Topic-Driven Detection and Analysis of Scholarly Data

Alfio Ferrara, Corinna Ghirelli, Stefano Montanelli, Eugenio Petrovich, Silvia Salini, and Stefano Verzillo

Abstract The chapter presents a topic mining approach that can used for a scholarly data analysis. The idea here is that research topics can emerge through an analysis of epistemological aspects of scholar publications that are extracted from conventional publication metadata, such as the title, the author-assigned keywords, and the abstract. As a first contribution, we provide a conceptual analysis of research topic profiling according to the peculiar behaviours/trends of a given topic along a considered time interval. As a further contribution, we define a disciplined approach and the related techniques for topic mining based on the use of publication metadata and natural language processing (NLP) tools. The approach can be employed within

A. Ferrara · S. Montanelli (✉)
Department of Computer Science, Data Science Research Center, Università degli Studi di Milano, Milan, Italy
e-mail: alfio.ferrara@unimi.it; stefano.montanelli@unimi.it

C. Ghirelli
Banco de Espana, DG Economics, Statistics and Research, Madrid, Spain
e-mail: corinna.ghirelli@bde.es

E. Petrovich
Department of Economics and Statistics, Università degli Studi di Siena, Siena, Italy
e-mail: eugenio.petrovich@unisi.it

S. Salini
Department of Economics, Management and Quantitative Methods, Data Science Research Center, Università degli Studi di Milano, Milan, Italy
e-mail: silvia.salini@unimi.it

S. Verzillo
European Commission, Joint Research Centre (JRC), Ispra, Italy
e-mail: stefano.verzillo@ec.europa.eu

D. Checchi et al. (eds.), *Teaching, Research and Academic Careers*,
https://doi.org/10.1007/978-3-031-07438-7_8

a variety of topic analysis issues, such as country-oriented and/or field-oriented research analysis tasks that are based on scholarly publications. In this direction, to assess the applicability of the proposed techniques for use in a real scenario, a case study analysis based on two publication datasets (one national and one worldwide) is presented.

Keywords Natural Language Processing · Scholarly Data Analysis · Topic Mining

1 Introduction

In contemporary science policy debates, one of the most heated discussions concerns the role and effects of research performance metrics in research assessment frameworks. According to the advocates of using these metrics, indicators based on citations and publications would be more objective than the traditional peer review system, hence allowing for breaking 'old boys circles' and hampering nepotism, cronyism, and other inappropriate academic practices (Geuna & Martin, 2003). Moreover, by setting measurable thresholds and benchmarks, performance metrics would stimulate both the quantity and quality of scientific production (Bonaccorsi, 2015; Geuna & Martin, 2003; Moed, 2017). Finally, research evaluation based on metrics would be less expensive than peer review, it would save taxpayers' money (Geuna & Piolatto, 2016), and as recent evidence shows, it may provide comparable results, at least when the need is to assess research performance at the institutional level (Checchi et al., 2021). In addition, at the individual level, the predictive power of bibliometrics is superior to peer review in almost all disciplines (except medicine). On the other hand, critics insist on the 'unintended consequences' of using metrics and on the 'constitutive effects' that their pervasive presence has on the behaviour of researchers (Dahler-Larsen, 2014). These effects include goal displacement (scoring high on the metrics becoming a target in and of itself), promotion of the unethical use of citations (excessive self-citation, creation of citation cartels, the strategic exchange of citations, etc.), task reduction (academic activities that are not considered in the calculation of the indicators, such as teaching and public engagement, being avoided), and an artificial increase in productivity by 'salami slicing' (dividing one scientific work into multiple publications) (Fochler et al., 2016; de Rijcke et al., 2015). Even the recent rise in retractions and research misconduct (e.g. fabrication of results, plagiarism, 'p-hacking', etc.) has been linked to the increasing pressure of metrics (Biagioli et al., 2019).

One of the most interesting criticisms raised against metrics in research evaluations is that they would not only affect the behaviour of researchers, but also the epistemic content of the science they produce (i.e. ideas, research themes, methods, etc.). In particular, the excessive weight of metrics would damage the pluralism of scientific enquiry, rewarding the mainstream approaches not because of their scientific merit, but only because of their (transient) popularity or connection

with academic power. For instance, in a recent joint declaration, the Académie des Sciences, Leopoldina, and the Royal Society (Académie des Sciences et al., 2017) write that 'undue emphasis on bibliometric indicators [...] may also hinder the appreciation of the work of excellent scientists outside the mainstream; it will also tend to promote those who follow current or fashionable research trends, rather than those whose work is highly novel and which might produce completely new directions of scientific research' (p. 2). Metrics have been blamed for inducing risk avoidance in science: the researchers, under the pressure of scoring well on indicators, would focus on topics, research programmes, and methods that are more likely to be rewarded. By contrast, they would avoid revolutionary ideas, out-of-the-box innovation, and interdisciplinarity because these would be deemed to be too risky enterprises. Thus, orthodoxy and conformism would be promoted at the expense of critical thinking, damaging scientific progress.

Until recently, the presence and magnitude of the effects of metrics on scientists and science have been debated more often than empirically investigated. In recent years, the empirical study of the effects of research evaluations on research practices has begun. An increasing body of literature has started documenting how researchers react under competitive conditions, which may affect their likelihood of promotion, particularly how research evaluation frameworks based on 'metrics' have induced a change in the publishing behaviour of researchers. In Italy, for instance, recent evidence shows that the introduction of research evaluation procedures has promoted strategic behaviours among researchers via the creation of 'citation clubs' that are aimed at artificially inflating bibliometric outcomes (Baccini et al., 2019; Scarpa et al., 2018; Seeber et al., 2019).

However, studying the impact of metrics and the evaluation on the epistemic content of research, that is, on the theories and ideas that are produced by the scientific community under the regime of metrics-based evaluation, is still in its infancy (Muller & de Rijcke, 2017). In particular, the accuracy of *the mainstream criticism* outlined above is still to be addressed by empirical studies.

In this chapter, we propose a mining approach for the detection and analysis of 'mainstream topics'. The proposed idea is that topics featuring mainstream research can emerge through an analysis of the epistemological aspects of scholarly publications extracted from conventional publication metadata, such as the title, the author-assigned keywords, and the abstract. As a first contribution, we provide a conceptual analysis of the notion of mainstream research that is exploited to enforce mainstream profiling based on peculiar behaviours/trends of research topics along a considered time interval. As a further contribution, we define a disciplined approach and the related techniques for topic mining based on the use of publication metadata and natural language processing (NLP) tools. Finally, a case study analysis is presented to assess i) the applicability of the proposed techniques to a real scenario based on a publication dataset of Italian scholars and ii) the scalability and reliability of some of the case study results when the proposed approach is based on a richer and comprehensive database of all international publications, as collected by Scopus Elsevier over 14 years.

The chapter is organised as follows: In Sect. 2, studies on the epistemic impacts of metrics and techniques for automatic topic extraction from scholarly publications are briefly presented. In Sect. 3, a conceptual analysis of mainstream and what it comprises is provided by highlighting the different and sometimes opposite meanings that are attached to this concept in the literature. In Sect. 4, we present our modelling considerations about mainstream profiling. The proposed approach and techniques to topic mining are illustrated in Sect. 5. In Sect. 6, the results obtained by applying the proposed techniques to both a real publication dataset of Italian scholars and to the whole Scopus publication set on the same disciplines are discussed. Finally, our concluding remarks are given in Sect. 7.

2 Literature Review

By the *epistemic impacts* of metrics, we are focusing on the array of changes induced by metrics-based evaluation regimes on the epistemic processes of knowledge production and their outputs (scientific ideas, theories, research programmes, etc.) (Muller & de Rijcke, 2017). Epistemic impacts should be analytically distinguished from the effects of metrics on the social structure of science and, specifically, on its reward system, even if, in concrete situations, both kinds of impacts are likely to occur together. The decline of interdisciplinary research and reduction of scientific pluralism are examples of the epistemic impacts of metrics. By contrast, the rise of self-citations and gift authorships are examples of changes in the reward system of science.

The reward system-related effects are relatively easier to capture using quantitative methods because they can be inferred from analysing publications and citations. Starting from the pivotal study of Butler (2003) on the effects of the Australian research evaluation system, most scientometric studies so far have focused on these quantitative indicators to investigate the changes in researcher behaviour under the pressure of metrics (Abramo et al., 2019; Abramo et al., 2021; Baccini et al., 2019). Epistemic impacts, on the other hand, are more difficult to track for three main reasons. First, epistemic concepts, such as interdisciplinarity, scientific pluralism, and scientific mainstream, do not have standard, uncontested definitions. Second, the quantitative operationalisations of these notions frequently run the risk of reducing complex phenomena to monodimensional measures that miss important epistemological nuances. Third, there is no consensus on what epistemic factors contribute the most to scientific progress. For instance, philosophers of science have long debated what degree and what kind of scientific pluralism is beneficial to scientific enquiry (see (Viola, 2018) for a detailed discussion of the literature on this topic). The epistemic deviations induced by metrics are difficult to point out because there is no universally accepted baseline normative epistemology that accounts for the correct functioning of science.

In light of these methodological and theoretical impasses, most of the research on the epistemic impacts of metrics so far has turned to methodologies, such as

surveys and interviews, showing how researchers themselves perceive the pressure of metrics on their epistemic practices. In one of the first studies of this kind, Muller and de Rijcke (2017) interviews 38 Dutch and Austrian post-docs and junior group leaders in the life sciences, finding that researchers pervasively 'think with indicators'. Indicators such as the journal impact factor do not intervene only in the evaluation of research after the fact, but also inform the entire research process, from the very conception of research projects to the choice of scientific collaborators and even the animal models used. Castellani et al. (2016) reach a similar conclusion after giving out a questionnaire to 12 Italian scientists from several disciplines. Their interviewees underline the risk that metrics in research evaluation can promote uniformity in the scientific community and discourage ground-breaking approaches. Also, the interviewees argue that metrics worsen the 'publish or perish' culture and induce scientists to publish low-quality material just to score better on productivity indicators. Feenstra and Lopez-Cozar (2021) interview 14 Spanish researchers in philosophy and ethics about the effects of metrics in their disciplines. Even though the interviewed researchers identify some positive effects, such as more transparent policies in the academic promotion process, they deem the impact on research agendas, publication language, and mental health as negative. In particular, metrics would hamper intellectual diversity in philosophical research and even lead to research misconduct. These studies highlight how metrics and indicators have gained a prominent place in the 'epistemic living space' of researchers, both in the natural and social sciences (Felt, 2009).

