

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

[View the peer-reviewed version](https://peerj.com/articles/cs-119) (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>

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

3 Angelo Antonio Salatino, Francesco Osborne, Enrico Motta

4 Knowledge Media Institute, The Open University, Milton Keynes, UK

5

6 Corresponding Author: Angelo Antonio Salatino

7 Email address: [angelo.salatino@open.ac.uk](mailto:angelo.salatino@open.ac.uk)

## 8 **ABSTRACT**

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

28

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

31

## 32 **INTRODUCTION**

33 Being aware of the rise of new research topics can bring significant benefits for anybody  
34 involved in the research environment. Academic publishers and editors can exploit this  
35 knowledge and offer the most up to date and interesting contents. Researchers might be  
36 interested in new trends related to their topics and in promising new research areas.  
37 Institutional funding bodies and companies need to be regularly updated on how the research  
38 landscape is evolving in order to make early decisions about critical investments.  
39 Nonetheless, considering the growth rate of research publications (Larsen & Von Ins 2010),  
40 keeping up with novel trends is a challenge even for expert researchers and traditional  
41 methods, such as the manual exploration of publications in significant conferences and  
42 journals, are no longer viable. This has led to the emergence of several approaches capable of  
43 detecting novel topics and research trends (Bolelli et al. 2009; Duvvuru et al. 2012; He et al.  
44 2009; Wu et al. 2016). However, these approaches focus on topics that are associated with a  
45 substantial number of publications or on which the scientific community has reached a  
46 consensus for a specific label. This limitation hinders the ability of aforementioned  
47 stakeholders to react promptly to new developments in the research landscape.

48 We thus need novel methods for identifying the appearance of new topics at a very early  
49 stage, assessing their potential and forecasting their trend. To this end, we must achieve a  
50 better understanding of the dynamics underlying the creation of new topics and how these can  
51 be detected using current knowledge bases.

52 Philosophy of Science offers a number of interesting theories about the emergence of new  
53 topics. Kuhn (2012) theorised that science evolves through paradigm shifts. According to  
54 him, scientific work is performed within a set of paradigms and when these paradigms cannot  
55 cope with certain problems, there is a paradigm shift that can lead to the emergence of a new  
56 scientific discipline. This happens often through the creation of novel scientific  
57 collaborations. In this context, Becher & Trowler (2001) explained that, even if science  
58 proceeds toward more specific disciplines and thus researchers in different communities  
59 become less compatible, they are still incline to collaborate for mutual benefit. Herrera et al.  
60 (2010), Sun et al. (2013), Nowotny et al. (2013) suggested that the development of new  
61 topics is actually encouraged by the cross-fertilisation of established research areas and  
62 recognised that multidisciplinary approaches foster new developments and innovative  
63 thinking. Sun et al. (2013) and Osborne et al. (2014) provided empirical evidence to these  
64 theories by analysing the social dynamics of researchers and their effect on research  
65 communities and topics.

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

71 Inspired by previous theories, we hypothesize the existence of an even earlier phase, that we  
72 label *embryonic phase*, in which a topic has not yet been explicitly labelled or recognized by  
73 a research community, but exists as a fuzzy entity which entices a number of researchers  
74 from a variety of fields to converge and collaborate, with the aim of defining the mission and  
75 the paradigms of this potential research area.

76 We also hypothesize that it is possible to detect topics in this particular stage by analysing the  
77 *dynamics* of established topics. In this context, we define *dynamics* as all the significant  
78 trends regarding a topic or the interaction between topics or between entities linked to these  
79 topics, such as publications, authors, venues. For example, the sudden appearance of a  
80 number of publications concerning a combination of previously uncorrelated topics may  
81 suggest that some pioneer researchers are investigating new possibilities and maybe shaping a  
82 new emerging area. In the same way, as pointed out in Salatino (2015), we can hypothesize a  
83 wide array of relevant dynamics that could anticipate the creation of a new research area,  
84 such as a new collaboration between two or more research communities (see for example  
85 Osborne et al. (2014)), the creation of interdisciplinary workshops, a rise in the number of  
86 experts working on a certain combination of topics, a significant change in the vocabulary  
87 associated with relevant topics (Cano Basave et al. 2016), and so on.

88 This paper presents a study of some dynamics preceding the creation of novel topics which  
89 supports our hypothesis. In particular, we analysed the topic co-occurrence graphs and found  
90 that the emergence of novel research topics can be anticipated by a significant increase of the  
91 pace of collaboration and density in the co-occurrence graphs of related topics.

92 This study was performed in the 2000-2010 interval on a sample of three million  
93 publications. It was conducted by selecting sections of the co-occurrence graph where a new  
94 topic is about to emerge and analysing their interactions in the previous five years versus a  
95 control group of subgraphs associated to established topics. The analysis was performed with  
96 two different approaches that integrate statistics and semantics. It was found that the pace of  
97 collaboration and density measured in the sections of the network that will give rise to a new  
98 topic are significantly higher ( $p < 0.0001$ ) than the one in the control group. These findings  
99 support our hypothesis about the existence of an embryonic phase, yield new empirical  
100 evidences to the aforementioned theories and confirm the strong benefits of an  
101 interdisciplinary environment. In addition, the identified dynamics could be used to build new  
102 automatic methods for suggesting the emergence of a research topic in a certain conceptual  
103 area, which could be complementary to the current methods for detecting research topics.  
104 Indeed, while these new methods may be unable to generate an accurate specification of the  
105 topic, they should however be able to detect its emergence at an earlier stage, since they  
106 would not require a minimum amount of publications directly associated with the topic.

107 The study presented in this paper is an extension of the one published in (Salatino & Motta  
108 2016). The new contribution of this paper are: 1) a larger sample (75 debutant topics and 100  
109 established ones), 2) a new technique for measuring the density of the topic graphs, 3) a more  
110 exhaustive statistical analysis, including the comparison of the different approaches, 4) a  
111 revised state of the art, and 5) a more comprehensive discussion of the findings.

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

## 116 **RELATED WORK**

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

122 The state of the art presents several works on research trend detection, which can be  
123 characterised either by the way they define a topic or the techniques they use to detect them  
124 (Salatino 2015). Blei et al. (2003) developed the well-known Latent Dirichlet Allocation  
125 (LDA), an unsupervised learning method to extract topics from a corpus, which models  
126 topics as a multinomial distribution over words. Since its introduction, LDA has been  
127 extended and adapted in several applications. For example, Blei & Lafferty (2006) introduced

128 the Correlated Topic Model using the logistic normal distribution instead of the Dirichlet one,  
129 to address the issue that LDA fails to model correlations between topics. Griffiths &  
130 Tenenbaum (2004) developed the *hierarchical LDA* where topics are grouped together in a  
131 hierarchy. Further extensions incorporate other kinds of research metadata. For example,  
132 Rosen-Zvi et al. (2004) presented the Author-Topic Model (ATM) which includes authorship  
133 information and then associates each topic to a multinomial distribution over words and each  
134 author to a multinomial distribution over topics. Bolelli et al. (2009) introduced the  
135 Segmented Author-Topic model which further extends ATM by adding the temporal ordering  
136 of documents to address the problem of topic evolution. In addition, Chang & Blei (2010)  
137 developed the *relational topic model* which combines LDA and the network structure of  
138 documents to model topics. Similarly, He et al. (2009) combined LDA and citation networks  
139 in order to address the problem of topic evolution. Their approach detects topics in  
140 independent subsets of a corpus and then leverages citations to connect topics in different  
141 time frames. In a similar way, Morinaga & Yamanishi (2004) employed a probabilistic model  
142 called Finite Mixture Model to represent the structure of topics and analyse the changes in  
143 time of the extracted components to track emerging topics. However, this was evaluated on  
144 an email corpus, thus it is not clear how it would perform on scientific corpus. A general  
145 issue affecting this kind of approaches is that is not always easy to associate specific research  
146 areas to the resulting topic models.

