KNOWLEDGE MEDIA

KMi

INSTITUTE

Early Detection and Forecasting of Research Trends

First Year Probation Report

Technical Report kmi-15-2
October 2015

Angelo Antonio Salatino

The Open University

Early Detection and Forecasting of Research Trends

Angelo A. Salatino

Knowledge Media Institute, The Open University

OU Personal Identifier: D2329820
Research Degree start date: December 1st 2014

Advisors:
Prof. Enrico Motta
Dr. Francesco Osborne

Contents

Abstract

Identifying and forecasting research trends is of critical importance for a variety of stakeholders, including researchers, academic publishers, institutional funding bodies, companies operating in the innovation space and others.

Currently, this task is typically performed by domain experts, with the assistance of tools for exploring research data. The overall increase of research data in the past decade makes the use of automatic approaches more suitable for this purpose. However, automatic methods still suffer from a number of limitations. In particular, they are unable to detect emerging and yet unlabelled research areas (e.g., Semantic Web before 2000) and moreover they usually quantify the popularity of a topic simply in terms of the number of related publications or authors for each year; hence they can provide forecasts only on trends which have existed for at least 3-4 years.

This report reviews the state of the art in methods for detecting research topics and forecasting their impact, highlights their main limitations, and provides a preliminary version of a novel approach for the early detection and forecasting of research trends that takes advantage of the rich variety of semantic relationships between research entities (e.g., authors, workshops, communities) as well as social media data (e.g., tweets, blogs).

Chapter 1 Introduction

The research world does not stand still for long, it changes and evolves rapidly: new potentially interesting research areas emerge regularly while others fade out. Keeping up with such dynamics is very difficult. The ability to recognise important new trends in research and forecasting their future impact is however critical not just for obvious stakeholders, such as researchers, institutional funding bodies, academic publishers, and companies operating in the innovation space, but also for any organization whose survival and prosperity depends on its ability to remain at the forefront of innovation. For this reason, experts and tools able to identify, make sense of and predict research trends are sought after.
In the following sections, I will discuss the limitations of current methods and the reasons why these approaches are not very apt to forecasting early research trends or detecting embryonic topics. I will also highlight the motivations that led to the decision of trying to create a novel approach, capable of supporting stakeholders both in academy and in industry.

1.1 Problem statement

Nowadays, a variety of datasets about research and computer generated analytics can support the task of understanding what are the main emergent research areas and estimating their potential. In this case, the task can be performed either in a fully automatic way or in a semi-automatic way, i.e., with human experts investigating trends with the help of analytics tools.
In the state of the art we can find several systems for exploring and making sense of research data. For example, some of the most widely used are Google Scholar[1], FacetedDBLP[2] and CiteSeerX[3] which provide good interfaces and built-in search engines to allow users in finding scientific papers, but they do not directly support identification of research trends.
Other tools such as Microsoft Academic Search[4], Rexplore[5], Arnetminer[6], and Saffron[7] provide a variety of visualizations that can be used for trend analysis, such as publication trends and co-authorship paths among researchers. Even with the support of these tools, it can be argued that the manual detection of

[1] https://scholar.google.co.uk/
[2] http://dblp.l3s.de/
[3] http://citeseerx.ist.psu.edu/
[4] http://academic.research.microsoft.com/
[5] http://technologies.kmi.open.ac.uk/rexplore/
[6] https://aminer.org/
[7] http://saffron.insight-centre.org/

research trends is still an intensive and time-consuming task. Moreover, the constant increase in the number of research data published every year makes the approach based on human experts less and less feasible. It is thus important to design and develop automatic and scalable methodologies able to perform this task in an automatic way.

In the state of the art, there are a number of approaches which exploit scholarly data aiming to detect topic trends in a fully automatic way. These are usually based on the statistical analysis of the impact of certain labels associated with a topic. However, these tools are unable to take full advantage of the variety of research data existing today and they need to examine a significant number of years (e.g., 3-4) before they are able to identify and forecast topic trends [1, 2]. Fig. 1 shows an example of a specific case regarding the "Semantic Web" and the graph line represents the amount of publications per year. Through this graph it is possible to discern three major phases involved in the evolution of the topic: *embryonic*, *early stage* and *recognized*.

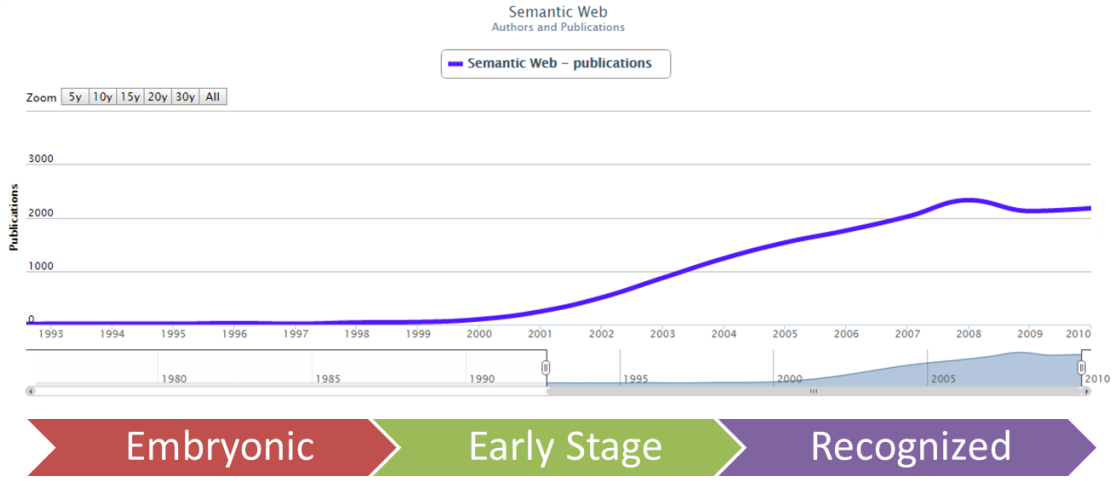

Fig. 1: Stages of a research topic.

As shown by Fig. 1, these three stages are characterized by different numbers of publications directly associated with the topics. Moreover, these stages correspond to a different awareness of authors about the topic.

In fact, it can be argued that a number of topics start to exist in an embryonic way, often as a combination of other topics, before being officially identified and then named by researchers. For example, the Semantic Web emerged as a common area for researchers working on Artificial Intelligence, WWW and Knowledge-Based Systems, before being acknowledged and labelled in the 2001 paper by Tim Berners-Lee et al. [3].

The early stage phase starts when a group of scientists agree with some theories related to the topic, build their own conceptual frameworks, and potentially give birth to a new scientific community.

Finally, in the recognized phase, many authors are aware of this topic and then they start to work on it, producing results and then publish research papers.

Most of the previously mentioned tools are able to identify only topics that have been explicitly labelled and recognized by researchers [4], since they identify topics by means of keywords or labels associated with publications. However, it can be argued that in many cases it is more interesting to detect and investigate the embryonic topics that are still forming and may shape the research landscape in the future, rather than already established topics. For this reason, my first aim is to develop an approach that will analyse a variety of research entities and knowledge bases for detecting any dynamics that may point to the creation of embryonic topics. After a topic is detected and analysed, it is of vital importance to foresee its potential and forecast its future trend.

Currently, we have a variety of approaches for forecasting trends or the dynamics of the growth of a topic. In particular, these approaches aim to estimate the number of publications in the near future using statistical approaches based on Single Moving Average [5] or polynomial interpolation [6]. However, all these techniques suffer from one main problem. They need a good amount of data to do any kind of statistical analysis, so they can usually be applied only after 3-4 years from the detection of the topics.
The doctoral work presented here aims to solve the aforementioned limitations and produce a novel approach to detect and forecast research topics by leaning on two main intuitions.
First, I believe that by analysing the various dynamics of research it should be possible to **detect a number of patterns that are correlated with the creation of new embryonic topics**, not yet labelled. For example, the fact that a number of authors from previously unrelated research communities or topics are starting to collaborate together may suggest the emergence of a new interdisciplinary research area. This theory is also reflected in the literature where it is claimed that the creation of a new discipline requires adventurous and talented scientists who are willing to leave their former discipline to move towards new areas [7]. Those scientists become leaders, provide a definition of the new discipline, describe its purposes and fundamental characteristics, and not less importantly inspire followers.
Secondly, I theorize that taking into account the rich variety of semantic relationships between research entities (e.g., authors, workshops and communities) and analysing their diachronic evolution, it should become possible to **forecast a topic impact in a much shorter timescale**, e.g., 6-18 months. This holistic and semantic-based analysis of the research environment is today made possible by the abundance of both scholarly data and other sources of evidence about research, including social networks, blogs, and so on.
Most datasets of scholarly data [8-10] usually contain metadata describing research papers, including information about title, authors, author's affiliation, keywords, publication venue, timestamp, abstract and content. However, it is important to note that the timestamp is usually available as year of publication. Hence, any method that is solely based on the statistical analysis of the number of publications or citations associated with a topic will have only one data point per year. Since most of these statistical methods require a good number of data points to return a sound result, it is very hard to use them to forecast the impact of a recently emerged topic. I plan to address this problem by i) considering also the information derived from information sources with shorter timescales, such as social media and preprint servers, and ii) taking into account a variety of features extracted from the semantically related research entities, such as authors, venues and research communities. For example, the fact that a new topic is investigated by a number of eminent researchers or seems to be the result of the inter-pollination of two growing research communities could indicate a higher probability of a significant future impact.

1.2 Motivation

In many real-world contexts, being aware of research dynamics can bring significant benefits. **Researchers** need to be updated regularly on the evolution of research environments because they are interested in new trends related to their topics and potentially interesting new research areas. **Institutional funding bodies** and **companies** need also to be aware of research developments and promising research trends. For example, being aware of the future research trends will allow them to make early decisions about investing in new topics on the basis of concrete evidence.
For **academic publishers** and **editors** knowing in advance new emerging topics is crucial for offering the most up to date and interesting contents. For example, an editor can gain a competitive advantage by being the first one to recognize the importance of a new trend and publish a special issue or a journal about it. Thus, an automatic approach to detect novel topics and estimate their potential will bring significant

advantages to a variety of stakeholders. Indeed financial support for this PhD project comes from Springer-Verlag, which is a global publishing company.

1.3 Structure of the report

Based upon the aforementioned problem statement and the motivations, this progress report illustrates what is currently available in the state of the art, the preliminary work I have carried out during my first year, and the overall research plan for my PhD. More in detail, it is organised as follows:
Chapter 2 provides an extended overview of the state of the art relevant to this research and discusses the current gaps and limitations. Chapter 3 presents the main core of this doctoral work, defining the research questions, discussing the main hypotheses, and describing the main research trajectory of this work. In Chapter 4, I will discuss what has been accomplished so far and illustrate the next steps.
Finally, Chapter 5 summarizes the key aspect of the report.

Chapter 2 Literature review

In this chapter I will: i) describe current approaches related to the detection of research trends, ii) discuss the concept of "topic" and provide an overview of approaches to identifying and linking topics to other research entities, iii) give an overview of current approaches to forecasting research trends, and iv) explain the limitations of existing approaches.

2.1 Background

As already mentioned, the state of the art presents several tools and approaches for the exploration of scholarly data. From the topic trend detection perspective, these systems can be either semi-automatic or fully automatic. Some systems for exploring the publication space can provide implicit support for semi-automatic trend detection. One of these is Google Scholar, which gives access to a comprehensive academic literature and is widely used by many researchers. FacetedDBLP [11], based on DBLP database, is a web interface which performs data exploration on authors, papers and venues by means of facets. CiteSeerX [12] is a digital library which provides a search engine for scientific papers including also a mechanism for suggesting relevant papers. Other systems offer instead an explicit support for semi-automatic trend detection. For example, Arnetminer [13], now called Aminer, offers a search engine, as well as support for expert search and trend analysis. Microsoft Academic Search (MAS) allows navigating into scholarly data through several visualization tools, such as co-authorship graphs and publication trends. Saffron [9], which is based on the Semantic Web Dog Food Corpus, provides insights in the research world associating topics to research communities and experts by exploiting Natural Language Process techniques. Rexplore [14] integrates statistical analysis, semantic technologies and visual analytics providing support in exploring and making sense of scholarly data.

While all the aforementioned systems are able to identify and visualize historical research trends, they do not provide support for the detection of future ones.

As expressed in the problem statement section, I believe that is important to take in consideration a variety of research elements. Typically, in the literature, the following entities have been identified as key elements when analysing the world of research: **authors** [15-17], **organizations** [15, 17-19], **communities** [15, 20, 21], **publications** [22, 23], **topics** [2, 18, 24, 25] and **venues** [16, 17, 23, 26].

In the next subsection I will illustrate the relationship between these research elements and in the following ones I will give an overview of how these entities are currently exploited by current methods for understanding the research domain.

2.1.1 Relationship between research entities

Basically, in the research environment, the central characters are the researchers, who are individuals undertaking activities to systematically acquire new knowledge [27]. Usually, research results are divulgated by research *papers*, written by *authors*. Authors are employed by *organizations*, including universities, research institutes and companies. An organization typically supports the activity of the author/researcher providing tools, data and human resources. Usually, authors are also members of one or more *research communities*, which are groups of researchers working in the same discipline [28]. In general, research communities are identified by means of collaboration or citation networks or by clustering authors according to their topics of interest [29].

A *research paper* is an essay presenting analysis, experiment, argument or evaluation on a certain discipline of interest [30]. It is usually published in a *venue*, such as a journal, a conference or workshop [31]. For an author, choosing the venue for its publication is not an easy decision, because many factors should be taken into account, such as the audience he or she is writing for, the topic and also the venue guidelines [32]. However, new forms of self-archiving, such as blogs or online repositories, present an interesting alternative, which is increasingly used in some research communities [33].

Research papers are also usually associated with a list of *topics*, which are typically inferred by the keywords stated by the authors or a third person [15, 30] (e.g., the editor), or extracted from the text with automatic methods [34]. Research topics are themes, that are investigated and analysed by researchers and their communities for a number of reasons, e.g., discovering new information, creating original methods [30].

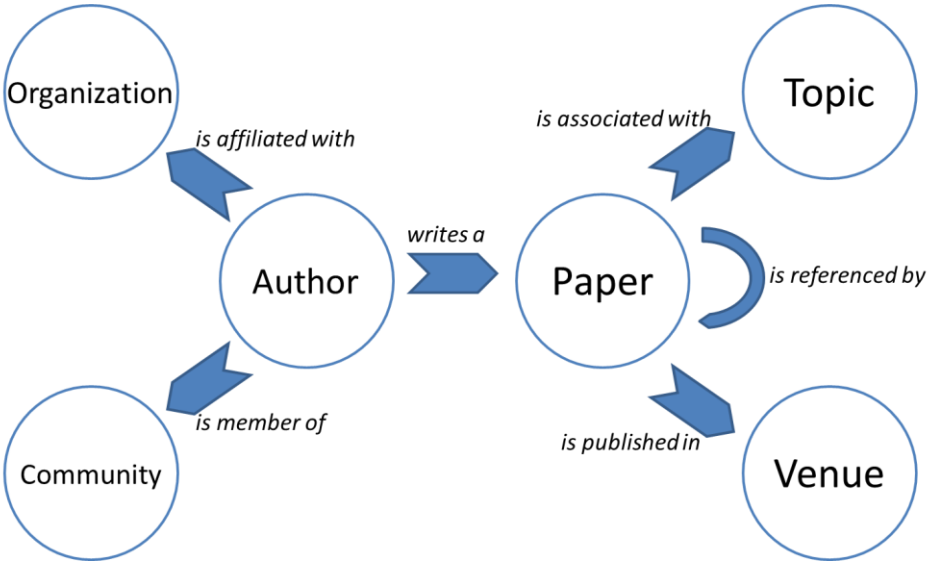

Fig. 2: Model representing the scholarly meta-data and their relationships.

