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

Angelo A. Salatino, Francesco Osborne and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom

ABSTRACT

The ability to promptly recognise new research trends is strategic for many stake-holders, including universities, institutional funding bodies, academic publishers and companies. While the literature describes several approaches which aim to identify the emergence of new research topics early in their lifecycle, these rely on the assumption that the topic in question is already associated with a number of publications and consistently referred to by a community of researchers. Hence, detecting the emergence of a new research area at an *embryonic stage*, i.e., before the topic has been consistently labelled by a community of researchers and associated with a number of publications, is still an open challenge. In this paper, we begin to address this challenge by performing a study of the dynamics preceding the creation of new topics. This study indicates that the emergence of a new topic is anticipated by a significant increase in the pace of collaboration between relevant research areas, which can be seen as the 'parents' of the new topic. These initial findings (i) confirm our hypothesis that it is possible in principle to detect the emergence of a new topic at the embryonic stage, (ii) provide new empirical evidence supporting relevant theories in Philosophy of Science, and also (iii) suggest that new topics tend to emerge in an environment in which weakly interconnected research areas begin to cross-fertilise.

INTRODUCTION

Early awareness of the emergence of new research topics can bring significant benefits to anybody involved in the research environment. Academic publishers and editors can exploit this knowledge and offer the most up to date and interesting contents. Researchers may not only be interested in new trends related to their areas but may also find it very useful to be alerted about significant new research developments in general. Institutional funding bodies and companies also need to be regularly updated on how the research landscape is evolving, so that they can make early decisions about critical investments. Considering the growth rate of research publications (*Larsen & Von Ins, 2010*), keeping up with novel trends is a challenge even for expert researchers. Traditional methods, such as the manual exploration of publications in significant conferences and journals, are no longer viable. This has led to the emergence of several approaches capable of detecting novel topics and

Corresponding author
Angelo A. Salatino,
angelo.salatino@open.ac.uk

research trends (*Bolelli, Ertekin & Giles, 2009*; *Duvvuru, Kamarthi & Sultornsanee, 2012*; *He et al., 2009*; *Wu, Venkatramanan & Chiu, 2016*). However, all of these approaches focus on topics that are already associated with a number of publications and consistently referred to by a community of researchers. This limitation hinders the ability of stakeholders to anticipate and react promptly to new developments in the research landscape.

Hence, there is a need for novel methods capable of identifying the appearance of new topics at a very early stage, assessing their potential and forecasting their trajectory. To this end, we need first to achieve a better understanding of the dynamics underlying the creation of new topics and then investigate whether such understanding can be exploited to develop computationally effective methods, which are capable of detecting the emergence of new topics at a very early stage.

The field of Philosophy of Science offers a number of interesting 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 context, *Becher & Trowler (2001)* explained that, even if science proceeds towards more specific disciplines, and thus researchers in different communities become less compatible, they are still inclined to collaborate for mutual benefit. *Herrera, Roberts & Gulbahce (2010)*, *Sun et al. (2013)* and *Nowotny, Scott & Gibbons (2013)* suggested that the development of new topics is 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, Scavo & Motta (2014)* provided empirical evidence to these theories by analysing the social dynamics of researchers and their effects on the formation and life-cycle of research communities and topics.

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

Inspired by these theories, we hypothesize the existence of an even earlier phase, which we name *embryonic phase*, in which a topic has not yet been explicitly labelled and recognized by a research community, but it is already taking shape, as evidenced by the fact that researchers from a variety of fields are forming new collaborations and producing new work, starting to define the challenges and the paradigms associated with the emerging new area.

We also hypothesize that it could be possible to detect topics at this stage by analysing the *dynamics* of already established topics. In this context, we use the term *dynamics* to refer to the significant trends associated with a topic, including the interactions between topics and those between entities linked to these topics, such as publications, authors, venues. For example, the sudden appearance of some publications concerning a combination of previously uncorrelated topics may suggest that some pioneer researchers are investigating

new possibilities and maybe shaping a new emerging area. In the same way, as pointed out by *Salatino (2015)*, we can hypothesize a wide array of relevant patterns of activity, which could anticipate the creation of a new research area. These may include a new collaboration between two or more research communities (*Osborne, Scavo & Motta, 2014*), the creation of interdisciplinary workshops, a rise in the number of experts working on a certain combination of topics, a significant change in the vocabulary associated with relevant topics (*Cano Basave, Osborne & Salatino, 2016*), and so on.

In this paper we present a study that aims to uncover key elements associated with the research dynamics preceding the creation of novel topics, thus providing initial evidence to support our hypotheses. In particular, our study provides evidence that the emergence of a novel research topic can be anticipated by a significant increase in the pace of collaboration between relevant research areas, which can be seen as the 'parents' of the new topic.

Our study was performed on a sample of three million publications in the 2000–2010 interval. It was conducted by comparing the sections of the co-occurrence graphs where new topics are about to emerge with a control group of subgraphs associated with established topics. These graphs were analysed by using two novel approaches that integrate both statistics and semantics. We found that the pace of collaboration and the density measured in the sections of the network that will give rise to a new topic are significantly higher than those in the control group. These findings support our hypothesis about the existence of an embryonic phase and also yield new empirical evidence consistent with the aforementioned theories in Philosophy of Science. In addition, the identified dynamics could be used as the starting point for developing new automatic methods, which could detect the emergence of a new research topic well before this becomes explicitly recognised and established.

The study presented in this paper is an extension of the work by *Salatino & Motta (2016)*. The new contributions of this paper are: (1) a larger sample including 75 debutant topics and 100 established ones, (2) a new technique for measuring the density of the topic graph, (3) a more exhaustive statistical analysis, including a 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 first review the literature regarding the early detection of topics, pointing out the existing gaps. We then describe the experimental approach used for the study, present the results and discuss their implications. Finally, we 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 recent years and has been applied to a variety of scenarios, such as social networks (*Cataldi, Di Caro & Schifanella, 2010*; *Mathioudakis & Koudas, 2010*), blogs (*Gruhl et al., 2004*; *Oka, Abe & Kato, 2006*), emails (*Morinaga & Yamanishi, 2004*) and scientific literature (*Bolelli, Ertekin & Giles, 2009*; *Decker et al., 2007*; *Erten et al., 2004*; *Lv et al., 2011*; *Osborne, Scavo & Motta, 2014*; *Sun, Ding & Lin, 2016*; *Tseng et al., 2009*).

The literature presents several works on research trend detection, which can be characterised either by the way they define a topic or the techniques they use detect it

(*Salatino (2015)*. *Blei, Ng & Jordan, (2003)* have developed the well-known Latent Dirichlet Allocation (LDA), an unsupervised learning method to extract topics from a corpus, which models topics as a multinomial distribution over words. Since its introduction, LDA has been extended and adapted to several applications. For example, *Blei & Lafferty (2006)* have introduced the Correlated Topic Model using the logistic normal distribution instead of the Dirichlet one, to address the issue that LDA fails to model correlations between topics. *Griffiths et al. (2004)* have 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)* present the Author-Topic Model (ATM), which includes authorship information and associates each topic to a multinomial distribution over words and each author to a multinomial distribution over topics. *Bolelli, Ertekin & Giles (2009)* introduce the Segmented Author-Topic model which extends ATM by adding the temporal ordering of documents to address the problem of topic evolution. In addition, *Chang & Blei (2010)* have developed the *relational topic model* which combines LDA and the network structure of documents to model topics. Similarly, *He et al. (2009)* have 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 leverages citations to connect topics in different time frames. In a similar way, *Morinaga & Yamanishi (2004)* employ 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, their evaluation rests on an email corpus, thus it is not clear how it would perform on scientific corpus. A general issue affecting this kind of approaches is that it is not always easy to associate clearly identifiable research areas to the resulting topic models.