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

158 Duvvuru et al. (2013) analysed the co-occurring network of keywords in a scholarly corpus  
159 and monitored the evolution in time of the link weights for detecting research trends and  
160 emerging research areas. However, as Osborne & Motta (2012) pointed out, keywords tend to  
161 be noisy and do not always represent research topics – in many cases different keywords even  
162 refer to the same topic. For example, Osborne et al. (2014) showed that the use of a semantic  
163 characterisation of research topics yields better results for the detection of research  
164 communities. To cope with this problem, some approaches rely on taxonomies of topics. For  
165 example, Decker et al. (2007) matched a corpus of research papers to a taxonomy of topics  
166 based on the most significant words found in titles and abstracts, and analysed the changes in  
167 the number of publications associated with topics. Similarly, Erten et al. (2004) adopted the  
168 ACM Digital Library taxonomy for analysing the evolution of topic graphs and monitoring  
169 research trends. However, human crafted taxonomy tend to evolve slowly and in a fast-

170 changing research field such as *Computer Science* (Pham et al. 2011) it is important to rely  
171 on constantly updated taxonomies. For this reason, in our experiment we adopted an ontology  
172 of Computer Science automatically generated and regularly updated by the Klink-2 algorithm  
173 developed by Osborne & Motta (2015).

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

## 179 **MATERIALS AND METHODS**

180 The aim of this study was to measure the association between the emergence of a new topic  
181 and the increase of pace of collaboration and density previously observed in the co-  
182 occurrence graphs of related topics. To this end, we represent topics and their relationships in  
183 a certain time frame as a graph in which nodes are topics whereas edges represent their co-  
184 occurrences in a sample of publications. This is a common representation for investigating  
185 topic dynamics (Boyack et al. 2005; Leydesdorff 2007; Newman 2001) and we will refer to it  
186 as *topic graph* or *topic network* in the following. To pursue our analysis, we analysed 75  
187 topics that debuted in the 2000-2010 period using 100 established topics as a control group.

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

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

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

### 196 **Semantic Enhanced Topic Network**

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

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

---

<sup>1</sup> <https://www.elsevier.com/solutions/scopus>

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

210 A second approach, used by a number of digital libraries and publishers is tagging  
211 publications with categories from a pre-determined taxonomy of topic. Some examples  
212 include the ACM computing classification system<sup>2</sup>, the Springer Nature classification<sup>3</sup>,  
213 Scopus subject areas<sup>4</sup>, and the Microsoft Academic Search classification<sup>5</sup>. This solution has  
214 the advantage of producing sound topics, agreed upon by a committee of experts. However,  
215 these taxonomies suffer from some common issues. First, building a large taxonomy requires  
216 a large number of experts; it is an expensive and lengthy process. Hence, they are seldom  
217 updated and grow obsolete very quickly. For example, the 2012 version of the ACM  
218 classification was finalized fourteen years after the previous version. In addition, these  
219 taxonomies are very coarse-grained and usually contain general fields rather than fine-  
220 grained research topics.

221 We address these issues by characterizing our topics according to the Klink-2 ontology of  
222 Computer Science, which describes the relationships between more than 15,000 research  
223 areas. Klink-2 is an algorithm which is able to generate very granular ontologies and update  
224 them regularly by analysing keywords and their relationships with research papers, authors,  
225 venues, and organizations and by taking advantage of multiple knowledge sources available  
226 on the web. Klink-2 is currently integrated in the Rexplore system (Osborne et al. 2013), a  
227 modern tool for exploring and making sense of scholarly data, which provides semantic-  
228 aware analytics. Klink-2 was run on a set of 35,983 keywords from a corpus of 16 million  
229 publications in the field of Computer Science, producing an ontology that contains 15,961  
230 terms, after filtering out 20,022 keywords that did not represent topics, were unrelated to any  
231 other topic in the taxonomy or were associated with a low number of publications (Osborne  
232 & Motta 2015).

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

239 We used the resulting semantic topics to build sixteen topic networks representing the topic  
240 co-occurrences in the 1995-2010 timeframe. Each network is a fully weighted graph  $G_{year} =$   
241  $(V_{year}, E_{year})$ , in which V is the set of topics while E is the set of links representing the topic

---

<sup>2</sup> <http://www.acm.org/publications/class-2012>

<sup>3</sup> <http://www.nature.com/subjects>

<sup>4</sup> <https://www.elsevier.com/solutions/scopus/content>

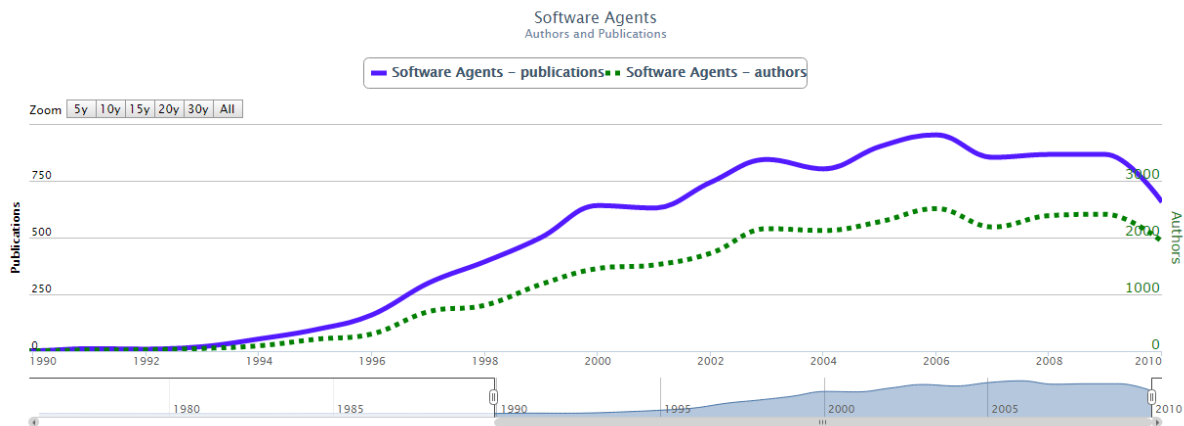
<sup>5</sup> <http://academic.research.microsoft.com/>



242 co-occurrences. The node weight represents the number of publications in which the topic  
243 appears in a certain year, while the link weight is equal to the number of publications in  
244 which two topics co-occur together in the same year.

## 245 Graph Selection

246 We randomly selected 75 topics that debuted in the period between 2000 and 2010 as  
247 treatment group (also referred as debutant group). A topic debuts in the year in which its label  
248 first appears in a research paper. The control group (also referred as non-debutant group), was  
249 obtained by selecting 100 well-established topics. We considered a topic as well-established if:  
250 i) it debuted before 2000, ii) it appears in the Klink Ontology, iii) it is associated each year  
251 with a substantial and consistent number of publications. As an example, Figure 1 shows the  
252 evolution through time of the well-established topic *Software Agents* in terms of number of  
253 active authors and papers published about it. The figure shows that the topic made its debut in  
254 1993 and in the year 2000 reached a rate of over 500 publications per year and more than  
255 1500 authors working on it. We can thus consider it established in the context of our study.