In this report I will focus mainly on these six key elements. These research entities are inherently interconnected; either by explicit relationships (e.g., a paper is connected with its authors) or by implicit relationships that involves a third element (e.g., authors are associated to the topics that are related to their papers). We define, according to a number of state of the art approaches, the **six basic explicit relationships** shown in Fig. 2 and Tab. 1.

Tab. 1: Explicit relations between research elements.

Entities	Label	Description	Examples in the SA
Author -> Papers	**writes**	the paper they wrote	[35, 36]
Author -> Organization	**is affiliated with**	the organisation with which they are affiliated	[37]
Author -> Communities	**is member of**	the communities in which they are involved	[20, 21, 38-40].
Paper -> Topic	**is associated with**	the topic they are associated with	[41]
Paper -> Venue	**is published in**	the publication venue (e.g., Journal, Conference, Workshop) on which they are published	[42]
Paper -> Paper	**is referenced by**	papers which they are referenced by	[42, 43]

These six basic relationships and their elements can be used to define a number of other relationships that potentially can connect any element of the domain to each other element. For example, authors have also (implicit) relationships with topics and venues through their publications. In the same way, topics are associated with publication venues (through related papers published on the venues or related authors active in the venues), with communities and with organizations (through their members/affiliated). These semantic relationships can be refined further by selecting a collection of the connected entities according to a metric or a procedure. For example, by selecting only the most high-impact papers of an author, rather than the complete set defined by the "*writes"* relationship, we can obtain a different semantic relationship, i.e., *mostImportantPapers*. Tab. 2, shows a partial list of this implicit relationships involving research topics, which are used by a number of approaches in the state of the art.

Tab. 2: Some implicit relationships between research elements with regards to topics.

Entities	Label	Description	Examples in the S.ofTheA.
Author -> Paper -> Topic	**topic of interest**	Important topics of the author inferred by its papers	[42, 44, 45]
Venue -> Paper -> Topic	**topics it covers**	The topic of the venue extracted from the published paper	[42]
Community -> Author -> Paper -> Topic	**connected with**	The topic of the community inferred by the papers of its authors	[46]
Organization -> Author -> Paper -> Topic	**Is related with**	The topic of the organization exploiting the paper of its employers	[47, 48]
…			

The next subsections contain a discussion of the aforementioned research components and related methods for detecting and forecasting research trends.

2.1.1 Communities

A community is a group of people sharing values or beliefs, whose social relations are characterized by mutually and emotional bond and frequent interaction [49].
In research, a scientific community is formed by a set of practitioners, who at a given time, are working on the same scientific area [28]. This scientific area, also referred as discipline, specifies the characteristics and the structures of the knowledge domain within which a group of academics place their attitudes, their values, activities and cognitive styles. Becher and Trowler, in their book "Academic Tribes and Territories" [50] concluded that those characteristics of scientific community are quite similar to the social aspects of a tribe. For this reason, they coined the term "academic tribes" indicating academic communities. In order to defend their conclusion, they cite the work of Clark [51] where it is claimed that nowadays as research has become more specialised, scientists in different disciplines possess fewer things in common, in their daily problems and their background. Since scientists, belonging to different disciplines, become less compatible they are less inclined to cooperate with each other.
However, Becher and Trowler also point out that the existing barriers among the tribes are not so high and then friendly relations may be established for mutual benefit. Indeed, many times we assisted to interdisciplinary approaches to solve problems that dealt with health, politics, engineering. Historians, for example, are quite famous for using any kind of source, in order to foster their research. To do so, they sometimes make use of other discipline techniques and that is why Harold Perkin [52] defined historians as "*a kind of licensed rustlers who wanders at will across his scholarly neighbours' fields, poaching their stock and purloining their crops and breaking down their hedges*".
Summarizing, the definition of scientific community differs from the traditional concept of community, in fact it is no longer values that hold the community together but knowledge. Usually, this knowledge is bounded by the education, beliefs, moral, professional initiations, symbolic forms of communication and the technical literature that scientists have absorbed. As Kuhn pointed out, all these factors mark the limits of a scientific subject matter, and each community ordinarily possesses its own subject matter [28].
By definition, a scientific community is a group of people who share common interests, but from a practical perspective, these communities can be either *explicit* or *implicit*.
A scientific community is defined explicit when the graph structure of the network and therefore the link between entities is considered. In this particular case the community is topology-based [29]. For example, explicit communities are the ones based on co-authorship relations or citation network [39, 40, 53].
On the contrary, an implicit scientific community is a network of authors in which links are not directly expressed, but they are obtained from other information associated with the nodes. A topic-based community is an example of implicit community which takes into account common interests of scientists [20, 29]. Hence, if two author share similar interests even if they do not cite each other, they can still belong to the same implicit community [20, 21].
Both implicit (topical) and explicit (topological) communities fall into Becher's vision of academic tribes, according to which communities are defined on the basis of a common characteristic [50]. Indeed, in the case of explicit communities the shared characteristic is being part of the same well-defined social network. While, for an implicit community and more specifically topic-based community, the topic is the feature they share.

In the state of the art it is possible to find several approaches aiming to identifying communities automatically according to both topical and topological criteria.

The *topological perspective*, as already mentioned, considers a graph in which nodes represent entities and edges represent the relationships that connect them in the real world. By partitioning the graphs it is possible to detect communities. For example, Girvan and Newman [54] introduce a well-known hierarchical approach for community detection, which iteratively computes the *betweenness* of all the edges in the network and removes the edge with highest betweenness until no edges remain. The output is a dendrogram, as shown in Fig. 3, which shows the hierarchical structures of the nodes.

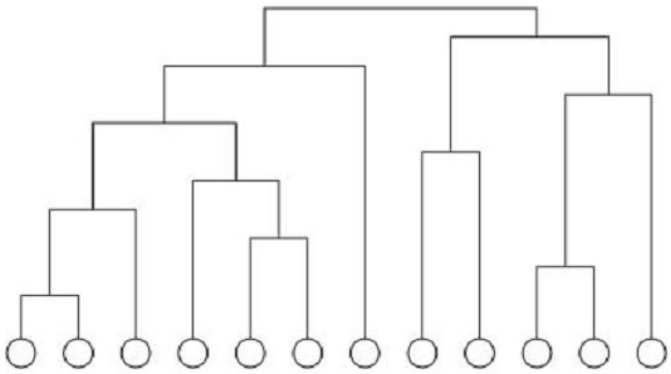

Fig. 3: An example of small dendrogram. The circles at the bottom represent the vertices in the network while the tree shows the order in which they are joined to form communities.

However, the algorithm proposed by Girvan and Newman (GN) is computationally costly. The approach by Radicchi et al. [55] is mainly based on the GN algorithm and introduces the *edge-clustering coefficient* which helps in filtering some useless connections between nodes, reducing the computational cost while keeping the same level of reliability.

Yang et al. [56] propose another technique, called *Probabilistically Mining Communities* (PMC), which uses a probabilistic approach for obtaining a good trade-off between efficiency and effectiveness. The PMC is based firstly on a heuristic phase in which it attempts to reduce the space of candidate community structures by means of random walks. It then goes through an optimization phase where it searches for the optimal structure by optimizing a constrained quadratic objective function.

De Meo et al. [57] suggest, instead, the use of clustering algorithms for the detection of communities on topological structures after a pre-processing step in which they assign a weight to each edge in order to perform cuts in the network. Edges are weighted according to their centrality that is estimated performing multiple random walks on the network.

Gong et al. [58] introduce an approach which uses a multi-objective evolutionary algorithm to optimize two objective functions: *Negative Ratio Association* (the ratio association can be considered as the sum of the density of intra-communities links) and *Ratio Cut* (the ratio cut can be considered as the sum of the density of the inter-communities links). The optimization of Negative Ratio Association tends to divide a network into small communities, while the optimization of Ratio Cut tends to divide a network into large communities.

In an explanatory study, Yan et al. [46] demonstrated that research topics and research communities are not disconnected from each other and actually they are interwoven and co-evolving. In order to detect communities they implemented the hierarchical agglomeration algorithm presented by Clauset et al. [59] which is based on the construction of the dendrogram, similarly to the Girvan-Newman method, but is more efficient in case of sparse networks.

Another approach, for topological community detection, was presented by Xia et al. [60], who analysed social interactions between users and used a clustering algorithm to understand the network structure. This technique proved to be effective on social networks and more specifically on Tianya[8] – a Chinese social network.

On the other hand, from the *topical perspective*, the graph modelling the real world contains links that are defined by the contents produced by network entities. For this reason, many approaches use two phases, firstly they create this different structure based on contents of entities that resemble a graph, and then they extract the community structures from it.

For instance, Osborne et al. [20] present the *Temporal Semantic Topic-Based Clustering* that is able to cluster researchers diachronically based on their research trajectory defined as a distribution of semantic topics. Topics have been provided by the Klink taxonomy [25], and the association topic-author is based on the topics of papers they have written.

Another topic oriented community detection has been proposed by Zhao et al. [21] in which they combine social object clustering and link analysis. The algorithm first clusters social object such as emails, blogs and citations into several topics with the *Entropy Weighting K-Means*, then it links authors to these clusters, and finally perform a link analysis to find the community structure.

An important aspect that has to be taken into account is *multiple community membership*. Usually a person has connections to several social groups like friends, family, colleges and so forth. In case of researchers, they may be active in more than one community. This assumption is supported from the fact that a researcher can collaborate with many other researchers belonging to different communities as well as he or she can be advisor of many doctoral students that may be working on different areas. Many algorithms are unable to take into account this common situation, since they perform a *disjoint* community detection, such as the Girvan-Newman that outputs a tree of communities where leafs – entities of the network – belong to only one community. A number of approaches to solve this problem have been recently proposed.

Xie et al. [61] propose the Speaker-listener Label Propagation Algorithm (SLPA) for overlapping community detection in large-scale networks. In this approach, nodes exchange their labels to their neighbourhood until the algorithm converges. At the end, each node is associated with a list of labels assigned during the computation, which represent the multiple communities to which it belongs.

Also the aforementioned approach by Osborne et al. [20] is able to detect multiple communities since the algorithm associates each author to a distribution of topics.

Nguyen et al. [62] propose instead the *Detecting Overlapping Community Algorithm* (DOCA). DOCA identifies all possible densely connected components of the analysed network, tries to merge highly overlapped communities and it either classifies unassigned nodes as outliers or it groups them into a community.

In conclusion, the state of the art from the community point of view already provides several approaches with different characteristics able to find different kind of communities.

[8] http://tianya.com/

2.1.2 Topics

A topic is a particular subject that someone writes about or discusses. From the research perspective, a topic is also known as academic discipline and it is focused on the study of a particular academic field. The word discipline comes from the Latin word "*disciplina*" which means instruction or knowledge, but it is also a technical term indicating the organization of learning and the systematic production of knowledge [7] and it is based on expertise, expert people, inquiry, projects and studies.

A great contribution to the definition of the notion of academic discipline has been given by Thomas Kuhn (1922-1996), who was a professor of Philosophy and History of Science at the MIT. In his influential book "*The structure of Scientific Revolution*", Kuhn introduces the concept of *paradigm*, that is "a universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of practitioners" [28]. Kuhn coined the term *paradigm* in order to express the idea that disciplines are organized around a certain way of thinking or a framework able to explain empirical phenomena in that discipline or field. However, Kuhn defines and uses the term paradigm several times as pointed out by Masterman [63], making it challenging for many readers to understand this concept. Indeed, Masterman discussed twenty-one possible meanings of the word paradigm. From her study, it is deduced that the comprehensive view of Kuhn about paradigms is a scientific achievement, a model or pattern, a source of tools, a constellation of questions and an epistemological viewpoint. Therefore it can be assumed that a scientific discipline can consist of one or more paradigms.

Kuhn also recalled the concept of scientific communities defining them as cohesive groups of scientists, in order to suggest the existence of a connection between a paradigm and a scientific community, stating that "*a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm*" [28].

Defining a scientific discipline has never been an easy business as pointed out in Becher's book [50]. In order to have a clear idea on what scientific disciplines are, it is useful to analyse them from various perspectives, such as philosophical, anthropological and sociological ones.

From a philosophical point of view, academic disciplines are merely particular branches of knowledge and taken together they form the whole or unity knowledge that has been created from the scientific endeavour. These academic disciplines have boundaries, within which it can be found some coherence in terms of theories, concepts and methodologies that allow the testing and validation of hypotheses according to the defined rules. In general, the kind of questions the discipline tries to answer, the problems it tries to solve, the explanations it attempts to provide and also the kind of scientific language it uses, define the boundaries of an academic discipline. These boundaries make disciplines different from each other or incommensurable as defined by Kuhn [28].

From an anthropological perspective, academic disciplines can consist of a cohesive group of scientists with a high degree of agreement about methods and contents which will have a stronger identity with well-defined boundaries. Indeed, an anthropologist sees academic discipline as a form of social segmentation in which he/she is mainly interested in understanding the cultural practices that produce and maintain them. Focusing on these aspects, an anthropologist comparing scientific communities would be able to find numerous cultural differences, the same differences detectable in tribes, as pointed out in the definition given by Becher [50].

From the sociological perspective, academic disciplines are seen as professions since they share similar characteristics: they have achieved a collegiate autonomy over professional training and certification of professional competence. They also have a community of practitioners who cultivate a distinct professional habitus, possess a professional ethics and also a set of knowledge and skills [64]. Sociologists look at academic disciplines and how they are linked to the world of work since they are interested in making

sense of what happens to academic professions, in knowing why academic disciplines enjoy a different reputation and why there is difference between more established and less established disciplines.
To sum up, providing a definition of topic is one of the main problems that philosophers and scientists are still facing.
An additional problem that many scientists are currently facing is how to extract topics from text documents for several purposes such as exploring and browsing large collections of documents.
In order to extract topic from documents, research papers, blog posts and text in general it is important to define exactly the model that will be used to represent a topic. In the state of the art, numerous ways to define a topic model can be found.
In the first instance, there is the probabilistic topic model and in particular the well-known Latent Dirichlet Allocation (LDA) method [34] developed by Blei, Ng and Jordan. LDA is based upon the intuition that documents show multiple topics. Moreover, from their perspective, a topic is a distribution over a fixed vocabulary.
A similar idea was behind the *probabilistic latent semantic analysis* (pLSA) developed by Hoffman [65]. The pLSA models words of documents as samples from a mixture model, then by means of the Expectation-Maximization algorithm, it performs the extraction of topics that are represented as multinomial random variables, thereby components of the mixture.
Blei and his colleagues developed the LDA aiming to fix some weaknesses of the pLSA, as i) the latter learns the topic mixtures only for documents seen in the training phase, so basically it is not capable of assigning probability to previously unseen documents; and ii) the number of parameters increase linearly with the size of the corpus, indicating that the model suffers from overfitting problems [34]. As an example, the *computer* topic has words about computer with high probability such as *network*, *software* and so on. Therefore, the aim of LDA is to discover the hidden structure which is words per topic and topic per document. In order to do so, it performs the conditional distribution of the hidden variables – topics – given the observed variables – words [41].
Since its introduction, LDA has been extended and adapted in several applications. This is possible just relaxing some assumptions. For example, the Correlated Topic Model [66] uses the logistic normal distribution instead of the Dirichlet, to solve the fact that LDA fails to model the correlation between topics. This approach fits the process of extracting topics from scientific text corpora since it is natural to expect that subset of the underlying latent topics will be highly correlated. For instance, a scientific paper about genetics is likely to be also about health and disease.
Other extensions of LDA are the *hierarchical LDA* [67] where topics are grouped together in a hierarchy and the *relational topic model* [68] which is a combination of topic model and network model for collections of linked documents.
A second general approach, adopted by applications dealing with scholarly data, is based on the use of keywords as proxies for research topics. In this case each keyword usually represents a single topic. This method can be defined as *keyword-based topic model*. Systems like MAS and Saffron [9] use this approach. In general, this method suffers from a number of problems. Firstly, keywords tend to be noisy and include some terms that are not topics at all, e.g., “case study” [10]. Secondly, while topics can have their own hierarchy based on macro areas having their own sub-areas, this does not apply for the keywords topic model in which the relationships among research topics are not expressed [10]. Moreover, the same keyword can have different meanings. This phenomenon, named polysemy, makes possible for two or more different concepts to be treated as one. For example, the keyword “Java” can represent a programming language, an island, or a variety of coffee. Finally, a keyword-based topic model usually does not handle synonyms, hence two or more keywords representing the same concept (e.g., “ontologies”, “ontology” and “ontology-based”) could be treated as different topics. These problems can be alleviated by

asking the authors to use keywords from an existing taxonomy, such as the Association for Computing Machinery (ACM) classification[9].

