A peer-reviewed version of this preprint was published in PeerJ on 19 June 2017.

View the peer-reviewed version (peerj.com/articles/cs-119), which is the preferred citable publication unless you specifically need to cite this preprint.

Salatino AA, Osborne F, Motta E. 2017. How are topics born? Understanding the research dynamics preceding the emergence of new areas. PeerJ Computer Science 3:e119
https://doi.org/10.7717/peerj-cs.119

PeerJ Preprints NOT PEER-REVIEWED

How are topics born? Understanding the research dynamics preceding the emergence of new areas

Angelo A Salatino Corresp., [1] , **Francesco Osborne** [1] , **Enrico Motta** [1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom

Corresponding Author: Angelo A Salatino
Email address: angelo.salatino@open.ac.uk

The ability to recognise new research trends early is strategic for many stakeholders, such as academics, institutional funding bodies, academic publishers and companies. While the state of the art presents several works on the identification of novel research topics, detecting the emergence of a new research area at a very early stage, i.e., when the area has not been even explicitly labelled and is associated with very few publications, is still an open challenge. This limitation hinders the ability of the aforementioned stakeholders to timely react to the emergence of new areas in the research landscape. In this paper, we address this issue by hypothesising the existence of an embryonic stage for research topics and by suggesting that topics in this phase can actually be detected by analysing diachronically the co-occurrence graph of already established topics. To confirm our hypothesis, we performed a study of the dynamics preceding the creation of novel topics. This analysis showed that the emergence of new topics is actually anticipated by a significant increase of the pace of collaboration and density in the co-occurrence graphs of related research areas. These findings are very relevant to a number of research communities and stakeholders. Firstly, they confirm the existence of an embryonic phase in the development of research topics and suggest that it might be possible to perform very early detection of research topics by taking into account the aforementioned dynamics. Secondly, they bring new empirical evidence to related theories in Philosophy of Science. Finally, they suggest that significant new topics tend to emerge in an environment in which previously less interconnected research areas start cross-fertilising.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2306v1 | CC BY 4.0 Open Access | rec: 20 Jul 2016, publ: 20 Jul 2016

PeerJ Preprints NOT PEER-REVIEWED

How Are Topics Born? Understanding the Research Dynamics Preceding the Emergence of New Areas.

Angelo Antonio Salatino[1], Francesco Osborne[1], Enrico Motta[1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK

Corresponding Author:

Angelo Antonio Salatino[1]

Email address: angelo.salatino@open.ac.uk

PeerJ Preprints NOT PEER-REVIEWED

ABSTRACT

The ability to recognise new research trends early is strategic for many stakeholders, such as academics, institutional funding bodies, academic publishers and companies. While the state of the art presents several works on the identification of novel research topics, detecting the emergence of a new research area at a very early stage, i.e., when the area has not been even explicitly labelled and is associated with very few publications, is still an open challenge. This limitation hinders the ability of the aforementioned stakeholders to timely react to the emergence of new areas in the research landscape. In this paper, we address this issue by hypothesising the existence of an embryonic stage for research topics and by suggesting that topics in this phase can actually be detected by analysing diachronically the co-occurrence graph of already established topics. To confirm our hypothesis, we performed a study of the dynamics preceding the creation of novel topics. This analysis showed that the emergence of new topics is actually anticipated by a significant increase of the pace of collaboration and density in the co-occurrence graphs of related research areas. These findings are very relevant to a number of research communities and stakeholders. Firstly, they confirm the existence of an embryonic phase in the development of research topics and suggest that it might be possible to perform very early detection of research topics by taking into account the aforementioned dynamics. Secondly, they bring new empirical evidence to related theories in Philosophy of Science. Finally, they suggest that significant new topics tend to emerge in an environment in which previously less interconnected research areas start cross-fertilising.

Keywords: Scholarly Data, Empirical Study, Research Trend Detection, Topic Emergence Detection, Topic Discovery, Digital Libraries, Ontology, Semantic Web

INTRODUCTION

Being aware of the rise of new research topics can bring significant benefits for anybody involved in the research environment. Academic publishers and editors can exploit this knowledge for offering the most up to date and interesting contents. Researchers might be interested in new trends related to their topics and in promising new research areas. Institutional funding bodies and companies need to be updated constantly on how the research landscape is evolving in order to make early decisions about critical investments. Nonetheless, considering the growth rate of research publications (Larsen & Von Ins 2010), keeping up with novel trends is a challenge even for expert researchers and traditional methods, such as the manual exploration of the publications in significant conference and journals, are no longer viable. This lead to the emergence of several approaches capable of detecting novel topics and research trends (Bolelli et al. 2009; Duvvuru et al. 2012; He et al. 2009; Wu et al. 2016). However, these approaches focus on topics that are associated with a substantial number of publications or on which the scientific community reached a consensus for a specific label. This limitation hinders the ability of aforementioned stakeholders to timely react to novelties in the research landscape.

An intriguing challenge is thus to identify in a very early phase the appearance of a new topic, assess its potential and forecast its trend. For this reason, we need a better understanding of the dynamics underlying the creation of new topics and how these can be detected using current knowledge bases.

The Philosophy of Science offers a number of intriguing theories about the emergence of new topics. Kuhn (2012) theorised that science evolves through paradigm shifts. According to him, scientific work is performed within a set of paradigms and when these paradigms cannot cope with certain problems, there is a paradigm shift that can lead to the emergence of a new scientific discipline. This happens often through the creation of novel scientific collaborations. In this regards, Becher & Trowler (2001) explained that, even if science proceeds toward more specific disciplines and thus researchers in different communities become less compatible, they are still incline to collaborate for mutual benefit. Herrera et al. (2010), Sun et al. (2013), Nowotny et al. (2013) suggested that the development of new topics is actually encouraged by the cross-fertilisation of established research areas and recognised that multidisciplinary approaches foster new developments and innovative thinking. Sun et al. (2013) and Osborne et al. (2014) provided empirical evidence to these theories by analysing the social dynamics of researchers and their effect on research communities and topics.

According to these theories, when a new scientific area emerges, it goes through two main phases. In the *initial stage* a group of scientists agree on some basic theories, build a conceptual framework and begin to establish a new scientific community. Afterwards, the area enters into a *recognised phase* in which a substantial number of authors start working on it, producing and disseminating results (Couvalis 1997).

Inspired by previous theories, we hypothesize the existence of an even earlier phase, that we label *embryonic phase*, in which a topic has not yet been explicitly labelled or recognized by a research community, but exist as a fuzzy entity which entices a number of researchers from a variety of fields to converge and collaborate, with the aim of defining the mission and the paradigms of this potential research area. We also hypothesize that it is possible to detect topics in this stage by analysing the dynamics of established topics, which should reflect the new collaborations of pioneer researchers shaping the new area.

This paper presents a study of the dynamics preceding the creation of novel topics which supports our hypothesis by showing that the emergence of novel research topics is actually anticipated by a significant increase of pace of collaboration and density in the co-occurrence graphs of related topics.

The study was conducted in the 2000-2010 interval on a sample of three million publications. It was conducted by selecting sections of the co-occurrence graph where a new topic is about to emerge and analysing their dynamics in the previous five years versus a control group of subgraphs related to established topics. The analysis was performed with two different approaches that integrate statistics and semantics. It was found that the pace of collaboration and density measured in the sections of the network that will give rise to a new topic are

significantly higher ($p < 0.0001$) than the one in the control group. These findings confirm the existence of an embryonic phase, yield new empirical evidences to aforementioned theories and confirm the strong benefits of an interdisciplinary environment.

The study presented in this paper is an extension of the one published in (Salatino & Motta 2016). The new contribution of this paper are: 1) a larger sample (75 debutant topics and 100 established ones), 2) a new technique for measuring the density of the topic graphs, 3) a more exhaustive statistical analysis, including the comparison of the different approaches, 4) a revised state of the art, and 5) a more comprehensive discussion of the findings.

The rest of the paper is organized as follows. We will first review the literature regarding the early detection of topics, pointing out the existing gaps. Then we will describe the experimental approach used for the study, present the results and discuss their implication. Finally, we will summarize the main conclusions and outline future directions of research.