256

257 Figure 1. Evolution of the topic Software Agents in terms of number of authors and number of publications per year. The  
258 chart has been produced by the Rexplore system.

259 We assume that a new topic will continue to collaborate with the topics that contributed to its  
260 creation for a certain time after its debut. This assumption was discussed and tested in  
261 previous work (Osborne & Motta 2012) where it was used for finding historical subsumption  
262 links between research areas. Hence, as summarized by Figure 2, for each debuting topic we  
263 extracted the portion of topic network containing its  $n$  most co-occurring topics from the year  
264 of debut until nowadays and analysed their activity in the five years preceding its year of  
265 debut. Since we want to analyse how the dimension of these subgraphs could influence the  
266 results, we tested different values of  $n$  (20, 40, and 60). For example, if a topic  $A$  made its  
267 debut in 2003, the portion of network containing its most co-occurring topics will be analysed  
268 in the 1998-2002 timeframe. We repeated the same procedure on the topics in the control  
269 group, assigning them a random year of analysis within the decade 2000-2010. In the  
270 previous study (Salatino & Motta 2016), we selected 50 established topics and we assigned a  
271 random *year of analysis* to each of them. For this study, we randomly assigned each  
272 established topic to two consecutive years within the decade 2000-2010, with the

273 consequence of doubling the control group and thus reducing the noise and smoothing the  
 274 resulting measures.

275 In brief, the selection phase associates to each topic from the treatment and the control groups  
 276 (also referred as *testing topics* or *topics under test*) a graph  $G^{topic}$  :

$$277 \quad G^{topic} = G_{year-5}^{topic} \cup G_{year-4}^{topic} \cup G_{year-3}^{topic} \cup G_{year-2}^{topic} \cup G_{year-1}^{topic} \quad (1)$$

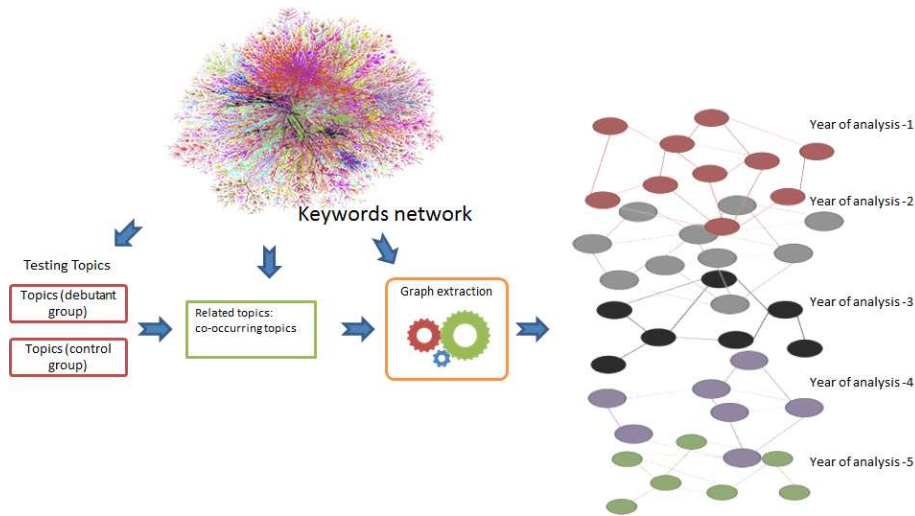
278 which corresponds to the collaboration network of a debutant topic in the five years prior to  
 279 its emergence (or year of analysis for non-debutant topics). In particular, each year  
 280 corresponds to the sub-graphs  $G_{year-i}^{topic}$  :

$$281 \quad G_{year-i}^{topic} = (V_{year-i}^{topic}, E_{year-i}^{topic}) \quad (2)$$

282 in which  $V_{year-i}^{topic}$  is the set of most co-occurring topics in a particular year and  $E_{year-i}^{topic}$  is the set of  
 283 edges that link nodes in the set  $V_{year-i}^{topic}$ .

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

286



287

288

Figure 2: Workflow representing all the steps for the selection phase.

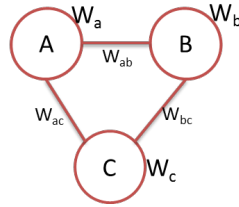
289

## 290 **Graph Analysis**

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

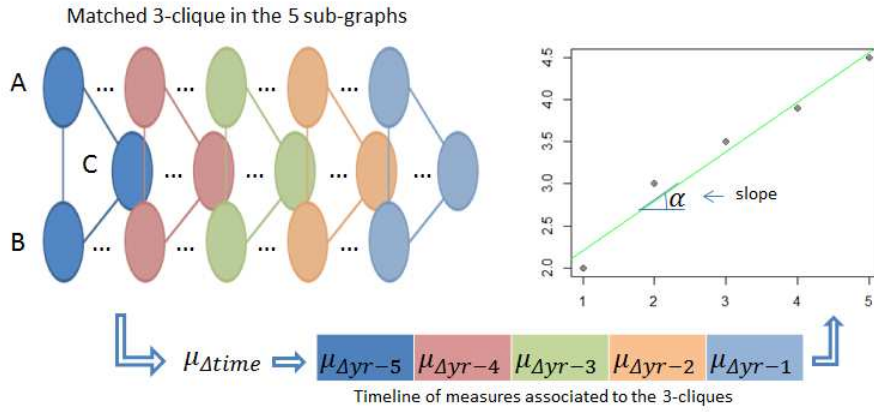
297 Cliques-based method

298 This approach is based on the intuition that we can measure the collaboration pace of a graph  
 299 by analysing the diachronic activity of triangles of collaborating topics. To this end, we first  
 300 extracted all 3-cliques from the five sub-graphs associated to each topic under analysis. A 3-  
 301 clique, as shown in Figure 3, is a complete sub-graph of order three in which all nodes are  
 302 connected to one another and is employed for modelling small groups of entities close to each  
 303 other (Luce & Perry 1949).



304  
305

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



306  
307

Figure 4. Main steps of the analysis phase.

308 To study the dynamics preceding the debut of each topic, we analysed the evolution of the  
 309 same 3-clique in subsequent years. Figure 4 summarizes the process. Considering a 3-clique  
 310 having nodes  $\{A, B, C\}$ , we quantify its collaboration index  $\mu_{\Delta}$  in a certain year by taking into  
 311 account both node weights  $\{W_a, W_b, W_c\}$  and link weights  $\{W_{ab}, W_{bc}, W_{ca}\}$ .

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

313 The index  $\mu_{\Delta}$  is computed by mean of Equation 3. The strength of collaboration  
 314  $\mu_{x-y}$  between two nodes of the topic network,  $x$  and  $y$ , is computed as the harmonic mean of  
 315 the conditional probabilities  $P(y|x)$  and  $P(x|y)$ , where  $P(x|y)$  is the probability that a  
 316 publication associated with a topic  $x$  will be associated also with a topic  $y$  in a certain year.  
 317 The advantage of using conditional probabilities over the number of co-occurrences is that

318 the resulting value  $\mu_{x-y}$  is already normalised according to the number of publications  
319 associated to each topic. Finally,  $\mu_{\Delta}$  is computed as the harmonic mean of the strength of  
320 collaboration of the three links of a 3-clique. This solution was adopted after testing  
321 alternative approaches during the *preliminary evaluation*, as will be discussed in the Findings  
322 section.