One limitation of these studies, however, is that they do not discuss the epistemic concepts used by the researchers to frame their experience of metrics. In this sense, then, they offer a valuable but partial perspective on epistemic impacts. By contrast, the present study is the first attempt to ground an investigation of an epistemic phenomenon, that is, the scientific mainstream, in a conceptual analysis of the related epistemic concept.

Further related work focuses on the methods and techniques for the classification of scholarly publications. Usually, a combination of automated procedures and manual activities/practices has been proposed (Glenisson et al., 2005). Solutions based on the use of human-assigned metadata, such as superimposed subject categories of articles and journals, represent a popular solution (Borner, 2010). This approach is effective when the choice of subject categories is shared by the final users and the classification results provide a scholarly picture in which the actors (i.e. the publication authors) can self-recognise the categorisation of their scientific products. However, manually defined subject categories are characterised by several well-known weaknesses. For instance, predefined categories are typically inadequate for dealing with publications about emerging topics characterised by recent formation and a new epistemic body (Suominen & Toivanen, 2016). Machine learning and unsupervised classification/clustering approaches have recently been proposed for overcoming such limitations. For instance, in Boyack et al. (2011) and Talley et al. (2011), topic modelling and clustering solutions are exploited to provide a visual, graph-based representation of a publication dataset extracted from the MEDLINE repository and the National Institutes of Health (NIH), respectively.

Similar approaches have been investigated (Nichols, 2014; Yan et al., 2012) for the information retrieval field and for the National Science Foundation awards, respectively. On the other hand, the construction of a map of science merely derived from scholarly data by using automated classification algorithms is characterised by possible limitations, as well. For instance, automated solutions are generally weak in capturing the minor trends within a discipline, even if they provide a relevant contribution from the historical and epistemic point of view. A recent comparison between unsupervised learning and human-assigned approaches to classification of scholarly data has been provided (Suominen & Toivanen, 2016); in the study, a topic modelling solution based on the latent Dirichlet allocation (LDA) algorithm is exploited. The results show that it is difficult to argue the superiority of one method (human-based scholarly data classification) over the other (algorithm-based scholarly data classification) (Suominen & Toivanen, 2016). However, it is well recognised that machine-generated scholarly data classifications provide a strong contribution in terms of practicality (Castano et al., 2018). This means that the capability to rapidly generate thematic, interactive views of an underlying (large) scholarly publication dataset can be considered as a result, but it also represents a worth support/contribution for experts that aim to further refine/revise the obtained results to provide their own data views.

3 Conceptual Analysis of the Mainstream Notion

Etymologically, the term ‘mainstream’ refers to the main current of a river or a stream. According to the dictionary, the mainstream is the ‘prevailing current of thought, influence or activity’. As an adjective, ‘mainstream’ means ‘representing the prevalent attitudes, values, and practices of a society or a group[1]’. The term usually belongs to the context of artistic and cultural phenomena, where it is mainly used to denote trends in popular and media culture.[2] Sometimes, it takes a pejorative sense by subcultures who view the mainstream culture as artistically inferior.

When it is employed in a discussion about science, ‘mainstream’ preserves its nature as a common language term. However, a precise and widely accepted definition of what ‘mainstream’ means in reference to science is missing, as is an operational definition of how to measure it.

The term can be used as a noun (‘the mainstream in economics’), as well as an adjective (‘mainstream science’). In both cases, mainstream is said of many different aspects of the scientific enquiry, from the most abstract to the most practical. Mainstream can be the following:

[1] American Heritage Dictionary of the English Language, Fifth Edition (2011).

[2] For an overview, see https://en.wikipedia.org/wiki/Mainstream

- A general theoretical framework or research programme (e.g. the neoclassical approach in economics)
- A specific theory (e.g. the big bang theory in cosmology)
- A position in a debate (e.g. the functionalism in the philosophy of mind)
- A research object or topic (e.g. the quark bottom in high energy physics)
- A research methodology (e.g. participant observation in cultural anthropology)
- A technique, a research protocol, or a procedure (e.g. the PCR in molecular biology).

In the empirical study of what is mainstream, such variability must be considered to set an appropriate level for the analysis. Different empirical methods will capture the mainstream at different levels of 'granularity', depending on the scientific aspect being considered. However, the most important feature about the term and its usage is that it assumes *different and sometimes opposite meanings* in the literature. 'Mainstream' is used to reference not only different things, but also different and sometimes incompatible ways. By surveying the literature, we can analytically distinguish six key meanings, whose differences can be appreciated better when they are compared with their opposites.[3]

1. **'Orthodox' vs. 'heterodox' or 'fringe'.** In this sense, the mainstream is the dominant school of thought within a certain discipline or field. The mainstream is characterised by adherence to certain scientific content (e.g., a theory, a research programme, a method, etc.). The nonmainstream schools, on the other hand, are characterised by their refusal of some of the mainstream's tenets. In this sense, the term is used both positively and negatively. In the positive sense, the mainstream represents the standard view of the scientific community, whereas heterodox schools represent the margins of science, dangerously bordering pseudo-science ('fringe science') (Gottfredson, 1997). By contrast, when the term is used negatively, the closed-mindedness or refusal of the 'pluralism' of the mainstream is stressed (Colander et al., 2004). An instance of this meaning can be found in economics, where heterodox schools (e.g., Marxists, post-Keynesians, feminists, Old Institutionalists, and Austrians) are distinguished from the neoclassical mainstream. In Anglo-American philosophy, analytic philosophy may be considered the mainstream, whereas Continental philosophy can be thought of as the heterodox approach (Katzav & Vaesen, 2017).
2. **'Normal' vs. 'Revolutionary', 'Ground-breaking'.** The second meaning of mainstream refers to the distinction between normal and revolutionary science (Kuhn, 1996). Mainstream science would be characterised by a step-by-step, cumulative nature, whereas nonmainstream science would be more revolutionary, ground-breaking, and frequently not understood by the mainstream because of this. Mavericks and misunderstood geniuses would be the typical makers of

[3] Note that the six meanings rarely appear in their pure form. Often, scholars and commentators mix two or more meanings together. The six meanings should be considered as ideal types for the analysis, not as simple descriptions of usage.

nonmainstream science (Heinze, 2013). Compared with the first meaning, the focus is on the mode of scientific progress rather than on the adherence to some specific theory.

3. **'Popular' vs. 'Niche'.** The third meaning is neutral with respect to the scientific content of mainstream and nonmainstream science. Here, mainstream is purely characterised in quantitative terms as the research that is currently done by most of the researchers in a field. No judgement on the orthodoxy or progress of the mainstream is implied. Nonmainstream science is not considered as heterodox or ground-breaking, just as topics that are addressed by fewer researchers. This meaning of mainstream can be compared with the concept of 'impact' in a citation analysis. When the number of citations received by a publication is not considered a proxy of its scientific quality (a normative concept), it can nonetheless be considered a measure of popularity or influence (a descriptive concept).
4. **'Trendy', 'Short-lived', 'Passing' vs. 'Stable', 'Long-Lasting'.** In this sense, the focus is on the temporal extension of mainstream science. The mainstream is equated with the currently 'hot' areas of research. However, the short life of these areas is also implied. From this point of view, a mainstream researcher is a researcher who follows the trends, doing what everybody does at that moment. Mainstream topics are the ones that are currently fashionable in the research community, for whatever reason. These mainstream topics have the highest chance of being published in the highest-ranked journals and produce a high impact in terms of citations. In the literature on the perverse effects of research metrics, mainstream is frequently intended in this sense (e.g. de Rijcke et al., 2015).
5. **'Supported by (academic) power' vs. 'Underground'.** This meaning focuses on the socioacademic dimension of the mainstream, stressing the connection between mainstream and power. Mainstream science is what is defended by academic elites in prestigious universities and supported by economic and industrial powers. By contrast, nonmainstream science is seen as resistant. The underground approaches are not published in mainstream journals and are unlikely to receive funding through normal channels, even though they might receive funding from alternative sources. This meaning echoes the distinction between the underground or independent labels and the 'majors' in the music industry.[4] In economics, it is not unusual to describe the difference between mainstream and heterodox schools by pointing out not only the theoretical divergences, but also the different relationships that they entertain with economic powers (Cedrini & Fontana, 2018; Colander et al., 2004).
6. **'Core', 'Western' vs. 'Periphery', 'Non-Western'.** This meaning of mainstream is eccentric compared with the others and is found only in the bibliometric literature on the scientific production of developing countries (e.g. Gasparyan et al., 2017). In this literature, mainstream is used as a synonym of Western, and

[4] https://en.wikipedia.org/wiki/Music_industry

Mainstream meaning	Opposite	Focus on	Example
1 Orthodox	Heterodox	Intellectual contents	[Gottfredson, 1997]
2 Normal	Revolutionary	Mode of scientific progress	[Heinze, 2013]
3 Popular	Niche	Quantity	-
4 Trendy, short-lived, passing	Stable, long-lasting	Temporal extension	[de Rijcke, 2015]
5 Supported by (academic) power	Underground	Power	[Colander, 2004]
6 Core	Periphery	Geopolitics	[Gasparyan, 2017]

Fig. 1 The six different meanings of 'mainstream' in reference to science. The notion of mainstream has no universal meaning in discussions about science and science policy. Specifically, six different meanings can be analytically distinguished in the literature, noting that sometimes two or more meanings are intended at the same time. In the first column of the table, the key terms that capture each meaning are presented, along with their opposites (second column) that contribute to specifying their semantic content. Each meaning stresses a different dimension of the notion of mainstream, focusing on various aspects of the scientific activity. In the last column of the table, examples of studies that employ each meaning of the notion are provided

mainstream science is the scientific research either produced by highly developed countries or published in international outlets. By contrast, nonmainstream science is the science produced in developing countries and published in local journals.