A third solution to the representation of a topic is based on a semantic topic model [2, 14, 25], in which topics are connected together by semantic relations, creating then a semantic network of research areas. An example of this semantic topic model has been provided by Osborne et al. [25] who developed an algorithm, named Klink, able to detect relationships and then build this semantic network of research areas from keywords associated with a collection of documents, exploiting heuristic rules, statistical methods and external knowledge. This approach is basically based on keywords like the previous topic model, but it adds a conceptual layer aiming to resolve some of the aforementioned drawbacks removing the keywords that seems to not represent a topic and introducing three relationships between keywords: "skos:broaderGeneric", "contributesTo" and "relatedEquivalent" [25].

To sum up, while probabilistic approaches can be applied to every kind of collection of documents, for the second and third approaches it is essential that documents are tagged by keywords. In the case of scientific papers, all three topic models can be adopted because the probabilistic model can be applied to the whole content of the paper and the keyword topic model as well as the semantic topic model can be applied using the keywords as shown in Fig. 4.

The state of the art proposes many approaches that use these three ways of representing topics.

In the category of probabilistic topic model, we have already examined the LDA developed by Blei et al. [34] and the pLSA developed by Hoffman [65]. Other approaches either use or extend the previous ones. For example, Gohr et al. [69] approach uses the pLSA for topic modelling in a window that slides across the stream of document to analyse the topic evolution. Also Mei et al. [70] uses the pLSA in order to create a network of topics. Instead, Nallapati et al. [71], as topic model combines pLSA and LDA.

On the other hand, Griffiths et al. [72] present a generative model like the same employed in the LDA but inferred with the Markov Chain Monte Carlo. Also, He et al. [73] adopt LDA and analyse the evolution of topics by means of citation network. Another usage of LDA is [74], in which the authors try to extract topics from an email corpus of researchers.

Other approaches using the LDA are the Author-Topic Model [75, 76] and Author-Conference-Topic Model [13, 46, 77] that will be explained with further details, respectively, in the authors and publications venues sections.

In the category of keywords topic model, the state of the art proposes different approaches, such as the one by Duvvuru et al. [18] which builds a network of keywords and subsequently performs statistical analysis by calculating degree, strength, clustering coefficient, end-point degree in order to create clusters to associate to research topics. Another approach is Erten et al. [40] in which they exploit the ACM taxonomy as taxonomy of subjects and based on the ACM corpus they visualize trends along time. In addition, there is the approach by Decker et al. [2], who generate paper-topic relationships by exploiting keywords and words extracted from the abstract in order to analyse the trends of topics on different timescales.

In the category of the semantic topic model, as already seen, the approach of Osborne et al. [25] is able to build a semantic network of research areas. Hybrid methods also are present, which combine LDA and semantic technologies. For example, the approach by Gou et al. [78] is based on the idea that exploiting dictionaries in the model can yield a better understanding of word semantics leading to a better model of the text.

[9] http://dl.acm.org/ccs_flat.cfm

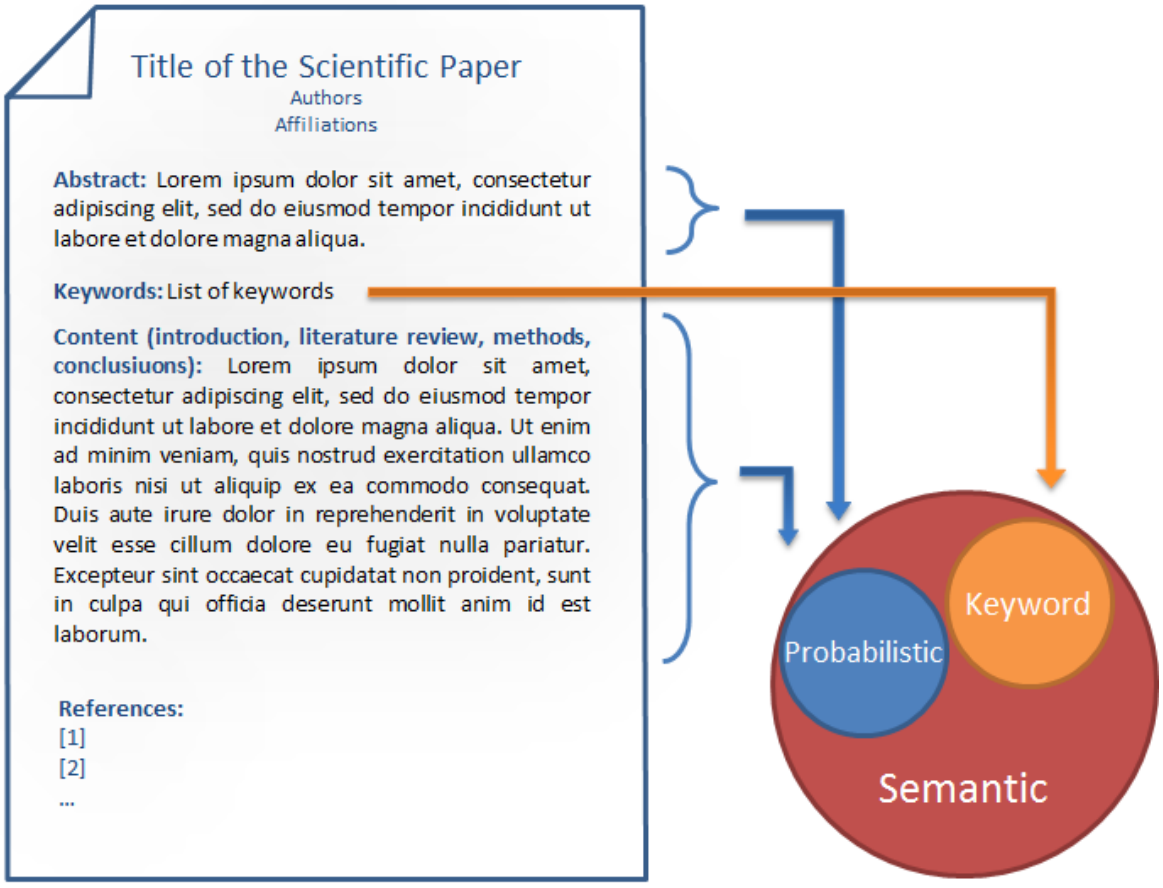

Fig. 4: Graphic explanation of the relation between topic models and their source of information.

2.1.3 Authors

In general, an author is the person who originates or gives existence to anything written. In the case of research, they are usually scientists or scholars who produce lines of research and scientific papers.
A number of indexes [79, 80] attempt to measure the productivity and the impact of the academic output of a research author. One of the most widely used is the *h-index, which depends on both publications and citations. A researcher with an h-index of h has published h papers with at least h citations* [81]. However, this index can be unfair to young researchers, who may have a small number of papers [80, 81]. Some other indexes attempt to overcome this disadvantage. For example, the m-index [79] divides the h-index by the years of activity of a researcher, and the g-index is calculated as the highest number of *g* papers which have g^2 citations [82].
Author networks, with links representing co-authorship relationships, are used by a variety of approaches to analyse research. For example, the author-topic (AT) model proposed by Rosen-Zvi et al. [75] improves LDA [34] to include authorship information. Starting from the LDA assumption that every document is a distribution of topics, it extends the concept of topics as distributions over both words and authors. As soon as the model has been trained, it is possible to understand the set of topics that appear in the corpus as well as identify which topics are relevant to which authors. However, since the AT model combines variables of topics and authors, it is not able to adapt its distribution over topics to the content of documents as LDA does. Another approach presented by the same research team is the Probabilistic Author-Topic model [76] which models the documents as being composed of multiple topics, topics as probability distribution over words and authors modelled as probability distribution over topics. However, instead of using the LDA, they train their model using a Markov chain Monte Carlo algorithm. Nevertheless, the probabilistic model is quite simple and disregards some aspects such as topic correlation and author interaction.

2.1.4 Publication venues

The Oxford Dictionaries[10] defines a venue as "the place where something happens, especially an organized event such as a concert, conference, or sports competition". In the case of research, a venue is characterised by a journal, a conference or a workshop that hosts research papers.
The dynamics of publication venues can influence the creation and evolution of research topics. Hence, the performance (e.g., number of publications/citations) of venues associated to a topic can yield a precious insight on its potential.
A number of state of the art approaches exploit venues in this way. For example, Tang et al. [13] presented an unified topic model for simultaneously modelling the topical distribution of papers, authors and conferences called *Author-Conference-Topic* (ACT) model. This model further extends the Author-Topic model [75] to include conference and journal information. In particular, they present three different implementation of the model, which differ from each other in the way they model the association between authors, distribution of topics and conference stamp. This same model has been employed in the already mentioned explanatory study conducted by Yan and his colleagues [46]. In this study, the authors wanted to demonstrate that research communities and research topic are interlaced and co-evolve, instead of being two different entities.

[10] http://www.oxforddictionaries.com/

However, Wang et al. [77] pointed out a limitation of this approach. Basically, this model extract topics for the corpus of documents and then it map them to the research areas promulgated by the "*call for papers*" of conferences. This operation is not always possible because the latent topics extracted with the LDA may not be equivalent with the conference topics. As an instance, the subjects of the Conference on Information and Knowledge Management (CIKM) that are knowledge management, information retrieval and databases can be easily associated with latent topics like biological database.

2.1.5 Trends

According to the Oxford Dictionaries[11] a trend is "a general direction in which something is developing or changing". Following this definition it is necessary to state clearly what is this "something" and what exactly this "developing or changing" involves. In addition to the previous, it is possible to find a different definition of trend that is "A topic that is the subject of many posts on a social media website within a short period of time". This last definition seems to be more specific than the former, because firstly it is related to posts on social media and secondly the "developing of something" is related to the amount of posts related to a specific subject. It seems that more posts are related to a particular subject, more "trendy" this subject is becoming. Therefore the trend is associated to the popularity.

Nevertheless, both definitions of trend fit the purposes of this doctoral work, because I want to investigate the development of or changes related to research topics and in particular I want to investigate which of them are becoming popular.

Taking into account the first definition of trend "that something is developing or changing" and that something is expressed as topic, indicates that there is an interest in understanding the dynamic behaviour and then the constant change behind the evolution of topics.

This dynamicity, defined as the constant change of topics, can be described by means of a mathematical model which binds different *states* in time.

In general, the concept of state [83] is always associated with an object or a being, and it is represented by a set of characteristics that define the condition or situation of that thing at any instant in time; in other words it is a sufficiently comprehensive description, a snapshot, of the system at a particular time.

To summarise, in order to detect topic trends, it is possible to define a topic state according to characteristics such as the number of related publications/citations [2, 17, 40, 84], the number of authors active in it [17], and so on, and then monitor their evolution in time.

As already discussed in previous sections, it is possible to take in consideration also the dynamics of other research elements, e.g., communities, publication venues, authors and so forth. For example, we can take in consideration a variety of characteristics associated to the related communities, such as the number of authors who become part of them each year, the total balance (outflow versus inflow) of authors migrations and so on. This general concept of dynamicity that involves all the research elements will lead to useful insights for the hypothesis formulation of this doctoral work.

Thomas Kuhn in his book described the dynamics involving a scientific discipline as a *paradigm shift*. He describes the paradigm shift as a revolution in ideas, knowledge and research project. This kind of phenomena occurs when a paradigm cannot cope with anomalies, which lead to a crisis that will persist until a new outcome redirects research through a new paradigm. Kuhn's intuition was that science is an alternating of progress and changes, therefore he designed a model, in particular a cycle model, which explains how science evolves. This model consists of three main stages:

[11] http://www.oxforddictionaries.com/

Pre-paradigm: in this phase, there are several incompatible and incomplete theories and there is no consensus on them. If a group of scientists concur with some theories, as a new scientific community they build a conceptual framework and ultimately they disseminate their methods, terminologies and also the kind of experiments that will contribute to the progress in knowledge;
Normal science: science progresses within the existing paradigm accumulating knowledge. In this stage, the research is firmly based upon one or more past scientific achievements, that are acknowledged from a particular scientific community, at that time, in order to provide the basis for further developments;
Revolution: it is a phase in which the boundaries of the fields are crossed, and then some unanswerable questions are discovered. Afterwards, there is a crisis derived from the fact that an old paradigm cannot explain some important observations and then the model is no longer capable to solve current problems. The revolution starts when a new paradigm challenges the previous one to encompass explanations and resolve some outstanding and generally recognized problems. The new paradigm settles in when it has few influential supporters and thus a new cycle begins all over again [28].
Some examples of scientific revolution which fit Kuhn's idea are the shift in physics from Newtonian mechanics to Einstein's special theory of relativity and also the shift in cosmology from Ptolemaic system to the Copernican heliocentrism. The paradigm shift in physics is quite popular because many books mention it and is recent. Basically, the two theories seem to use the same concept of mass, velocity and time, but the prediction of these theories start to diverge significantly if they are applied to objects travelling at high speed [28]. One can argue that Newtonian mechanics is an approximation of Einstein's theory but actually the two theories have different ontological assumptions which make them really different and in Kuhn's words: incommensurable.
In the state of the art it is possible to find several approaches for detecting trends. For example, Wu et al. [1] integrate bibliometric analysis, patent analysis and text-mining analysis in order to detect research trends. Some other models take in consideration the citation graph. For example, Bolelli et al. [85] propose a generative model that uses temporal ordering of documents in order to identify topic evolution and then use citations to evaluate the weights for the main terms in documents. He et al. [73] combine Latent Dirichlet Allocation and citation networks for detecting topics and understand their evolution. Instead, Gohr et al. [69] combine the PLSA for topic modelling in a window that slides across the stream of document to analyse the topic evolution. Duvvuru et al. [18, 86] propose a different approach which monitors the weights of the links of a keyword network to detect changes in the research environments.
On the other hand and from a different perspective, the approach by Decker et al. [2] monitors the changes in the number of publications associated with research topics. In studying these topic evolutions, their approach is also able to classify if a topic is growing suddenly – in a "bursty way" – or gradually, based on a defined dynamic thresholds. In the same way, approaches like [17, 84] compare the number of papers published per topic.
Erten et al. [40] propose a hybrid approach which detects the growth and the decline of research topics by analysing both the number of papers per topic and the co-occurrence between the topics.
To sum up, many current approaches are able to keep track of the evolution of topics, however they are able to detect trends only after the associated research areas are already recognised. Therefore, they do not provide any support to the early detection of research trends, which is still an open problem.

2.1.6 Forecasting

State of the art methods for forecasting the impact of research topics take usually into consideration the number of publications and authors associated with a topic [87]. In order to do so, many approaches analyse these time series either by means of statistical techniques, such as computing the slope of linear regression on time series [4], second degree polynomial interpolation [5], exponential smoothing which extends the Simple Medium average [6], and machine learning methods [88], yielding a prediction for the following years. However, these methods do not take advantage of the knowledge that can be extracted by analysing the dynamics of multiple research entities (e.g., communities, venues), and they ignore the growing mass of research data that today can be acquired from social networks. Moreover, as already mentioned, all these approaches define the impact of research topics in terms of number of publications and authors associated with topics. Arguably, a new definition of the impact based also on new data sources can improve the forecasting phase and can allow to intervene within a shorter timescale.