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

$$327 \quad \mu_{\Delta\text{time}}^{\text{clique-}i} = [\mu_{\Delta\text{yr}-5}, \mu_{\Delta\text{yr}-4}, \mu_{\Delta\text{yr}-3}, \mu_{\Delta\text{yr}-2}, \mu_{\Delta\text{yr}-1}] \quad (4)$$

328 Initially, we tried to determine the trend of a clique by simply taking the difference between  
329 the first and last values of the timeline ( $\mu_{\Delta\text{yr}-5} - \mu_{\Delta\text{yr}-1}$ ). However, this method ignores the  
330 other values in the timeline and can thus neglect important information. For this reason, we  
331 applied instead the linear interpolation method on the five measures using the least-squares  
332 approximation to determine the linear regression of the time series  $f(x) = a \cdot x + b$ . The  
333 slope  $a$  is then used to assess the increase of collaboration in a clique. When  $a$  is positive the  
334 degree of collaboration between the topics in the clique is increasing over time, while when is  
335 negative the number and intensity of collaborations are decreasing.

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

338 To summarize, for each testing topic we selected a subgraph of related topics in the five years  
339 preceding the year of debut (or *analysis* for topics in the control group). We then extracted  
340 the 3-cliques and associated each of them with a vector representing the evolution of their  
341 pace of collaboration. The trend of each clique was computed as the angular coefficient of the  
342 linear regression of these values. Finally, the increase in the pace of collaboration of a  
343 subgraph was obtained by averaging these values.

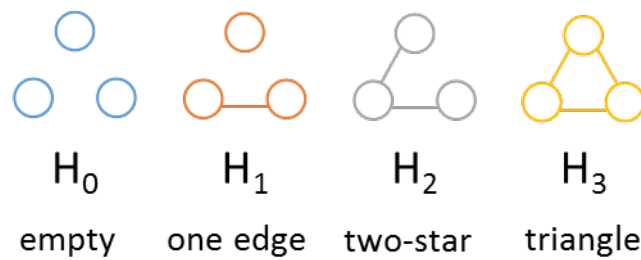
344

#### 345 Triad-based method

346 The triad-based method employs the triad census (Davis & Leinhardt 1967) to measure the  
347 change of topology and the increasing density of the subgraphs during the five year period.  
348 The triad census of an undirected graph, also referred as global 3-profiles, is a four  
349 dimensional vector representing the frequencies of the four isomorphism classes of triad, as  
350 shown in Figure 5.

351 The triad census summarises the structural information in networks and is useful to analyse  
352 structural properties in social networks. It has been applied to several scenarios, such as  
353 identifying spam (Kamaliha et al. 2008; O'Callaghan et al. 2012), comparing networks (Pržulj  
354 2007), analysing social networks (Faust 2010; Ugander et al. 2013) and so on.

355



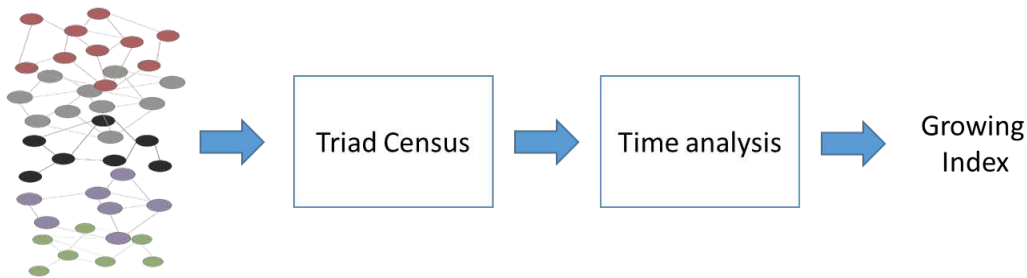
356

357 Figure 5. The four isomorphism classes of triad. The triad census consists in counting the frequencies of  $H_i$  of the input  
 358 graph.

359 In this study, we used triad census to describe all the sub-graphs  $G_{year-i}^{topic}$  associated to a  
 360 particular testing topic in terms of frequencies of  $H_i$  (see Figure 5) and then evaluate how the  
 361 frequencies of *empties* ( $H_0$ ), *one edges* ( $H_1$ ), *two-stars* ( $H_2$ ) and *triangles* ( $H_3$ ) changed in  
 362 time. Figure 5 illustrates the four classes of triads for an undirected graph as in the case of  
 363 topic network. Naturally an increase of the numbers of triangles suggests the appearance of a  
 364 number of new collaborations clusters between previous distant topics.

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

369 Figure 6 shows the workflow for analysing the evolution of topology of networks related to a  
 370 testing topic, in the five years preceding its debut.



371

372 Figure 6. Main step of the analysis phase for the triad census approach.

373 We first performed the triad census over the five graphs associated to each testing topic. For  
 374 example, Table 1 shows the results of the triad census over the five sub-graphs associated to  
 375 the debutant topic *Artificial Bee Colonies*.

376 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	1064
$G_{year-3}^{topic}$	125	486	967	1698

$G_{year-2}^{topic}$	100	410	908	1858
$G_{year-1}^{topic}$	68	486	849	2251

377

378 We then measured whether the collaboration graph was becoming denser by analysing the  
 379 change of frequencies associated with  $H_i$  (see Figure 7). To do so, we computed the  
 380 percentage growth of each  $H_i$  using Equation 5.

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

381

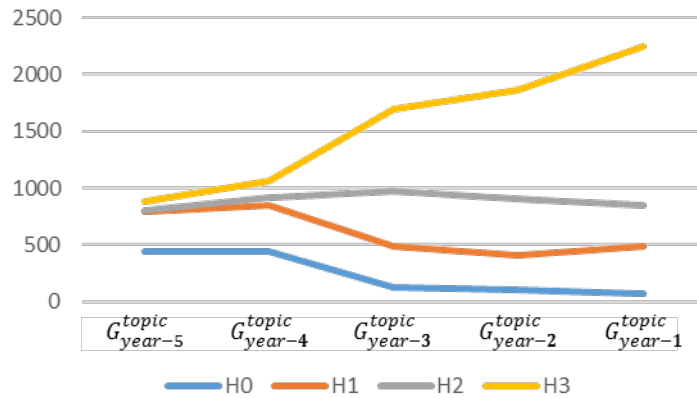
382 Then, we used Equation 6, which performs a weighted summation of all the contributions of  
 383 percentage of growth.

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

384

385 The *growth index* takes into account all the isomorphism classes, even if the number of  
 386 triangles ( $H_3$ ) can by itself be a fair indicator of the density. Indeed, previous studies by Faust  
 387 (2010) and Holland & Leinhardt (1976) showed that all four classes of triads are useful for  
 388 computing useful properties of the network, including transitivity, intransitivity and density.  
 389 In our case, taking in consideration only  $H_3$  might fail to detect some subtler cases,  
 390 characterized for example by a significant increase of  $H_2$  and decrease of  $H_1$  and  $H_0$ .

391



392

393 Figure 7: Development in time of the frequencies of  $H_i$  in the network related to the emergence of "Artificial Bee Colonies".

394 To summarize, the triad-based method received the same input of the clique-based method.  
 395 For each of these five subgraphs associated to a topic, we performed the triad census  
 396 obtaining the different frequencies  $H_i$  in different years. We then analysed them  
 397 diachronically to quantify the increase in density.