Note that the six meanings, even if they are closely related, should not be considered synonyms. In fact, they are not mutually implied. For instance, a molecular biologist can deal with a niche topic (nonmainstream according to meaning 3) by applying a standard experimental method (mainstream according to 1). As a further example, an astrophysicist can investigate a 'trendy' celestial object (mainstream according to meaning 4) but in the context of a heterodox cosmological model (nonmainstream according to meaning 1). In Fig. 1, a summary of the six meanings, their opposites and the aspect of *mainstreamness* they highlight are provided.

4 Modelling Mainstreams

Previous approaches to the mainstream definition have led to different operational definitions of it.

The meanings 1 and 5 ('orthodox' and 'supported by power') require considerable expert knowledge of the scientific fields to assess whether a publication belongs

to the mainstream. To empirically investigate the mainstream that is intended in this sense implies gathering the opinion of several experts, with evident limitations in the number of publications that can be considered. Meaning 6 ('core') is easier to treat with quantitative methods because the geographical information can be retrieved automatically from the publications' metadata. However, this meaning of mainstream is less interesting from the point of view of the debate on research metrics.

Hence, we remain with meanings 2, 3, and 4, that is, 'normal science', 'popular', and 'trendy'. With relative ease, meanings 3 and 4 can be translated into quantitative measures. Popularity can be measured by the number of publications addressing a topic, whereas the trendiness of a topic can be measured by its temporal extension. Meaning 3 is particularly interesting because it refers to the epistemological concepts of normal versus revolutionary science advanced by Kuhn. Some observations by Kuhn and Lakatos can help us translate (partially) these notions into measures. According to Kuhn, during the normal science period, a paradigm is 'articulated' by the researchers, that is, it is expanded in different directions. Lakatos calls these paradigm articulations 'progressive research programmes' (1978). A progressive research programme can be recognised by its capacity to produce new research lines, that is, by its fruitfulness (Ivani, 2019). Thus, meaning 2 can be measured as a factor of productivity or the fruitfulness of a topic.[5]

The proposed approach to mainstream detection integrates the following three meanings of the term: *popularity* (meaning 3), *trendiness* (meaning 4), and *fruitfulness* (meaning 2). They constitute the three dimensions of what is mainstream that will be considered in our study. Based on them, several profiles of mainstream can be outlined (Fig. 2):

1. **Spot**. This profile corresponds to a short-lived topic characterised by a short burst of attention from the research community that is focused in a limited interval of time. This profile mostly relies on meaning 4 ('trendy').
2. **Persistent**. This profile of mainstream is based on meaning 3, popularity. It describes a topic that enjoys stable attention from the research community but that has low productivity in terms of new research lines.
3. **Impasse**. This profile describes the behaviour of a research programme that progressively decreases in importance until it becomes marginally important.
4. **Boosting**. This mainstream profile corresponds to a fruitful research programme of the normal scientific phase, hence relying mostly on meaning 2. It is characterised by a long life with a high number of descendant topics.

[5] Clearly, both Kuhn's and Lakatos' theories of scientific change are far more complex and richer than the sketchy picture offered in this report. In fact, we do not aim to offer a full operationalisation of these theories. Our limited goal is to draw on some epistemological topics to better design our methodology.

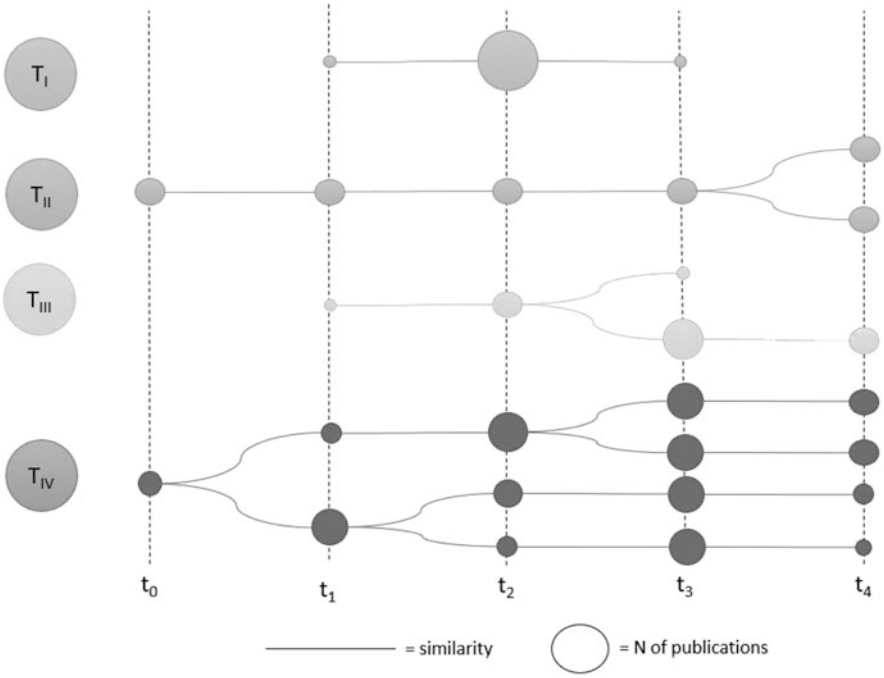

Fig. 2 Mainstream profiles. By combining the three meanings or aspects of the notion of mainstream that can be quantified (i.e. popularity, trendiness,, and fruitfulness), it is possible to delineate the various temporal profiles of a mainstream topic, that is, the various modes in which a mainstream topic may develop over time. The figure shows four of these modes. From the top to the bottom of the figure, they are as follows: spot topic (a short-lived topic that attracts a burst of attention in the research community), persistent topic (a topic that enjoys stable attention in the community but does not produce new research lines), impasse topic (a topic that branches in research lines, some of which decay), and boosting topic (a topic characterised by high fruitfulness that produces several new research lines). In the figure, the relation of filiation within a topic is represented by lines, whereas the size of the research lines (quantified in terms of publications) that form a topic is represented by circles

In different ways, each of these ideal profiles of mainstream integrates the three core aspects of meanings 2, 3, and 4. Our method aims at individuating instances of such profiles into the scientific production of our case studies.

5 Semiautomatic Topic Detection

Consider a dataset of scholarly publications $P = \{p_1, p_2, \ldots, p_n\}$. For topic detection in P, we propose the approach shown in Fig. 3 based on a pipeline characterised by *dataset acquisition*, *keyword extraction*, *keyword graph construction*, *topic discovery*, *topic filtering*, and *topic analysis*. In the following, we first present

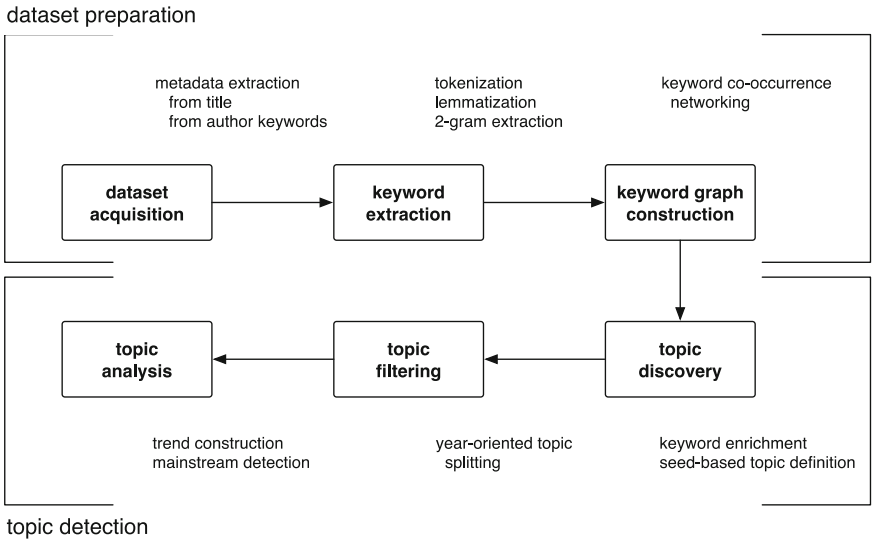

Fig. 3 The proposed mining approach to topic detection. The proposed topic mining approach is based on a pipeline where the initial publication dataset with related metadata is first submitted to a *keyword extraction* stage aimed at extracting relevant tokens. The tokens are then organised in a graph based on keyword co-occurrences within publications (i.e. *keyword graph construction*). The subsequent steps of *topic discovery* and *topic filtering* are applied to generate the set of topics emerging from the publications. Finally, a *topic analysis* is enforced to determine trends over topics and mainstream behaviours

dataset acquisition, keyword extraction, and keyword graph construction as the preparation steps; then, we focus on the subsequent activities related to topic discovery, filtering, and analysis.

5.1 Dataset Preparation

Dataset preparation has the goal of extracting keywords from publications that are representative of the study's focus. Moreover, once those keywords have been extracted, preparation aims at explicitly representing the distribution of keywords over publications so that the co-occurrence of the same keywords in publications is highlighted.

An initial step of *dataset acquisition* is extracting the metadata of each publication $p \in P$, namely the title and the author-assigned keywords. The *keyword extraction* step is then executed on the publication metadata by applying conventional NLP techniques, such as tokenisation, lemmatisation, and 2-gram recognition based on mutual information (Manning et al., 2008). A *keyword set* K_p is associated with each publication $p \in P$ as a result. The step of *keyword graph construction* is finally

executed to highlight when keywords co-occur in the publication descriptions, namely in the associated keyword sets. The result is a graph $G = (N, E)$, where $N = \bigcup_{i=1}^{i=n} K_{p_i}$ is the set of nodes constituted by the overall set of publication keywords extracted from the metadata and E is the set of graph edges connecting pairs of keyword nodes. An edge $e_{ij} = (n_i, n_j, w_{ij})$ denotes that the keyword represented by the node n_i co-occurs with the keyword of node n_j in the keyword sets of the publication descriptions. The weight w_{ij} denotes the strength/relevance of the n_i, n_j co-occurrence, namely the number of publications in which n_i, n_j co-occur.