In addition to LDA, the Natural Language Processing (NLP) community have 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 performed clustering analysis to discover patterns in the evolution of science *Jo, Lagoze & Giles (2007)* developed an approach that correlates the distribution of terms extracted from a text with the distribution of the citation graphs related to publications containing these terms. Their work assumes that if a term is relevant to a topic, documents containing that term will have a stronger connection than randomly selected ones. 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) have analysed the network of co-occurring keywords in a scholarly corpus and monitored the evolution in time of the link weights, to detect 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—in many cases different keywords even refer to the same topic. For example, *Osborne, Scavo & Motta (2014)* showed that a semantic characterisation of research topics yields better results than keywords 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 such 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 taxonomies tend to evolve slowly and, in a fast-changing research field, such as *Computer Science* (*Pham, Klamma & Jarke, 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 literature comprises a wide collection of approaches for detecting research trends. However, they focus on already recognised topics, which are either already associated with recognized label or, in the case of probabilistic topics models, with a set of terms that have previously appeared in a good number of publications. Detecting research trends at an embryonic stage remains an open challenge.

MATERIALS AND METHODS

The aim of this study was to measure the association between the emergence of a new topic and the increase of the pace of collaboration and density previously observed in the co-occurrence graphs of related topics. To this end, we represent topics and their relationships in a certain time frame 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, Klavans & Börner, 2005*; *Leydesdorff, 2007*; *Newman, 2001*). In the following we will refer to it as *topic graph* or *topic network*. We analysed 75 topics that debuted in the 2000–2010 period using 100 established topics as a control group.

In our previous work (*Salatino & Motta, 2016*), we conducted a similar analysis on a smaller sample. The sample analysed in this paper was selected by iteratively adding new topics until we reached data saturation (*Fusch & Ness, 2015*), i.e., the results of the analysis did not vary significantly with the inclusion of new data points.

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 results of this study are available at https://osf.io/bd8ex/.

Semantic enhanced topic network

We use as dataset the metadata describing three million papers in the field of *Computer Science* from a dump of the well-known Scopus dataset (https://www.elsevier.com/solutions/scopus). In this dataset each paper is associated to a number of keywords that can 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 do not represent topics (e.g., *case study*) and multiple keywords can refer to the same topic (e.g., *ontology mapping* and *ontology matching*).

The literature offers several methods for characterizing research topics. Probabilistic topic models, such as LDA, are very popular solutions, which however are most effective in scenarios where fuzzy classification is acceptable, there is no good domain knowledge and

it is not important for users to understand the rationale of a classification. However, these tenets do not apply to this study. Furthermore, it is not easy to label the topics produced by a probabilistic topic model with specific and distinct research areas. Conversely, in this study it is important to be able to associate topics with well-established research areas.

A second approach, used by several digital libraries and publishers is tagging publications with categories from a pre-determined taxonomy of topic. Some examples include the ACM computing classification system (http://www.acm.org/publications/class-2012), the Springer Nature classification (http://www.nature.com/subjects), Scopus subject areas (https://www.elsevier.com/solutions/scopus/content), and the Microsoft Academic Search classification (http://academic.research.microsoft.com/). This solution has the advantage of producing sound topics, agreed upon by a committee of experts. However, these taxonomies suffer from a number of issues. First, building large-scale taxonomies requires a sizable number of experts and it is an expensive and time-consuming process. Hence, they are seldom updated and grow obsolete very quickly. For example, the 2012 version of the ACM classification was finalized fourteen years after the previous version. In addition, these taxonomies are very coarse-grained and usually contain general fields rather than fine-grained research topics.

We addressed these issues by characterizing our topics according to the Computer Science Ontology (CSO) produced by Klink-2 (*Osborne & Motta, 2015*), which describes the relationships between more than 15,000 research areas extracted from a corpus of 16 million publications. Klink-2 is an algorithm that is able to generate very granular ontologies and update them regularly by analysing keywords and their relationships with research papers, authors, venues, and organizations, and by taking advantage of multiple knowledge sources available on the web. Klink-2 is currently integrated in the Rexplore system (*Osborne, Motta & Mulholland, 2013*), a platform for exploring and making sense of scholarly data, which provides semantic-aware analytics.

We used the CSO ontology to semantically enhance the co-occurrence graphs by removing all keywords that did not refer to research areas and by aggregating keywords representing the same concept, i.e., keywords linked by a *relatedEquivalent* relationship in the ontology (*Osborne, Motta & Mulholland, 2013*). For example, we aggregated keywords such as "semantic web", "semantic web technology" and "semantic web technologies" in a single semantic topic and we assigned it to all publications associated with these keywords.

We built sixteen topic networks representing 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 represents the number of publications in which a topic appears in a year, while the link weight is equal to the number of publications in which two topics co-occur in the same year.

Graph selection

We randomly selected 75 topics that debuted in the period between 2000 and 2010 as treatment group (also referred to as debutant group). A topic debuts in the year in which its label first appears in a research paper. The control group (also referred to as non-debutant
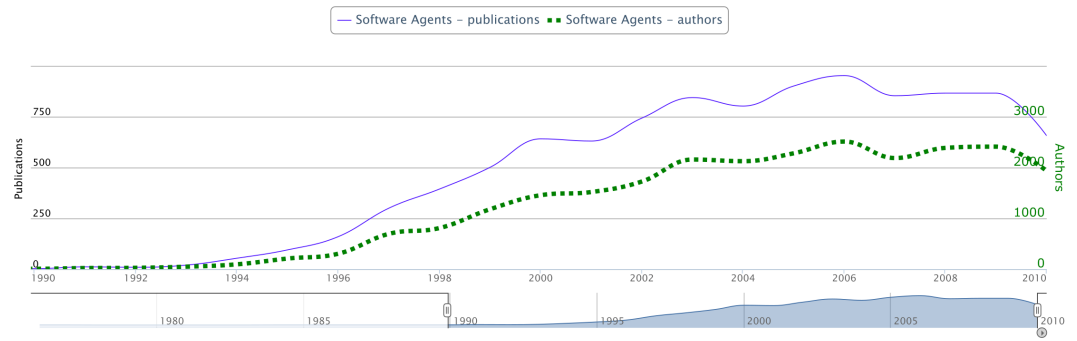
Figure 1 **Evolution of the topic Software Agents in terms of number of authors and number of publications per year.** The chart has been produced using the Rexplore system.
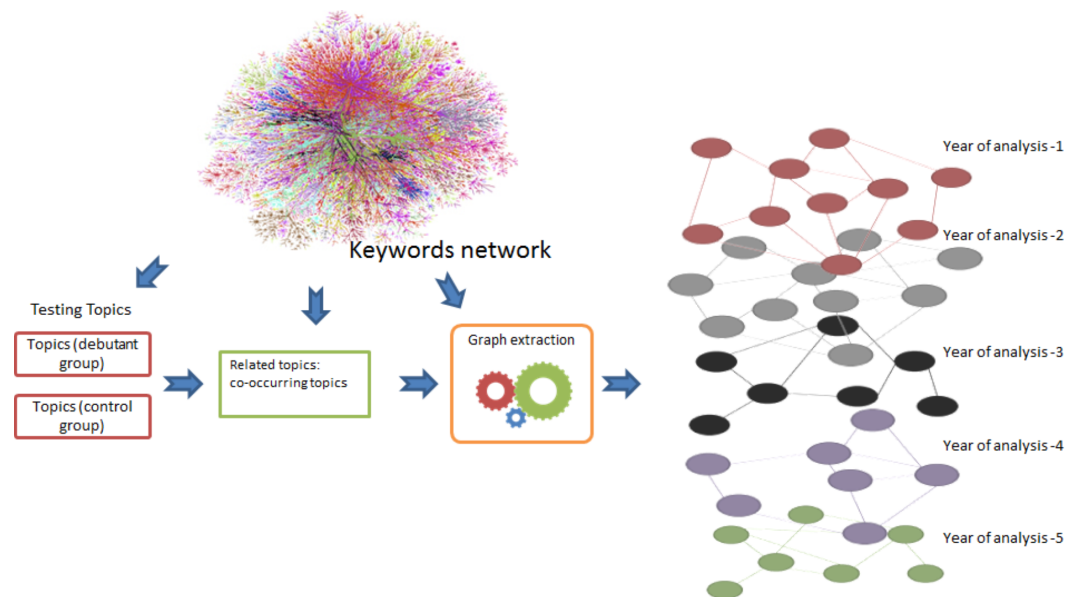
Figure 2 **Workflow representing all the steps for the selection phase.**

group), was obtained by selecting 100 well-established topics. We considered a topic as well-established if: (i) it debuted before 2000, (ii) it appears in the CSO Ontology, (iii) it is associated each year with a substantial and consistent number of publications. As an example, Fig. 1 shows the evolution through time of the well-established topic *Software Agents,* in terms of number of active authors and publications. The figure shows that the topic made its debut in 1993 and in the year 2000 reached a rate of over 500 publications per year with more than 1,500 authors working on it. It can thus be considered established in the context of our study.