2.2 Defining the gap

The previous sections presented an overview of the main approaches and directions for the detection of research trends and the forecasting of their research impact. As discussed, a good number of methods deals with the detection of research trends, but can be applied only on already recognised topics, associated with a label or, in the case of probabilistic topics models, with a set of terms. Hence, there are not any comprehensive solutions to perform the detection of topics in their embryonic phase. Moreover, these approaches do not exploit the mass of relevant information that can currently be extracted from social media and other useful web sources.

Similarly, the methods for forecasting the impact of research topics suffer from two main limitations. First, they can be only applied on topics that have existed for a good number of years to accurately assess the topic future impact, since the granularity of academic data is yearly. I hypothesise that using a number of additional features (e.g., the track history of authors supporting the topic, the presence of related workshop) and data from more granular sources (e.g., social media, preprint servers) it may be possible to forecast the future impact of a topic in a shorter time. Indeed, it has already been shown in the literature review (e.g., the Author-Conference-Topic), that the introductions of new features can yield a significant improvement in topic detection.

Secondly, these approaches are based on a very simplistic definition of the impact itself, such as the number of publications or the number of authors associated with the topic. It can be argued that we need more comprehensive metrics to assess a topic impact. For this reason further work needs to be done, in order to provide a better definition and therefore better indexes for evaluating of the impact of a topic.

To sum up, the limitations identified in the state of the art are:

- the inability of detecting embryonic research trends;
- the inability of forecasting the impact of research topics in their early life;
- a simplistic definition of topics impact;
- a limited use of informative data sources such as social media, preprint server analytics and so on.

In the following two chapters I will elaborate on these limitations for formulating the research questions and propose an approach for addressing them.

Chapter 3 Research plan

The main goal of my research program is to identify and predict the impact of new research trends. In order to achieve this goal the work has been divided into two main phases: i) detecting new research areas in their early stage and ii) forecast their future impact. The following chapter presents the main hypotheses, formulates the research questions and describes the approach that will be used to answer them. The research questions are based on the limitations and gaps identified in the literature review as well as on the scenarios presented in the introduction. Finally, I will discuss how I plan to evaluate the results of my approach.

3.1 Research questions

The main research question is:

> *"How is it possible to detect the early emergence of new research topics and forecast their future impact?"*

This question entails two different challenges: i) detecting research topics at a very early stage, taking into account that a novel research topic may not be associated with a definitive label, ii) figuring out which characteristics define the impact of a topic and how they can be exploited to improve the accuracy of the forecast.
The main research question can be better investigated by defining a related set of sub-questions. In the following I will discuss them.

3.1.1 Question 1 – Identifying the relevant data

William Edwards Deming once said "*Without data you're just another person with an opinion*". Thus, it is important to investigate which kind of data is more apt for the purposes of my doctoral work. Scholarly data, such as the metadata of research publications (e.g., MAS, DBLP, Semantic Web Dog Food), are a good starting point since they allow us to infer knowledge about many elements of the research environment. Indeed, many applications already showed in the literature review use this kind of data. However, for the

purposes of this research, are scholarly data enough? Are there any other sources of information that can augment scholarly data to facilitate the early detection and forecasting of research trends?
Considering that nowadays many researchers use social media (e.g., Twitter) to communicate and publicize their work, social media data can also be a precious source of additional knowledge. However, in social media, ideas and information are shared and exchanged for general purposes. Therefore, what kind of approach or model can be used for identifying the data that are relevant to the research domain? Once this research-oriented collection of posts is available, how can this information be linked to research entities and then integrated with other scholarly data?

3.1.2 Question 2 – Detection of new emerging research topic

How is it possible to understand if a new topic is emerging? I hypothesize that in many cases a new discipline grows when two or more already existing disciplines start co-operating and sharing their knowledge in order to break their boundaries and achieve new knowledge. For example, as previously discussed, the communities interested in Artificial Intelligence, World Wide Web and Knowledge Based Systems started to collaborate on novel ideas, giving rise to a novel research area later labelled Semantic Web by Tim Berners Lee et al. [3]. I intend to confirm this hypothesis with empirical evidences by investigating the dynamics of research areas associated with new topics. This will lead to the definition of the typical patterns, which tend to anticipate the creation of new topics (e.g., a significant increment in the collaborations between research areas). Additional questions are the following. What kind of features should be examined to investigate such patterns? How can we learn and represent them? How can we build an algorithm able to use this pattern for forecasting the creation of new topics? How can we label a still unnamed embryonic topic? Is it possible to design a general approach able to consider the peculiarities of different fields, such as Computer Science, Business, Medicine and so on?

3.1.3 Question 3 – Forecasting of research trends

As seen in the literature review, there are a number of approaches that aim to forecast the impact of topics by analysing the number of publications/citations in time. Is the number of citations and publications sufficient for defining the impact of a research topic? If not, how can the definition of impact be improved? Which other features can be employed in addition to the aforementioned ones?
As soon as we have a proper way to define the impact and therefore a list of features to exploit, which kind of approach will be able to forecast the impact of research topics? How will it be possible to assess the accuracy of the forecasted impacts?
Similarly to the detection phase, also for the forecasting it is important to intervene at an early stage. Will it be possible to perform a valid estimation of the potential impact of embryonic topics? What kinds of issues are related to it? Can the same patterns found for the early detection help to address this task? How can social media data contribute to this process? What additional features can be extracted from them to support this task?

3.2 Hypotheses

From a philosophical point of view, academic disciplines can be seen as specific branches of knowledge which together create the unity of knowledge that has been produced by the scientific endeavour. When

two or more disciplines start to cooperate they share knowledge, theories, concepts, tools and methods. The results of this cooperation may lead either to the creation of a new interdisciplinary research area or simply to a contribution in knowledge from one area to another.
The basic hypothesis is that the creation of a topic is thus anticipated by a number of dynamics derived from scholarly data. For example, as already seen for the Semantic Web example, these dynamics can involve the co-operation between two or more existing research areas. Additionally, the involvement of dynamics of other research entities, such as research communities, authors, venues and so on, might facilitate a very early detection of emerging topics.
Scholarly data can be used to analyse research entities such as papers, authors, affiliations, venues, topic and communities [20]. As already seen in the literature review, all these research entities are inherently interconnected by relations that can be defined as either explicit or implicit. For example, a topic is also associated with publication venues through relevant papers published in venues. These relationships can be analysed diachronically to derive the dynamics that led to the emergence of a topic and to estimate how they may affect its future impact. For example, if two communities start to share research interests or authors, this may lead to the fact that a common new topic is developing.
In a nutshell, the fundamental hypothesis at the basis of this doctoral work is that by exploiting the large variety of scholarly data which are now available, as well as modelling their semantic relationships, it may be possible to extract patterns leading to the creation of new research trends, even in a relative small interval of time. In addition, from the same source of data it may also be possible to extract patterns allowing us to estimate its future impact.
Finally, we also hypothesise that, since many researchers are actively involved on social networks, social media data can provide an effective input to this analysis.

3.3 Approach

The approach is structured according to the proposed research questions. Basically, it is organised in four main tasks, which do not necessarily introduce a temporal sequence.

3.3.1 Data integration

In this task I plan to use the datasets integrated in Rexplore, which include scholarly data from Microsoft Academic Search, Springer and Scopus[12]. I will also evaluate if social media can provide further support for the detection of trends and the forecast of their future impact. Since social media are used for general purposes, it is important to understand how information and entities related to research could be extracted and filtered from them. Moreover, it is crucial to understand how to integrate a variety of heterogeneous data sources, such as tweets, blogs post, slides and so forth, in the already existing scholarly database. However, considerable steps in this direction have already been accomplished thanks to Altmetrics, introduced by Priem et al. [89]. Altmetrics, which stands for 'alternative metrics', is a new research area which studies the research environment using data from the social web such as discussion forums, Tweets, Facebook pages, Mendeley, blog posts and so forth. However, many of these metrics can actually be influenced by a number of factors (e.g., likes and mentions can actually be bought [90]), thus I plan to also investigate alternative methods for assessing impact via social media.

[12] http://www.scopus.com/

The output will be a comprehensive knowledge base containing both the research entities from Fig. 2 and entities from social media, such as authors' profiles, number of followers, analytics, etc. Topics and communities will be identified by extending state of the art techniques. In particular, it is planned to treat topics semantically, by describing their relationships using the topic networks produced by the Klink algorithm [25]. Moreover, it is also planned to use the approach for detecting topic-based research communities described in [20], since it explicitly links communities and topics.
The rich network of semantic relationship between the research elements will be described by an ontology and it will be populated by semi-automatic statistical methods. To build it, I plan to extend the topic network created by Klink with the research entities and their relationships. The analysis of these relationships and how they change in time will support the next steps of the approach.

3.3.2 Exploration of the Research Dynamics

During this task, it is planned to verify a number of hypotheses about the development of topics, such as the dynamics that helped the evolution of the Semantic Web. Moreover, dynamics involving other research entities correlated with the emergence of new topics will be investigated. Subsequently, exploiting the same idea of combining topics with other research entities, such as Osborne et al [20], Rosen-Zvi et al. [75] and Tang et al. [13], I plan to create a series of models connecting each research entity to topics. These models will aim to map the evolution of particular research entities with the status of the associated topics.
Therefore, I will analyse a number of topics which appear in the 2000-2010 interval in the Rexplore dataset and verify if their emergence is correlated with a number of dynamics, such as the raise of co-publications of related research areas, the increase of collaborations between authors of related areas, shifts of interests or migration phenomena in related communities, transfer of topics between related venues, and so on.
In particular, I have already built co-occurrence graphs with nodes representing topics and links representing the number of co-occurrences between them. I am now conducting a diachronic analysis on these graphs to confirm if the creation of novel topics is actually correlated to an increase in the pace of collaboration of already existing ones. I plan to use a community detection algorithm to further analyse these temporal activities in graph-dynamics. I am also planning to analyse some important characteristics extracted from these communities (e.g., structural cohesion [91], clustering coefficient, degree distribution and so forth) to verify if they can be associated with the creation or evolution of a topic.
The output of this analysis will be a collection of patterns of knowledge flows associated with the creation of a new research area.

3.3.3 Early topic detection

This task aims to exploit the patterns identified in the previous task for the early detection of research trends. To this end, I will build a number of distinct graphs, in which nodes are a kind of research entity (e.g., topics, communities) and the links represent one of the elements of the dynamics, which were found in the previous phase – e.g., the increase in the number of collaborations between authors from two distinct topics. I will then analyse highly connected sub-graphs, representing the area in which multiple entities exhibit the identified dynamics for detecting emergent disciplines.
In order to produce more robust evidence, I will use the semantic network of research entities to confirm that the emergence of a new topic is supported by a number of different dynamics, among the ones discovered in the previous phase, and research entities. For example, if a set of topics suggests that a

correlated research area is emerging, the dynamics of the set of communities and venues related to these topics will also be checked.
The intuition is that, while the evidence coming from a single dynamics or a single kind of entity could be biased or noisy, their combination should yield a more accurate result. The result will be a number of sets of linked entities, each one anticipating the emergence of a new topic. Different combinations of entities and metrics will be tested, aiming to find the best approach to derive sets that are strongly correlated with the creation of new topics.
In this phase, another challenge is how to label a research topic that is still in the embryonic stage and without name. Basically, the choice will be either to generate a pseudonym on the basis of the topics from which it is developing and then leave the final labelling to experts or to try and develop a semantic approach that is able to label future research topics.

3.3.4 Trend forecasting

Initially, I will investigate a number of baseline techniques to estimate the impact of topics, taking in consideration basic metrics, such as the number of publications and citations. I will then try to improve on current methods and investigate novel indexes and a variety of features in order to provide a better representation of the impact of topics. If social media data will prove to be relevant, I will also incorporate them in the definition of a number of indexes for measuring the impact of a topic. In particular, I plan to extract from social media a variety of features, such as number of posts, number of share per post, number of favourites per post received, and others, which, as mentioned before, can be easily manipulated.
In contrast with current approaches, [1, 2], I aim to develop a method which will be able to work also on relatively short time series (6-18 months). In order to do so, I will take advantage of a wide variety of features associated with a topic, representing both the performances of related entities (e.g., the track record of significant authors) and the previously discussed dynamics. Hence, I will conduct a comprehensive analysis on historical data, looking for the correlations between these features and the topic impact in the following years. For example, I will analyse how the performance of related authors, communities, workshops and so on, can influence the previously defined impact metrics. It is hypothesised that such abundance and diversity of the features will compensate for the small interval of time in which early topics will be analysed. Moreover, data from the social web and other real-time information, such as the number of views and downloads on the publisher sites and open access repositories, will offer a more granular timeline for the analysis of the topics, measured in weeks, rather than in years.
I will then exploit the extracted features in order to forecast the performance of a topic using statistical techniques and machine learning methods. In particular, I plan to test different machine learning approaches, such as Artificial Neural Networks, Support Vector Machines and Deep Belief Networks to identify the techniques more apt for this task.

3.4 Evaluation plan

I plan to conduct an iterative evaluation during the different phases of my work using both quantitative and qualitative approaches.
From a quantitative point of view, I will evaluate both the ability of the system to identify novel topics and its accuracy to assess their impact in the following years. The discussed approaches will be compared with current methods and the difference between their performances will be measured via statistical tests. I will evaluate the detection of emerging trends in terms of recall, precision and F-measure using cross-validation

on historical data. Similarly, I will assess the agreement between the estimated and the real impact of a research area.

In the qualitative evaluation, the achieved results will be compared with experts' opinions in order to measure its reliability. I will prepare a number of surveys for domain experts with questions both about the past - such as the main topics recently emerged in their area of expertise - and about the future - such as the research areas which seem on the verge of being created and an estimation of their likely impact.

Chapter 4 Piece of Work

In this chapter I will report on the preliminary accomplishments of my PhD program. In particular, I will show that my initial experiments seem to confirm that the emergence of a topic can be anticipated by the dynamics of a topic network.

4.1 Current Progress

The initial goal of this research is to understand which dynamics give birth to a new topic. As already stated, one of the main hypotheses is that a new topic arises when two or more already existing disciplines *start to get close to each other*, creating this new interdisciplinary field that implies the sharing of techniques, assumptions, and methodologies to solve a problem or answer a scientific question.

Using scholarly data is possible to find several ways to measure the growing closeness of topics. First of all, it is possible to study the co-occurrence graph of topics and other research entities, such as authors or venues, to monitor if the collaboration pace is increasing. Furthermore, we can study the evolution of the topic distribution associated with each of these research entities. For example, it is possible to monitor the increment in collaborations for authors working in different fields, as well as how their topic distribution is changing. This will allow us to detect how two or more previously uncorrelated topics are starting to interact. Similarly, it is possible to analyse the structural changes of research communities. For example, the fact that a good number of authors from a certain community start to migrate towards another one or that authors belonging to different communities come together to create a new interdisciplinary one. In the same way, it is possible to monitor the introduction or change of the topics of interest in publication venues.

For this analysis, the Rexplore system will play a crucial role since it contains a significant amount of scholarly data. In particular, the Rexplore dataset already contains up to date keyword networks describing how keywords extracted from papers interact in subsequent years. Each keyword network can be represented by means of a graph structure, G = (V, E), in which V is the set of keywords while E is the set of links representing co-occurrences between keywords in a certain year. The node weight is given by the number of publications in which the keyword appeared, while the link weight is equal to the numbers of times two keywords co-occurred. As an example, in 2008 the tag semantic web was used in 1583 papers while artificial intelligence was used in 9657 papers. As shown in Fig. 5, only 100 papers are tagged with both of them.

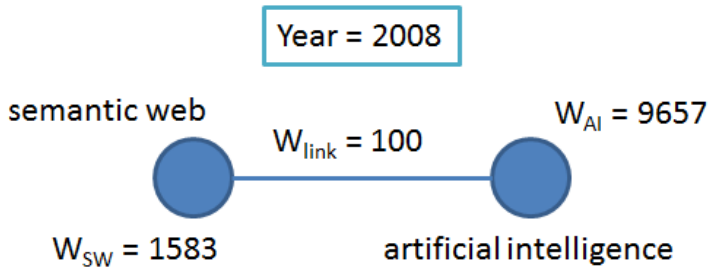

Fig. 5: Example of keywords network in the 2008 taking into account only two keywords and their co-occurrence.