Example As an example of data preparation, we consider publication p1 with the title 'Bologne et le Cardinal Légat Bertrand du Pouget' and the following author-assigned keywords: avignon, bertrand du pouget, bologne, cardinal légat. The following keyword set K_{p1} is extracted for the publication p1:

$$K_{p1} = \left\{\underline{\text{avignon}}, \text{bertrand}, \text{bertrand du pouget}, \underline{\text{bologne}}, \text{cardinal}, \text{cardinal légat}, \text{légat}, \text{pouget}\right\}$$

Similarly, consider a further publication p2 characterised by the following keyword set K_{p2}:

$$K_{p2} = \left\{\underline{\text{avignon}}, \text{bertrand du pouget}, \underline{\text{bologne}}, \text{histoire de l'Église}, \text{jean xxii}\right\}$$

In the keyword graph construction step, each item of the sets K_{p1} and K_{p2} becomes a node of the graph $G = (N, E)$. Call n_a the graph node for the keyword avignon and n_b the node for the keyword bologne. The edge $e_{ab} = (n_a, n_b, 2)$ is defined in G to denote that the keywords avignon and bologne co-occur in two publications (i.e. p1 and p2); thus, the weight of the edge between their respective nodes is 2.

5.2 Topic Discovery

The keywords used for describing publications are characterised by *sparseness*, meaning that the terms appearing in the keyword set K_p of a publication p are usually highly focused and are rarely employed in the keyword set of other publications. To reduce the impact of keyword sparseness and capture possible topic overlaps among publications, we exploit the idea that the keywords of a publication can be enriched with the keywords of other publications when these keywords are frequently co-occurring and, thus, when they are used within the same terminological context. For topic discovery, each publication $p \in P$ is associated with an *enriched keyword set* $\overline{K_p}$ that has the goal of describing the publication p with keywords that are general enough to reveal the publication topic instead of the publication focus.

For a publication $p \in P$, the construction of the set $\overline{K_p}$ is described in the following way: Consider the keyword graph $G = (N, E)$ built during dataset preparation and consider a keyword $k_i \in K_p$. We call *keyword co-occurrence context*

the set $K_i^* = \left\{k_j : \exists e_{ij}\left(n_i, n_j, w_{ij}\right) \in E\right\}$ such that there is at least one co-occurrence relation between k_i and k_j in G (i.e. the two keywords co-occur in the description of at least one publication). Given the publication p, we call the *publication co-occurrence context* the set $K_p^* = \bigcup_{k_i \in K_p} K_i^*$. The set K_p^* contains keywords that are not directly used to describe the publication p but that co-occur with the keywords of K_p in other publications. Each keyword $k_j \in K_p^*$ is associated with a weight ω_j to denote the relevance of the keyword k_j in describing the topic of the publication p. For a keyword $k_j \in K_p^*$, the weight ω_j is calculated as follows:

$$\omega_j = \frac{1}{\max\limits_{k_z \in K_p^*} \omega_z} \sum_{k_i \in K_p} \sum_{k_j \in K_i^*} \alpha + w_{ij}$$

where $\alpha \in \mathbb{N}$ is a constant parameter and w_{ij} is the weight associated with the edge e_{ij} in the graph G, which denotes the number of co-occurrences in the publications of the keywords $k_i \in K_p$ and the keyword $k_j \in K_p^*$. The α parameter is introduced to support a flexible definition of the weight ω_j associated with a keyword $k_j \in K_p^*$. In particular, the value of α is added to the weight ω_j each time a keyword $k_i \in K_p$ co-occurs with a keyword $k_j \in K_p^*$. When low values of α are considered (i.e. $\alpha = 0$ or $\alpha = 1$), the weight ω_j mostly depends on the weight w_{ij} of the co-occurrences of the keyword k_j with the keywords of $k_i \in K_p$. When high values of α are considered, the weight ω_j is increased each time a co-occurrence of the keyword k_j is found with the keywords of $k_i \in K_p$, despite the strength of the weight w_{ij}. This means that when α is high, we assign more importance to the keywords $k_j \in K_p^*$ that have numerous co-occurrences with the keyword $k_i \in K_p$ and give less importance to the weight w_{ij} of such co-occurrences.

Finally, the enriched keyword set of a publication p is defined as $\overline{K_p} = \left\{k_j : k_j \in K_p^* \wedge \omega_j \geq th\right\}$, where *th* is a prefixed threshold to distinguish relevant versus nonrelevant keywords to include in K_p. Finally, a new graph $\overline{G} = \left(N, \overline{E}\right)$ is generated according to the enriched keyword sets $\overline{K}$. In $\overline{G}$, the edges $\overline{E}$ denote the keyword co-occurrence in the enriched keyword sets $\overline{K}$ of the publications.

According to the enriched co-occurrence graph $\overline{G} = \left(N, \overline{E}\right)$, we provide the following topic definition:

Topic A topic T_s is a set of featuring keywords that describes a common research argument. A topic T_s is defined around a *seed keyword* k_s that represents the label/name of the research argument. Given a seed keyword k_s associated with a corresponding keyword node $n_s \in N$ in $\overline{G}$, the topic T_s corresponds to the set of keywords associated with the nodes $N_s \subseteq N$ connected with k_s in the enriched co-occurrence graph $\overline{G} = \left(N, \overline{E}\right)$, namely $N_s = \left\{n_j : \exists \overline{e_{sj}}\left(n_s, n_j, w_{sj}\right) \in \overline{E}\right\}$.

We say that a publication p is about a topic T_s when at least one common keyword exists between the enriched keyword set K_p and topic T_s, namely $K_p \bigcap T_s \neq \varnothing$.

	K_p				α value		ω_j values	
K_p^*	papacy	xiv century	avignon	K_p^*	$\alpha = 1$	$\alpha = 4$	$\alpha = 1$	$\alpha = 4$
modern era	5			modern era	5	9	0.62	0.56
papacy		1	2	papacy	3	11	0.37	0.69
xiv century	1			xiv century	1	5	0.12	0.31
avignon	2			avignon	2	6	0.25	0.37
history		3		history	3	7	0.37	0.44
church history	1	2	1	church history	4	16	0.50	1.00
middle ages	2	6		middle ages	8	16	1.00	1.00

Fig. 4 Example of keywords and topics. An example of topic discovery. Given a publication p with keywords K_p and context K_p^*, we show the keyword co-occurrences in the publications of the dataset (left side) and the weight ω_j of each keyword $k_j \in K_p^*$ with two different setting of the α parameter (right side)

Example As an example of topic discovery, in Fig. 4, we show the excerpt of a keyword set K_p and related co-occurrence context K_p^*:

$$K_p \subseteq \{\text{avignon, papacy, xiv century}\}$$

$$K_p^* \subseteq \{\text{avignon, church history, history, middle ages, modern era, papacy, xiv century}\}$$

On the left side of Fig. 4, we show the number of co-occurrences between any pair of keywords $k_i \in K_p$ and $k_j \in K_p^*$. For instance, given $k_i =$ papacy and $k_i =$ modern era, the value $w_{ij} = 5$ is shown in Fig. 4 as denoting that papacy and modern era co-occur in five publications, namely $e_{ij} = (n_i, n_j, 5)$ is set in the graph G.

On the right side of Fig. 4, we show the weight ω_j of each keyword $k_j \in K_p^*$ when two different settings of the α parameter are considered. When $\alpha = 1$, the keywords $k_j \in K_p^*$ with a higher weight ω_j are middle ages and modern era, which are the keywords with highest w_{ij} value. It is interesting to note that modern era has a high ω_j weight, even if this keyword only co-occurs with papacy in K_p. When $\alpha = 4$, the keywords $k_j \in K_p^*$ with a higher weight ω_j are church history and middle ages, which are the keywords that co-occur with most of the publication keywords in K_p. It is interesting to note that the weight ω_j of modern era is strongly reduced when $\alpha = 4$. According to this example, by considering a threshold $th = 0.8$, the enriched keyword set $\overline{K_p}$ is defined as follows:

$$\overline{K_p} = \{\text{middle ages}\} \ \text{(when } \alpha = 1)\,;$$

$$\overline{K_p} = \{\text{church history, middle ages}\} \ \text{(when } \alpha = 4)\,.$$

5.3 Topic Filtering and Analysis

The ultimate goal of the proposed approach to topic detection is to analyse topics over time to highlight possible mainstream behaviours. To this end, *topic filtering* is executed to split the co-occurrence graph $\overline{G}$ into a set of subgraphs $\overline{G_Y} = \left(N_Y, \overline{E_Y}\right)$, where each one is related to keyword co-occurrences in a specific year Y. A graph $\overline{G_Y} \subseteq \overline{G}$ is constituted by *i*) the nodes $N_Y = \bigcup_{i=1}^{i=k} \overline{K_{p_i}}$, where the sets $\overline{K_{p_i}}$ are the enriched keyword sets of the publications from the year Y and *ii*) the edges $\overline{E_Y}$, where an edge $\overline{e_{ij}} \in \overline{E_Y}$ is defined as $\overline{e_{ij}} = \left(n_i, n_j, w_{ij}\right)$ and connects two keyword nodes n_i, $n_j \in N_Y$. The weight w_{ij} denotes the number of publications from the year Y in which the keywords n_i, n_j co-occur. We note that the number of publications per year can be (very) different from one year to another. As a result, for comparison of keyword co-occurrence weights across consecutive years, given an edge as $\overline{e_{ij}} = \left(n_i, n_j, w_{ij}\right)$ in the year Y, the number of co-occurrences w_{ij} is normalised by the overall number of publications from the year Y.