We assume that a new topic will continue to collaborate with the topics that contributed to its creation for a certain time after its debut. This assumption was discussed and tested in a previous study (*Osborne & Motta, 2012*), where it was used to find historical subsumption links between research areas. Hence, as summarized in Fig. 2, for each debutant 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 the year of debut. Since we want to analyse how the dimension of these subgraphs influences the results, we tested different values of n (20, 40, and 60). For example, if a topic A made its debut in 2003, the portion of network containing its most co-occurring topics is 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 selected 50 established topics and assigned a random *year of analysis* to each of them. For this study, we randomly assigned each established topic to two consecutive years within the decade 2000–2010, with the consequence of doubling the control group, thus reducing noise and smoothing the resulting measures.

In brief, the selection phase associates to each topic in the treatment and control groups (also referred as *input topics*) a graph:

$$G^{\text{topic}} = G^{\text{topic}}_{\text{year}-5} \cup G^{\text{topic}}_{\text{year}-4} \cup G^{\text{topic}}_{\text{year}-3} \cup G^{\text{topic}}_{\text{year}-2} \cup G^{\text{topic}}_{\text{year}-1}. \tag{1}$$

This graph corresponds to the co-occurrence network of a debutant topic in the five years prior to its emergence (or year of analysis for non-debutant topics). In particular, each year corresponds to the sub-graphs:

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

in which $V^{\text{topic}}_{\text{year}-i}$ is the set of most co-occurring topics in a year and $E^{\text{topic}}_{\text{year}-1}$ is the set of edges linking the nodes in the set.

The graphs associated to the debutant topics included 1,357 unique topics, while the ones associated to the control group included 1,060 topics.

Graph analysis

We assess the dynamics in the graphs with two main approaches: clique-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 relevant topics and then averages the results over all 3-cliques. The second measures the increase in the topic graph density using the triad census technique (*Davis & Leinhardt, 1967*). In the following two sections we describe both methods in details.

Clique-based method

We measure the collaboration pace of a graph by analysing the diachronic activity of triangles of collaborating topics. To this end, we first extract all 3-cliques from the five sub-graphs associated to each topic under analysis. A 3-clique, as shown in Fig. 3, 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*).

To study the dynamics preceding the debut of each topic, we analyse the evolution of the same 3-clique in subsequent years. Figure 4 summarizes the process. Considering a 3-clique having nodes $\{A, B, C\}$, we quantify its collaboration index μ_Δ in a year by taking
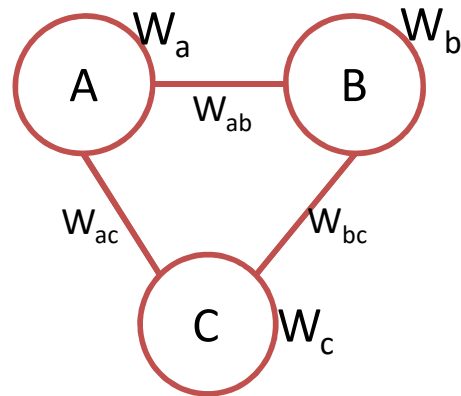
Figure 3 An instance of a 3-clique containing node and link weights.
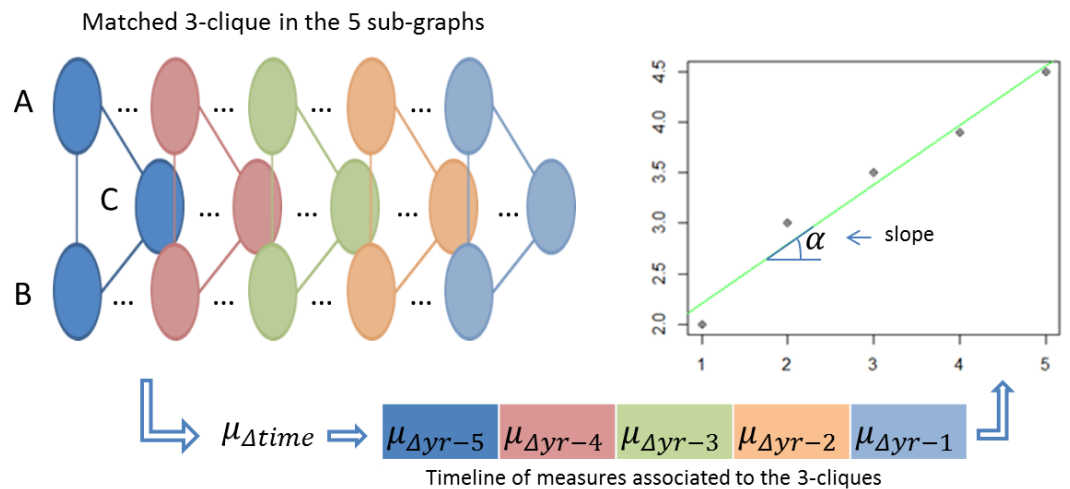
Figure 4 Main steps of the analysis phase.

into account both node weights $\{W_a, W_b, W_c\}$ and link weights $\{W_{ab}, W_{bc}, W_{ca}\}$.

$$
\begin{aligned}
\mu_{A-B} &= \text{mean}(P(A|B), P(B|A)) \\
\mu_{B-C} &= \text{mean}(P(B|C), P(C,B)) \\
\mu_{C-A} &= \text{mean}(P(C|A), P(A|C)) \\
\mu_\Delta &= \text{mean}(\mu_{A-B}, \mu_{B-C}, \mu_{C-A}).
\end{aligned}
\tag{3}
$$

The index μ_Δ is computed by aggregating the three coefficients μ_{A-B}, μ_{B-C} and μ_{C-A} as illustrated by Eq. (3). The strength of collaboration μ_{x-y} between two nodes of the topic network, x and y, is computed as the mean of the conditional probabilities $P(y|x)$ and $P(x|y)$, where $P(y|x)$ is the probability that a publication associated with a topic x will be associated also with a topic y in a certain year. The advantage of using conditional probabilities instead of the number of co-occurrences is that the value μ_{x-y} is normalised with respect to the number of publications associated to each topic. Finally, μ_Δ is computed as the mean of the strengths of collaboration of the three links in a 3-clique. This solution
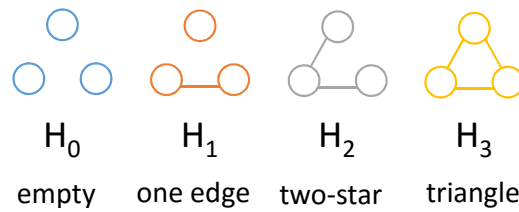
Figure 5 **The four isomorphism classes of triad.** The triad census counts the frequencies of *Hi* in the input graph.

was adopted after testing alternative approaches during the *preliminary evaluation*, as discussed in the Results section.

The evolution of the 3-clique collaboration pace can be represented as a timeline of values in which each year is associated with its collaboration pace, as in Eq. (4). We assess the increase of the collaboration pace in the period under analysis by computing the slope of the linear regression of these values.

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

Initially, we tried to determine the increase in the collaboration pace exhibited by a clique by simply taking the difference between the first and last values of the timeline ($\mu_{\Delta\text{yr}-5} - \mu_{\Delta\text{yr}-1}$). 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 it is negative, the number and intensity of collaborations are decreasing.

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

To summarize, for each input topic we select a subgraph of related topics in the five years preceding the year of debut (or *analysis* for topics in the control group). We then extract the 3-cliques and associate each of them with a vector representing the evolution of their pace of collaboration. The trend of each clique is computed as the angular coefficient of the linear regression of these values. Finally, the increase in the pace of collaboration of a subgraph is obtained by averaging these values.

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 to as *global 3-profiles*, is a four-dimensional vector representing the frequencies of the four isomorphism classes of triad, as shown in Fig. 5.