RELATED WORK

Topic detection and tracking is a task that has drawn much attention in the last years and has been applied to a variety of scenarios, such as social networks (Cataldi et al. 2010; Mathioudakis & Koudas 2010), blogs (Gruhl et al. 2004; Oka et al. 2006), emails (Morinaga & Yamanishi 2004) and scientific literature (Bolelli et al. 2009; Decker et al. 2007; Erten et al. 2004; Lv et al. 2011; Osborne et al. 2014; Sun et al. 2016; Tseng et al. 2009).

The state of the art presents several works on research trend detection, which can be characterised either by the way they define a topic or the techniques they use to detect them (Salatino 2015). Blei et al. (2003) developed the well-known Latent Dirichlet Allocation (LDA) which is an unsupervised learning method to extract topics from a corpus and models topics as a multinomial distribution over words. Since its introduction, LDA has been extended and adapted in several applications. For example, Blei & Lafferty (2006) introduced the Correlated Topic Model using the logistic normal distribution instead of the Dirichlet one, to solve the fact that LDA fails to model the correlation between topics. Griffiths & Tenenbaum (2004) developed the *hierarchical LDA* where topics are grouped together in a hierarchy. Further extensions, incorporate other kinds of research metadata. For example, Rosen-Zvi et al. (2004) presented the Author-Topic model (ATM) which includes authorship information and then associates each topic to a multinomial distribution over words and each author to a multinomial distribution over topics. Bolelli et al. (2009) introduced the Segmented Author-Topic model which further extends ATM by adding the temporal ordering of documents to address the problem of topic evolution. In addition, Chang & Blei (2010) developed the *relational topic model* which combines LDA and the network structure of documents to model topics. Similarly, He et al. (2009) combined LDA and citation networks in order to address the problem of topic evolution. Their approach detects topics in independent subsets of a corpus and then leverages citations to connect topics in different time frames. In a similar way, Morinaga & Yamanishi (2004) employed a probabilistic model called Finite Mixture Model to represent the structure of topics and analyse the changes in time of the extracted components to track emerging topics. However, it was evaluated on an email corpus, thus it is not clear how it could perform on scientific corpus. A general issue of

this kinds of approaches is that is not always easy to associate specific research areas to the resulting topics.

In addition to LDA, the Natural Language Processing (NLP) community proposed a variety of tools for identifying topics. For example, Chavalarias & Cointet (2013) used CorText Manager to extract a list of 2000 n-grams representing the most salient terms from a corpus and derived a co-occurrence matrix on which they perform clustering analysis to discover patterns in the evolution of science. Jo et al. (2007) developed an approach that correlates the distribution of terms extracted from the text with the distribution of the citation graph related to publications containing those terms. Their work is based on the assumption that if a term is relevant to a particular topic, documents containing that term will have a stronger connection than randomly selected ones. However, this approach is not suitable for topics in their very early stage since it takes time for the citation network of a term to become tightly connected.

Duvvuru et al. (2013) analysed the co-occurring network of keywords in a scholarly corpus and monitored the evolution in time of the link weights for detecting research trends and emerging research areas. However, as Osborne & Motta (2012) pointed out, keywords tend to be noisy and do not always represent research topics and in many cases different keywords can represent the same topic. For example, Osborne et al. (2014) showed that the use of a semantic characterisation of research topics yields better results for the detection of research communities. To cope with this problem, some approaches rely on taxonomies of topics. For example, Decker et al. (2007) matched a corpus of research papers to a taxonomy of topics based on the most significant words found in titles and abstracts, and analysed the changes in the number of publications associated with topics. Similarly, Erten et al. (2004) adopted the ACM Digital Library taxonomy for analysing the evolution of topic graphs and monitoring research trends. However, human crafted taxonomy tend to evolve slowly and in a fast-changing research field such as Computer Science (Pham et al. 2011) it is important to rely on constantly updated taxonomies. For this reason, in our experiment we adopted an ontology of Computer Science automatically generated and regularly updated by the Klink-2 algorithm developed by Osborne & Motta (2015).

In brief, the state of the art provides a wide collection of approaches for detecting research trends. However, these focus on already recognised topics, associated with either a label or, in the case of probabilistic topics models, with a set of terms that should have previously appeared in a good number of publications. Therefore, detecting research trends in a very early stage is still an open challenge.

EMPIRICAL STUDY

The aim of this study was to explore whether the emergence of new topics is anticipated by a significant increase of pace of collaboration and density in the co-occurrence graphs of related topics. To this end, we represented topics and their relationships in certain time interval as a graph in which nodes are topics and edges represent their co-occurrences in a sample of publications. This is a common representation for investigating topic dynamics (Boyack et al. 2005; Leydesdorff 2007; Newman 2001) and we will refer to it as *topic graph* or *topic network* in the following. We then selected 75 topics debuting between 2000 and

2010 and extracted the sub-graphs containing their *n* most co-occurring topics in the five years before their debut. Finally, we measured the collaboration pace and the density of these graphs, comparing it to those of a control group of established topics.

In the following sections we will describe the dataset, the semantically enhanced topic graph and the methods used to measure the pace of collaboration and the density of the subgraphs.

The raw data and the outcomes of this study are publically available at http://technologies.kmi.open.ac.uk/rexplore/peerj2016/.

Semantic Enhanced Topic Network

We used as dataset the metadata describing 3 million papers in the field of Computer Science from a dump of the well-known Scopus dataset[1]. In this dataset each paper is associated to a number of keywords that could be used to build the topic graph. However, as pointed out in (Osborne & Motta 2012), the use of keywords as proxies for topics suffers from a number of problems: some keywords tend to be noisy and do not represent topics (e.g., "case study") and multiple keywords can refer to the same topic (e.g., "ontology mapping" and "ontology matching").

We address this issue by building a semantically enhanced topic graph with the Klink-2 ontology of Computer Science, which describes the relationship between almost 15000 research areas. Klink-2 is an algorithm which analyses keywords and their relationships with research papers, authors, venues, and organizations and takes advantage of multiple knowledge sources available on the web in order to produce an ontology of research topics linked by three different semantic relationships. Klink-2 is integrated in the Rexplore system (Osborne et al. 2013), a modern tool for exploring and making sense of scholarly data, which adopts novel solutions in large-scale data mining, semantic technologies and visual analytics. In particular, Rexplore uses the large Computer Science ontology to craft semantic-aware analytics.

We took advantage of the Klink-2 ontology by filtering from our dataset the keywords that did not represent specific research areas and aggregating keywords representing the same concept, i.e., linked by a *relatedEquivalent* relationship in the ontology (Osborne et al. 2013). For example, we aggregated keywords such as "semantic web", "semantic web technology" and "semantic web technologies" in a single semantic topic and assigned it to all publications associated with these keywords.

We used the resulting semantic topics to build sixteen topic networks representing the topic co-occurrences in the 1995-2010 timeframe. Each network is a fully weighted graph $G_{year} = (V_{year}, E_{year})$, in which V is the set of topics while E is the set of links representing the topic co-occurrences. The node weight is given by the number of publications in which the keyword appears, while the link weight is equal to the number of publications in which two topics co-occur together in a year.

[1] https://www.elsevier.com/solutions/scopus

Graph Selection

We randomly selected 75 topics that debuted in the period between 2000 and 2010 as treatment group (also referred as debutant group) and 100 well-established topics that made their debut at least in the previous decade as control group (also referred as non-debutant group). A topic debuts in the year in which its label first appears in a research paper. The debutant topic graphs included 1357 topics, while the control group ones 1060 topics. The fact that the set of graph is larger is due to the fact that novel topics tend to collaborate with a larger variety of research areas.

We assume that after a new topic emerges it will continue to collaborate with the topics that contributed to its creation for a certain time. This assumption was discussed and tested in previous work (Osborne & Motta 2012) where it was used for finding historical subsumption links between research areas. Hence, for each debuting topic we extracted the portion of topic network containing its *n* most co-occurring topics from the year of debut until nowadays and analysed their activity in the five years preceding its year of debut. Since we want to analyse how the dimension of these subgraphs could influence the results, we tested different values of *n* (20, 40, and 60).