Given a seed keyword $k_s$with the associated keyword node n_s, a topic T_{Y_s} in the year Y corresponds to the set of keyword nodes $N_{Y_s} \subseteq N_Y$ in the subgraph $\overline{G_{Y_s}} \subseteq \overline{G_Y}$ where $N_{Y_s} = \left\{n_j : \exists \overline{e_{sj}} \left(n_s, n_j, w_{sj}\right) \in \overline{E_Y}\right\}$.

As a result of topic filtering, a topic T_s can change over time because the set of keywords T_{Y_s} that characterises the topic can vary from one year to another. Moreover, when a topic is associated with a stable pair of keyword nodes n_i, n_j in two consecutive years Y and $Y + 1$, it is possible that the co-occurrence weight w_{ij} is different in the two considered years. As a result, the *topic analysis* step is executed to observe the behaviour of topics along time/years. In particular, the goal of this step is to recognise the possible mainstream topics according to the following definition:

Mainstream Topic A mainstream topic M is a topic whose trend within a certain time interval of years $[Y_1, Y_2]$ follows one of the mainstream profiles presented in Sect. 4, namely spot, persistent, impasse, or boosting.

Example As an example, we consider the following enriched keyword sets associated with six publications in the time interval from 2015 to 2017:

$-\overline{K_0}$(2015): church history, middle ages
$-\overline{K_1}$(2015): christianity, middle ages, history
$-\overline{K_2}$(2016): christianity, catholic church, philosophy
$-\overline{K_3}$(2016): middle ages, philosophy
$-\overline{K_4}$(2017): catholic church history, middle ages
$-\overline{K_5}$(2017): philosophy, middle ages

In Fig. 5, we show a tabular representation of the graph $\overline{G}$ built according to the above keyword sets $\overline{K}$. Consider a seed keyword k_s=middle ages. All the keywords of Fig. 5 have at least one co-occurrence with the seed keyword; thus, they are all belonging to the considered topic T_s about 'Middle Ages'. The strength of the

	church history	middle ages	christianity	history	catholic church	philosophy
church history	-	1	0	0	0	0
middle ages	1	-	1	2	1	2
christianity	0	1	-	1	1	1
history	0	2	1	-	1	0
catholic church	0	1	1	1	-	1
philosophy	0	2	1	0	1	-

Fig. 5 Example of co-occurrence graph $\overline{G}$for a set of enriched keywords. As an example in the framework of topic discovery, the figure reports the number of co-occurrences within the publications for each pair of the considered keywords

$T_{middleages}$,	2015	2016	2017
church history	0.5	0	0
christianity	0.5	0	0
history	0.5	0	0.5
catholic church	0	0	0.5
philosophy	0	0.5	0.5

Fig. 6 Example of keyword weight in the years 2015–2017 about the topic 'Middle Ages'. As an example of topic trend, we show the weight of the keywords associated with the topic 'Middle Ages' in the time interval of the years 2015–2017

co-occurrences between the seed keyword and keywords of Fig. 5 in the years 2015–2017 is shown in Fig. 6 (normalised value by the number of publications in each considered year). By observing the keyword strength in the considered time interval, we can envisage possible mainstream topic profiles. In particular, for the topic $T_{\text{middle ages}}$, the keyword history denotes a *persistent topic* behaviour (despite showing little fluctuation in 2016). We also note that the keywords church history and christianity denote an *impasse topic* behaviour, while the keyword philosophy denotes a *boosting topic* behaviour for the topic $T_{\text{middle ages}}$. As a final consideration of the observed mainstreams, we could claim that in the context of the Middle Ages studies, an initial interest in the History of the Church and Christianity shifted towards more philosophical studies about Catholicism.

6 Case Study Analysis

In this section, we present the results obtained by applying the proposed approach and related techniques for topic mining on a real publication dataset taken from selected institutional research archives of Italian universities. The main idea is to provide a clear description of the results we obtained, here by focusing on a

few disciplines and using institutional publications data provided by four Italian universities. Regarding the case study analysis, an Online Appendix is provided with complementary figures and comments. The Appendix is available for download at the following link: http://islab.di.unimi.it/content/maverick_data/appendix.pdf.

6.1 Dataset Description

The proposed case study is based on a publication dataset collected from selected Italian universities. In the early 2000s, most Italian universities started to populate and maintain institutional research archives for persistently storing publications and research products. In particular, each university supported the creation of its own repository based on products published (and compulsorily uploaded) by its affiliated scholars. In selecting both universities and research areas to consider for building the dataset of the case study, we relied on the following recommendations: i) choose large, representative Italian universities, ii) choose a few selected research areas, and iii) compose a dataset that is representative of both bibliometric and nonbibliometric research areas according to the Italian regulation for research evaluation. As a result, the following four Italian universities have been selected: UNIBO—University of Bologna (with 2896 academic researchers—data consulted on April 15, 2021, from https://cercauniversita.cineca.it/php5/docenti/cerca.php), UNIMI—University of Milan (with 2258 academic researchers), UNIRM—University of Rome 'La Sapienza' (with 3350 academic researchers), and UNITO—University of Turin (with 2086 academic researchers). Moreover, among all the available disciplines, we focus on those publications authored by all scholars of the following research areas (as defined by The Italian National University Council—CUN): A01 (mathematics and informatics), A11 (history, philosophy, pedagogy, and psychology), and A13 (economics and statistics).

A summary view of the collected dataset is provided in Fig. 7. The dataset contains 123,504 publications labelled with 124,820 author-defined keywords.

	UNIBO	UNIMI	UNIRM	UNITO	Total
A01 pubs	4,831	7,227	8,889	7,705	28,652
A11 pubs	13,313	8,196	24,737	17,007	63,253
A13 pubs	11,817	3,455	12,436	6,762	34,470
Total pubs	27,223	18,805	46,002	31,474	123,504
Keywords	39,624	22,634	38,105	24,457	124,820

Fig. 7 Summary picture of the Maverick dataset. Number of publications and keywords in the Italian case study by university (Univ. of Bologna, Univ. of Milan, Univ. of Rome, and Univ. of Turin) and discipline (scientific area of study as classified by ANVUR)

For topic mining purposes, the keywords and publication titles are exploited. It is important to note that 58,585 publications of the dataset do not provide any keywords. For these publications, only the keywords extracted from the title are then used for topic mining. As a further remark, we observe that the number of publications per year is not constant. In fact, at the beginning of the 2000s, only a few publications were inserted into the selected archives by their authors, and this practice became a regular—and often compulsory—routine only around the year 2005. For this reason, the considered publications in this empirical exercise cover the years from 2005 to 2018. It is also important to stress that the number of publications per year is continuously increasing throughout the whole observed period because of the increasing role of performance-based exercises in Italy; thus, a normalisation step is required when the analysis focuses on the consistency of topics across different years.

6.2 General Results in Italian Academia

The results obtained on the considered dataset are briefly discussed by separately exploiting the publications of each considered research area, namely A01, A11, and A13.

First, to identify the mainstream profiles defined in Sect. 4 using the bibliometric data collected for this case study, we started defining a couple of synthetic operators to describe a topic's behaviour over time. For any given topic k belonging to its G graph, we first generate a matrix of the number of links with all the other existing topics within the same discipline during the observed years, where in each of its cells, we have the corresponding number of papers. Then, we compute two simple correlation coefficients, $\rho_{kt,j}$, for any identified topic k: a pair of ρ-s (namely $\rho_{k,t}$ and $\rho_{k,j}$) that represents the correlation coefficients of the topic's number of publications published over time (t) and of the number of topic links that the selected topic (k) establishes with the other topics (j) in the discipline, respectively.

Figure 8 shows how a topic may behave according to different combinations of the values defined by its pair of ρ-s coefficients. Using the computed ρ coefficients, it is possible to broadly map the mainstream topic profiles defined in Sect. 4 in the area defined by the two ρ pairs.

A *spot topic*, for example, corresponds to a short-lived topic that, after a burst of attention from the research community in the past, is now abandoned. A 'trendy' topic within a discipline appears in the bottom left area of Fig. 8 (e.g. grey circle).

An *impasse topic* describes the development of a research programme having some topic links that died in recent years; this is in the bottom middle area of the graph (e.g. below the big yellow circle).

A *persistent topic* identifies a mainstream topic that enjoys stable attention from the research community but with low productivity in terms of links with new research lines. This topic appears in the bottom right part of Fig. 8 (e.g. purple circle).