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_{year-5}^{topic}	446	790	807	882
G_{year-4}^{topic}	443	854	915	1,064
G_{year-3}^{topic}	125	486	967	1698
G_{year-2}^{topic}	100	410	908	1,858
G_{year-1}^{topic}	68	486	849	2251

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

In this study, we use triad census to describe all the sub-graphs associated to a input topic in terms of frequencies of H_i (see Fig. 5) and we then evaluate how the frequencies of *empties* (H_0), *one edges* (H_1), *two-stars* (H_2) and *triangles* (H_3) change in time. Figure 5 illustrates the four classes of triads for an undirected graph in the case of topic networks. An increase in the number of triangles suggests the appearance of new collaboration clusters among previously distant topics.

In contrast with the 3-cliques approach, the triad census does not consider the weight of the links, but only their existence. Hence, it is useful to assess how the inclusion of links with different strengths affects 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.

We initially perform the triad census over the five graphs associated to each input topic. For example, Table 1 shows the results of the triad census over the five sub-graphs associated with the debutant topic *Artificial Bee Colonies*.

Next, we check whether the co-occurrence graph is becoming denser by analysing the change of frequencies associated with H_i (see Fig. 6). We first calculate the percentage growth of each H_i (Eq. (5)) and then compute their weighted summation (Eq. (6)). We label the resulting metric the *growth index*. We empirically tested other solutions for aggregating the various contributions (e.g., considering only H_3, summing the values, weighting the sum in a variety of ways) and found that this definition of growth index provides the best discrimination between the two classes of graphs.

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

$$Growing\ Index_{topic} = \sum_{i=0}^{3} i \cdot \%GrowthH_i. \qquad (6)$$

The *growth index* takes into account the contributions from H_1, H_2 and H_3. Although, the number of triangles (H_3) can by itself be a fair indicator of the density, previous studies showed that all four classes of triads are useful for computing network properties,
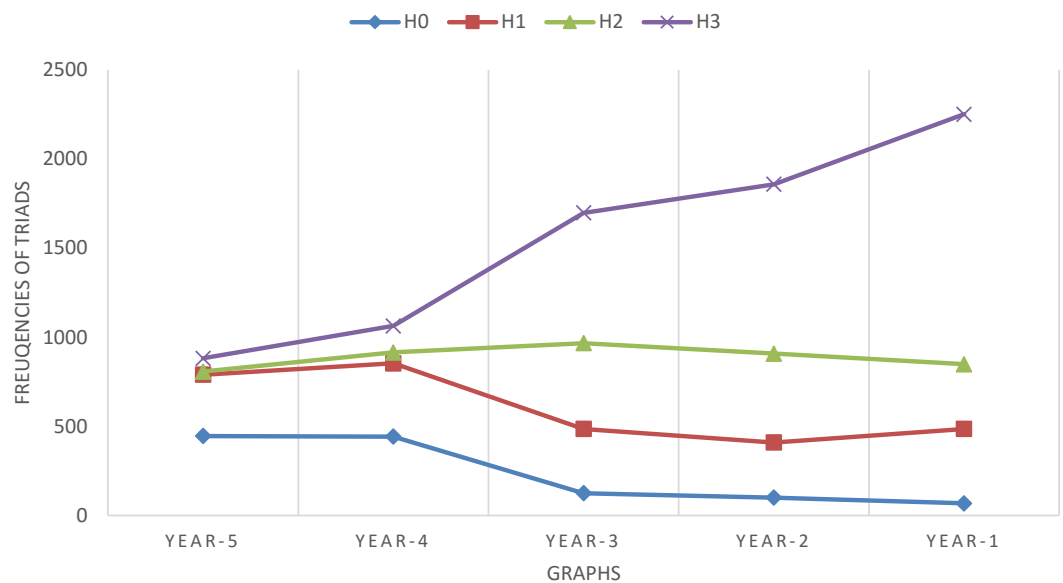
Figure 6 Development in time of the frequencies of *Hi* in the network related to the emergence of "Artificial Bee Colonies".

including transitivity, intransitivity and density (*Faust, 2010*; *Holland & Leinhardt, 1976*). Taking into consideration only H_3 might fail to detect some subtler cases, characterized for example by a contemporary increase of H_2 and decrease of H_1.

To summarize, the triad-based method receives the same input as the clique-based method. For each of the five subgraphs associated to a topic, we perform the triad census obtaining the different frequencies, H_i, in different years. We then analyse them diachronically to quantify the increase in density.

RESULTS

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

- The preliminary evaluation performed on a reduced dataset for assessing the metrics used in the clique-based method;
- The full study using the clique-based method;
- The full study using the triad-based method.

Preliminary evaluation with alternative clique-based methods

We conducted a preliminary evaluation aiming at choosing the most effective clique-based method for assessing the pace of collaboration. This test focused on the subgraph of the 20 most co-occurring topics associated with the topic *Semantic Web* (debuting in 2001) and *Cloud Computing* (2006) versus a control group of 20 subgraphs associated to a group of non-debutant topics. We tested on this dataset two techniques to compute the weight of a clique (harmonic mean and arithmetic mean) and two methods to evaluate its trend
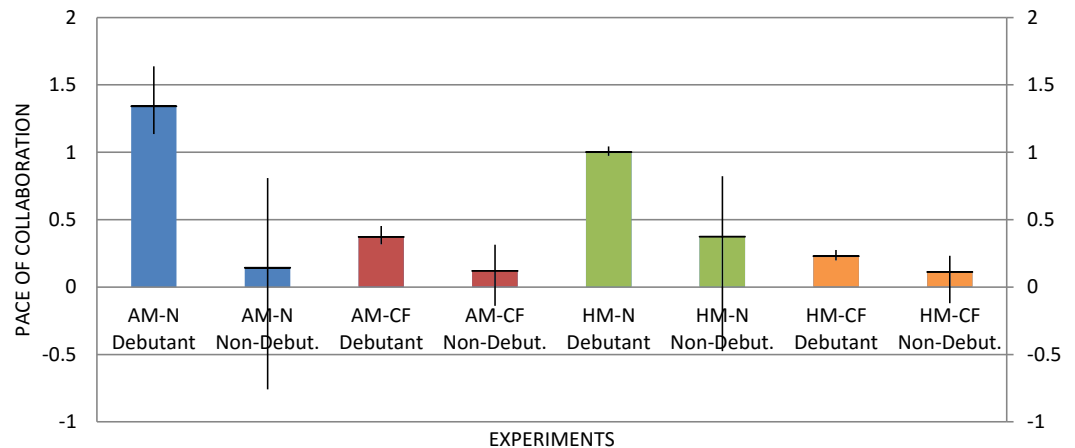
Figure 7 Overall directions of the sub-graphs related to input topics in both debutant and control group with all four approaches.

(computing the difference between the first and the last values and linear regression). 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 regression coefficient;
- **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 regression coefficient.

Figure 7 illustrates the average pace of collaboration for the sub-graphs associated with each topic according to these methods and the range of their values (thin vertical line). The results support the initial hypothesis: 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.

The approaches based on the simple difference (AM-N and HM-N) exhibit the larger gaps between the two groups in terms of average pace of collaboration. However, the ranges of values 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 Fig. 9), we can see that the overall collaboration paces of the debutant topics (DB) are always significantly higher than the control group (NDB).