Figure 1 summarizes this process. For example, if a topic *A* had its debut in 2003, the portion of network containing its most co-occurring topics will be analysed in the 1998-2002 timeframe. We repeated the same procedure on the topics in the control group, assigning them a random year of analysis within the decade 2000-2010. In the previous study (Salatino & Motta 2016) we assigned each established topic to a random year of analysis, while for this study we randomly assigned each established topic to two consecutive years with the aim of reducing the noise and smoothing the resulting measures.

At the end of the selection phase we associated to each topic in the two groups a graph G^{topic}:

$$G^{topic} = G^{topic}_{year-5} \cup G^{topic}_{year-4} \cup G^{topic}_{year-3} \cup G^{topic}_{year-2} \cup G^{topic}_{year-1} \qquad (1)$$

which corresponded to its collaboration network in the five years prior to its emergence. This graph contained five sub-graphs G^{topic}_{year-i} and each one corresponded to:

$$G^{topic}_{year-i} = (V^{topic}_{year-i}, E^{topic}_{year-i}) \qquad (2)$$

in which V^{topic}_{year-i} is the set of most co-occurring topics in a particular year and E^{topic}_{year-i} is the set of edges that link nodes in the set V^{topic}_{year-i}.

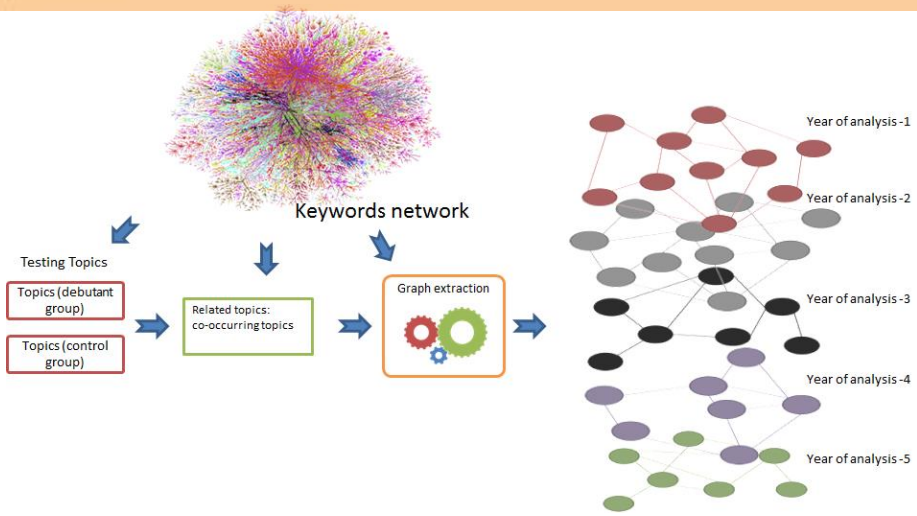

Figure 1: Workflow representing all the steps for the selection phase.

Graph Analysis

We assess the dynamics in the graphs with two main approaches: cliques-based and triad-based. The first transforms the graph in 3-cliques, associates to each of them a measure reflecting the increase in collaboration between the relevant topics and then averages the results over all 3-cliques. The second measures the increase in the topics graph density via the triad census technique (Davis & Leinhardt 1967). In the following two sections we will describe both methods in details.

Cliques-based method

This method is based on the intuition that we can measure the collaboration pace of a graph by analysing the diachronic activity of triangles of collaborating topics. Hence, we first transformed the graphs in sets of 3-cliques. A 3-clique, as shown in Figure 2, is a complete sub-graph of order three in which all nodes are connected to one another and is employed for modelling small groups of entities close to each other (Luce & Perry 1949). We then extracted the 3-cliques from the five sub-graphs associated to each topic and created timelines of cliques in subsequent years. In order to assess the increase of collaboration between nodes $\{A, B, C\}$ in a 3-clique we adopted Equation 3, which takes in consideration both node weights $\{W_a, W_b, W_c\}$ and link weights $\{W_{ab}, W_{bc}, W_{ca}\}$. It does so by computing the conditional probability $P(y|x) = W_{xy}/W_x$ that a publication associated with a topic x will be also associated with a topic y in a certain year. The advantage of using the conditional probability over the number of co-occurrences is that the resulting value is already normalised according to the number of publications associated to each topic.

We computed the weight associated to each link between topic x and y by using the harmonic mean of the conditional probabilities $P(y|x)$ and $P(x|y)$ and then computed the final index μ_Δ as the harmonic mean of all the weights of the clique. This solution was adopted after testing a variety of alternative approaches (e.g., arithmetic mean) during a preliminary evaluation discussed in the Findings section.

$$\begin{aligned} \mu_1 &= harmmean\left(P(A\,|\,B), P(B\,|\,A)\right) \\ \mu_2 &= harmmean\left(P(B\,|\,C), P(C,B)\right) \\ \mu_3 &= harmmean\left(P(C\,|\,A), P(A\,|\,C)\right) \\ \mu_\Delta &= harmmean\left(\mu_1, \mu_2, \mu_3\right) \end{aligned} \tag{3}$$

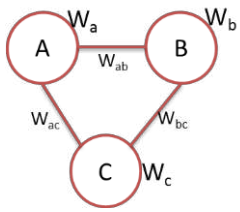

Figure 2. An instance of a 3-clique containing nodes and links weights.

At this stage, each clique was reduced to the timeline of measures showed in Equation 4. We then studied the evolution of these values for determining whether the collaboration pace of a clique was increasing or decreasing in the time interval, as illustrated in Figure 3.

$$\mu_{\Delta\text{time}}^{clique-i} = [\mu_{(\Delta\text{yr}-5)}, \mu_{(\Delta\text{yr}-4)}, \mu_{(\Delta\text{yr}-3)}, \mu_{(\Delta\text{yr}-2)}, \mu_{(\Delta\text{yr}-1)}] \tag{4}$$

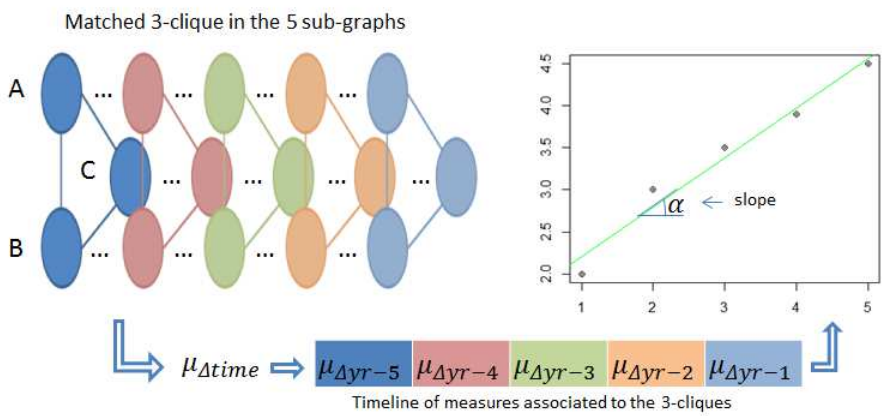

Figure 3. Main steps of the analysis phase: from 3-cliques matching to slope processing.

We first tried to determine the trend of a clique by simply taking the difference between the first and last values of the timeline. However, this method ignores the other values in the timeline and can thus neglect important information. For this reason, we applied instead the linear interpolation method on the five measures using the least-squares approximation to determine the linear regression of the time series $f(x) = a \cdot x + b$. The slope a is then used to assess the increase of collaboration in a clique. When a is positive the degree of collaboration between the topics in the clique is increasing over time, while when is negative the number and intensity of collaborations are decreasing.

Finally, the collaboration pace of each sub-graph was measured by computing the mean of all the slopes associated with the 3-cliques.

Triad-based method

The triad-based method employs the triad census (Davis & Leinhardt 1967) to measure the change of topology and the increasing density of the subgraphs during the five year period. The triad census of an undirected graph, also referred as global 3-profiles, is a four dimensional vector representing the frequencies of the four isomorphism classes of triad, as shown in Figure 4.