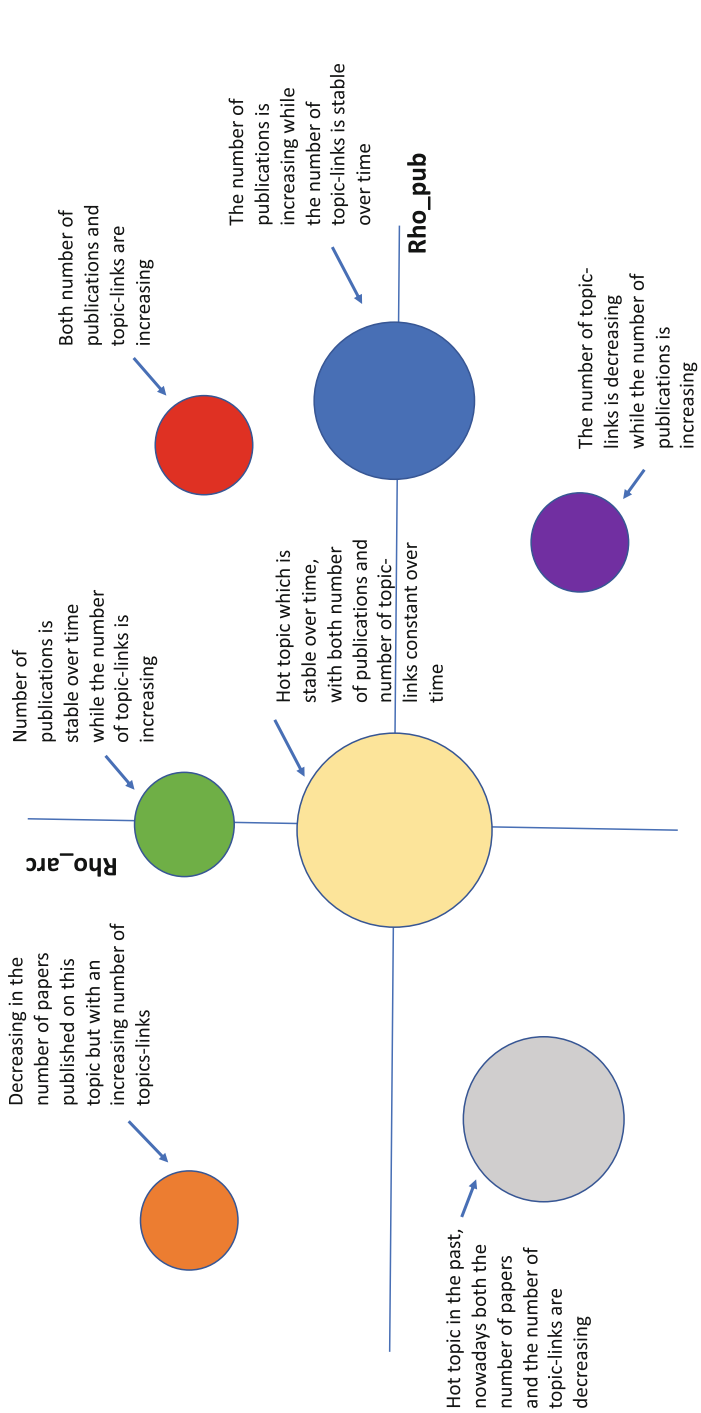

Fig. 8 Topic behaviour in the Italian case study according to ρ-s coefficients. Figure 8 shows how a topic may behave according to different combinations of the values defined by its pair of ρ-s coefficients. Examples of topic characteristics are provided in different zones of the provided space to show how the mapping of mainstream topic profiles defined in Sect. 4 can be obtained in the ρ-s coefficients space

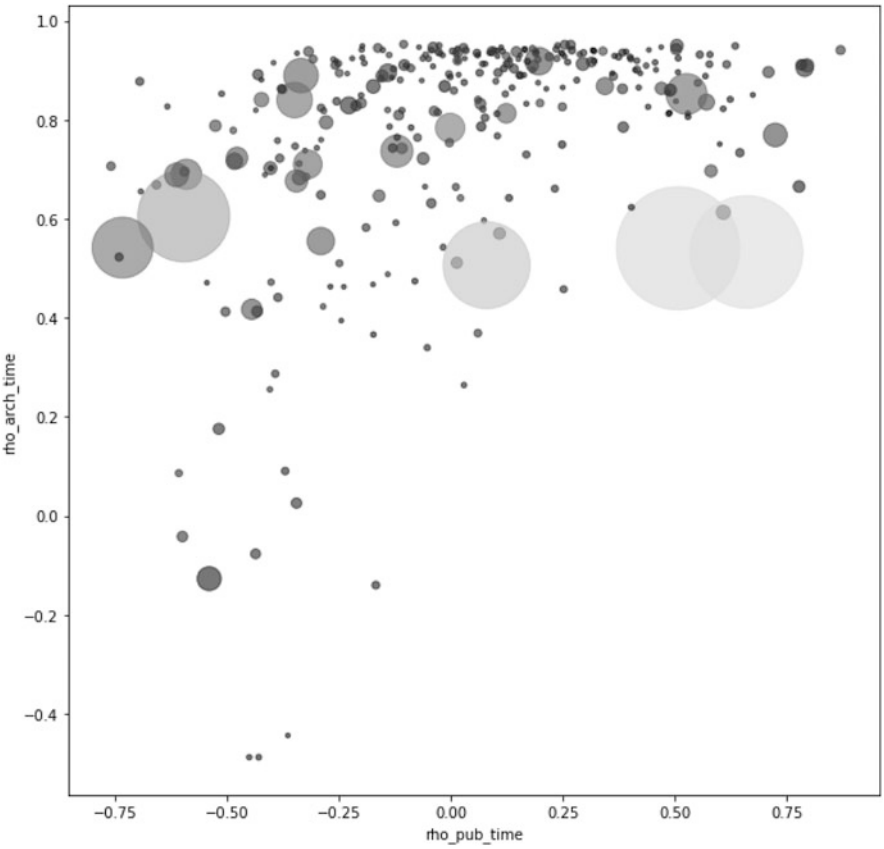

Fig. 9 Mathematics and informatics. Each of the considered disciplines in this empirical study is visualised on a map (Mathematics and informatics only is reported above), showing a circle for each topic characterising the discipline during the period under investigation. Circle size and colour represent the number of topic links and topic size (e.g., number of publications), respectively

A *boosting topic*, which has been described in Sect. 4 as a topic characterised by a long life and a high number of connections with other topics, is the top centre or top right corner of Fig. 8 (e.g. green and/or red circles).

In addition to this, Fig. 8 may be useful to identify *niche* topics, like the orange-like circle in the top left area, which is characterised by a decreasing number of papers published in the past few years along with an increasing number of topic links.

For each one of the disciplines considered in this empirical study, a figure has been created that visualises a map similar to Fig. 8, with a circle for each topic characterising the discipline during the period under investigation. Circle size and colour represent the number of topic links and topic size (e.g. number of publications), respectively. Each map describes the corresponding discipline using the proposed topic approach (Fig. 9).

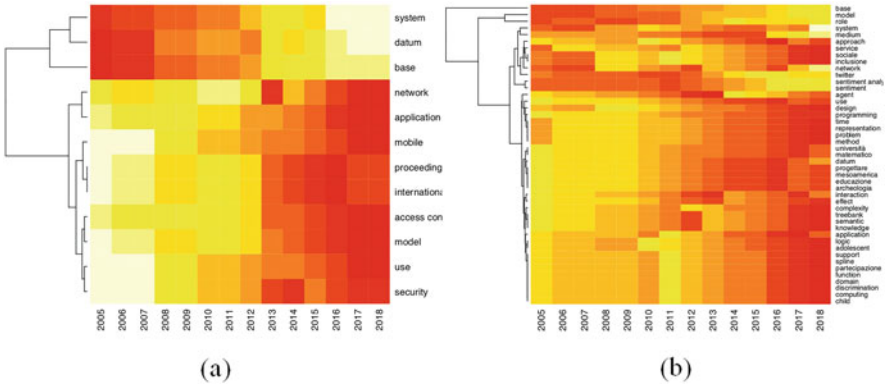

Fig. 10 Examples of a heatmap for mathematics and informatics. In (a) the topic *privacy*, in (b) the topic *social.* Examples of a heatmap for mathematics and informatics. Each measure of topic link evolution over time is standardised by its dimension to generate a comparable heatmap that clearly visualises the temporal topic's dynamics in the discipline

Then, for each of the relevant topics of the identified mainstream categories, we compute a matrix describing the evolution over time of all the topic links generated by the topic itself, here based on the structure described in Fig. 6. Each measure of the evolution of a topic link over time is standardised by its dimension to generate a comparable heatmap that clearly visualises the temporal topic's dynamics. The case study on A01 is reported here, including the figures mentioned above. For areas A11 and A13, the figures are reported and described in the Online Appendix (see Figs. 1–4 in the Appendix).

Area 1—Mathematics and Informatics As a first example, Fig. 10a provides the heatmap of the topic '**privacy**'. This is an *impasse topic* with negative values for both the correlation coefficients (ρ-s). Each row of the heatmap represents the topics with which the topic 'privacy' reports links over time, and the colour indicates the intensity of such links. Red means few links, whereas white means many links. According to this heatmap, the topic 'privacy' used to be linked to topics like 'access', 'security', and 'network' in the past, whereas recently, it started to be associated with different topics such as 'data' and 'systems'. This dynamic seems to be pretty much in line with the current increasing availability of new sources of (individual) data, for example, hospital individual data or credit bank transactions, which consequently challenges new issues related to data 'privacy' concerns.

Figure 10b provides the heatmap of the topic '**social**'. Both ρ-s are positive, which makes it a *boosting topic*. The corresponding heatmap suggests that this topic is now (in 2016–2018) very much linked with topics like 'sentiment analysis' and 'social networks' (e.g. Twitter).

This is again very much in line with the new and fast-growing literature that uses 'big data' to extract indicators to summarise, for example, users' opinions. By contrast, at the beginning of the sample period, the topic 'social' was associated

with more traditional topics like 'education', 'participation', 'university', and 'discrimination'.

Area 11—History, Philosophy, Pedagogy, and Psychology Turning to Area 11, the topic '**ageing**', represented in detail in Fig. 2a of the Online Appendix, provides another interesting example of an *impasse topic* with both negative correlation coefficients (ρ-s).

In the most recent years, this topic has very much been associated with 'experience', 'activity', 'creativity', 'life', and 'health', while in the previous years, it used to be linked with discussions and studies more focused on the past (i.e. 'history' and 'wars').

Because the problem of the ageing of the population is increasingly and extremely relevant, the topic 'ageing' seems to be now more associated with discussions related to aged people's quality of life (both in terms of health and wealth) and their occupations rather than their past historical memories.

In addition, the topic '**female studies**' is a good example of a *boosting topic* (both rho-s are positive and high). According to graph Fig. 2b in the online supplementary material, this topic has been recently associated with topics like 'child' and 'adolescent' (which suggests an emerging focus on the relationship between mothers and children), 'male' (which points to gender-related studies), or 'patient' and 'effect' (which relates to the literature of causal analysis of health issues, which often may provide heterogeneous effects by gender).

By contrast, in the past, 'female studies' have been associated with topics related to women's mental status (e.g. 'mental health', 'stress', 'personality', 'attention', 'perception', 'memory', 'brain', and 'neuro'). In addition, it used to be associated with 'work' and 'quality of life', which may refer to work–life balance issues that appeared commonly in the literature.