In the first phase of the preliminary analysis (discussed in section 4.2-4.5), I focused on engineering the methodology to verify my initial hypothesis, and I adopted a simple keyword-based topic model, as discussed in Chapter 2. I then enhanced the model (see section 4.6), by considering also the semantic relationships between research topics.

As previously discussed, I expected that by investigating the keywords network it would be possible to detect two kinds of behaviours. In particular, I expected to find a significant increment in the number of co-occurrences between keywords in the areas in which new embryonic topics are emerging while the pace of co-occurrences is regular elsewhere.

I ran a preliminary test on this hypothesis by selecting twenty significant novel keywords (e.g., semantic web) and measuring the increase in cooperation between related keywords (e.g., AI, WWW, knowledge-based systems) during a period of five years before the appearance of the novel keywords. The same analysis was also conducted on a control group of mature keyword, to prove that the dynamics exhibited in the portion of graph in which a new topic emerges are statistically different from the ground noise.

I will now describe the main steps of the process.

4.1.1 Graph selection

The graph selection phase is intended to select a set of keywords from the whole network to test my hypothesis. I decided to first focus on the topics “semantic web” (debuting in 2001) and “cloud computing” (2006), since they are well-known research areas and I wanted to confirm the intuition that topics such as artificial intelligence and world wide web contributed to the birth of Semantic Web.

I also randomly selected ten keywords that closely resemble the debutant ones in terms of characteristics as *control group*. I labelled this set also as *non-debutant group*. All the *control group* keywords debuted some year before the debut of the main keywords and thus are associated to mature topics. Since different years can exhibit different dynamics of the keyword in the *non-debutant group* were analysed both in 2001 and 2006. In order to avoid further confusion between the keywords that will be employed and analysed in the following experiments, the keywords selected for these two particular groups will be referred as *generating keywords*.

Tab. 3: Group of debutant keywords used for the analysis.

Keyword	Year of debut
semantic web	2001
cloud computing	2006

Tab. 4: Group of non-debutant keywords used for the analysis.

Keyword	Year of non-debut (analysis)
automata theory	2001 and 2006
computer vision	2001 and 2006
constraint theory	2001 and 2006
cryptography	2001 and 2006
data structures	2001 and 2006
forecasting	2001 and 2006
knowledge management	2001 and 2006
model checking	2001 and 2006
multimedia systems	2001 and 2006
scheduling	2001 and 2006

To analyse the dynamics preceding the creation of the debutant keywords I selected, for each of these keywords the sub-graph including the twenty most co-occurring keywords, during the entire period of their activity, as showed in Fig. 6.

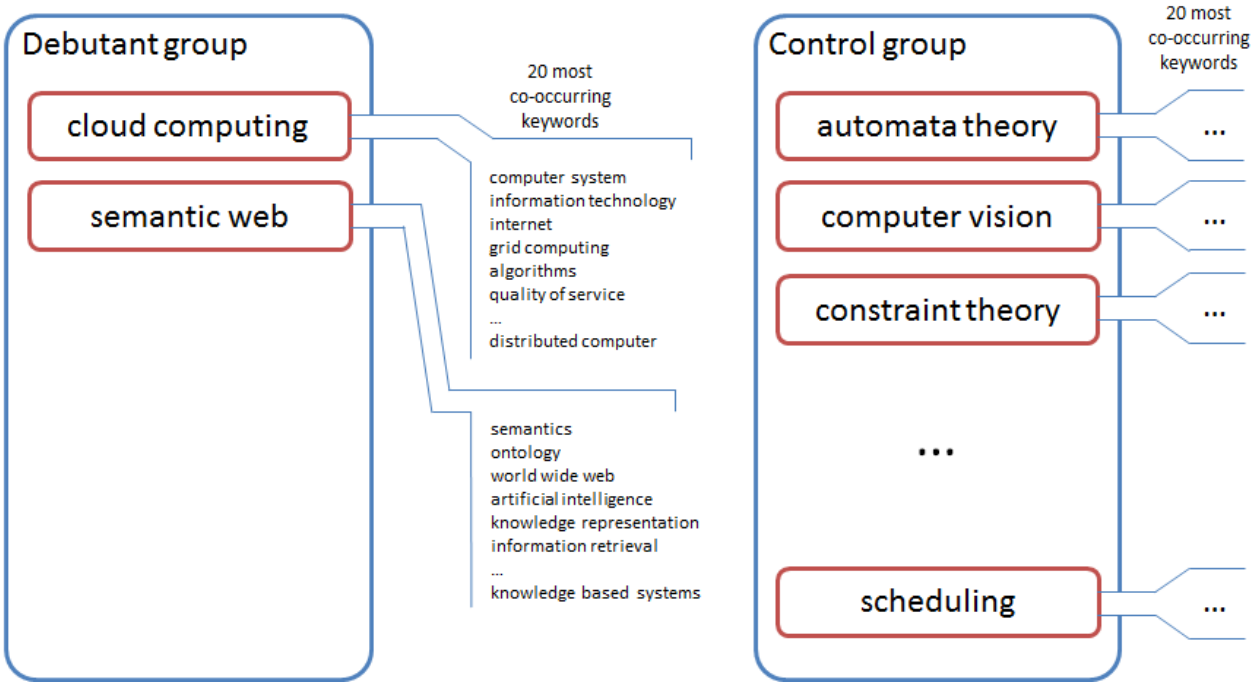

Fig. 6: Selection of keyword in the debutant group and in the control group with their twenty most co-occurring keywords.

Since the analysis was focused on the five years prior to the debut of the new topic, for each *generating keyword* I extracted five sub-graphs representing the related portion of the network in those years. For example, in the case of "semantic web", I selected the sub-graph including keywords such as "semantics", "ontology", "world wide web" and so forth. Considering that the keyword "semantic web" appeared as a keyword in 2001, the extraction phase collected five sub-graphs associated to the years 1996-2000, as shown in Fig. 7.

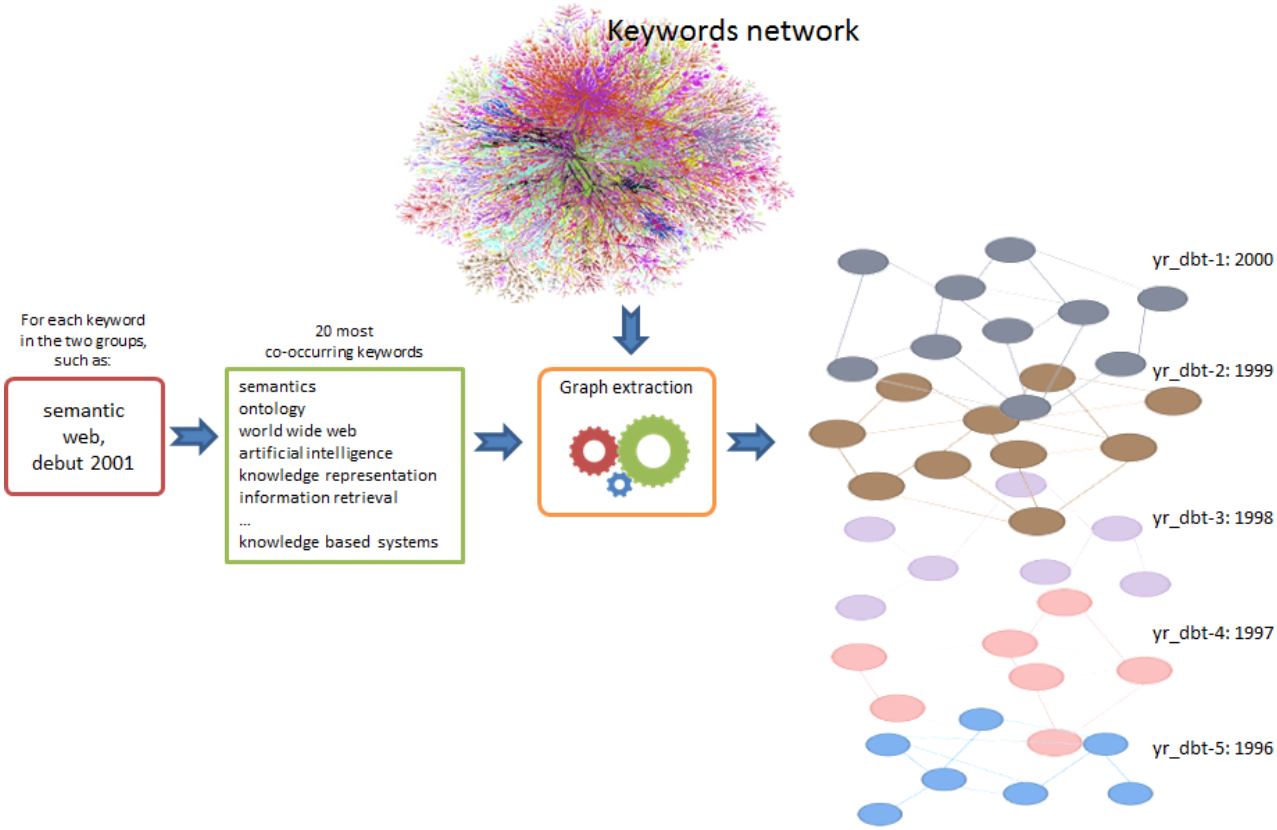

Fig. 7: An example of the entire workflow of the graph extraction phase.

4.1.2 Graph analysis

In the graph analysis phase, all the previous extracted graphs for both groups are analysed. In particular, the purpose of this analysis is to discern differences in the temporal evolution of the graphs associated to the two groups. As already stated, it is expected that an analysis of the extracted sub-graphs would highlight two different behaviours. From the sub-graph associated to the new topics, I expect a significant increase of collaboration between keywords. Meanwhile from the non-debutant group, it is expected that the pace in which the keywords co-occur would remain roughly constant.

The idea underneath this analysis, which is schematised in Pseudocode 1, is that the evolution of graphs can be tracked analysing how the weight associated to nodes and links evolve in subsequent years. I did so by analysing the changes of relevant 3-cliques of the graphs. Cliques and in general, k-cliques are complete sub-graphs of order *k* in which all the nodes are connected to each other. The order *k* defines the number of nodes in the complete sub-graph. For example, a 3-clique (*k=3*) is a sub-graph with three nodes and all the nodes are connected each other, which makes it look like a triangle. Other examples of cliques are in Fig. 8.

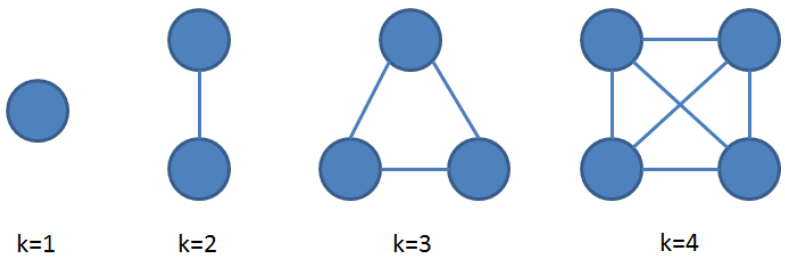

Fig. 8: First four example of the k-cliques family.

Analysing the evolution of the weights associated to nodes and links is thus equivalent to analysing 1-cliques and 2-cliques, respectively. However, using 3-cliques allows us to perform an analysis of the network at a higher level of abstraction. Arguably, also 4-cliques can be taken into account to perform this analysis however in this case the number of cliques extracted from the sub-graphs can be significantly lower than the number of 3-cliques and therefore yield less data points.

In order to perform this analysis, the sub-graphs were converted in adjacency matrices and then the 3-cliques were extracted using a modified implementation of the Bron–Kerbosch algorithm [92]. The Bron–Kerbosch algorithm is able to find maximal cliques in a graph and thus cliques with any order *k*. The alterations from the original version of the algorithm consisted in taking the output of the matrix and converting it in a series of 3-cliques, because as it is known from graph theory, a clique with high order contains also low order cliques.

For each generating keyword the related sub-graphs are converted into sets of cliques, containing the 3-cliques associated to the same keywords in the five years. Then the approach I implemented measures the increase in collaboration of the three keywords represented by the timelines of 3-cliques. Considering the clique example showed in Fig. 9, which shows three interconnected nodes *{A,B,C}*; it is possible to devise an index for measuring the increment in the collaboration of *{A, B, C}*, by exploiting both node weights (W_a, W_b, and W_c) and link weights (W_{ab}, W_{bc} and W_{ac}).

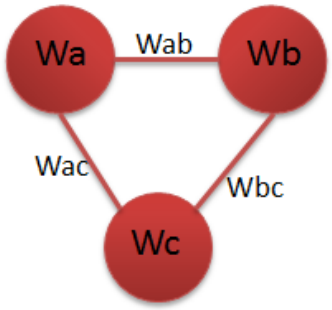

Fig. 9: Example of 3-clique with its associated weights to nodes and links.

I did so by using two different approaches, summarized in Eq. 1and Eq. 2. Both approaches were based on the computation for each couple of keywords A and B of the conditional probability P(B|A)= W_{AB}/W_A that a paper tagged with keyword A would also be tagged with keyword B. The main difference between these two approaches is how I combined these probability values. The first approach, showed in Eq. 1, first compute the strength of each link *{A,B}* as the harmonic mean of the conditional probabilities *P(A|B)* and *P(B|A)* and then computes μ_Δ as the harmonic mean of the three link strengths. The second approach uses the arithmetic mean instead of the harmonic mean, as showed in Eq. 2.

$$\mu_1 = harmmean\left(\frac{W_{AB}}{W_A}, \frac{W_{AB}}{W_B}\right)$$
$$\mu_2 = harmmean\left(\frac{W_{BC}}{W_B}, \frac{W_{BC}}{W_C}\right)$$
$$\mu_3 = harmmean\left(\frac{W_{CA}}{W_A}, \frac{W_{CA}}{W_C}\right)$$
$$\mu_\Delta = harmmean(\mu_1, \mu_2, \mu_3)$$

Eq. 1: Measure associated to the clique using the harmonic mean.

$$\mu_1 = mean\left(\frac{W_{AB}}{W_A}, \frac{W_{AB}}{W_B}\right)$$
$$\mu_2 = mean\left(\frac{W_{BC}}{W_B}, \frac{W_{BC}}{W_C}\right)$$
$$\mu_3 = mean\left(\frac{W_{CA}}{W_A}, \frac{W_{CA}}{W_C}\right)$$
$$\mu_\Delta = mean(\mu_1, \mu_2, \mu_3)$$

Eq. 2: Measure associated to the clique using the arithmetic mean.

Hence, as showed in Fig. 10, the timeline of cliques were reduced to timeline of measures, which I then analysed diachronically to detect an increment in the pace of co-occurrences. In order do so, I tested two main solutions. First, I simply evaluated the increase in collaborations by subtracting the extreme values as showed in Eq. 3.

$$direction = yr_{-1} - yr_{-5}$$

Eq. 3: Naïve approach to analyse the direction of a timeline of measures.

In this case, if the direction value was positive and therefore the measure associated to the year before the debut was higher than the measure associated to the fifth year prior to the debut, it means that the number of co-occurrences and the number of paper is increasing. Conversely, if the direction is negative, it would mean that the collaboration between the three keywords is getting weaker in time. The second and less naive approach exploits a linear interpolation method that for each cliques takes the five measures in input and uses least-squares approximation to determine the equation representing their linear regression (Eq. 4).

$$f(x) = \alpha x + \beta$$

Eq. 4: Function for a geometric line.

The slope α is then used as index for assessing the trend of the clique in time. In particular, if the slope is positive it means that the measure is increasing over time and therefore the co-occurrence between the keywords associated to that clique are increasing as well. This implies that the topics *A*, *B* and *C* are getting 'closer' to each other in term of collaborations. On the other hand, if the slope is negative, it means that the measures in time are decreasing and thus the keywords are growing apart.