398

## 399 RESULTS

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

- 402 • The preliminary evaluation performed on a reduced dataset for assessing the metrics  
403 used in the Cliques-based method;
  - 404 • The full study using the Cliques-based method;
  - 405 • The full study using the Triads-based method.
- 406

### 407 Preliminary evaluation with alternative cliques-based methods

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

- 416 • **AM-N**, which uses the arithmetic mean and the difference between first and last  
417 value;
- 418 • **AM-CF**, which uses the arithmetic mean and the linear interpolation;
- 419 • **HM-N**, which uses the harmonic mean and the difference between first and last value;
- 420 • **HM-CF**, which uses the harmonic mean and the linear interpolation.

421 Figure 8 illustrates the average pace of collaboration for the sub-graphs associated to each  
422 topics according to these methods (thick horizontal black lines) and the range of their values  
423 (thin vertical line). The results support the initial hypothesis: according to all methods, the  
424 pace of collaboration of the cliques within the portion of network associated with the  
425 emergence of new topics is positive and higher than the ones of the control group.  
426 Interestingly, the pace of collaboration of the control group is also slightly positive. Further  
427 analysis revealed that this behaviour is probably caused by the fact that the topic network  
428 becomes denser and noisier in time. Figure 9 confirms this intuition illustrating the fast  
429 growth of the number of publications per year in the dataset during the time window 1970-  
430 2013.

431  
432  
433

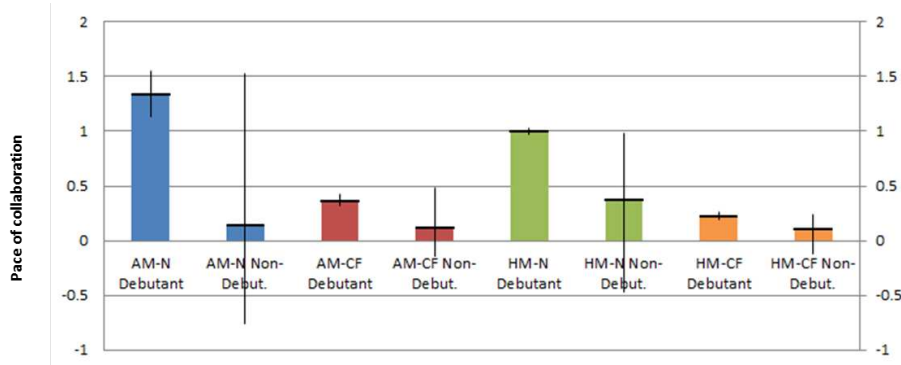


Figure 8. Overall directions of the sub-graphs related to testing topics in both debutant and control group with all the four approaches.

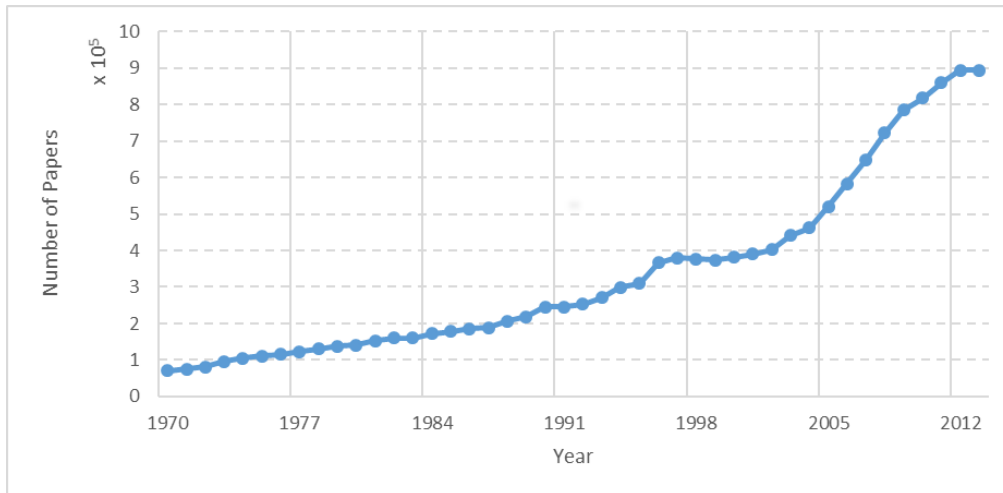
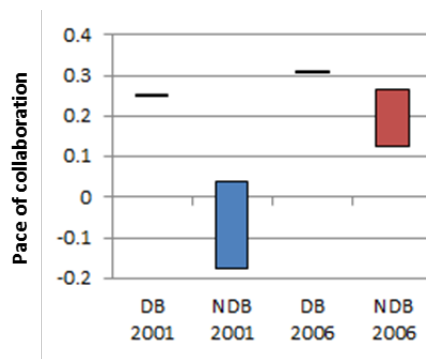


Figure 9. Number of papers each year in period 1970-2013

434  
435

436 The approaches based on the simple difference (AM-N and HM-N) exhibit the larger gaps  
437 between the two groups in terms of the average pace of collaboration. However, the ranges of  
438 values actually overlap, making it harder to assess if a certain sub-group is incubating a novel  
439 topic. The same applies to AM-CF. HM-CF performs better and even if the values slightly  
440 overlap when averaging the pace over different years they do not when considering single  
441 years. Indeed, analysing the two ranges separately in 2001 and 2006 (see Figure 10), we can  
442 see that the overall collaboration paces of the debutant topics (DB) are always significantly  
443 higher than the control group (NDB).



444

445  
446

Figure 10. Overall directions of the sub-graphs related to testing topics in both debutant and control group in HM-CF approach



447 We ran the Student's t-test on the sample of data provided by the HM-CF approach, to verify  
 448 whether the two groups belong to different populations. The test yielded  $p < 0.0001$ , which  
 449 allowed us to reject the null hypothesis that the differences between the two distributions  
 450 were due to random variations<sup>6</sup>. On the basis of this result, we could further confirm that the  
 451 HM-CF approach performs better compared to the other approaches. For this reason, we  
 452 selected the combination of the *harmonic mean* and the *linear interpolation* as the approach  
 453 for the full study using the clique-based method.

454 The results of HM-CF show also interesting insights on the creation of some well-known  
 455 research topics. Table 2 and Table 3 list the cliques which exhibited a steeper slope for  
 456 semantic web and cloud computing. We can see that *Semantic Web* was anticipated in the  
 457 1996-2001 timeframe by a significant increase in the collaborations of the *World Wide Web*  
 458 area with topics such as *Information Retrieval*, *Artificial Intelligence*, and *Knowledge Based*  
 459 *Systems*. This is consistent with the initial vision of the semantic web, defined in the 2001 by  
 460 the seminal work of Tim Berners-Lee (Berners-Lee et al. 2001). Similarly, *Cloud Computing*  
 461 was anticipated by an increase in the collaboration between topics such as *Grid Computing*,  
 462 *Web Services*, *Distributed Computer Systems* and *Internet*. This suggests that our approach  
 463 can be used both for forecasting the emergence of new topics in distinct subsections of the  
 464 topic network and for identifying the topics that gave rise to a research area.