With the null hypothesis: "*The differences in the pace of collaboration between the debutant topics and topics in the control group result purely from chance*", we ran Student's *t*-test on the sample of data provided by the HM-CF approach, to verify whether the two
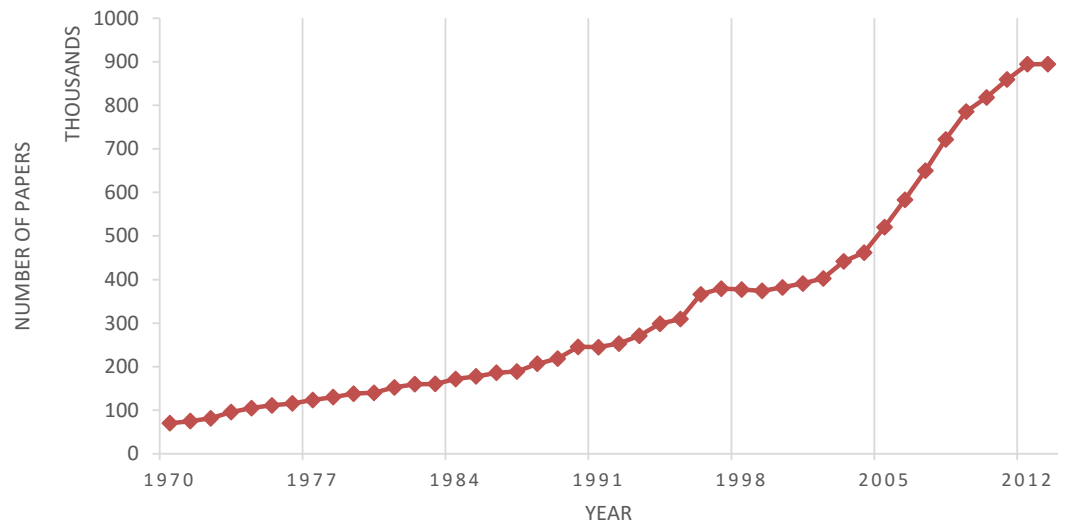
Figure 8 Number of papers each year in the period 1970–2013 in the dataset under analysis.
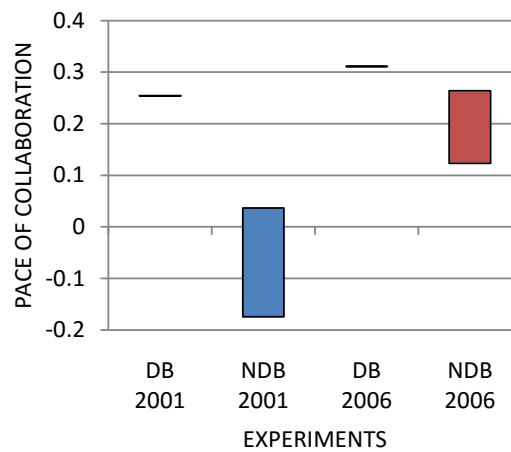
Figure 9 Overall directions of the sub-graphs related to input topics in both debutant and control group in HM-CF approach.

groups belong to different populations. The test yielded $p < 0.0001$, which allowedt us to reject the null hypothesis that the differences between the two distributions were due to random variations.[1] Based on this result, we could further confirm that the HM-CF approach performs better compared to the other approaches. For this reason, we selected the combination of *harmonic mean* and *linear regression* as the approach for the full study using the clique-based method.

The results of HM-CF give interesting insights on the creation of some well-known research topics. Tables 2 and 3 list the cliques that 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 collaboration of the *World Wide Web* area with topics such as, *Information Retrieval*, *Artificial Intelligence*, and *Knowledge Based Systems*. This is

[1] We consider $p < 0.0001$ as a conventional statistical representation to indicate an extremely high statistical significance (>500 times stronger than the conventional 0.05 threshold for claiming significance). It includes all mathematical outcomes below 0.0001, which are essentially equivalent in assessing excellent significance.

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

consistent with the initial vision of the semantic web, defined in the 2001 by the seminal work of *Berners-Lee, Hendler & Lassila (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 also for identifying the topics that gave rise to a research area.

Clique-based method study

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

Since the pace of collaboration shows significant changes within the period considered, we studied its behaviour across the 2000–2010 interval. Figures 11A– 11C, show the average yearly collaboration pace 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 suffered a fall for both debutant and non-debutant topics. This may be due to the fact that the topic network became denser and noisier in the final years of the interval. Moreover, the most recent debutant topics often have an underdeveloped network of co-occurrences, which may result in a suboptimal selection of the group of topics to be analysed in the previous years. Therefore, simply selecting the 20 most co-occurring topics may not allow us to highlight the real dynamics preceding the creation of a new topic.
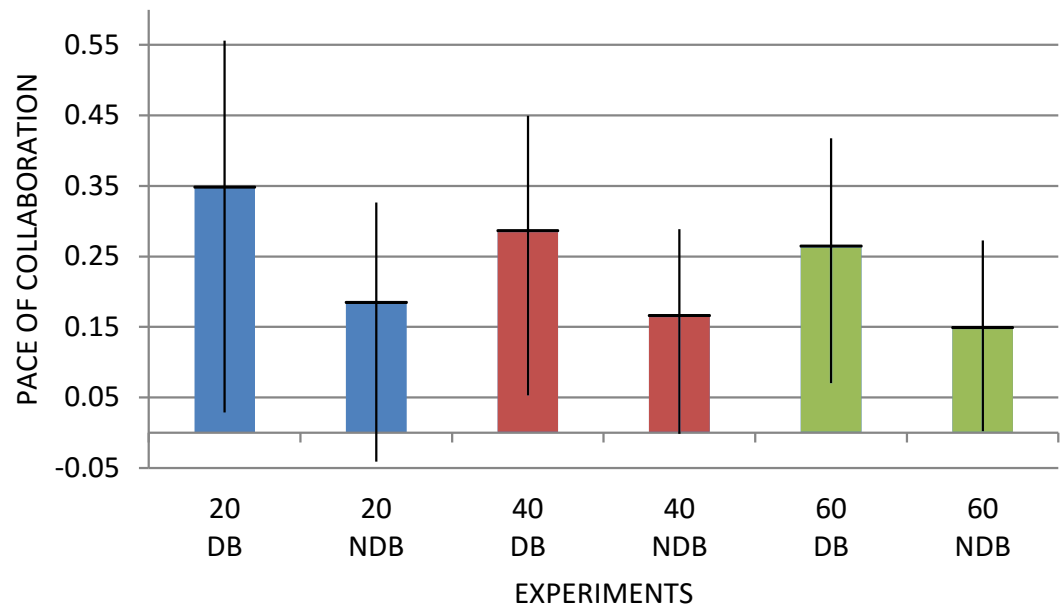
Figure 10 Average collaboration pace of the sub-graphs associated to the treatment (DB) and control group (NDB), when selecting the 20, 40 and 60 most co-occurring topics. The thin vertical lines represent the ranges of values.

Table 4 compares the collaboration pace of 24 debutant topics with 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, which emerged in the last decade, was anticipated by the dynamics of the topic network. The Student's t-test confirmed that the debutant and established topics do not belong to the same population ($p < 0.0001$). The results of the t-test also suggest that the experiment involving the 60 most co-occurring topics, represented in Fig. 11C, provides a better discrimination of debutant topics from non-debutant ones. For the sake of completeness, in Table 5 we report the p-values yielded by each experiment.

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

Triad-based method study

We applied the triad-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, to understand how collaboration strength influences the outcome.