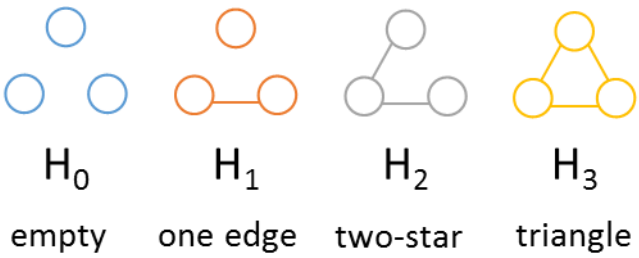

Figure 4. The four isomorphism classes of triad. The triad census consists in counting the frequencies of *Hi* of the input graph.

The triad census summarises the structural information in networks and is useful to analyse structural properties in social networks. In has been applied to several scenarios, such as identifying spam (Kamaliha et al. 2008; O'Callaghan et al. 2012), comparing networks (Pržulj 2007), analysing social networks (Faust 2010; Ugander et al. 2013) and so on.

In this study, we used triad census to describe all the sub-graphs G_{year-i}^{topic} associated to a particular testing topic in terms of frequencies of H_i (see Figure 4) and then evaluate how the frequencies of *empties* (H_0), *one edges* (H_1), *two-starts* (H_2) and *triangles* (H_3) changed in time. Figure 4 illustrates the four classes of triad for an undirected graph as in the case of topic network. Naturally an increase of the numbers of triangles suggests the appearance of a number of new collaborations clusters between previous distant topics.

Differently from the previous approach, the triad census does not consider the weight of links, but only their existence. Hence, it is useful to assess how including links with different strength might influence the analysis. To this end, we performed three experiments in which we considered only links associated with more than 3, 10 and 20 topic co-occurrences.

Figure 5 shows the workflow for analysing how the topology of networks related to a testing topic evolved in the five years preceding its debut. We first performed the triad census over the five graphs associated to each testing topic. For example, Table 1 shows the results of the triad census over the five sub-graphs associated to the debutant topic "Artificial Bee Colonies". We then measured whether the collaboration graph was becoming denser by analysing how the frequencies associated to H_i evolved (see Figure 6). To do so, we computed the percentage growth of each H_i using Equation 5.

$$\%GrowthH_i = \frac{(H_i^{Yr-1} - H_i^{Yr-5}) * 100}{H_i^{Yr-5}} \quad (5)$$

Then we used Equation 6, which performs a weighted summation of all the contributions of percentage of growth.

$$GrowingIndex_{topic} = \sum_{i=0}^{3} i \cdot \%GrowthH_i \quad (6)$$

As a result, we expect that the growing index of the portion of network related to debutant topics would be significantly higher than the one associated to non-debutant topics.

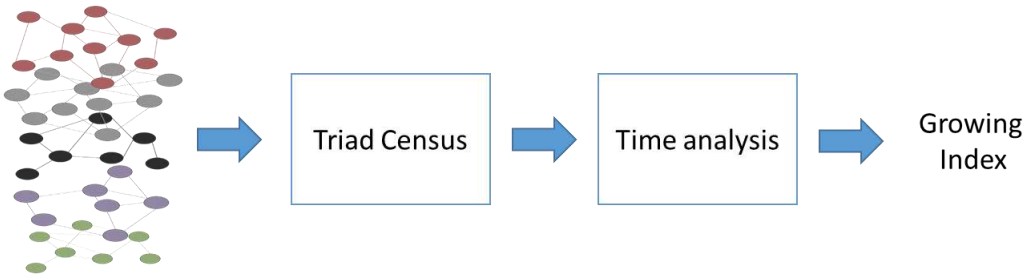

Figure 5. Main step of the analysis phase for the triad census approach.

Table 1. Frequencies of H_i obtained performing triad census on the debutant topic "Artificial Bee Colonies"

Graph	H_0	H_1	H_2	H_3
G^{topic}_{year-5}	446	790	807	882
G^{topic}_{year-4}	443	854	915	1064
G^{topic}_{year-3}	125	486	967	1698
G^{topic}_{year-2}	100	410	908	1858
G^{topic}_{year-1}	68	486	849	2251

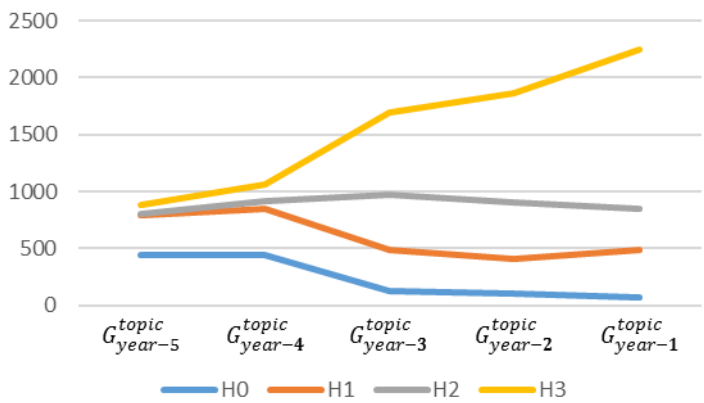

Figure 6: Development in time of the frequencies of H_i in the network related to the emergence of "Artificial Bee Colonies".

Findings

In this section we report the results obtained by analysing the debutant and the control groups with the previously discussed methods. We will describe:

- The preliminary evaluation performed on a reduced dataset for assessing the metrics used in the Cliques-based method;

- The full study using the Cliques-based method;
- The full study using the Triads-based method.

Preliminary evaluation with alternative cliques-based methods

We initially conducted a preliminary evaluation with the aim of choosing the most effective Cliques-based method for assessing the pace of collaboration. This test focused on the subgraph of the 20 most co-occurring topics associated to the Semantic Web (debuting in 2001) and Cloud Computing (2006) versus a control group of 20 subgraphs associated to a non-debutant group. We tested on this dataset two techniques to compute the weight of a clique (i.e., harmonic mean and arithmetic mean) and two methods to evaluate its trend (i.e., computing the difference between the first and the last values and linear interpolation). Hence, we evaluated the following four approaches:

- **AM-N**, which uses the arithmetic mean and the difference between first and last value;
- **AM-CF**, which uses the arithmetic mean and the linear interpolation;
- **HM-N**, which uses the harmonic mean and the difference between first and last value;
- **HM-CF**, which uses the harmonic mean and the linear interpolation.

Figure 7 illustrates the average pace of collaboration for the sub-graphs associated to each topics according to these methods (thick horizontal black lines) and the range of their values (thin vertical line). The results support the initial hypothesis: according to all methods, the pace of collaboration of the cliques within the portion of network associated with the emergence of new topics is positive and higher than the ones of the control group. Interestingly, the pace of collaboration of the control group is also slightly positive. Further analysis revealed that this behaviour is probably caused by the fact that the topic network becomes denser and noisier in time. Figure 8 confirms this intuition illustrating the fast growth of the number of publications per year in the dataset during the time window 1970-2013.

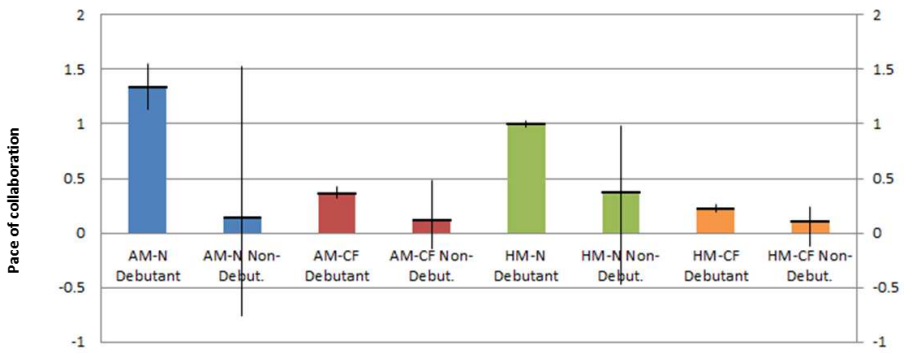

Figure 7. Overall directions of the sub-graphs related to testing topics in both debutant and control group with all the four approaches.

