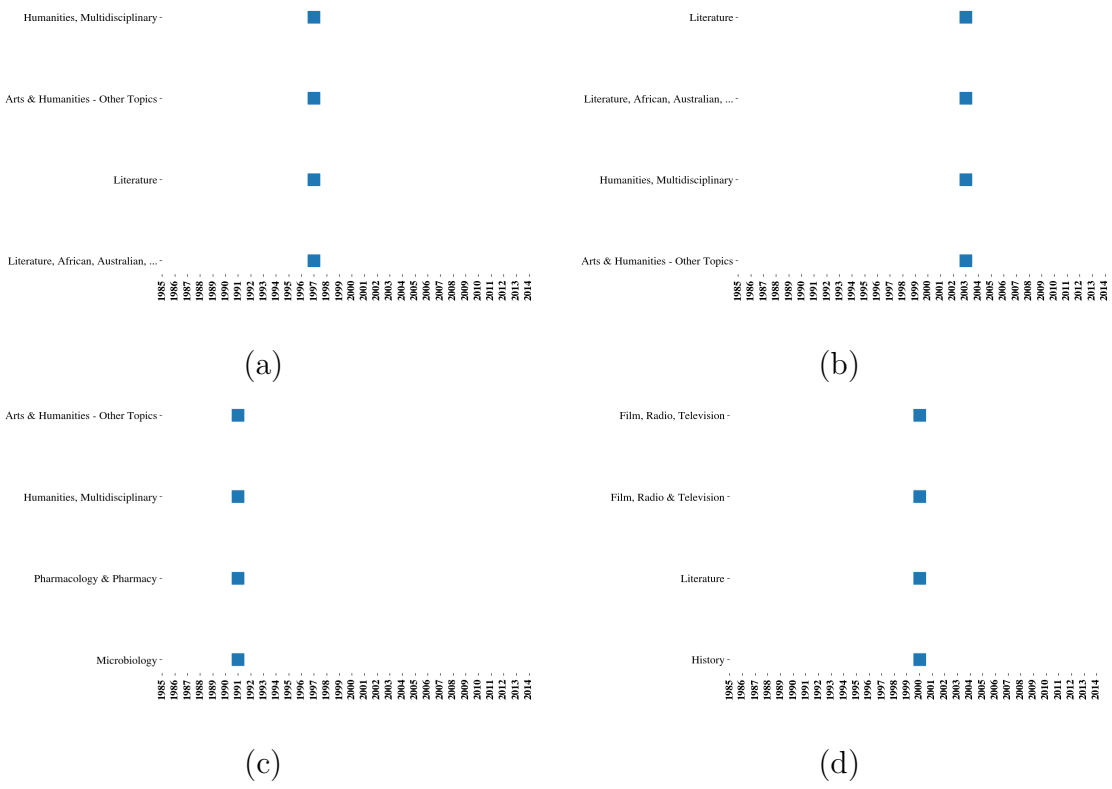


Figure B.7: Humanities groups

Bibliography

Bettencourt, L. M., Kaiser, D. I., and Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210 – 221. Science of Science: Conceptualizations and Models of Science.

Bohlin, L., Edler, D., Lancichinetti, A., and Rosvall, M. (2014). Community detection and visualization of networks with the map equation framework. In Ding, Y., Rousseau, R., and Wolfram, D., editors, *Measuring Scholarly Impact: Methods and Practice*, pages 3–34. Springer International Publishing, Cham.

Edler, D. and Rosvall, M. (2017). The mapequation software package. http://www.mapequation.org.

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

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).

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

Golosovsky, M. and Solomon, S. (2014). Uncovering the dynamics of citations of scientific papers. *arXiv:1410.0343*.

Kawamoto, T. and Rosvall, M. (2015). Estimating the resolution limit of the map equation in community detection. *Physical Review E*, 91.

Klavans, R. and Boyack, K. W. (2016). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4):984–998.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117.

Martin, T., Ball, B., Karrer, B., and Newman, M. E. J. (2013). Coauthorship and citation patterns in the physical review. *Phys. Rev. E*, 88:012814.

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.

Rosvall, M., Axelsson, D., and T. Bergstrom, C. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1):13–23.

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.

Shi, F., Foster, J. G., and Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73 – 85.

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

Sinatra, R., Wang, D., Deville, P., Song, C., and Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312).

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.