Figure 12A reports the average value of the growth indexes when discarding links with less than 3 co-occurrences. The approach allows us to discriminate well the portion of networks related to debutant topics from the ones related to the control group. In particular, the density of network associated with the debutant topics is always higher than its counterpart. Figsures 12B and 12C report the results obtained by removing links with less than 10 and 20 co-occurrences. As in the previous experiment, we adopted the   Student's t-test to
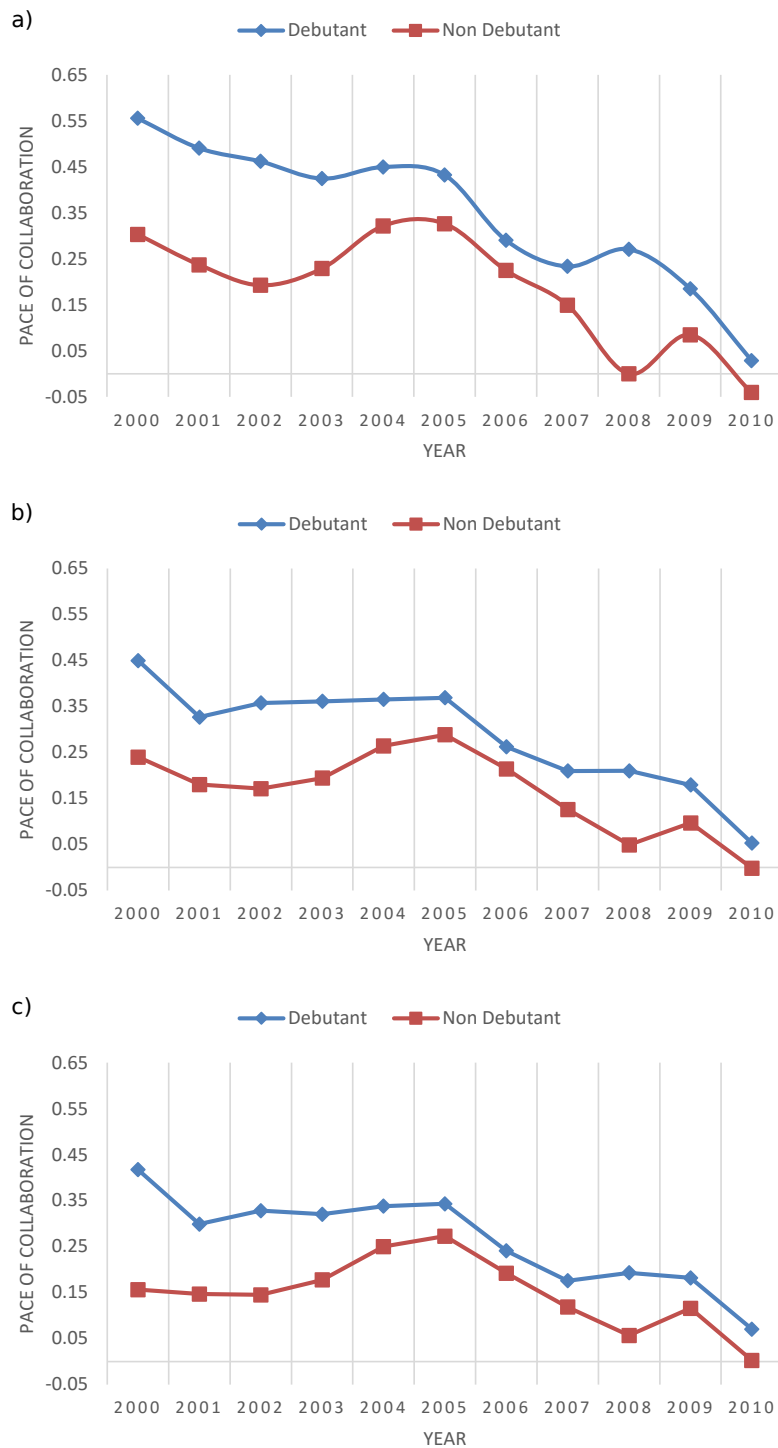
Figure 11 **Average collaboration pace per year of the sub-graphs related to input topics in both debutant and control groups considering their 20 (A), 40 (B) and 60 (C) most co-occurring topics.** The year refers to the year of analysis of each topic.

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 (year of debut)	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
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

Table 5 *P*-values obtained performing the Student's *t*-test over the distributions of both debutant and control groups considering their 20, 40 and 60 most co-occurring topics. The best result is bolded.

Experiment	*p*-value	Associated chart
20 most co-occurring topics	$4.22 \cdot 10^{-2}$	Fig. 11A
40 most co-occurring topics	$6.84 \cdot 10^{-2}$	Fig. 11B
60 most co-occurring topics	$\mathbf{4.64 \cdot 10^{-45}}$	**Fig. 11C**

understand among the three tests which one could provide better discrimination between the two classes of topics. The results of the *t*-test suggest that the experiment in which we discard links with less than three co-occurrences provides a better discrimination of debutant topics from non-debutant ones. This suggests that considering weak connections is more beneficial for discriminating the two groups. The 2004 peak is caused by the debut of a number of topics associated with particularly strong underlying dynamics, such as *Linked Data*, *Pairing-based Cryptography*, *Microgrid* and *Privacy Preservation*.

Table 6 reports as an example the triad census performed over the subgraph associated with the topic *Semantic Web Technologies* (SWT) debuting in 2001. We can see an increase

Table 6 The results of the triad census performed on the network associated with the debutant topic "semantic web technology" removing links associated with less than 3 (left), 10 (right) and 20 (bottom) publications.

	Removing links <3				Removing links <10				Removing links <20			
Graph	H_0	H_1	H_2	H_3	H_0	H_1	H_2	H_3	H_0	H_1	H_2	H_3
1996	1,124	1,157	658	337	641	676	316	138	796	509	174	618
1997	928	1,237	670	441	1,022	828	315	135	632	432	204	62
1998	1,255	1,353	657	389	585	705	300	181	525	418	145	52
1999	1,307	1,431	861	461	1,222	1,098	413	192	569	497	187	77
2000	913	1,399	1,043	705	1,482	1,361	554	257	842	618	228	83

in the number of triangles (H_3) and two-stars (H_2), mirroring the increasing density of the topic network. Again, this phenomenon is more evident when also using 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).

Table 7 shows a selection of debutant topics and their growth indexes compared with the growth index of the control group in the same year. If we can compare this table to Table 4, we can see that the two methods used in this study reflect the same dynamics.

With the null hypothesis "*The differences in growth index between the debutant topics and topics in the control group result purely from chance*", we ran Student's t-test over the two distributions of growth indexes, for all three experiments. It yielded $p < 0.0001$ for all the experiments. More details about the computed p-values per each experiment performed in this triad-based study can be found in Table 8. Figure 13 shows, as an example, the distributions associated to the two groups of topics obtained in the first test.

Hence, the results from this second experiment confirm our initial hypothesis too. In addition, per Table 8, the results from the t-test also suggest that the first experiment, which ignores the links associated with less than three publications, better discriminates the two populations.

DISCUSSION

We analysed the topic network with the aim of experimentally confirming our hypothesis that the emergence of new research areas is anticipated by an increased rate of interaction of pre-existing topics. We examined the pace of collaboration (via the clique-based method) and the change in topology (via the triad-based method) in portions of the network related to debutant topics, showing that it is possible to effectively discriminate areas of the topic graph associated with the future emergence of new topics. The first experiment showed that the subgraphs associated with the emergence of a new topic exhibit a significantly higher pace of collaboration than the control group of subgraphs associated with established topics. Similarly, the second experiment showed that the graphs associated with a new topic display a significantly higher increase in their density than the control group. We can thus confirm that these two aspects can play a key role in the context of defining methods for detecting embryonic topics.

a)

b)

c)
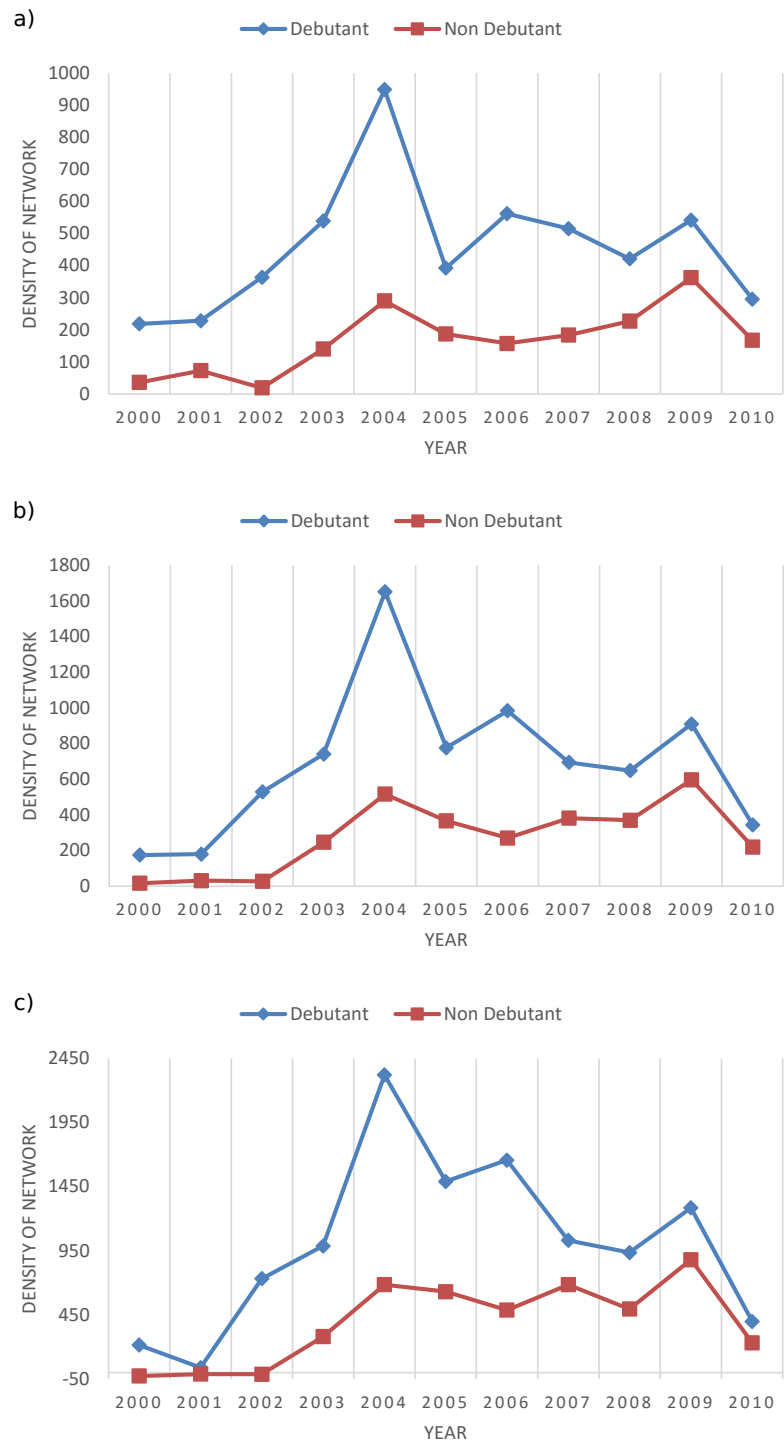
Figure 12 Average growth index per year of the sub-graphs related to the topics in both debutant and non-debutant groups considering their 60 most co-occurring topics and filtering links having with less than 3 (A), 10 (B) and 20 (C) publications.