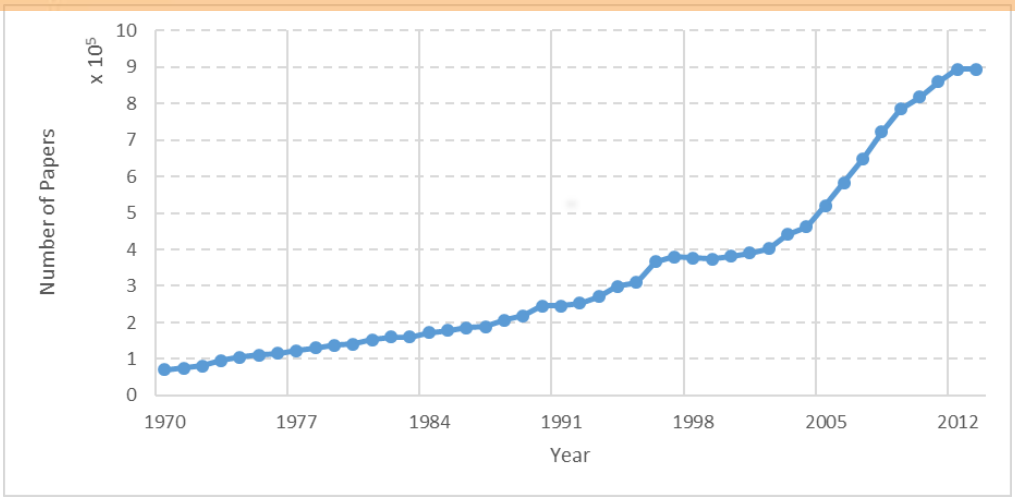

Figure 8. Number of papers each year in period 1970-2013

The approaches based on the simple difference (AM-N and HM-N) exhibit the larger gaps between the two groups in terms of the average pace of collaboration. However, the ranges of values actually overlap, making it harder to assess if a certain sub-group is incubating a novel topic. The same applies to AM-CF. HM-CF performs better and even if the values slightly overlap when averaging the pace over different years they do not when considering single years. Indeed, analysing the two ranges separately in 2001 and 2006 (see Figure 9), we can see that the overall collaboration paces of the debutant topics (DB) are always significantly higher than the control group (NDB).

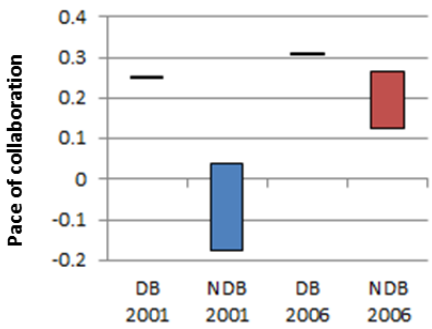

Figure 9. Overall directions of the sub-graphs related to testing topics in both debutant and control group in HM-CF approach

We ran Student's t-test on the HM-CF approach in order to verify whether the two groups belong to different populations. The test yielded $p < 0.0001$, which allowed us to reject the null hypothesis that the differences between the two distributions were due to random variations[2]. For this reason, we selected the combination between the harmonic mean and the

[2] $p<0.0001$ is the conventional statistical representation to indicate an extremely high statistical significance (> 500 times stronger than the conventional 0.05 threshold for

linear interpolation (HM-CF) as the approach for the full study using the clique-based method.

The results of HM-CF show also interesting insights on the creation of some well-known research topics. Table 2 and Table 3 list the cliques which exhibited a steeper slope for semantic web and cloud computing. We can see that Semantic Web was anticipated in the 1996-2001 timeframe by a significant increase in the collaborations of the world wide web area with topics such as information retrieval, artificial intelligence, and knowledge based systems. This is consistent with the initial vision of the semantic web, defined in the 2001 by the seminal work of Tim Berners-Lee (Berners-Lee et al. 2001). Similarly, Cloud Computing was anticipated by an increase in the collaboration between topics such as grid computing, web services, distributed computer systems and internet. This suggests that our approach can be used both for forecasting the emergence of new topics in distinct subsections of the topic network and for identifying the topics that gave rise to a research area.

Table 2. Ranking of the cliques with highest slope value for the "semantic web".

Topic 1	Topic 2	Topic 3	Slope
world wide web	information retrieval	search engines	2.529
world wide web	user interfaces	artificial intelligence	1.12
world wide web	artificial intelligence	knowledge representation	0.974
world wide web	knowledge based systems	artificial intelligence	0.850
world wide web	information retrieval	knowledge representation	0.803

Table 3. Ranking of the cliques with highest slope value for the "cloud computing".

Topic 1	Topic 2	Topic 3	Slope
grid computing	distributed computer systems	web services	1.208
web services	information management	information technology	1.094
grid computing	distributed computer systems	quality of service	1.036
internet	quality of service	web services	0.951
web services	distributed computer systems	information management	0.949

Cliques-based method study

We applied the cliques-based methods on the subgraphs associated to both topics in the treatment and control groups. Figure 10 reports the results obtained by using subgraphs composed by the most 20, 40 and 60 co-occurring topics. Each bar shows the mean value of the average pace of collaboration for the debutant (DB) and non-debutant (NDB) topics. As before, the average pace computed in the portion of topic network related to debutant topics is higher than the one of the control group.

claiming significance). It includes all mathematical outcomes from 0 to below 0.0001, which are essentially equivalent in assessing the excellent significance.

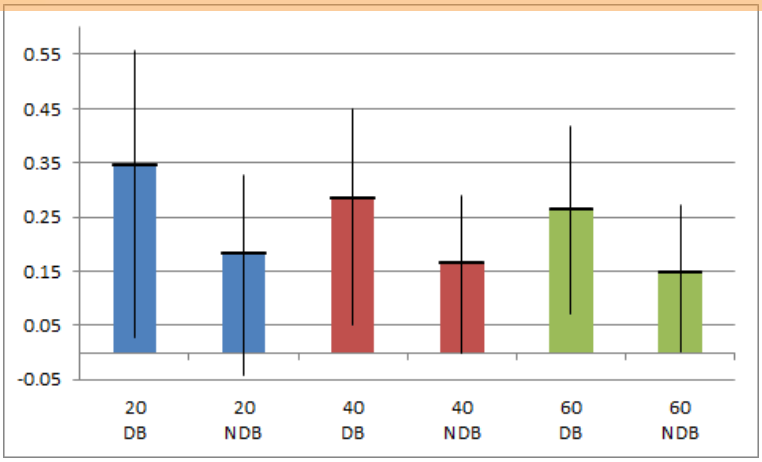

Figure 10. Average collaboration pace of the sub-graphs associated to the treatment (DB) and control group (NDB), when selecting the 20, 40 and 60 most co-occurring topics. The thin vertical lines represent the ranges of values.

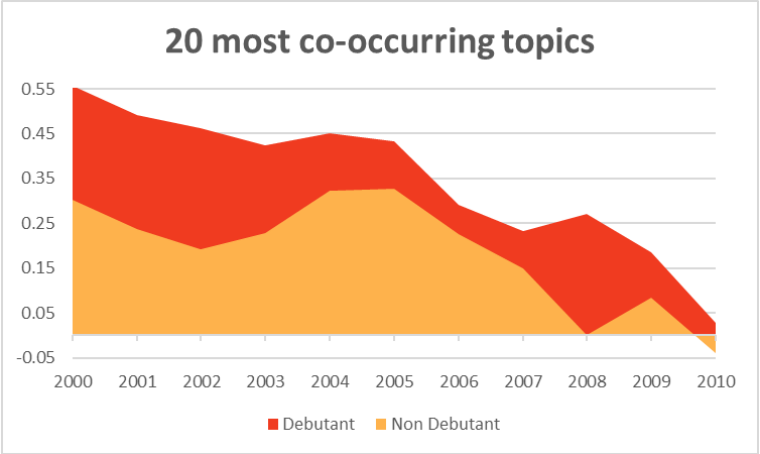

Figure 11. Average collaboration pace per year of the sub-graphs related to testing topics in both debutant and control group considering their 20 most co-occurring topics. The year refers to the year of analysis of each topic.