465 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

466

467 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

468

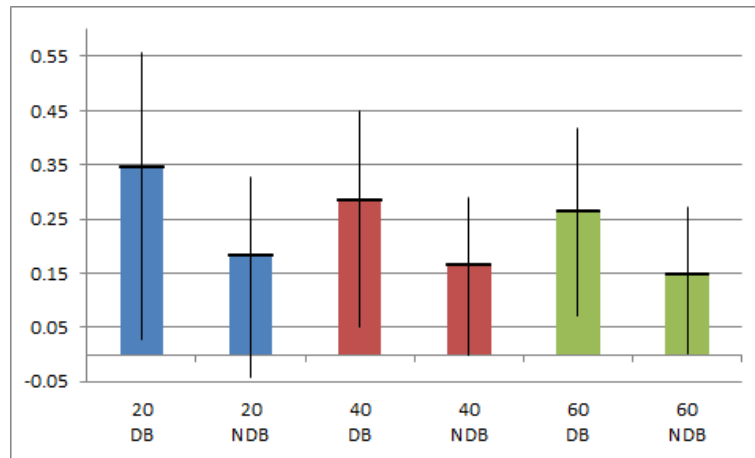
### 469 Cliques-based method study

470 We applied the cliques-based methods on the subgraphs associated to both topics in the  
 471 treatment and control groups. Figure 11 reports the results obtained by using subgraphs  
 472 composed by the most 20, 40 and 60 co-occurring topics. Each bar shows the mean value of  
 473 the average pace of collaboration for the debutant (DB) and non-debutant (NDB) topics. As

---

<sup>6</sup>  $p < 0.0001$  is the conventional statistical representation to indicate an extremely high statistical significance ( $> 500$  times stronger than the conventional 0.05 threshold for claiming significance). It includes all mathematical outcomes from 0 to below 0.0001, which are essentially equivalent in assessing the excellent significance.

474 before, the average pace computed in the portion of topic network related to debutant topics  
 475 is higher than the one of the control group.

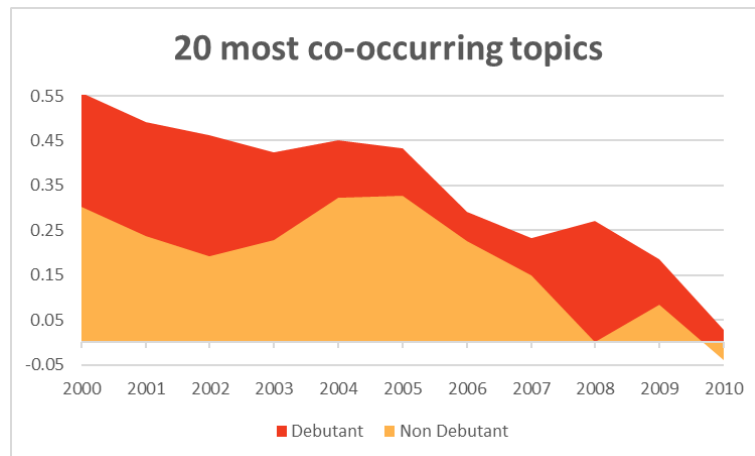


476

477 Figure 11. Average collaboration pace of the sub-graphs associated to the treatment (DB) and control group (NDB), when  
 478 selecting the 20, 40 and 60 most co-occurring topics. The thin vertical lines represent the ranges of values.

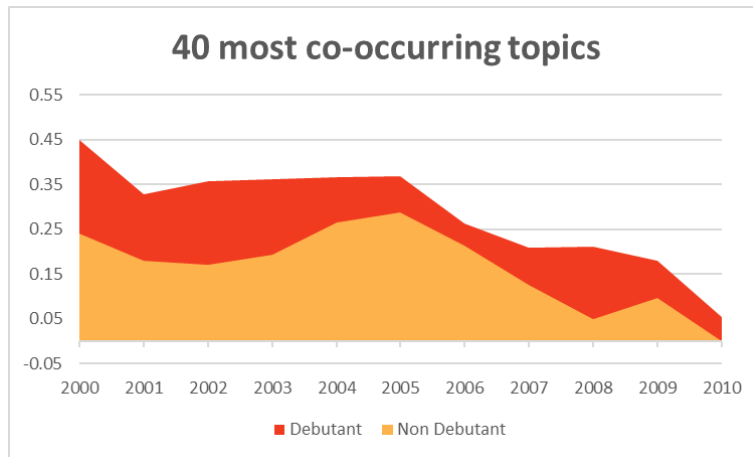
479 Since the pace of collaboration changes significantly according to the period considered, it is  
 480 useful to study it across different years. Figure 12, Figure 13 and Figure 14, show the average  
 481 collaboration pace for each year when considering the 20, 40 and 60 most co-occurring  
 482 topics. In all cases the collaboration pace for the debutant topics is higher than the one for the  
 483 control group. We can also notice that in the last five years the overall pace of collaboration  
 484 for both debutant and non-debutant topics suffered a significant fall. This is due to the fact  
 485 that the topic network became denser and noisier in recent years.

486



487

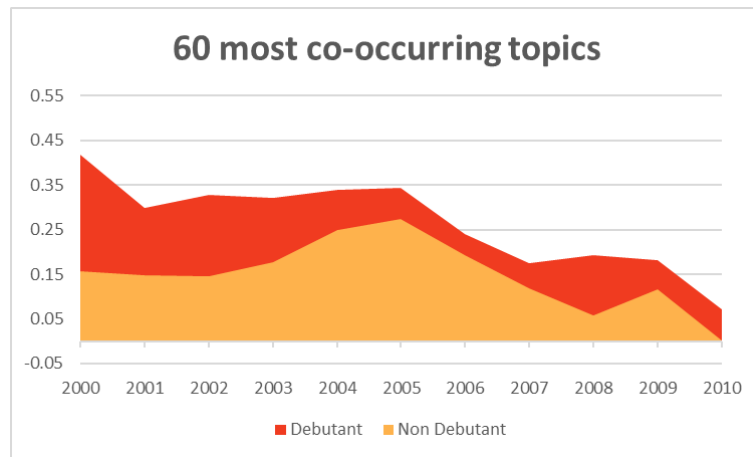
488 Figure 12. Average collaboration pace per year of the sub-graphs related to testing topics in both debutant and control group  
 489 considering their 20 most co-occurring topics. The year refers to the year of analysis of each topic.



490

491  
492

Figure 13. 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.



493

494  
495

Figure 14. 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.

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

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

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

509

510  
511

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

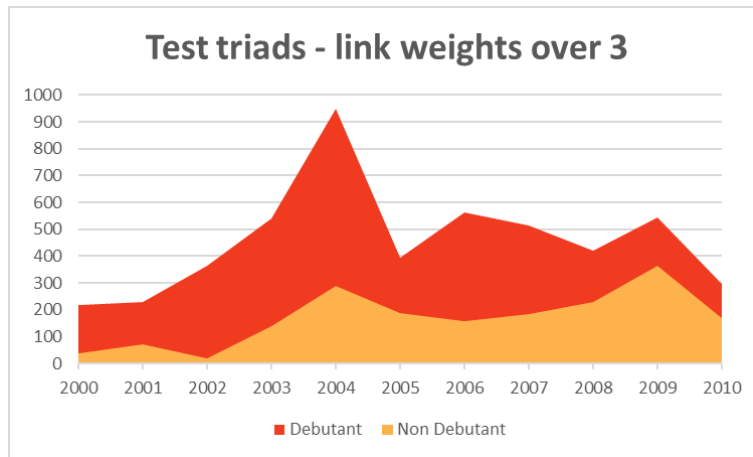
512

### 513 **Triads-based method study**

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

518 Figure 15 reports the average value of the growing indexes when discarding links with less  
519 than 3 co-occurrences. The approach allows to discriminate well the portion of networks  
520 related to debutant topics from the ones related to the control group and the collaboration  
521 pace associated with the debutant topics is always higher than its counterpart. Figure 16 and  
522 Figure 17 report the results obtained by removing links with less than 10 and 20 co-  
523 occurrences. The gap between the groups in these two last experiments is reduced in  
524 comparison with the first experiment. This suggest that considering weak connections is more  
525 beneficial for discriminating the two groups. Nonetheless, the indexes associated with  
526 debutant topics are always higher than the ones associated to non-debutant ones. The 2004  
527 peak is caused by the debut of number of topics associated with particularly strong  
528 underlying dynamics, such as *Linked Data*, *Pairing-based Cryptography*, *Microgrid* and  
529 *Privacy Preservation*.