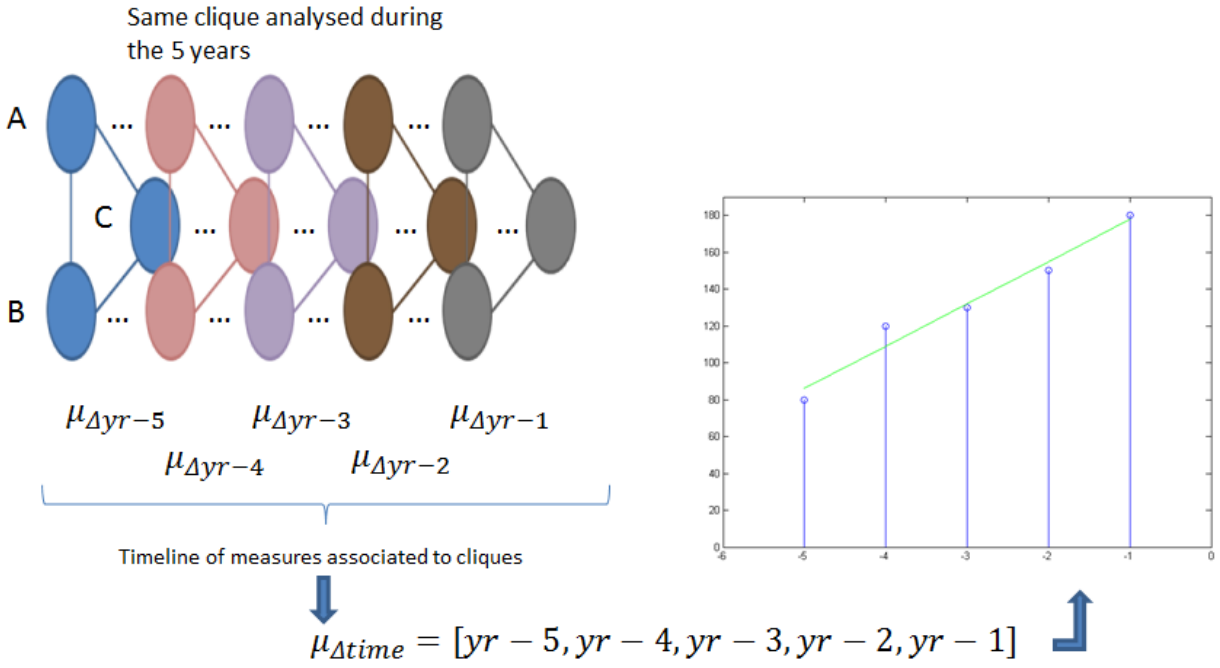

Fig. 10: Timeline of measures associated to a timeline of cliques.

Pseudocode:
input: sub-graphs of keywords network
output: direction indexes

for each generating keyword **in** [debutant group, control group] do
    read graphs in the five years prior its debut
    create adjacency matrices
    extract 3-cliques
    collect all similar cliques in time creating timelines of cliques
    for each clique
        associate measure
    end for each
    for each timeline of measures
        extract the slope of the linear interpolation
    end for each
    analyse the slopes taking statistical values
end for each

Pseudocode 1: Workflow representing the graph analysis phase.

4.2 Experiment zero

The experiment zero, which is considered the pilot experiment, has been useful to understand some properties of the data and then design the approach described in section 4.1.
The workflow adopted in this experiment is similar to the one just described except for some variations. In particular the measure associated to the clique included only the weight of the links and not the weights of the nodes showed in Fig. 11. The final measure associated to the clique was computed as the harmonic mean of the weights, as shown in Eq. 5.

$$\mu_{\Delta} = harmmean(W_{AB}, W_{BC}, W_{CA})$$

Eq. 5: Measure associated to the clique using the harmonic mean on the weights of links.

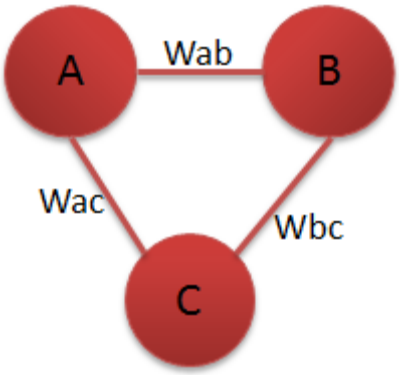

Fig. 11: Example of 3-clique with its associated weights to links.

This first experiment did not show any significant difference between the two classes of generating keywords, but allowed a better understanding of the data bias. In particular, I discovered that researchers seemed to tag their papers with a higher number of keywords in more recent years and thus as side effect the keywords network becomes more dense. This bias needs to be taken into account for a comprehensive analysis of the keyword network.

To alleviate this problem I introduced two improvements to the initial approach. The first one consisted of pruning some of the links in the sub-graphs extracted. In particular, links between keywords with weight equal or less than *n* (*n=3* in the prototype) are cut out with the aim of reducing the noise created by the increasing number of keywords for each paper. This solution allows also to speed up the computation, since it reduces the number of cliques to be analysed. I also changed the way I used to compute the tendency of a clique, to the one previously described in section 4.1, which uses the weights of the nodes to normalize co-occurrence values. This new metric reduces the bias since the number of co-occurrences and total publications tend to rise in time with a similar pace.

4.3 Experiment one

This new experiment follows all the phases described in section 4.1. In particular, the generating keywords for the debutant group and the control group are respectively the ones showed in Tab. 3 and

Tab. 4. Then, using their twenty most co-occurring keywords, the respective sub-graphs in the five year prior to the analysis year (debut) have been extracted from the keywords network. The extracted sub-graphs were converted in adjacency matrices for the cliques extraction and subsequently timelines of cliques have been converted in timelines of measures to understand the overall trend of the graphs.
In this experiment, I tried all possible combinations of approaches for associating a measure to a clique and the methodologies to understand the direction of the timelines of measures discussed in section 4.1, with the aim of understanding which combination of techniques is most effective.
First I adopted the harmonic mean showed in Eq. 1 and the linear interpolation showed in Eq. 4.
As discussed previously, the process of computing all the cliques and timelines returns per each generating keyword a series of slopes that can be used to derive the degree of collaboration of related keywords. For example, for the keyword "semantic web" the process returned 120 timelines of cliques to which a direction (slopes) value had been identified. To understand the overall behaviour of the extracted sub-graphs which are linked to the "semantic web", the mean value for these timeline directions was computed, yielding 0.219. The positive value confirms that, as hypothesised, the keywords which are related to a new topic accelerate their collaboration ratio.
Tab. 5 shows the result of this analysis for "semantic web" and "cloud computing" while

Tab. 6 shows the results obtained in the control group. The mean values are higher in the debutant group than the control group. Furthermore,

Tab. 6 shows also that for the same keyword the mean value increase from 2001 to 2006. This effect it is due to the fact that keyword networks become denser and the amount of co-occurrences increases in time. Therefore, introducing the node weights for normalizing the link weights has been beneficial, but it did not remove this effect entirely. However, even if the time effect still exists, comparing the mean value of the "semantic web" as debutant against the generating keywords of the 2001 control group allows us to distinguish the two classes of generating keywords. The same happens when comparing the mean value of "cloud computing" as debutant against the control group in the 2006.

Tab. 5: Overall directions of the sub-graphs related to the generating keywords in the debutant group.

Generating keyword analysed	Mean value
cloud computing_2006	0.257
semantic web_2001	0.219

Tab. 6: Overall directions of the sub-graphs related to the generating keywords in the control group.

Generating keyword analysed	Mean value
automata theory_2001	0.018
automata theory_2006	0.192
computer vision_2001	-0.099
computer vision_2006	0.166
constraint theory_2001	-0.045
constraint theory_2006	0.152
cryptography_2001	0.146
cryptography_2006	0.275
data structures_2001	0.038
data structures_2006	0.077
forecasting_2001	0.127
forecasting_2006	0.207

knowledge management_2001	0.091
knowledge management_2006	0.205
model checking_2001	0.001
model checking_2006	0.219
multimedia systems_2001	0.067
multimedia systems_2006	0.132
scheduling_2001	0.004
scheduling_2006	0.170

Fig. 12 depicts the distribution of the debutant group versus the distribution of the control group. The abscissa represents the slopes of the linear function associated to the evolution of the cliques, while the ordinate represents the percentage number of cliques which fall in each slope interval. We can see that the two distributions seem to differ from each other and that the distribution referred to the debutant group is shifted towards high slope values.

In order to verify that the two groups (debutant and control) effectively belong to different populations and thus that the initial hypothesis is supported by empirical evidence, we ran on the two distributions the Student's t-test, which yielded $4.02*10^{-10}$ as p-value. We can thus reject the hypothesis that the differences between the two distributions are due to chance or random variations.

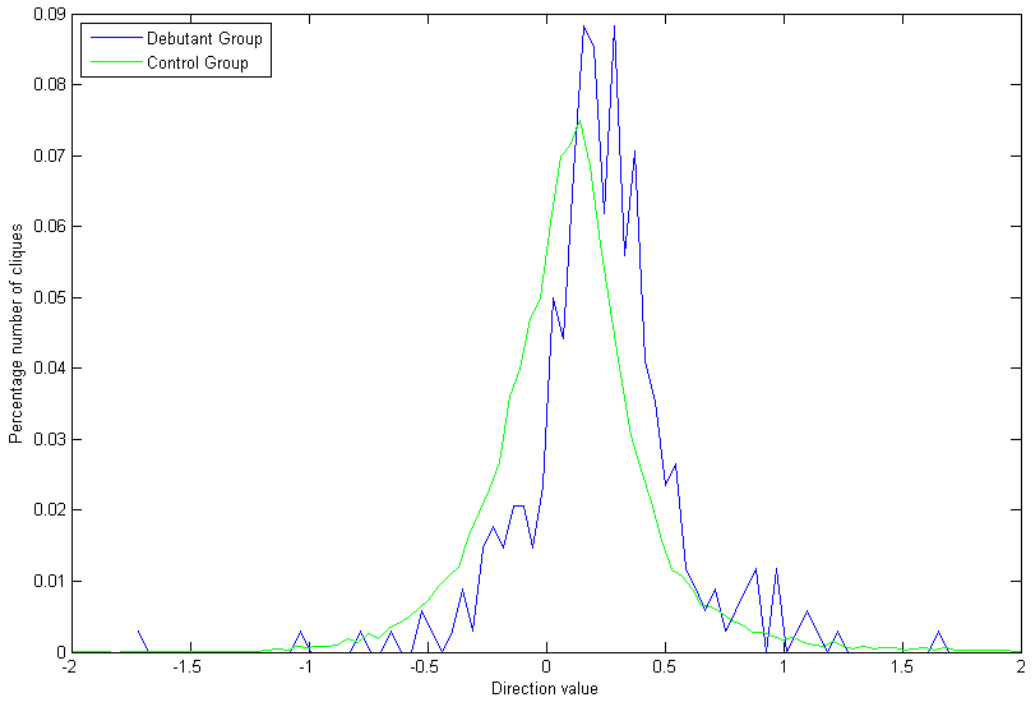

Fig. 12: Distribution (histogram) of the direction values of both the debutant group and the non-debutant group.

The Student's t-test has been performed also on the specific case of "semantic web" direction distribution against the generating keywords of the control group analysed in 2001 and the case of "cloud computing" against the generating keywords of the control group analysed in 2006.

In particular, for "semantic web" the p-value returned by the Student's t-test is $1.75*10^{-10}$ while for "cloud computing" the p-value returned is 0.0028. Hence, also in these two cases the distributions differ

significantly from each other. Fig. 13 and Fig. 14 show respectively the distribution of slopes in 2001 and 2006.

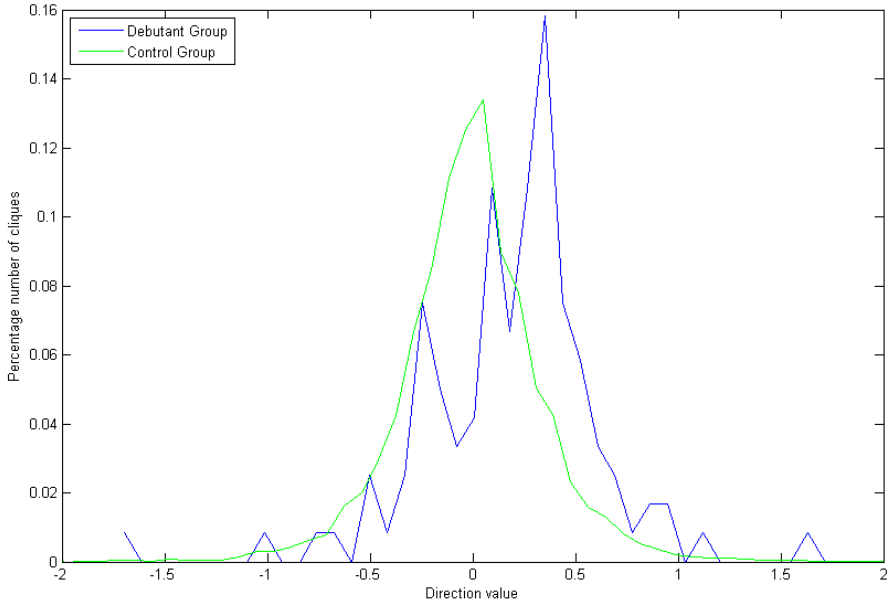

Fig. 13: Distribution (histogram) of the direction values of the "semantic web" for the debutant group and the non-debutant group in the 2001.

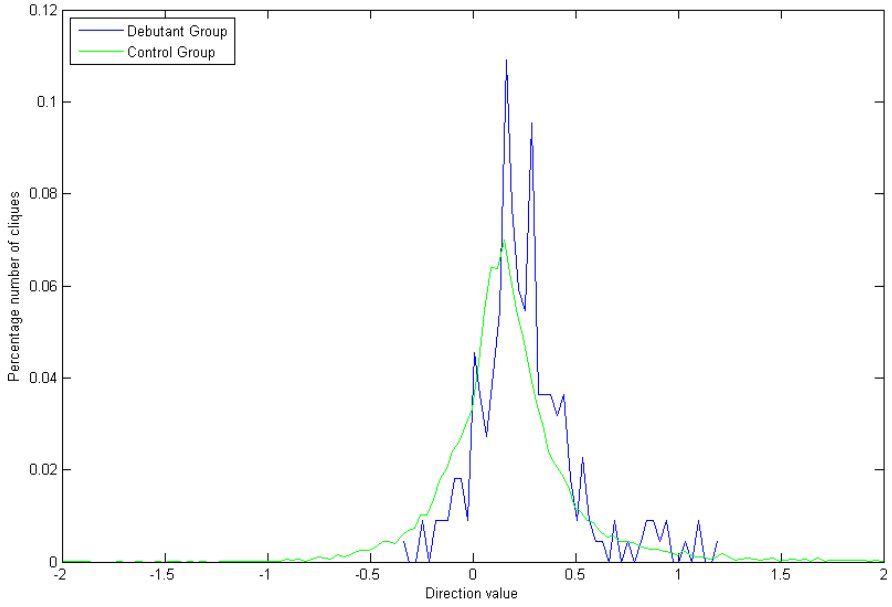

Fig. 14: Distribution (histogram) of the direction values of the "cloud computing" for the debutant group and the non-debutant group in the 2006.

Another interesting aspect emerged from this analysis. Observing carefully Fig. 13 and Fig. 14, some spikes towards high values of slopes can be spotted. Basically, these spikes represent cliques with maximum value of slopes found in the graph and therefore they had the highest growth.

Ranking the cliques based on the computed slope we can find some interesting aspects. For the “semantic web”, as also shown in Tab. 7, it is possible to appreciate the involvement of keywords like the “artificial intelligence”, the “world wide web” and the “knowledge based systems” which proves the aforementioned hypotheses about the creation of the Semantic Web.

Tab. 7: Ranking of the cliques with highest slope value for the "semantic web".