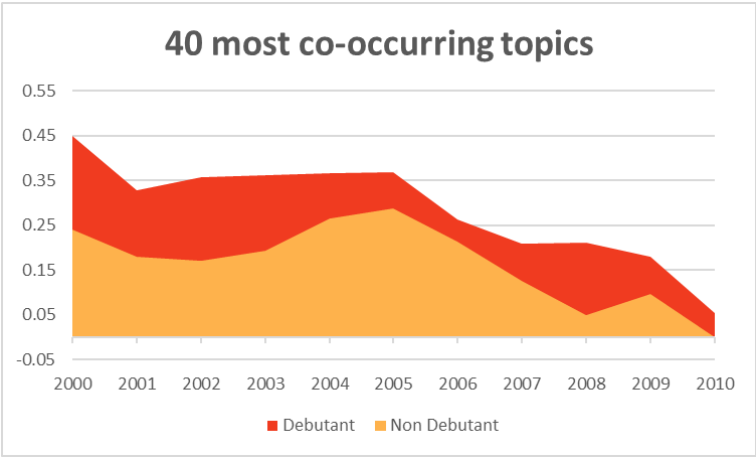

Figure 12. Average collaboration pace per year of the sub-graphs related to testing topics in both debutant and control group considering their 40 most co-occurring topics. The year refers to the year of analysis of each topic.

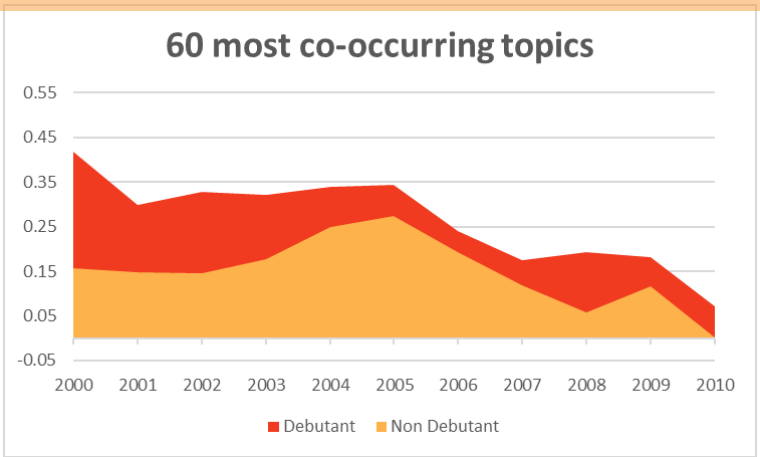

Figure 13. Average collaboration pace per year of the sub-graphs related to testing topics in both debutant and control group considering their 60 most co-occurring topics. The year refers to the year of analysis of each topic.

Since the pace of collaboration changes significantly according to the period considered, it is useful to study it across different years. Figure 11, Figure 12 and Figure 13, show the average collaboration pace for each year when considering the 20, 40 and 60 most co-occurring topics. In all cases the collaboration pace for the debutant topics is higher than the one for the control group. We can also notice that in the last five years the overall pace of collaboration for both debutant and non-debutant topics suffered a significant fall. This is due to the fact that topic network became denser and noisier in recent years.

Table 4 shows as example a number of debutant topics and their collaboration pace versus the collaboration pace of the control group in the same year. We can see how the appearance of a good number of well-known topics that emerged in the last decade was actually anticipated by the dynamics of the topic network.

Table 4. Collaboration pace of the sub-graphs associated to selected debutant topics versus the average collaboration pace of the control group in the same year of debut.

Topic	Collaboration Pace	Standard Collaboration pace
service discovery (2000)	0.455	0.156
ontology engineering (2000)	0.435	0.156
ontology alignment (2005)	0.386	0.273
service-oriented architecture (2003)	0.360	0.177
smart power grids (2005)	0.358	0.273
sentiment analysis (2005)	0.349	0.273
semantic web services (2003)	0.349	0.177
linked data (2004)	0.348	0.250
semantic web technology (2001)	0.343	0.147
vehicular ad hoc networks (2004)	0.342	0.250
mobile ad-hoc networks (2001)	0.342	0.147
p2p network (2002)	0.340	0.145
location based services (2001)	0.331	0.147
service oriented computing (2003)	0.331	0.177
ambient intelligence (2002)	0.289	0.145
social tagging (2006)	0.263	0.192
wireless sensor network (2001)	0.258	0.147
community detection (2006)	0.243	0.192
cloud computing (2006)	0.241	0.192

user-generated content (2006)	0.240	0.192
information retrieval technology (2008)	0.231	0.057
web 2.0 (2006)	0.224	0.192
ambient assisted living (2006)	0.224	0.192
Internet of things (2009)	0.221	0.116

We ran Student's t-test on the groups in different years, in order to confirm that the two distributions belong to different populations. In all cases it yielded $p < 0.0001$ in all years. However, the experiment containing 60 most co-occurring topics allows to better discriminate debutant topics from non-debutant ones. Indeed, the p-values obtained by this solution are lower than the one yielded by the other two experiments for every single year of the period under analysis.

In conclusion, the results confirm that the portions of the topic network in which a novel topic will appear exhibit a measurable fingerprint, in terms of increased collaboration pace, well before the topic is recognized and labelled by researchers.

Triads-based method study

We applied the triads-based methods on the subgraphs composed by the 60 most co-occurring topics, since this configuration provided the best outcomes in previous tests. We performed multiple tests by filtering links associated with less than 3, 10 and 20 co-occurrences, for understanding how the collaboration strength influences the outcome.

Figure 14 reports the average value of the growing indexes when discarding links with less than 3 co-occurrences. The approach allows to discriminate well the portion of networks related to debutant topics from the ones related to the control group and the collaboration pace associated with the debutant topics is always higher than its counterpart. Figure 15 and Figure 16 report the results obtained by removing links with less than 10 and 20 co-occurrences. The gap between the groups in these two last experiments is reduced in comparison with the first experiment. This suggest that considering weak connections is more beneficial for discriminating the two groups. Nonetheless, the indexes associated with debutant topics are always higher than the ones associated to non-debutant ones. The 2004 peak is caused by the debut of number of topics associated with particularly strong underlying dynamics, such as Linked Data, Pairing-based Cryptography, Microgrid and Privacy Preservation.

Table 5 reports as an example the triad census performed over the subgraph associated to the "semantic web technologies" (SWT) debuting in the 2001. We can see an increase in the number of triangles (H_3) and two-stars (H_2), mirroring the increasing density of the topics network. Again, this phenomenon is more evident when using also weak links (< 3). The percentage of growth of full triangles is 109% in the first test and then it decreases to 86% (< 10) and 36 % (< 20).

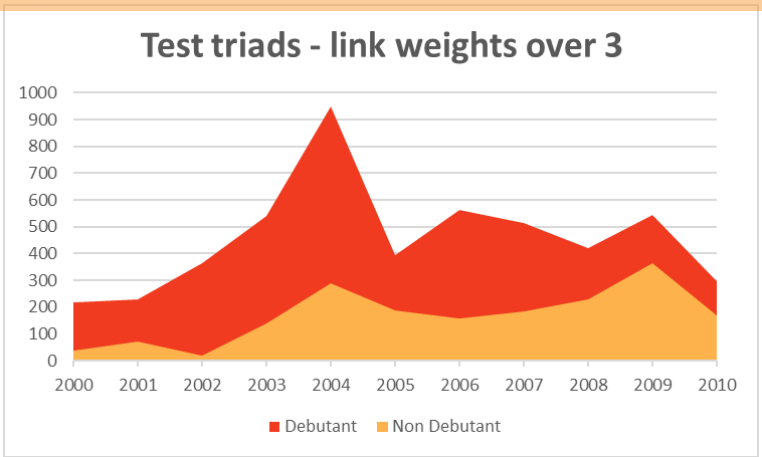

Figure 14. Average growing index per year of the sub-graphs related to the topics in both debutant and non-debutant group considering their 60 most co-occurring topics and filtering links associated with less than 3 publications.