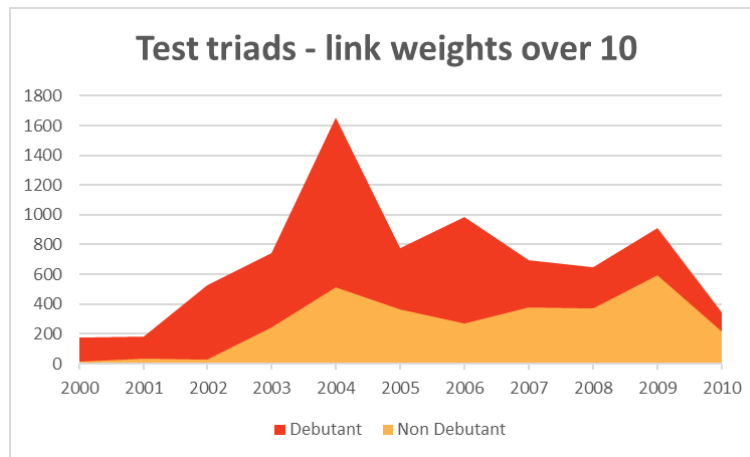
530



531

532  
533

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 3 publications.

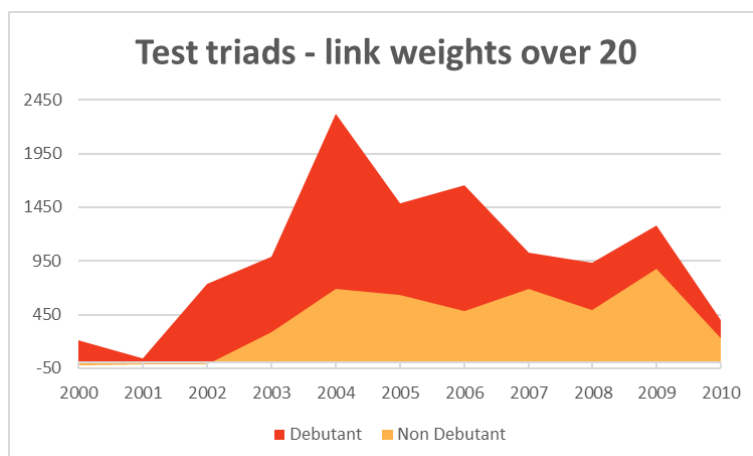


534

535  
536

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 10 publications.

537



538

539  
540

Figure 17. 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.

541

542

Table 5 reports as an example the triad census performed over the subgraph associated to the topic *Semantic Web Technologies* (SWT) debuting in the 2001. We can see an increase in the

543 number of triangles ( $H_3$ ) and two-stars ( $H_2$ ), mirroring the increasing density of the topic  
 544 network. Again, this phenomenon is more evident when using also weak links ( $< 3$ ). The  
 545 percentage of growth of full triangles is 109% in the first test and then it decreases to 86% ( $<$   
 546 10) and 36 % ( $< 20$ ).

547

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

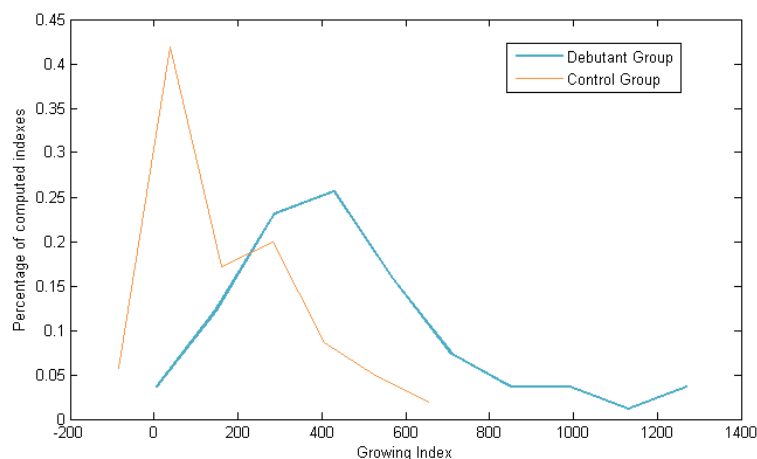
<i>Graph</i>	Removing links $< 3$				Removing links $< 10$			
	$H_0$	$H_2$	$H_2$	$H_3$	$H_0$	$H_2$	$H_2$	$H_3$
1996	1124	1157	658	337	641	676	316	138
1997	928	1237	670	441	1022	828	315	135
1998	1255	1353	657	389	585	705	300	181
1999	1307	1431	861	461	1222	1098	413	192
2000	913	1399	1043	705	1482	1361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

<i>Graph</i>	Removing links $< 20$			
	$H_0$	$H_2$	$H_2$	$H_3$
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

550

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



554

555 Figure 18. Distributions of growing indexes for both groups when filtering links associated with less than 3 publications.

556 As before, we ran Student’s t-test over the two distributions of growing indexes, for all the  
 557 three experiments. It yielded  $p < 0.0001$  for all the experiments. Figure 18 shows as an  
 558 example the distribution obtained in the first test.

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

562 Table 6. Growing indexes of sub-graphs associated to selected debutant topics versus the average growing index of the  
 563 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

## 564 DISCUSSION

565 In this study, we analysed the topic network with the aim of confirming the hypothesis that  
 566 the emergence of new research areas is anticipated by the interaction of already existing  
 567 topics. We examined the pace of collaboration (via the cliques-based method) and the change  
 568 in topology (via the triads-based method) in portions of network related to debutant topics,  
 569 showing that is possible to effectively discriminate areas of the topic graph associated to the  
 570 future emergence of new topics. In particular, the first experiments showed that the subgraphs  
 571 associated with the emergence of a new topic exhibit a significant higher pace of  
 572 collaboration than the control group of subgraphs associated with established topics ( $p$   
 573  $<0.0001$ ). Similarly, the second experiment showed that the graphs associated with a new  
 574 topic display a significant higher increase in their density than the control group ( $p$   
 575  $<0.0001$ ). We can thus confirm that these two dynamics can play a key role when performing the  
 576 detection of embryonic topics.

577 However, the ability of these two approaches in discriminating the debutant graph from the  
 578 control group varies according to the period. Looking at the best results, reported in Figure 14  
 579 and Figure 15, it appears that the cliques-based approach works better (according to the  
 580 resulting p-values) in the first years of the decade (2000-2004) while the triads-based  
 581 approach performs in the last years (2005-2010). This may indicate that the second approach

582 works better when the topic network is nosier and denser, as it is in the second period. In this  
 583 sense, the two approaches are complementary and the choice of the best one will depend on  
 584 the characteristics of the topic graph under analysis. We plan to study other dynamics,  
 585 regarding authors, venues, citations and so on, with the aim of further understanding the  
 586 patterns that precede the emergence of research topic.