Keyword 1	Keyword 2	Keyword 3	Score
world wide web	information retrieval	search engines	2.529
world wide web	user interfaces	artificial intelligence	1.127
world wide web	knowledge based systems	knowledge representation	0.982
world wide web	artificial intelligence	knowledge representation	0.974
world wide web	user interfaces	knowledge representation	0.885
world wide web	knowledge based systems	artificial intelligence	0.850
world wide web	information retrieval	knowledge representation	0.803

Instead, for the "cloud computing", as shown in Tab. 8, it is possible to see involvement of the "web services", "grid computing" and the "distributed computer systems".

Tab. 8: Ranking of the cliques with highest slope value for the "cloud computing".

Keyword 1	Keyword 2	Keyword 3	Score
grid computing	distributed computer systems	web services	1.208
information technology	information management	web services	1.094
grid computing	quality of service	distributed computer systems	1.036
internet	quality of service	web services	0.951
web services	information management	distributed computer systems	0.949
grid computing	virtual reality	distributed computer systems	0.902
computer systems	information technology	information management	0.888
information technology	distributed computer systems	web services	0.874
grid computing	quality of service	web services	0.848
information technology	internet	web services	0.841
internet	distributed computer systems	web services	0.805
quality of service	information management	web services	0.762
information technology	quality of service	web services	0.697
internet	information management	web services	0.680
computer systems	information technology	distributed computer systems	0.629

The other three combinations of the approaches for computing the measure associated to each clique and the tendency of cliques in time yielded less good results and did not fully succeed in discriminating the debutant from the control group for two main reasons. The overall directions (mean values) of the generating keywords of both groups have the tendency to overlap and some distributions of directions do not succeed the Student's t-test. Tab. 9 and Tab. 10 show respectively the overall direction of the clique timelines for the debutant group and the control group for all these three combinations of approaches.

Tab. 9: Overall directions of the sub-graphs related to the generating keywords in the debutant group with the other three combinations of approaches.

Generating keyword analysed	Harmonic mean and Naïve approach	Arithmetic mean and Interpolation	Arithmetic mean and Naïve approach
cloud computing_2006	1.021	0.274	0.836
semantic web_2001	0.854	0.190	0.475

Tab. 10: Overall directions of the sub-graphs related to the generating keywords in the control group with the other three combinations of approaches.

Generating keyword analysed	Harmonic mean and Naïve approach	Arithmetic mean and Interpolation	Arithmetic mean and Naïve approach
automata theory_2006	0.659	0.293	0.721
computer vision_2006	0.526	0.131	0.080
constraint theory_2006	0.424	0.224	0.385
cryptography_2006	1.067	0.203	0.510
data structures_2006	0.266	0.155	0.315
forecasting_2006	0.931	0.209	0.733
knowledge management_2006	0.931	0.193	0.821
model checking_2006	0.732	0.360	0.842
multimedia systems_2006	0.371	0.096	-0.140
scheduling_2006	0.477	0.295	0.499
automata theory_2001	0.154	-0.019	-0.165
computer vision_2001	-0.356	0.003	-0.016
constraint theory_2001	-0.196	-0.073	-0.403
cryptography_2001	0.903	0.551	1.920
data structures_2001	0.525	0.190	0.407
forecasting_2001	-0.092	0.884	15.387
knowledge management_2001	0.472	-0.156	-0.612
model checking_2001	0.060	-0.061	-0.255
multimedia systems_2001	0.315	0.097	0.318
scheduling_2001	0.023	0.007	-0.057

4.4 Experiment two: hard pruning

As discussed earlier, in *experiment one* I introduced a pruning phase in which links between keywords having a weight equal or less than 3 were cut out. For this experiment, I cut out all the links having weight equal or less than 10, on the basis of the hypothesis that this would further reduce the noise derived from the dense keyword network and improve the discriminatory power of my approach, and also because it may significantly reduce the computational cost. However, this hypothesis proved to be false. Actually, the overall direction of the sub-graphs associated to the debutant group decreased meanwhile the overall direction of the sub-graphs associated to the control group increased, making the two distributions more overlapped, as shown in Fig. 15.
Moreover, performing the Student's t-test on the two distributions of directions of timelines returns 0.27 which does not allow us to reject the null hypothesis.

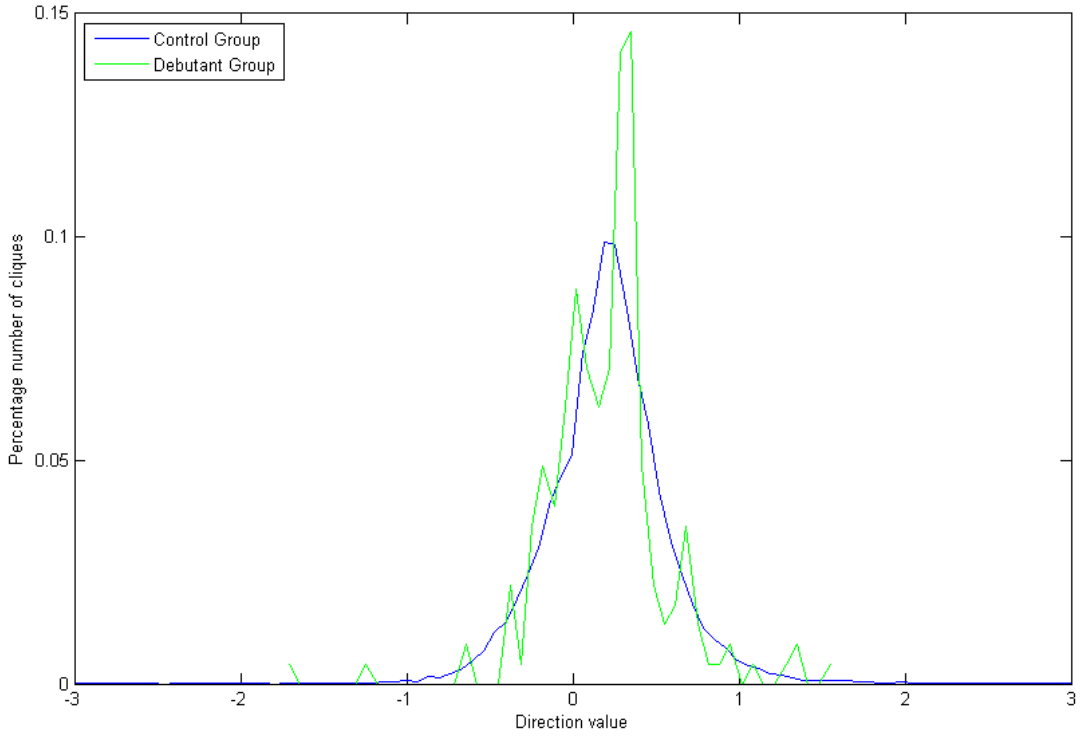

Fig. 15: Distribution (histogram) of the direction values of both the debutant group and the non-debutant group for experiment two.

Analysing the number of cliques resulting from pruning the links <=3 versus pruning the links <=10, we see that in the second case we lose about 43% of cliques for the debutant group and 15% for the control group. Thus, it seems that the hard pruning affected too much the debutant keyword networks, causing a significant loss of information. This can be explained by the fact that in the control group many links between nodes are stable and characterized by a high number of co-occurrences while for the debutant group many informative links are still quite weak and associated with a low number of co-occurrences. Pruning these networks simply on the basis of the number of co-occurrences is thus an ineffective approach.

4.5 Experiment three: timeline analysis

Since the hard pruning was not effective, I tried a different approach for alleviating the noise and improve computation time. I noticed that a good number of the time series exhibit an intermittent behaviour and some of them were actually not useful to the discriminating process. I thus implemented a new filtering module for pruning uninformative time series. This module analyses all the timelines of measures and prunes them according to a defined heuristic. Basically, this process discard the timelines of measures containing too many zeros or sequences of values that can lead to an erroneous estimation of the direction. For example, timelines containing four zeros and a non-zero value will be discarded.

Tab. 11 shows some common examples of timelines of measures and the strategies applied to them. Each year in the timeline can be associated either with a zero or a number (indicated by a *X* in the table). It should be noted that zero value can occur in two cases: i) when the mean value associated to the clique produces zero or ii) when in that specific year the clique did not actually exist.

Tab. 11: Standards for timeline of measures filtering.

Debut year -5	Debut year -4	Debut year -3	Debut year -2	Debut year -1	Strategy
0	0	0	0	X	Removed, too few information
0	0	X	0	X	Removed, one link is close to zero
X	0	0	X	X	Retained, the first element is zeroed
0	0	X	X	X	Retained
0	X	X	X	X	Retained
X	X	X	X	X	Retained
0	0	0	X	X	Retained
0	X	0	0	X	Removed, sparse values

Performing the Student's t-test on the distribution returned by this new experiment yields a p-value equal to $1.22*10^{-8}$. However, Tab. 12 and Tab. 13 show the overall directions of both the debutant group have the tendency to overlap. An analysis of the data showed that also in this case the effect of removing timelines of cliques introduced an unfair bias on the debutant because of the weaker connections associated to them.

Tab. 12: Overall directions of the sub-graphs related to the generating keywords in the debutant group for the timeline analysis.

Generating keyword analysed	Mean value
cloud computing_2006	0.275
semantic web_2001	0.270

Tab. 13: Overall directions of the sub-graphs related to the generating keywords in the control group for the timeline analysis.

Generating keyword analysed	Mean value
automata theory_2006	0.225
computer vision_2006	0.193
constraint theory_2006	0.169
cryptography_2006	0.317
data structures_2006	0.097
forecasting_2006	0.208
knowledge management_2006	0.238
model checking_2006	0.232
multimedia systems_2006	0.150
scheduling_2006	0.178
automata theory_2001	0.021
computer vision_2001	-0.086
constraint theory_2001	-0.032
cryptography_2001	0.239
data structures_2001	0.201
forecasting_2001	0.130
knowledge management_2001	0.136
model checking_2001	-0.006
multimedia systems_2001	0.103
scheduling_2001	0.010

4.6 Experiment four: introducing semantics in the keyword networks

As discussed in the literature review, the use of keywords as a proxy for topics brings several drawbacks. Usually keywords tend to be noisy because some of them do not represent a topic, such as “case study”. Moreover, they also suffer of synonymy and polysemy since many keywords can refer to the same topic or a single keyword can represent more topics. A way to address this limitation is using a semantic topic model. Osborne et al. [25] do so by means of Klink, an algorithm which analyses networks of research entities (including papers, authors, venues, and technologies) to infer three kinds of semantic relationships between topics. Recently a new version of Klink was introduced, Klink-2 [93], which takes advantage of multiple knowledge sources available on the web. Klink-2 outputs a semantic topic network that can support many kinds of analytics on the research environment.

I thus decided to integrate this semantic network with the keyword network used in the previous experiments. The resulting experiment follows the same workflow used in *experiment one*, but it introduces two further improvements based on the use of this semantic knowledge source. First, I discarded all the keywords which do not represent a proper research topic according to the taxonomy of topics produced by Klink-2. For example this new approach would discard keywords like “data centers” and “clouds”. Secondly, keywords which refer to the same topic are grouped together and treated as a single semantic topic. To this end I exploited the *relatedEquivalent* relationships inferred by Klink-2. For example, keywords like “multi agent systems”, “multiagent system”, “multi-agent systems”, “multi-agent system” and “multi agent system (mas)” will be treated as a single entity.

I then performed two sub-experiments, implementing these new techniques. The first one used as usual graphs of 20 keywords, whereas the second one uses 40 keywords. The aim was to understand if extending the graph could be beneficial for the analysis.

Tab. 14 and Tab. 15 show the overall directions of the graphs related to the two sub-experiments. Focusing on the first experiment, it can be seen that the gap between the direction of the graph generated for the "semantic web" and the overall directions of the graphs extracted for the control groups in the 2001 increased. Actually many graphs in 2001 have a negative or almost zero direction, which implies that these graphs seems to be regular in time and also have the tendency to decrease their collaboration. The only exception in this case is "cryptography" which presents an overall direction comparable to the "semantic web". This exception can be justified by considering that even if cryptography was not a debutant in those years, it was actually a hot topic considering the security issue related to communication protocols. However, if we extend the sub-graph extracted with the 40 most co-occurring keywords this phenomenon is less relevant.

Meanwhile, the gap between the direction of the graph extracted for the "cloud computing" and the overall direction obtained for the generating keywords of the non-debutant group is still close. Instead, using the 40 most co-occurring keywords the overall directions between the previously compared kinds of graphs tend to diverge. To conclude the analysis, in both cases, the Student's t-test has been performed returning values of less than 5% indicating that the null hypothesis can be rejected.

This experiment presents two main interesting outcomes. The first is that the introduction of semantic technologies seems effective in discriminating the debutant group from the control group. The second one is that increasing the number of keywords in the examined graph is also beneficial. Therefore, an important question is: how large should be the set of co-occurring keywords to provide an optimal measure to discern the two groups? Moreover, since 40 most co-occurring keywords solve the problem in the 2001 scenario, as reported in the two tables, but it only slightly improves the analysis for the 2006 one, it seems that this number may also depend on characteristics of keyword graphs which may change over time.

Tab. 14: Overall directions of the sub-graphs related to the generating keywords in the debutant group for the experiment four with 20 and 40 most co-occurring keywords.

Generating keyword analysed	Mean value [20 most co-occ]	Mean value [40 most co-occ]
cloud computing_2006	0.315	0.311
semantic web_2001	0.428	0.254

Tab. 15: Overall directions of the sub-graphs related to the generating keywords in the control group for the experiment four with 20 and 40 most co-occurring keywords.

Generating keyword analysed	Mean value [20 most co-occ]	Mean value [40 most co-occ]
automata theory_2006	0.329	0.236
computer vision_2006	0.106	0.123
constraint theory_2006	0.229	0.193
cryptography_2006	0.218	0.263
data structures_2006	0.265	0.210
forecasting_2006	0.251	0.217
knowledge management_2006	0.225	0.179
model checking_2006	0.368	0.231
multimedia systems_2006	0.099	0.144
scheduling_2006	0.305	0.222
automata theory_2001	0.018	-0.117
computer vision_2001	-0.085	-0.106
constraint theory_2001	-0.110	-0.096
cryptography_2001	0.479	0.036
data structures_2001	-0.137	-0.106
forecasting_2001	-0.087	-0.088
knowledge management_2001	-0.035	-0.174
model checking_2001	-0.139	-0.072
multimedia systems_2001	0.077	-0.147
scheduling_2001	-0.022	-0.090

4.7 Future plans

The experiments have shown that analysing the evolution of some portions of the keyword network provides useful insights in understanding the dynamics that lead to the appearance of a new topic. Even though these are still preliminary experiments, the approach seems to be very promising. Moreover, it revealed some important properties of data, such as the fact that these keyword networks tend to become denser in time and some other issues, like noise.

As a next step, I plan to do an additional evaluation on the current method and to experiment with different techniques for selecting graphs of research entities. Initially, I plan to replicate the described experiment with a larger number of keywords, to gain a better understand of the performance on larger numbers and to collect more evidence on the aforementioned dynamics. In second instance, I will investigate other techniques to enrich the current approach for detecting embryonic topics. For instance, I plan to test community detection algorithms with the aim of highlighting sub-graphs that exhibit dynamics associated to the emergence of novel research areas. I plan to analyse these sub-graphs using a set of significant features, such as degree distribution, clustering coefficient, betweenness centrality and so forth [94], aiming to find a correlation between them and the probability that a new topic would emerge from a certain region of the graph.

Finally, I plan to improve my approach by allowing it to exploit a variety of other research entities (e.g., authors, publications venues) with the goal of making the inference process more robust. For example, I intend to analyse different kinds of topic networks that adopt alternative metrics for assessing the weights, such as the number of new collaborations between authors or the impact of related venues.