Table 7 Growth indexes of sub-graphs associated to selected debutant topics versus the average growth index of the control group in the same year of debut (Standard Growth Index).

Topic (year of debut)	Growth index	Standard growth 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

Table 8 *P*-values obtained performing the Student's *t*-test over the distributions of both debutant and control groups considering their 60 most co-occurring topics filtering links having with less than 3, 10 and 20 publications. The best result is bolded.

Experiment	*p*-value	Associated chart
Less than 3 publications	$\mathbf{6.43 \cdot 10^{-16}}$	**Fig. 12A**
Less than 10 publications	$1.69 \cdot 10^{-11}$	Fig. 12B
Less than 20 publications	$3.52 \cdot 10^{-10}$	Fig. 12C

Interestingly, the ability of the two approaches in discriminating the debutant group from the control group varies with the time interval considered. It appears that the clique-based approach (see Fig. 11) discriminates them more effectively in the initial period, whereas the triad-based method (Fig. 12) seems to perform better in the central years (2004–2007). We intend to investigate in future work if these behaviours are associated with specific characteristics of the network.

The results of these two experiments allow us to effectively discriminate specific sections of the topic graph and suggest that a significant increase in the rate of collaboration
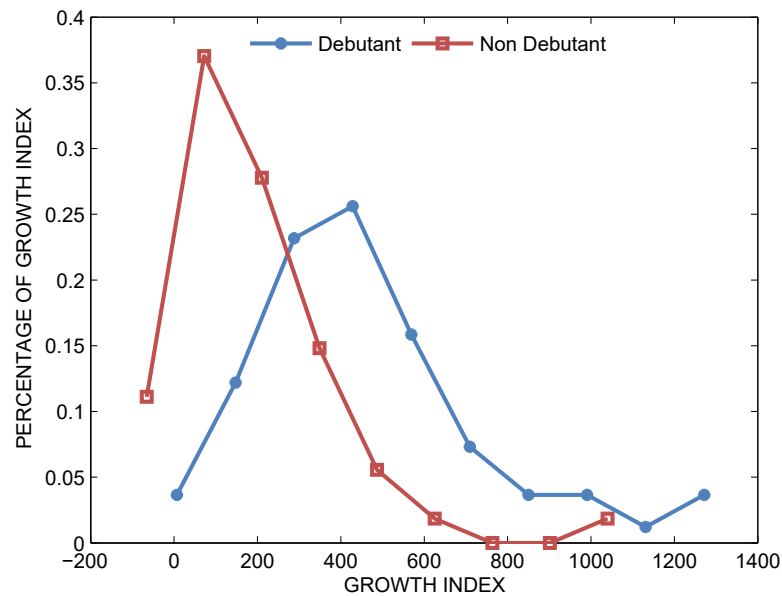
Figure 13 Distributions of growth indexes for both groups when filtering links associated with less than three publications.

between existing topics provides a strong indicator for predicting the emergence of new research areas. Even a simple threshold over the indexes introduced in this study allows us to discriminate well the subgraphs that will produce new research areas. For example, Table 9 reports the pace of collaboration obtained for both debutant and non-debutant topics in 2004. Here we can appreciate that a 0.41 threshold corresponding to a 100% precision is able to retrieve 8 out of 9 debutant topics. Table 10 displays other cases in which it is possible to obtain a very good recall when choosing a threshold corresponding to 100% precision. The application of this technique in a realistic setting would however require a scalable method for identifying promising topic graphs.

While these results are satisfactory, our analysis presents some limitations, which we shall address in future work.

In particular, we identified the relevant subgraph during the selection phase simply by selecting the *n* most co-occurrent topics of the topic under analysis. This solution allows us to compare graphs of the same dimension, however it introduces two issues.

First of all, it assumes that all topics derive from the same number of research areas, which is an obvious simplification. Emerging topics may have a different nature, based on their origin, development patterns, interactions of pioneer researchers, and so on. Therefore, each of them will be linked to a different number of established research areas. A manual analysis of the data suggests that using a constant number of co-occurring topics is one of the reasons why the overall pace of collaboration and growth index associated with the emergent topics are not much higher than the ones of the control group. When

Table 9 List of topics, both debutant and non-debutant with their pace of collaboration analysed in the 2004.

Testing topic	Pace of collaboration	Debutant/Control
Linked data	0.538	D
Bilinear pairing	0.499	D
Wimax	0.488	D
Separation logic	0.463	D
Phishing	0.446	D
Micro grid	0.433	D
Privacy preservation	0.426	D
Vehicular ad hoc networks	0.416	D
Mobile computing	0.409	C
Electromagnetic dispersion	0.401	C
Online learning	0.357	C
Wavelet analysis	0.326	C
Program interpreters	0.325	C
Zigbee	0.313	D
Natural sciences computing	0.308	C
Knowledge discovery	0.300	C
Fuzzy neural networks	0.298	C
Three term control systems	0.250	C

Table 10 Precision and Recall when choosing particular thresholds for distinguish the classes of topics.

Year	2001	2004	2006
Threshold	0.35	0.41	0.23
Recall	8/9	8/9	11/14
Precision	8/8	8/8	11/11

selecting too many co-occurring topics, we may include less significant research areas or, alternatively, research areas that started to collaborate with the topic in question only after its emergence. Conversely, when selecting too few topics, the resulting graph may exclude some important ones.

A second limitation is that the selection phase performed in our study cannot be reused in a system capable of automatically detecting embryonic topics, since it requires knowledge of the set of topics with which the embryonic topic will co-occur in the future. However, this could be fixed by developing techniques that are able to select promising subgraphs according to their collaboration pace and density. For this purpose we are currently developing an approach that generates a topic graph in which (i) links are weighted according to the acceleration in the pace of collaboration between the two relevant topics and (ii) community detection algorithms are applied to select portions of the network characterized by an intense collaboration between topics. We expect that this solution will be able to detect at a very early stage that 'something' new is emerging in a certain area of

the topic graph, even if it may not be able to accurately define the topic itself. It would thus allow relevant stakeholders to react very quickly to developments in the research landscape.

The findings of this analysis also provide contributions of potential value to research in Philosophy of Science. Firstly, they appear to support our hypothesis about the existence of an embryonic phase in the lifecycle of research topics. Secondly, they bring new empirical evidence to fundamental theories in Philosophy of Science, which are concerned with the evolution of scientific disciplines, e.g., *Herrera, Roberts & Gulbahce (2010)*, *Kuhn (2012)*, *Nowotny, Scott & Gibbons (2013)*, and *Sun et al. (2013)*. Finally, they highlight that new topics tend to be born in an environment in which previously less interconnected research areas start to cross-fertilise and generate new ideas. This suggests that interdisciplinarity is one of the most significant forces that drives innovation forward, allowing researchers to integrate a diversity of expertise and perspectives, and yield new solutions and new scientific visions. Hence the results of our analysis could be used to support policies that promote interdisciplinary research.