Area 13—Economics and Statistics Figure 3 in the Online Appendix shows a pretty different scenario for Area 13 compared with Area 11 and Area 1.

In fact, Fig. 4 in the Online Appendix shows three heatmaps for the three following topics: 'development', 'taxation', and 'network analysis'.

'**Network analysis**' can be defined as "*a set of integrated techniques to depict relations among actors and to analyse the social structures that emerge from the recurrence of these relations"* (see Smelser & Baltes, 2001).

From our analysis, it may be characterised as a boosting topic that exhibits positive values of rho-s. Although in the past it focused on theory (being related with abstract analysis) and empirical analysis, it has been recently applied among economists, econometricians, and statisticians to topics such as 'sentiment analysis', 'Twitter', and 'social media', generating a new strand of literature studying a 'network analysis' taking advantage of the new sources of (big) data now available.

On the contrary, a clear example of an impasse topic in economics and statistics is represented by the topic '**taxation**'. The economics of taxation mainly collects studies regarding both the effects and consequences of taxes on economic decisions, as

well as on how to efficiently design tax systems (e.g. income, capital, environmental taxes).

For this topic, both *rho*-s are negative, meaning that there has been a decreasing interest in this topic over the past decade. However, looking carefully at its development over the past few years (see Fig. 4c in the Online Appendix), it seems quite reasonable to identify dead links with topics such as 'literature review', 'inequality measures', 'country taxation', and 'equity' in favour of new emerging trends with topics like 'income distribution', 'evidence', and 'effects', which are very much in line with recent works on the global evolution of inequality, taxation top income dynamics, progressive wealth taxation, and so forth.

Finally, an example of a persistent topic is also identified in the economics and statistics area when looking at the topic '**development**', for which both *rho*-s are almost close to zero. Development economics is a branch of economics that focuses on studies of economic, health, education, and social conditions in developing countries (especially low-income ones) compared with developed ones.

The heatmap for this topic, as represented by Fig. 4b in the Online Appendix, makes evident how this topic turns from being historically related with topics like 'global', 'sustainability and growth', or 'industry' in the past to new research frontiers aiming to explore how to estimate the 'effects of policies' and field interventions in emerging countries, often following—also in Italian academia—the studies winning Nobel Prizes in 2019 on the use of randomised clinical trials (RCTs) in this field to measure their 'performances', as well as on 'innovation' and 'new perspectives' in general.

6.3 Robustness of the Proposed Approach Using an International Dataset over 14 Years

To show the ability of the proposed approach to identify the existing publication topics, their evolution over time and their topic links in a broader (not only restricted to the national context as in the Italian case described before) and international context, we rely on different data sources: Scopus Elsevier. Through the Elsevier API service, we downloaded all the Scopus research products published between 2005 and 2018 that are classified as instances of at least one of the following subject areas (each journal may belong to more than one subject area): business, economics and econometrics, decision sciences, statistics and probability, and demography.

The dataset contains 1,700,286 unique papers published as articles (articles in press, editorial, erratum, and business articles), chapters, books, conference papers, notes, reviews, letters, and short surveys between 2005 and 2018 written by 1,433,297 different authors and labelled with 1,168,680 author-defined keywords. The obtained dataset is 12 times larger than the database analysed in the Italian case and covers almost all papers published in the selected disciplines by all the authors who are active in these research fields around the world. This database, even if not

perfect in terms of its coverage for all the existing disciplines (it is well known how social sciences and humanities or medicine are not perfectly represented by Scopus, see for example Archambault et al., 2006), provides a good set of information to explore in depth the ability of the approach to identify and group the topics in the literature.

From these data, we have selected two topics ('**development**' and '**taxation**')—out of the several topics analysed before for the Italian case—to demonstrate the scalability of the proposed approach to a broader data source and the ability of the method to go deeper into the identification of the relevant topic links. As a matter of fact, the main contribution of this chapter is to propose a methodological approach to topic mining, showing its applicability to different disciplines and its reliability—in terms of the obtained results—when datasets of different richness and size are considered.

Obviously, the topic description may be slightly different depending on the different sets of authors and publications considered. For example, consider that in a given discipline, the Italian authors could have a different publishing behaviour in the 14 years analysed when compared with the authors who are active in the international literature. In this case, the two analyses may not perfectly overlap.

As for the '**development'** topic, the approach applied to the Scopus dataset can identify several different topics within the 'development' field. From the scholarly publications in 'development' journals, the proposed approach identifies eight (sub)topics. A first topic, which is the most relevant in terms of publications, is broadly named 'development' and is classified as a boosting topic (both rho-s are positive and high) in the Scopus dataset. That is, it has around 3000 papers published (increasing over time) with an increasing number of links with other topics. This topic is a persistent one in the Italian case study. In addition to this, seven additional subfields in the 'development' area have been identified, describing a relevant heterogeneity within the field. Topics like 'development finance' and 'development funding' are both classified as impasse topics (with negative rho-s), exhibiting ended links with other topics and reduced attention from the researchers publishing in the discipline over the considered years. At the same time, this approach provides evidence of some new boosting topics (with both rho-s positive and large) in subfields like 'development economics' and 'development strategies'. Moreover, a 'development' heatmap (Fig. 6 in the Online Appendix) shows how the emerging topics over the past few years of the analysed sample have focused on studying cultural, educational, agricultural, trade, and migration issues, along with managerial, institutional, and governance strategies, with special attention given to sustainability, climate change, and the evaluation of the effects of policies in the context of African and Asian countries (like China and India). This more detailed description of the field is in line with the one offered in the Italian case study but with an improved degree of available details on both the thematic issues and specific countries.

As for the '**taxation'** topic, we have now a richer set of subtopics identified by our approach. Although the overall 'taxation' topic has received decreasing interest over the past decade in the Italian case study (as shown in the previous section),

the Scopus dataset shows how this field is highly heterogenous. Some topics show a decreasing interest from scholars (like 'tax competition'), while other topics are clearly emerging ('boosting topics') in terms of both the number of papers and the number of topic links. Examples of these boosting topics are 'tax incentives', 'tax havens', 'tax compliance', and 'tax morale', which are identified as new emerging topics over the past 15 years.

A first heatmap on 'taxation' as a whole shows how the evolution of the international literature spans from links with topics like 'regulation', 'redistribution', and 'welfare' analysis in the early years to more recent developments in the field focused first on 'income inequality' (from 2011 to 2015) and then on 'income distribution', 'tax reforms', and 'policy evaluation' of interventions in countries like USA, Australia, and United Kingdom (Fig. 7 in the Online Appendix). Taking advantage of the richness of the considered Scopus dataset, we can go deeper in the analysis of these topics by looking at the heatmaps representing the links that even smaller subtopics in 'taxation' have established over time. For example, if we focus on the smaller topics identified by the proposed approach within the taxation field, we can identify 'tax havens' as a new emerging topic (which exhibits positive and large rho-s) with a rising number of publications from 2008 onwards. In addition, this subtopic has been interlinked since the very beginning (2008/2010) with topics such as 'tax competition', while from 2014 to 2018, it has been associated with topics like 'tax avoidance' and 'tax evasion', probably reflecting very recent contributions in the literature following the international debate on tax havens (e.g., the 'Panama papers' debate). Note that a similar degree of precision in describing the emerging or declining fields of research within a more general discipline like taxation studies could not be found when dealing with national subsamples of publications like the ones described in the Italian case study.

To conclude, the representation and discussion of some selected topics of the three disciplines analysed here shows how the proposed approach may be useful in describing both the geography of topics and the evolution and interlinkages of topics within a discipline by means of two datasets: i) institutional publications provided by four Italian universities and ii) international publications over 14 years. Moreover, the examples show how having a more comprehensive database of the worldwide production of papers is essential to prove the scalability of the proposed approach and reliability of the obtained results. All in all, richer and sizable datasets allow clearer and in-depth analyses to be provided. In particular, for a given number of papers available from the literature, the larger the number of the analysed topics, the sparser the topic links matrix. The cell size of this matrix is crucial to obtain reliable information on the temporal evolution of detailed topics and their interlinkage with new emerging or declining ones. Therefore, large bibliometric databases with millions of records can provide enough information to make this possible.

7 Concluding Remarks

The results from the case study show that the proposed approach to topic mining is capable of revealing trends of publication keywords and changes of these trends over time both when country-level data are available and when—even better—the larger international literature in a given field is considered. In combination with the contribution about mainstream modelling, this result represents a promising achievement, allowing us to recognise topic behaviours that can be associated with one of the profiles of the mainstream research defined in the project (i.e. spot, persistent, impasse, and boosting). In the following, we provide some considerations about possible extensions and applications based on our results.

Possible Research Extensions Future research activities could focus on i) the extension of the case study dataset in the Italian context to get complete coverage of the topic evolution of a given discipline in Italian academia, ii) the use of a discipline-specific keyword dictionary for a more refined topic cleaning, and iii) the comparison of case study results against third-party datasets similar to the case described in Sect. 6.2. The extension of the Italian case study dataset requires including the institutional research archive of additional Italian universities and may be a powerful tool to comprehensively analyse the topic's evolution and interlinkages across the different disciplines of Italian academia. On this point, we note that institutional research archives started to be populated at the beginning of the 2000s for almost all Italian universities. This means that i) the dataset adopted for the case study cannot be improved in terms of the size of the considered time interval and ii) the initial years of the considered time interval (i.e. the years from 2000 to 2004) are marginally useful for topic mining because few publications are present in the archives. As a result, through the extension of the available data, we aim to improve the richness of the publication corpus and increase the relevance of the case study in providing meaningful insights into the Italian picture in the period 2005–2018. A progressive inclusion of very recent publications from the considered universities is also required to keep the case study up to date. This may allow researchers from the various fields in the social sciences to study the evolution of their disciplines and the temporal changes that occurred in relation to a number of features, for example, new generations of researchers being more open towards international academia, the introduction of the research assessment exercises on the studied topics, and so forth. Moreover, the proposed approach applied to a complete country-specific bibliometric database may enable policy makers, such as ANVUR (Italian Agency for Evaluation of the University and Research System) or MIUR (Italian Ministry for Education, University, and Research) to design new policy interventions (which is their institutional mission) based on a solid analysis of 'what worked' (or not) in the past (as described in more detail in the possible applications to research evaluation provided below). A further issue for future research activities is the specification of a keyword dictionary for topic cleaning, possibly with a more detailed approach for each specific discipline. Sometimes, very general and poorly relevant keywords are included in the results of topic mining

activities. A manually defined dictionary of keywords can be set up to refine the results of keyword extraction and improve the quality of the discovered topics. Finally, a comparison of the obtained results against a third-party dataset can be also envisaged, here following the lines described in Sect. 6.2, to compare the topics found in the case study with the Italian community against a dataset that is representative of the international academic community. The goal is to observe possible similarities and/or peculiar behaviours of the Italian community compared with a larger, international group of scholars.