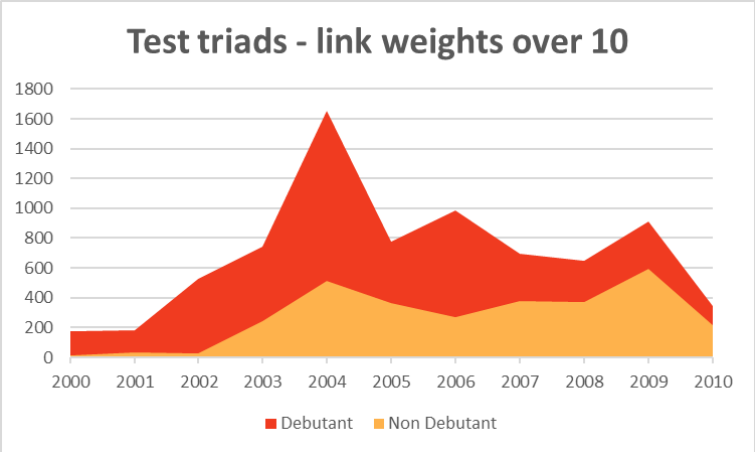

Figure 15. Average growing index per year of the sub-graphs related to the topics in both debutant and non-debutant group considering their 60 most co-occurring topics and filtering links associated with less than 10 publications.

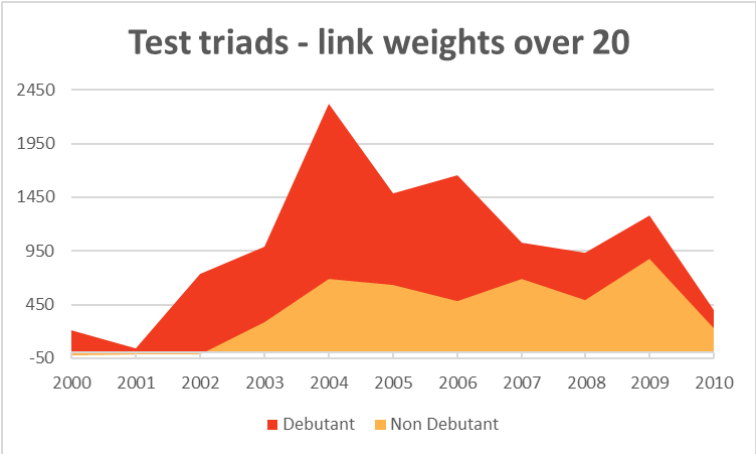

Figure 16. Average growing index per year of the sub-graphs related to the topics in both debutant and non-debutant group considering their 60 most co-occurring topics and filtering links associated with less than 20 publications.

481

482 Table 5. The results of the triad census performed on the network associated with the debutant topic "semantic web
483 technology" removing links associated with less than 3 (left), 10 (right) and 20 (bottom) publications.

	Removing links < 3				Removing links < 10			
Graph	H_0	H_2	H_2	H_3	H_0	H_2	H_2	H_3
1996	1124	1157	658	337	641	676	316	138
1997	928	1237	670	441	1022	828	315	135
1998	1255	1353	657	389	585	705	300	181
1999	1307	1431	861	461	1222	1098	413	192
2000	913	1399	1043	705	1482	1361	554	257

Removing links < 20

Graph	H_0	H_2	H_2	H_3
1996	796	509	174	61
1997	632	432	204	62
1998	525	418	145	52
1999	569	497	187	77
2000	842	618	228	83

484

485

486 Table 6 shows a selection of debutant topics and their growing index compared with the
487 growing index of the control group in the same year. We can compare this table to Table 4 to
488 appreciate how the two methods used in this study reflect the same behaviour.

489 Table 6. Growing indexes of sub-graphs associated to selected debutant topics versus the average growing index of the
490 control group in the same year of debut.

Topic	Growing Index	Standard Growing Index
service discovery (2000)	290.29	35.97
ontology engineering (2000)	207.22	35.97
ontology alignment (2005)	399.60	186.89
service-oriented architecture (2003)	628.07	140.17
smart power grids (2005)	637.53	186.89
sentiment analysis (2005)	354.10	186.89
semantic web services (2003)	439.85	140.17
linked data (2004)	590.81	289.94
semantic web technology (2001)	465.53	72.71
vehicular ad hoc networks (2004)	859.44	289.94
mobile ad-hoc networks (2001)	87.31	72.71
p2p network (2002)	305.28	18.92
location based services (2001)	595.90	72.71
service oriented computing (2003)	422.92	140.17
ambient intelligence (2002)	308.34	18.92
social tagging (2006)	429.77	157.69
community detection (2006)	583.21	157.69
cloud computing (2006)	695.79	157.69
user-generated content (2006)	485.89	157.69
information retrieval technology (2008)	552.14	227.02
web 2.0 (2006)	387.42	157.69
ambient assisted living (2006)	940.79	157.69
Internet of things (2009)	580.33	167.86

491

As before, we ran Student's t-test over the two distributions of growing indexes, for all the three experiments. It yielded p <0.0001 for all the experiments. Figure 17 shows as an example the distribution obtained in the first test.

Hence, also the results of this second experiment confirm our initial hypothesis. In addition, if we use the p-values for measuring the relative distance between the sample means, the technique which include weaker links performs better in discriminating the two populations.

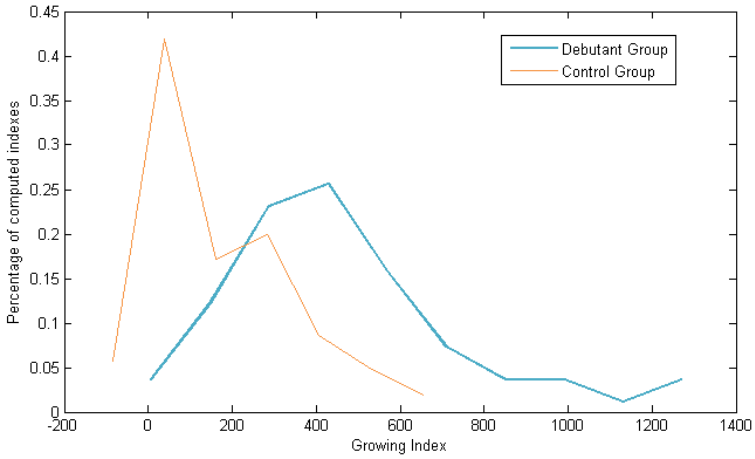

Figure 17. Distributions of growing indexes for both groups when filtering links associated with less than 3 publications.

Discussion

In this study, we analysed the topic network with the aim of confirming that the emergence of new research areas is anticipated by the dynamics of established topics. We examined the pace of collaboration (via the cliques-based method) and the change in topology (via the triads-based method) in the portion of network related to debutant topics, confirming that is possible to effectively discriminate the subgraphs associated to the future emergence of a debutant topic from the ones in the control group. In particular, the first experiment showed that the portion of the topics network in which a new topic will arise exhibits a significant increase in the pace of collaboration. The second experiment suggested that also the topology of networks tends to anticipate the appearance of a topic. In particular, it highlighted that the density of the network is higher in the portions which will give birth to new topics.

The ability of the two approaches of discriminating the debutant graph from the control group varies according to the period. Looking at their best results, reported in Figure 13 and Figure 14, it appears that the cliques-based approach works better (according to the resulting p-values) in the first years of the decade (2000-2004) while the triads-based one approach yielded better performance in the last years (2005-2010). This indicates that the second approach seems to work better when the topic network is nosier and denser, as it does happen for the second period. In this sense, the two approaches are complementary and the best one will depends on the characteristics of the topic graph under analysis.

The findings of this study are relevant to a number of research communities and stakeholders. Firstly, they confirm the existence of an embryonic phase in the development of research topics and suggest that it might be possible to perform very early detection of research trends by taking into account the aforementioned dynamics. Secondly, they bring new empirical evidences to related theories in philosophy of science, such as Herrera et al. (2010), Kuhn (2012), Nowotny et al. (2013), and Sun et al. (2013). Finally, they highlight that most new topics actually tend to be born in an environment in which previously less interconnected research areas start cross-fertilising and generating original ideas. This suggests that interdisciplinarity is one of the most significant forces that push research forward, allowing to integrate a diversity of expertizes and perspectives to come up with new solutions and new visions. The results of our analysis may thus support relevant research policies.