587 The results of this study should allow us to develop new methods for detecting the  
 588 aforementioned dynamics in specific sections of the topics graph and suggesting that a new  
 589 research area may emerge from a combination of other topics. Indeed, even simply using a  
 590 threshold over the indexes introduced in this study allows us to discriminate effectively the  
 591 subgraphs in which a new topic will shortly emerge from the ones of the control group. For  
 592 example, Table 7 reports the pace of collaboration obtained for both debutant and non-  
 593 debutant topics in 2004. If we consider a threshold of 0.41, our approach is able to select 8  
 594 out of 9 debutant topics, obtaining 89% precision and 100% recall. It should also be noted  
 595 that since both the pace of collaboration and the density are time-dependent the threshold  
 596 should also be set accordingly. Similarly, Table 8 shows precision and recall obtained by  
 597 using a threshold over the collaboration pace in the years 2001, 2004 and 2006. In future  
 598 work we plan to adopt more sophisticated statistical methods for detecting these topic graphs.

599

600

Table 7. List of topics, both debutant and non-debutant with their pace of collaboration analysed in the 2004.

<b>Testing topic</b>	<b>Pace of collaboration</b>	<b>Debutant/Control</b>
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

601

602



Table 8. Precision and Recall when choosing particular thresholds for distinguish the classes of topics

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

604

605 While these results are satisfactory, our analysis presents some limitations that we shall  
606 address in future work. In particular, we identified the relevant subgraph during the selection  
607 phase by simply selecting the  $n$  most co-concurrent topics of the topic under analysis. This  
608 solution allows us to compare graphs of the same dimension but presents two issues. In the  
609 first instance, it assumes all topics will derive from the same number of research areas, which  
610 is an obvious simplification. Indeed, emerging topics may have different natures, based on  
611 their origin, development patterns through time, interactions of pioneer researchers, and so  
612 on. Therefore, each of them will actually be linked to a different number of established  
613 research areas. A manual analysis on the data suggests that using a constant number of co-  
614 occurring topics is one of the reasons why the overall pace of collaboration and growth index  
615 associated to the emergent topics are not much higher than the ones of the control group.  
616 When selecting too many co-occurring topics, we may include some less significant research  
617 areas or some research area that started to collaborate with the topic only after its emergence.  
618 Conversely, when selecting too few topics, the resulting graph may exclude some important  
619 ones.

620 A second limitation is that the selection phase performed in our study could not be directly  
621 reused in a system to automatically detect embryonic topics, since it requires knowledge of  
622 the set of topics with which the embryonic topic will co-occur in the future. However, this  
623 could be fixed by developing techniques to select promising subgraphs according to their  
624 collaboration pace and density. Indeed, we are currently developing an approach to do so that  
625 first generates a topic graph in which the links are weighted according to the collaboration  
626 pace and then exploits community detection techniques for selecting candidate sub-graphs to  
627 further analyse the dynamics discussed in this paper. This solution should be able to detect at  
628 a very early stage that ‘something’ new is emerging in a certain area of the topic graph, even  
629 if it may not be able to accurately define the topic itself. It would thus allow relevant  
630 stakeholders to react very quickly to novelties in the research landscape.

631 The findings of this study are also relevant to a number of research communities. Firstly, they  
632 appear to support our hypothesis about the existence of an embryonic phase in the lifecycle of  
633 research topics. Moreover, they bring new empirical evidences to related theories in  
634 philosophy of science, such as (Herrera et al. (2010)), Kuhn (2012), Nowotny et al. (2013),  
635 and Sun et al. (2013). Finally, they highlight that new topics actually tend to be born in an  
636 environment in which previously less interconnected research areas start cross-fertilising and  
637 generating original ideas. This suggests that interdisciplinarity is one of the most significant  
638 forces that push research forward, allowing to integrate a diversity of expertise and  
639 perspectives to come up with new solutions and new visions. The results of our analysis may  
640 thus support relevant research policies.

## 641 CONCLUSIONS

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

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

## 660 REFERENCES

- 661 Becher T, and Trowler P. 2001. *Academic tribes and territories: Intellectual enquiry and the culture of*  
662 *disciplines*: McGraw-Hill Education (UK).
- 663 Berners-Lee T, Hendler J, and Lassila O. 2001. The semantic web. *Scientific american* 284:28-37.
- 664 Blei D, and Lafferty J. 2006. Correlated topic models. *Advances in neural information processing*  
665 *systems* 18:147.
- 666 Blei DM, Ng AY, and Jordan MI. 2003. Latent dirichlet allocation. *J Mach Learn Res* 3:993-1022.
- 667 Bolelli L, Ertekin Ş, and Giles CL. 2009. Topic and trend detection in text collections using latent  
668 dirichlet allocation. *Advances in Information Retrieval*: Springer, 776-780.
- 669 Boyack KW, Klavans R, and Börner K. 2005. Mapping the backbone of science. *Scientometrics* 64:351-  
670 374.
- 671 Cano Basave A, Osborne F, and Salatino A. 2016. Ontology Forecasting in Scientific Literature:  
672 Semantic Concepts Prediction based on Innovation-Adoption Priors.
- 673 Cataldi M, Di Caro L, and Schifanella C. 2010. Emerging topic detection on twitter based on temporal  
674 and social terms evaluation. Proceedings of the Tenth International Workshop on  
675 Multimedia Data Mining: ACM. p 4.
- 676 Chang J, and Blei DM. 2010. Hierarchical relational models for document networks. *The Annals of*  
677 *Applied Statistics*:124-150.
- 678 Chavalarias D, and Cointet J-P. 2013. Phylomemetic patterns in science evolution—the rise and fall  
679 of scientific fields. *PloS one* 8:e54847.
- 680 Couvalis G. 1997. *The philosophy of science: science and objectivity*: Sage.
- 681 Davis JA, and Leinhardt S. 1967. The structure of positive interpersonal relations in small groups.  
682 *ERIC*.
- 683 Decker SL, Aleman-Meza B, Cameron D, and Arpinar IB. 2007. Detection of bursty and emerging  
684 trends towards identification of researchers at the early stage of trends. University of  
685 Georgia.

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

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

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

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

696 Fusch PI, and Ness LR. 2015. Are we there yet? Data saturation in qualitative research. *The*  
697 *Qualitative Report* 20:1408.

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

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

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

705 Herrera M, Roberts DC, and Gulbahce N. 2010. Mapping the evolution of scientific fields. *PloS one*  
706 5:e10355.

707 Holland PW, and Leinhardt S. 1976. Local structure in social networks. *Sociological methodology* 7:1-  
708 45.

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

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

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

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

717 Leydesdorff L. 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific  
718 journals. *Journal of the American Society for Information Science and Technology* 58:1303-  
719 1319.

720 Luce RD, and Perry AD. 1949. A method of matrix analysis of group structure. *Psychometrika* 14:95-  
721 116.

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

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

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

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

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

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

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

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

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

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

745 Osborne F, Scavo G, and Motta E. 2014. A hybrid semantic approach to building dynamic maps of  
746 research communities. *Knowledge Engineering and Knowledge Management*: Springer, 356-  
747 372.

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

750 Pržulj N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics*  
751 23:e177-e183.

752 Rosen-Zvi M, Griffiths T, Steyvers M, and Smyth P. 2004. The author-topic model for authors and  
753 documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*:487-  
754 494.

755 Salatino A. 2015. Early Detection and Forecasting of Research Trends. *ISWC-DC 2015 The ISWC 2015*  
756 *Doctoral Consortium*:49.

757 Salatino AA, and Motta E. 2016. Detection of Embryonic Research Topics by Analysing Semantic  
758 Topic Networks. *2016 Workshop on Semantics, Analytics, Visualisation: Enhancing Scholarly*  
759 *Data (SAVE-SD 2016)*.

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

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

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

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

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

772