Possible Applications to Research Evaluation The topic trends that have emerged by applying the proposed techniques can be exploited to analyse changes in the publication practices of researchers along the temporal dimension. It is possible to apply these techniques to a publication dataset that is representative of the overall Italian Academy, meaning that almost all the institutional research archives of the Italian universities will be considered. A possible application scenario is to consider a 'median scholar' and the corresponding set of authored publications. By extracting the featured keywords from the 'median scholar' publications, it is possible to compare and correlate their research production against the topic trends associated with the scholarly keywords. In this way, shifts in the 'median scholar' interests can be tracked, as well as possible changes in terms of publication practices over time so that it is possible to observe whether the scholar's behaviour endorses a topic whose trend can be recognised as mainstream according to specific time intervals. Similarly, one can focus on identifying heterogeneous publication patterns along the ability distribution, for example, studying if 'top scholars' behave differently than median or bottom ones. As a further application scenario, a similar approach can be enforced to analyse the changes that occur over time regarding a reference publication source (e.g. top journals) within a specific research area. In this way, it is possible to observe the evolution of 'hot research topics' in certain publication sources in correlation with the topic trends emerging from the already available results in that research area.

In addition to this, having access to the relevant data for the worldwide production of papers belonging to a specific discipline (as collected by standard bibliometric sources such as Scopus or Web of Science) may also enable a comparison of the national evolution of a discipline in a specific country (e.g. in Italy in our case) with respect to its own international benchmark. Moreover, a similar approach may also be adopted to study the effects of introducing a performance-based assessment exercises—as has happened in several countries around the world over the past decades—on the topics' evolution in different disciplines at the local level.

Does the system of incentives provided by a performance-based assessment exercise have an impact on the evolution and choice of topics studied by academic scholars in their disciplines? Is there any evidence of a temporal shift towards international mainstream research (e.g. leaving niche topics aside) following the introduction of this type of assessment exercise? If so, is it socially optimal? All

these research questions (and probably many others) will be part of the future research agenda in this strand of the literature.

Author Contributions All the chapter authors equally contributed to the conceptualization, data curation/analysis, and interpretation of the results of this study. Sections 1 and 2 were primarily written by Eugenio Petrovich, Stefano Verzillo, and Stefano Montanelli; Sects. 3 and 4 by Eugenio Petrovich; Sect. 5 by Stefano Montanelli and Alfio Ferrara; Sect. 6 by Silvia Salini, Stefano Verzillo, and Corinna Ghirelli; Sect. 7 by Stefano Montanelli and Stefano Verzillo.

References

Abramo, G., D'Angelo, C. A., & Costa, F. (2019). When research assessment exercises leave room for opportunistic behavior by the subjects under evaluation. *Journal of Informetrics, 13*, 830–840.

Abramo, G., D'Angelo, C. A., & Grilli, L. (2021). The effects of citation-based research evaluation schemes on self-citation behavior. *Journal of Informetrics, 15*(4), 101204. abs/2102.05358.

Académie des Sciences, Leopoldina, & Royal Society. (2017). Statement by three national academies on good practice in the evaluation of researchers and research programmes. *Vestnik Rossiĭskoĭ Akademii Nauk, 88*(11), 979–981.

Archambault, É., et al. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342.

Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLos One, 14*(9), e0221212.

Biagioli, M., Kenney, M., Martin, B. R., & Walsh, J. P. (2019). Academic misconduct, misrepresentation and gaming: A reassessment. *Research Policy, 48*(2), 401–413.

Bonaccorsi, A. (2015). *La Valutazione Possibile: Teoria e Pratica della Valutazione della Ricerca*. Il Mulino.

Borner, K. (2010). *Atlas of science: Visualizing what we know*. MIT Press.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., & Borner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), 1–11.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation, 12*, 39–46.

Castano, S., Ferrara, A., & Montanelli, S. (2018). Topic summary views for exploration of large scholarly datasets. *Journal on Data Semantics, 7*, 155–170.

Castellani, T., Pontecorvo, E., & Valente, A. (2016). Epistemic consequences of Bibliometrics-based evaluation: Insights from the scientific community. *Social Epistemology, 30*(4), 398–419.

Cedrini, M., & Fontana, M. (2018). Just another niche in the wall? How specialization is changing the face of mainstream economics. *Cambridge Journal of Economics, 42*(2), 427–451.

Checchi, D., Ciolfi, A., De Fraja, G., Mazzotta, I., & Verzillo, S. (2021). Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm. *Economica, 88*(352), 1107–1129. https://doi.org/10.1111/ecca.12373

Colander, D., Holt, R., & Rosser, B. (2004). The changing face of mainstream economics. *Review of Political Economy, 16*(4), 485–499.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review, 16*(7), 969–986.

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2015). Evaluation practices and effects of indicator use — A literature review. *Research Evaluation, 25*(2), 161–169.

Feenstra, R. A., & Lopez-Cozar, E. D. (2021). The footprint of a metrics-based research evaluation system on Spanish philosophical scholarship: an analysis of researchers perceptions. *ArXiv*. abs/2103.11987. https://doi.org/10.48550/arXiv.2103.11987

Felt, U. (Ed.). (2009). *Knowing and living in academic research: Convergences and heterogeneity in research cultures in the European context*. Institute of Sociology of the Academy of Sciences of the Czech Republic.

Fochler, M., Felt, U., & Muller, R. (2016). Unsustainable growth, hyper-competition, and worth in life science research: Narrowing evaluative repertoires in doctoral and postdoctoral scientists' work and lives. *Minerva, 54*(2), 175–200.

Gasparyan, A. Y., Nurmashev, B., Udovik, E. E., Koroleva, A. M., & Kitas, G. D. (2017). Predatory publishing is a threat to non-mainstream science. *Journal of Korean Medical Science, 32*(5), 713.

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*(4), 277–304.

Geuna, A., & Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). *Research Policy, 45*(1), 260–271.

Glenisson, P., Glanzel, W., & Persson, O. (2005). Combining Full-text Analysis and Bibliometric Indicators. a Pilot Study. *Scientometrics, 63*(1), 163–180.

Gottfredson, L. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23.

Heinze, T. (2013). Creative accomplishments in science: Definition, theoretical considerations, examples from science history, and bibliometric findings. *Scientometrics, 95*(3), 927–940.

Ivani, S. (2019). What we (should) talk about when we talk about fruitfulness. *European Journal for Philosophy of Science, 9*(1), 4.

Katzav, J., & Vaesen, K. (2017). Pluralism and peer review in philosophy. *Philosophers' Imprint, 17*(19), 1–20.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge University Press.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Moed, H. F. (2017). *Applied evaluative Informetrics*. Springer International Publishing, Cham.

Muller, R., & de Rijcke, S. (2017). Thinking with indicators. Exploring the Epistemic Impacts of Academic Performance Indicators in the Life Sciences. *Research Evaluation, 26*(3), 157–168.

Nichols, L. (2014). A topic model approach to measuring Interdisciplinarity at the National Science Foundation. *Scientometrics, 100*(3), 741–754.

Scarpa, F., Bianco, V., & Tagliafico, L. A. (2018). The impact of the National Assessment Exercises on self-citation rate and publication venue: An empirical investigation on the engineering academic sector in Italy. *Scientometrics, 117*(2), 997–1022.

Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy, 48*(2), 478–491.

Smelser, N. J., & Baltes, P. B. (Eds.). (2001). *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 11). Elsevier.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476.

Talley, E. M., et al. (2011). Database of NIH Grants using machine-learned categories and graphical clustering. *Nature Methods, 8*(6), 443–444.

Viola, M. (2018). Evaluation of research(ers) and its threat to epistemic pluralisms. *European Journal of Analytic Philosophy, 13*(2), 55–78.

Yan, E., Ding, Y., Milojevic, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics, 6*(1), 140–153.

Alfio Ferrara Alfio Ferrara (Ph.D., University of Milan) is a Professor of Computer Science at the University of Milan. His research interests are focused on data science methods for natural language processing, information retrieval, and text mining.

Corinna Ghirelli Corinna Ghirelli (Ph.D., University of Ghent) is a research economist at the Bank of Spain. Her research interests are applied econometrics, labour economics, policy evaluation, and textual analysis.

Stefano Montanelli Stefano Montanelli (Ph.D., University of Milan) is an Associate Professor at the University of Milan. His main research interests include semantic web, data matching, web data classification and summarisation, and crowd-collaborative data management.

Eugenio Petrovich Eugenio Petrovich (Ph.D., University of Milan) is a post-doctoral researcher at the University of Siena. He works on scientometrics and quantitative science studies.

Silvia Salini Silvia Salini (Ph.D., University of Milano-Bicocca) is an Associate Professor of Statistics at the University of Milan. Her main research interests focus on statistical models for social science, multivariate statistics, statistical learning methods, robust statistics, and scientometrics.

Stefano Verzillo Stefano Verzillo (Ph.D., University of Milan) is a Senior Research Scientist at the Joint Research Centre of the European Commission. His research interests are education, labour and health economics, and evaluation of public policies.