Chapter 5 Summary

This report presented the results of the first year of my doctoral work, whose aim is to provide an approach to the early detection and forecast of research trends. This approach will be based on a semantic characterization of research entities, on the statistical analysis of research dynamics and on the integration of scholarly and social media data. As previously discussed in the problem statement some systems can partially accomplish part of this task, but they are not able to detect embryonic topic and do not provide any support to forecast the impact of topics in their early phase. A number of stakeholders, such as funding bodies, journals editors and researchers would find great benefits in an approach that would address these limitations and allow a better insight in the future of research.

In the literature review section, I presented an overview of technologies and currently available approaches that are related to the goal of my doctoral work, also identifying some open issues that this area is still facing.

Based on the defined gaps, I elaborated the main research question driving this work. Subsequently, the question was divided into several sub-questions in order to make the problem more granular, so to better define a strategy.

Consequently, I formulated hypotheses about the evolution of topics as well as the research entities that can influence this evolution, thanks to their relation to research topics. Building upon these hypotheses, I proposed a research plan that defines some key points of this work: i) data integration, ii) exploration of the research dynamics, iii) topic detection, and iv) impact forecasting. I also outlined a methodology for answering the research questions.

The initial experiments on Rexplore data and in particular on the semantic topic network have been useful to verify some initial hypotheses. Moreover, the same experiments proposed very promising results in terms of recognising whether a temporal portion of the graph of keywords is developing a new topic.

My work during the next two years will focus on building a comprehensive model, which takes into account the semantic relationships between research elements and is able to perform the detection of research trends at their early stage. Equally important will be the design of a methodology able to forecast with high accuracy the impact of research topics.

References

1. Wu, F.-S., et al., *A systematic approach for integrated trend analysis—The case of etching.* Technological Forecasting and Social Change, 2011. **78**(3): p. 386-407.
2. Decker, S.L., et al., *Detection of bursty and emerging trends towards identification of researchers at the early stage of trends*. 2007: (Doctoral dissertation, University of Georgia).
3. Berners-Lee, T., J. Hendler, and O. Lassila, *The semantic web.* Scientific american, 2001. **284**(5): p. 28-37.
4. Tseng, Y.-H., et al., *A comparison of methods for detecting hot topics.* Scientometrics, 2009. **81**(1): p. 73-90.
5. Behrens, H. and P. Luksch, *Mathematics 1868–2008: a bibliometric analysis.* Scientometrics, 2010. **86**(1): p. 179-194.
6. Krampen, G., A. von Eye, and G. Schui, *Forecasting trends of development of psychology from a bibliometric perspective.* Scientometrics, 2011. **87**(3): p. 687-694.
7. Krishnan, A., *What are academic disciplines.* Some observations on the disciplinarity vs interdisciplinarity debate [NCRM working paper series 03/09]. Southampton: University of Southampton National Centre for Research Methods, 2009.
8. Tang, J., et al. *Arnetminer: extraction and mining of academic social networks*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.
9. Monaghan, F., et al. *Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food*. in *SW Challenge - ISWC*. 2010. Citeseer.
10. Osborne, F., E. Motta, and P. Mulholland, *Exploring scholarly data with rexplore*, in *The Semantic Web–ISWC 2013*. 2013, Springer. p. 460-477.
11. Diederich, J., W.-T. Balke, and U. Thaden. *Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp*. in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. 2007. ACM.
12. Li, H., et al. *CiteSeerx: an architecture and web service design for an academic document search engine*. in *Proceedings of the 15th international conference on World Wide Web*. 2006. ACM.
13. Tang, J., et al. *Arnetminer: extraction and mining of academic social networks*. in *Proceedings of the 14th ACM SIGKDD*. 2008. ACM.
14. Osborne, F. and E. Motta. *Rexplore: Unveiling the dynamics of scholarly data*. in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*. 2014. IEEE.
15. Maurer, H. and M. Salman Khan, *Research trends in the field of e-learning from 2003 to 2008.* Interactive Technology and Smart Education, 2010. **7**(1): p. 5-18.
16. Lv, P., et al., *Bibliometric trend analysis on global graphene research.* Scientometrics, 2011. **88**(2): p. 399-419.
17. Jin, J.H., S.C. Park, and C.U. Pyon, *Finding research trend of convergence technology based on Korean R&D network.* Expert Systems with Applications, 2011. **38**(12): p. 15159-15171.
18. Duvvuru, A., et al., *Analyzing Structural & Temporal Characteristics of Keyword System in Academic Research Articles.* Procedia Computer Science, 2013. **20**: p. 439-445.

19. Eysenbach, G., *Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact.* Journal of medical Internet research, 2011. **13**(4).
20. Osborne, F., G. Scavo, and E. Motta, *Identifying diachronic topic-based research communities by clustering shared research trajectories*, in *The Semantic Web: Trends and Challenges*. 2014, Springer. p. 114-129.
21. Zhao, Z., et al., *Topic oriented community detection through social objects and link analysis in social networks.* Knowledge-Based Systems, 2012. **26**: p. 164-173.
22. Garfield, E., *Scientography: Mapping the tracks of science.* Current Contents: Social & Behavioural Sciences, 1994. **7**(45): p. 5-10.
23. Sitarz, R., *Identification of research trends in the field of separation processes. Application of epidemiological model, citation analysis, text mining, and technical analysis of the financial markets.* Acta Universitatis Lappeenrantaensis, 2013.
24. Chen, T.-t. and M.R. Lee. *Revealing the research themes and trends in Knowledge Management studies*. in *Machine Learning and Cybernetics, 2008 International Conference on*. 2008. IEEE.
25. Osborne, F. and E. Motta, *Mining semantic relations between research areas*, in *The Semantic Web-ISWC 2012*. 2012, Springer. p. 410-426.
26. Van den Besselaar, P. and L. Leydesdorff, *Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence.* Journal of the American Society for Information Science, 1996. **47**(6): p. 415-436.
27. OECD, *Frascati Manual 2002*. 2002: OECD Publishing.
28. Kuhn, T.S., *The structure of scientific revolutions*. First Edition ed. 1962: Chicago: University of Chicago Press.
29. Ding, Y., *Community detection: topological vs. topical.* Journal of Informetrics, 2011. **5**(4): p. 498-514.
30. Valiela, I., *Doing Science-Design, Analysis, and Communication of Scientific Research.* Doing Science-Design, Analysis, and Communication of Scientific Research, by Ivan Valiela, pp. 304. Foreword by Ivan Valiela. Oxford University Press, Jan 2001. ISBN-10: 0195134133. ISBN-13: 9780195134131, 2001. **1**.
31. Yan, S. and D. Lee, *Toward alternative measures for ranking venues: a case of database research community*, in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. 2007, ACM: Vancouver, BC, Canada. p. 235-244.
32. Lewallen, L.P. and P.B. Crane, *Choosing a Publication Venue.* Journal of Professional Nursing, 2010. **26**(4): p. 250-254.
33. Harnad, S. and T. Brody, *Comparing the impact of open access (OA) vs. non-OA articles in the same journals.* D-lib Magazine, 2004. **10**(6).
34. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* the Journal of machine Learning research, 2003. **3**: p. 993-1022.
35. Giles, C.L., K.D. Bollacker, and S. Lawrence. *CiteSeer: An automatic citation indexing system*. in *Proceedings of the third ACM conference on Digital libraries*. 1998. ACM.
36. Lawrence, S., C. Lee Giles, and K. Bollacker, *Digital libraries and autonomous citation indexing.* Computer, 1999. **32**(6): p. 67-71.
37. Baskurt, O., *Time series analysis of publication counts of a university: what are the implications?* Scientometrics, 2011. **86**(3): p. 645-656.
38. Liu, X., et al., *Co-authorship networks in the digital library research community.* Inf. Process. Manage., 2005. **41**(6): p. 1462-1480.
39. Erten, C., et al., *GraphAEL: Graph Animations with Evolving Layouts*, in *Graph Drawing*, G. Liotta, Editor. 2004, Springer Berlin Heidelberg. p. 98-110.
40. Erten, C., et al., *Exploring the computing literature using temporal graph visualization*, in *Electronic Imaging 2004*. 2004. p. 45-56.
41. Blei, D.M., *Probabilistic topic models.* Communications of the ACM, 2012. **55**(4): p. 77-84.

42. Tang, J., et al., *ArnetMiner: extraction and mining of academic social networks*, in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, ACM: Las Vegas, Nevada, USA. p. 990-998.
43. Redner, S., *How popular is your paper? An empirical study of the citation distribution.* The European Physical Journal B-Condensed Matter and Complex Systems, 1998. **4**(2): p. 131-134.
44. Steyvers, M., et al., *Probabilistic author-topic models for information discovery*, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, ACM: Seattle, WA, USA. p. 306-315.
45. Rosen-Zvi, M., et al., *The author-topic model for authors and documents*, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. 2004, AUAI Press: Banff, Canada. p. 487-494.
46. Yan, E., et al., *Topics in dynamic research communities: An exploratory study for the field of information retrieval.* Journal of Informetrics, 2012. **6**(1): p. 140-153.
47. Faust, K., *Centrality in affiliation networks.* Social Networks, 1997. **19**(2): p. 157-191.
48. Holme, P., et al., *Korean university life in a network perspective: Dynamics of a large affiliation network.* Physica A: Statistical Mechanics and its Applications, 2007. **373**: p. 821-830.
49. Gläser, J., *'Producing Communities' as a Theoretical Challenge.* Proceedings of The Australian Sociological Association, 2001: p. 1-11.
50. Becher, T. and P. Trowler, *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. 2001: McGraw-Hill International.
51. Clark, B.R., *Faculty culture*. 1962: Center for the Study of Higher Education, University of California.
52. Perkin, H., *1. THE HISTORICAL PERSPECTIVE.* Perspectives on higher education: Eight disciplinary and comparative views, 1984: p. 17.
53. Liu, X., T. Jiang, and F. Ma, *Collective dynamics in knowledge networks: Emerging trends analysis.* Journal of Informetrics, 2013. **7**(2): p. 425-438.
54. Girvan, M. and M.E.J. Newman, *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences, 2002. **99**(12): p. 7821-7826.
55. Radicchi, F., et al., *Defining and identifying communities in networks.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(9): p. 2658-2663.
56. Yang, B., et al., *Hierarchical community detection with applications to real-world network analysis.* Data & Knowledge Engineering, 2013. **83**: p. 20-38.
57. De Meo, P., et al., *Enhancing community detection using a network weighting strategy.* Information Sciences, 2013. **222**: p. 648-668.
58. Gong, M., et al., *Community detection in networks by using multiobjective evolutionary algorithm with decomposition.* Physica A: Statistical Mechanics and its Applications, 2012. **391**(15): p. 4050-4060.
59. Clauset, A., M.E. Newman, and C. Moore, *Finding community structure in very large networks.* Physical review E, 2004. **70**(6): p. 066111.
60. Xia, Z. and Z. Bu, *Community detection based on a semantic network.* Knowledge-Based Systems, 2012. **26**: p. 30-39.
61. Xie, J. and B. Szymanski, *Towards Linear Time Overlapping Community Detection in Social Networks*, in *Advances in Knowledge Discovery and Data Mining*, P.-N. Tan, et al., Editors. 2012, Springer Berlin Heidelberg. p. 25-36.
62. Nguyen, N.P., et al. *Overlapping community structures and their detection on social networks*. in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. 2011. IEEE.
63. Masterman, M., *The nature of a paradigm. Criticism and the growth of knowledge, eds. I. Latakos, and A. Musgrave*. 1970, Cambridge: Cambridge University Press.
64. Beck, J. and M.F. Young, *The assault on the professions and the restructuring of academic and professional identities: a Bernsteinian analysis.* British journal of sociology of education, 2005. **26**(2): p. 183-197.
65. Hofmann, T. *Probabilistic latent semantic indexing*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM.

66. Blei, D. and J. Lafferty, *Correlated topic models.* Advances in neural information processing systems, 2006. **18**: p. 147.
67. Griffiths, D. and M. Tenenbaum, *Hierarchical topic models and the nested Chinese restaurant process.* Advances in neural information processing systems, 2004. **16**: p. 17.
68. Chang, J. and D.M. Blei, *Hierarchical relational models for document networks.* The Annals of Applied Statistics, 2010: p. 124-150.
69. Gohr, A., et al., *Topic evolution in a stream of documents*. 2009, SIAM.
70. Mei, Q., et al. *Topic modeling with network regularization*. in *Proceedings of the 17th international conference on World Wide Web*. 2008. ACM.
71. Nallapati, R.M., et al. *Joint latent topic models for text and citations*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.
72. Griffiths, T.L. and M. Steyvers, *Finding scientific topics.* Proceedings of the National Academy of Sciences, 2004. **101**(suppl 1): p. 5228-5235.
73. He, Q., et al. *Detecting topic evolution in scientific literature: how can citations help?* in *Proceedings of the 18th CIKM*. 2009. ACM.
74. McCallum, A., X. Wang, and A. Corrada-Emmanuel, *Topic and role discovery in social networks with experiments on enron and academic email.* Journal of Artificial Intelligence Research, 2007: p. 249-272.
75. Rosen-Zvi, M., et al. *The author-topic model for authors and documents*. in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. 2004. AUAI Press.
76. Steyvers, M., et al. *Probabilistic author-topic models for information discovery*. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004. ACM.
77. Wang, J., et al., *Author-conference topic-connection model for academic network search*, in *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012, ACM: Maui, Hawaii, USA. p. 2179-2183.
78. Guo, W. and M. Diab. *Semantic topic models: Combining word distributional statistics and dictionary definitions*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011. Association for Computational Linguistics.
79. Bornmann, L. and H.D. Daniel, *The state of h index research.* EMBO reports, 2009. **10**(1): p. 2-6.
80. Rousseau, R. and K. Leuven, *Reflections on recent developments of the h-index and h-type indices.* COLLNET Journal of Scientometrics and Information Management, 2008. **2**(1): p. 1-8.
81. Hirsch, J.E., *An index to quantify an individual's scientific research output.* Proceedings of the National academy of Sciences of the United States of America, 2005. **102**(46): p. 16569-16572.
82. Egghe, L., *An improvement of the h-index: the g-index.* ISSI newsletter, 2006. **2**(1): p. 8-9.
83. Zadeh, L.A., *The concept of state in system theory.* 1964.
84. Tho, Q., S. Hui, and A. Fong, *Web Mining for Identifying Research Trends*, in *Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access*. 2003, Springer Berlin Heidelberg. p. 290-301.
85. Bolelli, L., Ş. Ertekin, and C.L. Giles, *Topic and trend detection in text collections using latent dirichlet allocation*, in *Advances in Information Retrieval*. 2009, Springer. p. 776-780.
86. Duvvuru, A., S. Kamarthi, and S. Sultornsanee, *Undercovering research trends: Network analysis of keywords in scholarly articles*, in *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on*. 2012. p. 265-270.
87. Budi, I., R.F. Aji, and A. Widodo, *Prediction of Research Topics on Science & Technology (S&T) using Ensemble Forecasting.* International Journal of Software Engineering and Its Applications, 2013. **7**(5): p. 253-268.
88. Jun, S. and D. Uhm, *Technology forecasting using frequency time series model: Bio-technology patent analysis.* Journal of Modern Mathematics and Statistics, 2010. **4**(3): p. 101-104.
89. Priem, J., et al., *Altmetrics: A manifesto.* 2010.
90. Bucknell, T. *Making sense and making use of Altmetrics in research evaluation*. in *Septentrio Conference Series*. 2014.
91. Moody, J. and D.R. White, *Structural cohesion and embeddedness: A hierarchical concept of social groups.* American Sociological Review, 2003: p. 103-127.

92. Bron, C. and J. Kerbosch, *Algorithm 457: finding all cliques of an undirected graph.* Commun. ACM, 1973. **16**(9): p. 575-577.
93. Osborne, F. and E. Motta, *Klink-2: integrating multiple web sources to generate semantic topic networks*, in *14th International Semantic Web Conference 2015*. 2015: Bethlehem, PA (forthcoming).
94. Newman, M.E., *The structure and function of complex networks.* SIAM review, 2003. **45**(2): p. 167-256.