CONCLUSIONS

In this paper, we hypothesised the existence of an embryonic stage for research topics, in which they are not yet been labelled or associated with a considerable number of publications, and suggest that it is possible to detect topics in this stage by analysing the dynamics between already existent topics. To confirm this hypothesis, we performed an experiment on 75 debutant topics in Computer Science, which brought to the extraction and analysis of topic networks including about 2000 topics, from a sample of 3 million papers in the 2000-2010 interval. The results confirmed that the creation of novel topic is anticipated by a significant ($p < 0.0001$) raise in the pace of collaboration and density of the portion of network in which they will appear. These findings confirm the existence of an embryonic phase, potentially allowing for a very early detection of research topics, bring new empirical evidence to related theories in philosophy of science and suggest that an interdisciplinary environment is the most fertile ground for the creation of novel topics.

We now plan to exploit the dynamics discovered in this study for creating a fully automatic approach for detecting embryonic topics. We also intend to study and integrate a number of additional dynamics involving other research entities, such as authors and venues. The aim is to produce a robust approach that relies on multiple dynamics correlated with the emergence of new topics such that it could be used by researchers and companies alike for gaining a better understanding of where research is heading.

ACKNOWLEDGEMENTS

We would like to thank Springer Nature for partially funding this research and Elsevier B.V. for providing us with access to their large repositories of scholarly data.

REFERENCES

Becher T, and Trowler P. 2001. *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*: McGraw-Hill Education (UK).

Berners-Lee T, Hendler J, and Lassila O. 2001. The semantic web. *Scientific american* 284:28-37.

Blei D, and Lafferty J. 2006. Correlated topic models. *Advances in neural information processing systems* 18:147.

Blei DM, Ng AY, and Jordan MI. 2003. Latent dirichlet allocation. *J Mach Learn Res* 3:993-1022.

Bolelli L, Ertekin Ş, and Giles CL. 2009. Topic and trend detection in text collections using latent dirichlet allocation. *Advances in Information Retrieval*: Springer, 776-780.

Boyack KW, Klavans R, and Börner K. 2005. Mapping the backbone of science. *Scientometrics* 64:351-374.

Cataldi M, Di Caro L, and Schifanella C. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. Proceedings of the Tenth International Workshop on Multimedia Data Mining: ACM. p 4.

Chang J, and Blei DM. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*:124-150.

Chavalarias D, and Cointet J-P. 2013. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one* 8:e54847.

Couvalis G. 1997. *The philosophy of science: science and objectivity*: Sage.

Davis JA, and Leinhardt S. 1967. The structure of positive interpersonal relations in small groups.

Decker SL, Aleman-Meza B, Cameron D, and Arpinar IB. 2007. Detection of bursty and emerging trends towards identification of researchers at the early stage of trends. University of Georgia.

Duvvuru A, Kamarthi S, and Sultornsanee S. 2012. Undercovering research trends: Network analysis of keywords in scholarly articles. *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on*:265-270.

Duvvuru A, Radhakrishnan S, More D, Kamarthi S, and Sultornsanee S. 2013. Analyzing Structural & Temporal Characteristics of Keyword System in Academic Research Articles. *Procedia Computer Science* 20:439-445.

Erten C, Harding PJ, Kobourov SG, Wampler K, and Yee G. 2004. Exploring the computing literature using temporal graph visualization. *Electronic Imaging 2004*:45-56.

Faust K. 2010. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks* 32:221-233. http://dx.doi.org/10.1016/j.socnet.2010.03.004

Griffiths D, and Tenenbaum M. 2004. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems* 16:17.

Gruhl D, Guha R, Liben-Nowell D, and Tomkins A. 2004. Information diffusion through blogspace. *Proceedings of the 13th international conference on World Wide Web*:491-501.

He Q, Chen B, Pei J, Qiu B, Mitra P, and Giles L. 2009. Detecting topic evolution in scientific literature: how can citations help? *Proceedings of the 18th ACM conference on Information and knowledge management*:957-966.

Herrera M, Roberts DC, and Gulbahce N. 2010. Mapping the evolution of scientific fields. *PloS one* 5:e10355.

Jo Y, Lagoze C, and Giles CL. 2007. Detecting research topics via the correlation between graphs and texts. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*:370-379.

Kamaliha E, Riahi F, Qazvinian V, and Adibi J. 2008. Characterizing Network Motifs to Identify Spam Comments. 2008 IEEE International Conference on Data Mining Workshops. p 919-928.

Kuhn TS. 2012. *The structure of scientific revolutions*: University of Chicago press.

Larsen PO, and Von Ins M. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84:575-603.

Leydesdorff L. 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology* 58:1303-1319.

Luce RD, and Perry AD. 1949. A method of matrix analysis of group structure. *Psychometrika* 14:95-116.

Lv PH, Wang G-F, Wan Y, Liu J, Liu Q, and Ma F-c. 2011. Bibliometric trend analysis on global graphene research. *Scientometrics* 88:399-419.

Mathioudakis M, and Koudas N. 2010. Twittermonitor: trend detection over the twitter stream. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*:1155-1158.

Morinaga S, and Yamanishi K. 2004. Tracking dynamics of topic trends using a finite mixture model. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*:811-816.

Newman ME. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98:404-409.

Nowotny H, Scott PB, and Gibbons MT. 2013. *Re-thinking science: Knowledge and the public in an age of uncertainty*: John Wiley & Sons.

O'Callaghan D, Harrigan M, Carthy J, and Cunningham P. 2012. Identifying discriminating network motifs in YouTube spam. *arXiv preprint arXiv:12025216*.

Oka M, Abe H, and Kato K. 2006. Extracting topics from weblogs through frequency segments. Proceedings of WWW 2006 Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics.

Osborne F, and Motta E. 2012. Mining semantic relations between research areas. *The Semantic Web–ISWC 2012*: Springer, 410-426.

Osborne F, and Motta E. 2015. Klink-2: integrating multiple web sources to generate semantic topic networks. *The Semantic Web–ISWC 2015*: Springer, 408-424.

Osborne F, Motta E, and Mulholland P. 2013. Exploring scholarly data with rexplore. *The Semantic Web–ISWC 2013*: Springer, 460-477.

Osborne F, Scavo G, and Motta E. 2014. A hybrid semantic approach to building dynamic maps of research communities. *Knowledge Engineering and Knowledge Management*: Springer, 356-372.

Pham MC, Klamma R, and Jarke M. 2011. Development of computer science disciplines: a social network analysis approach. *Social Network Analysis and Mining* 1:321-340.

Pržulj N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23:e177-e183.

Rosen-Zvi M, Griffiths T, Steyvers M, and Smyth P. 2004. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*:487-494.

Salatino A. 2015. Early Detection and Forecasting of Research Trends.

Salatino AA, and Motta E. 2016. Detection of Embryonic Research Topics by Analysing Semantic Topic Networks.

Sun X, Ding K, and Lin Y. 2016. Mapping the evolution of scientific fields based on cross-field authors. *Journal of Informetrics* 10:750-761.

Sun X, Kaur J, Milojević S, Flammini A, and Menczer F. 2013. Social Dynamics of Science. *Scientific Reports* 3:1069. 10.1038/srep01069

Tseng Y-H, Lin Y-I, Lee Y-Y, Hung W-C, and Lee C-H. 2009. A comparison of methods for detecting hot topics. *Scientometrics* 81:73-90.

Ugander J, Backstrom L, and Kleinberg J. 2013. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. Proceedings of the 22nd international conference on World Wide Web: International World Wide Web Conferences Steering Committee. p 1307-1318.

Wu Y, Venkatramanan S, and Chiu DM. 2016. Research collaboration and topic trends in Computer Science based on top active authors. *PeerJ Computer Science* 2:e41.