CONCLUSIONS

We hypothesised the existence of an embryonic phase for research topics, where, while they have not yet been consistently labelled or associated with a considerable number of publications, they can nonetheless be detected through an analysis of the dynamics of already existing topics. To confirm this hypothesis, we performed an experiment on 75 debutant topics in Computer Science, which led to the analysis of a topic network comprising about 2,000 topics extracted from a sample of three million papers in the 2000–2010 time interval. The results confirm that the creation of novel topics is anticipated by a significant increase in the pace of collaboration and density of the portions of the network in which they will appear. These findings provide supporting evidence for the existence of an embryonic phase for research topics and can be built on to foster further research to develop new techniques for the detection of topics at this stage. They also bring new empirical evidence to theories in Philosophy of Science. Finally, they suggest that an interdisciplinary environment provides a fertile ground for the creation of novel research topics.

We now plan to exploit the dynamics discussed in this study to create a fully automatic approach for detecting embryonic topics. We also intend to study and integrate additional dynamics involving other research entities, such as authors and venues. The aim is to produce a robust approach to be used by researchers and companies alike for gaining a better understanding of where research is heading.

Funding

PhD studentship for Angelo A. Salatino is funded by Springer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Springer.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Angelo A. Salatino conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Francesco Osborne conceived and designed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.
- Enrico Motta wrote the paper, reviewed drafts of the paper.

Data Availability

Open Science Framework: https://osf.io/bd8ex/.

REFERENCES

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

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

Blei DM, Lafferty JD. 2006. Correlated topic models. In: Weiss Y, Schölkopf PB, Platt JC, eds. *Advances in neural information processing systems*, vol. 18, 147–154.

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**:993–1022.

Bolelli L, Ertekin Ş, Giles CL. 2009. Topic and trend detection in text collections using latent dirichlet allocation. In: *Proceedings of the 31th European conference on IR research on advances in information retrieval, ECIR'09*. Berlin, Heidelberg: Springer-Verlag, 776–780.

Boyack KW, Klavans R, Börner K. 2005. Mapping the backbone of science. *Scientometrics* **64**:351–374 DOI 10.1007/s11192-005-0255-6.

Cano Basave AE, Osborne F, Salatino AA. 2016. Ontology forecasting in scientific literature: semantic concepts prediction based on innovation-adoption priors. In: *Knowledge engineering and knowledge management: 20th international conference, EKAW 2016, Bologna, Italy, Proceedings*. New York: Springer International Publishing, 51–67.

Cataldi M, Di Caro L, Schifanella C. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In: *Proceedings of the tenth international workshop on multimedia data mining*. ACM, p 4.

Chang J, Blei DM. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics* **4(1)**:124–150.

Chavalarias D, Cointet J-P. 2013. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLOS ONE* **8**:e54847 DOI 10.1371/journal.pone.0054847.

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

Davis JA, Leinhardt S. 1967. *The structure of positive interpersonal relations in small groups.* Washington, D.C.: Institute of Education Sciences.

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

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

Duvvuru A, Radhakrishnan S, More D, Kamarthi S, Sultornsanee S. 2013. Analyzing structural & temporal characteristics of keyword system in academic research articles. *Procedia Computer Science* **20**:439–445 DOI 10.1016/j.procs.2013.09.300.

Erten C, Harding PJ, Kobourov SG, Wampler K, 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 DOI 10.1016/j.socnet.2010.03.004.

Fusch PI, Ness LR. 2015. Are we there yet? Data saturation in qualitative research. *The Qualitative Report* **20(9)**:1408–1416.

Griffiths TL, Jordan MI, Tenenbaum JB, Blei DM. 2004. Hierarchical topic models and the nested Chinese restaurant process. In: Thrun S, Saul LK, Schölkopf PB, eds. *Advances in neural information processing systems*, vol. 16. Cambridge: MIT Press, 17–24.

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

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

Herrera M, Roberts DC, Gulbahce N. 2010. Mapping the evolution of scientific fields. *PLOS ONE* **5**:e10355 DOI 10.1371/journal.pone.0010355.

Holland PW, Leinhardt S. 1976. Local structure in social networks. *Sociological Methodology* **7**:1–45 DOI 10.2307/270703.

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

Kamaliha E, Riahi F, Qazvinian V, Adibi J. 2008. Characterizing network motifs to identify spam comments. In: *2008 IEEE international conference on data mining workshops.* Piscataway: IEEE, 919–928.

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

Larsen PO, 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 DOI 10.1007/s11192-010-0202-z.

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 DOI 10.1002/asi.20614.

Luce RD, Perry AD. 1949. A method of matrix analysis of group structure. *Psychometrika* **14**:95–116 DOI 10.1007/BF02289146.

Lv PH, Wang G-F, Wan Y, Liu J, Liu Q, Ma F-C. 2011. Bibliometric trend analysis on global graphene research. *Scientometrics* **88**:399–419 DOI 10.1007/s11192-011-0386-x.

Mathioudakis M, Koudas N. 2010. Twittermonitor: trend detection over the twitter stream. In: *Proceedings of the 2010 ACM SIGMOD international conference on management of data.* New York: ACM, 1155–1158.

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

Newman ME. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America* **98**:404–409 DOI 10.1073/pnas.98.2.404.

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

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

Oka M, Abe H, Kato K. 2006. Extracting topics from weblogs through frequency segments. In: *Proceedings of WWW 2006 annual workshop on the weblogging ecosystem: aggregation, analysis, and dynamics.*

Osborne F, Motta E. 2012. Mining semantic relations between research areas. In: Cudré-Mauroux P, Heflin J, Sirin E, Tudorache T, Euzenat J, Hauswirth M, Xavier Parreira J, Hendler J, Schreiber G, Bernstein A, Blomqvist E, eds. *The Semantic Web—ISWC 2012. ISWC 2012. Lecture notes in computer science*, vol. 7649. Berlin, Heidelberg: Springer.

Osborne F, Motta E. 2015. Klink-2: integrating multiple web sources to generate semantic topic networks. In: Arenas M, Corcho O, Simperl E, Strohmaier M, d'Aquin M, Srinivas K, Groth P, Dumontier M, Heflin J, Thirunarayan K, Staab S, eds. *The Semantic Web—ISWC 2015. Lecture notes in computer science*, vol. 9366. Cham: Springer.

Osborne F, Motta E, Mulholland P. 2013. Exploring scholarly data with rexplore. In: *The Semantic Web—ISWC 2013.* Berlin, Heidelberg: Springer.

Osborne F, Scavo G, Motta E. 2014. A hybrid semantic approach to building dynamic maps of research communities. In: Janowicz K, Schlobach S, Lambrix P, Hyvönen E, eds. *Knowledge engineering and knowledge management. EKAW 2014. Lecture notes in computer science*, vol. 8876. Berlin, Heidelberg: Springer.

Pham MC, Klamma R, Jarke M. 2011. Development of computer science disciplines: a social network analysis approach. *Social Network Analysis and Mining* **1**:321–340 DOI 10.1007/s13278-011-0024-x.

Pržulj N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**:e177–e183 DOI 10.1093/bioinformatics/btl301.

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

Salatino A. 2015. Early detection and forecasting of research trends. In: *ISWC-DC 2015 The ISWC 2015 doctoral consortium. Available at* http://ceur-ws.org/Vol-1491/paper_5.pdf.

Salatino AA, Motta E. 2016. Detection of embryonic research topics by analysing semantic topic networks. In: *2016 Workshop on semantics, analytics, visualisation: enhancing scholarly datae (SAVE-SD 2016)*. Cham: Springer.

Sun X, Ding K, Lin Y. 2016. Mapping the evolution of scientific fields based on cross-field authors. *Journal of Informetrics* **10**:750–761 DOI 10.1016/j.joi.2016.04.016.

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

Tseng Y-H, Lin Y-I, Lee Y-Y, Hung W-C, Lee C-H. 2009. A comparison of methods for detecting hot topics. *Scientometrics* **81**:73–90 DOI 10.1007/s11192-009-1885-x.

Ugander J, Backstrom L, Kleinberg J. 2013. Subgraph frequencies: mapping the empirical and extremal geography of large graph collections. In: *Proceedings of the 22nd international conference on World Wide Web: international World Wide Web conferences steering committee*. 1307–1318.

Wu Y, Venkatramanan S, Chiu DM. 2016. Research collaboration and topic trends in Computer Science based on top active authors. *PeerJ Computer Science* **2**:e41 DOI 10.7717/peerj-cs.41.