Journal of Informetrics 11 (2017) 1158–1174

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Research portfolio analysis and topic prominence

Richard Klavans[a], Kevin W. Boyack[b,*]

[a] SciTech Strategies, Inc., Wayne, PA 19087, USA
[b] SciTech Strategies, Inc Albuquerque, NM 87122, USA

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 4 October 2017
Accepted 5 October 2017

Keywords:
Research portfolio analysis
Direct citation
Research topics
Prominence
Project-level grant data

ABSTRACT

Stakeholders in the science system need to decide where to place their bets. Example questions include: Which areas of research should get more funding? Who should we hire? Which projects should we abandon and which new projects should we start? Making informed choices requires knowledge about these research options. Unfortunately, to date research portfolio options have not been defined in a consistent, transparent and relevant manner. Furthermore, we don't know how to define demand for these options. In this article, we address the issues of consistency, transparency, relevance and demand by using a model of science consisting of 91,726 topics (or research options) that contain over 58 million documents. We present a new indicator of topic prominence – a measure of visibility, momentum and, ultimately, demand. We assign over $203 billion of project-level funding data from STAR METRICS® to individual topics in science, and show that the indicator of topic prominence, explains over one-third of the variance in current (or future) funding by topic. We also show that highly prominent topics receive far more funding per researcher than topics that are not prominent. Implications of these results for research planning and portfolio analysis by institutions and researchers are emphasized.

1. Introduction

Research portfolio analysis should be a key activity for all stakeholders in the current science system. Funding bodies allocate resources among topics, administrators choose which researchers to hire and which projects to support internally, while researchers (for the most part) choose the topics they want to work on (Fisher, 2005; Foster, Rzhetsky, & Evans, 2015; Zuckerman, 1978). The notions of research portfolios and portfolio analysis, once largely confined to the corporate R&D world, are now being increasingly considered in academic and agency settings (Wallace & Rafols, 2015).

Portfolio-related choices are, however, difficult to make in the science system because the potential choices themselves are often not well defined or understood. In the industrial world, research portfolio choices are typically governed by perceptions of (long-term) supply and demand. We suggest that the concepts of supply and demand can also provide a useful framework for research portfolio analysis in the science system. However, to use these terms we must take care to define them properly as they can be defined in different ways.

For instance, Sarewitz and Pielke (2007) define supply and demand at a high level in terms of the interplay between scientific results and their providers (supply) and specific societal goals (demand). At a more detailed level, Sarewitz & Pielke define demand in terms of the information used by a wide variety of stakeholders to address a broad set of challenges

* Corresponding author.
E-mail addresses: rklavans@mapofscience.com (R. Klavans), kboyack@mapofscience.com (K.W. Boyack).

https://doi.org/10.1016/j.joi.2017.10.002
1751-1577/

(p. 12). Dalrymple (2006) defines supply as the output of the public research process and demand as reflecting the interests of the users. These interests – the inputs to demand – can vary. The demand for science to address social and health needs is prevalent, and can be represented by metrics such as disease burden (Evans, Shim & Ioannidis, 2014). Some science, particularly that invested in by industry, responds to economic motives (Klavans & Boyack, 2017a). These diverse demands are important, and they (and their advocates) obviously play a role in priority setting within governments, agencies, and other funding bodies. It is also true that these types of demand are very difficult to measure. Rather than considering each type of demand separately, we simply note that the ultimate result of these varying demands is that research priorities are set and funding is made available to address these priorities. Accordingly, we make explicit this assumption that funding amounts represent an aggregate (though undoubtedly crude and incomplete) measure of demand for each priority or topic in science. Thus, our definition of demand differs from those of Sarewitz & Pielke and Dalrymple in that it assumes that interests and goals are codified in a fiscal sense. The scientific topics themselves – the outputs of science mentioned by Sarewitz & Pielke and Dalrymple – represent the scientific supply.

Other frameworks are possible. For instance, another common framework uses the language of investments and outputs (Wang & Shapira, 2015). Funders invest in research, while researchers create outputs (articles) that may influence social outcomes. While this framework is valid (it is commonly used to analyse securities as well as grant investments), we prefer the more common definition of portfolio analysis associated with commerce (http://www.businessdictionary.com/definition/portfolio-analysis.html). Research is viewed as a product (publications represent supply) while funders purchase that research (grants represent demand).

With this framework in place, the gap that remains is the lack of a detailed model of scientific topics that represents supply and to which demand can be linked. Accordingly, this study introduces a topic-level model of the scientific literature. We define a topic as a collection of documents with a common focused intellectual interest, such as the work on a specific research problem. We then go further by developing a new indicator of topic prominence, and show that this indicator is a good predictor of current and future funding at the topic level, and is thus an indicator of demand. Unlike previous bibliometric studies, where the emphasis has only been on identifying emerging topics (Small, Boyack, & Klavans, 2014) or research fronts (Clarivate, 2016), our goal is to look at the entire portfolio of choices. As such, this is the first large-scale test of a highly detailed portfolio model of research.

The article is organized as follows. The background section starts by describing the theory and practice underlying the identification and evaluation of all possible research topics in the scientific literature. Critical to this discussion are the issues of coverage, granularity, accuracy and stability. We also briefly discuss potential indicators of topic impact and relevance. Additional sections then detail the four main contributions of the article: 1) creation and description of the topic-level model of science, 2) formulation of an indicator of topic prominence, 3) the assignment of project-level grant data to individual topics and 4) the use of the prominence indicator to explain and predict topic funding levels. Each of these sections combines methods and results. The article closes with a discussion of weaknesses and limitations of the study, and with implications for research planning by funding agencies, research institutions and individual researchers.

2. Background

The structure of science, as a whole or by parts, can be represented in many different ways including journal subject categories, controlled vocabularies such as MeSH, or clusters based on citation or textual characteristics. The funding of science has been linked to its structure, but this has been done only at very high levels. For instance, R&D expenditures have been reported at the level of eight S&E fields (National Science Board, 2016), by agency or sub-agency, and by disease category (Evans et al., 2014). We are unaware of any studies that report funding at much more detailed levels, perhaps because more detailed classification systems are not commonly available.

Creation of a detailed topic-level model of science requires many design choices, the most important of which are related to coverage, granularity, accuracy and stability. Accordingly, this section focuses on these choices and on the theoretical and historical bases that give us reasonable guidance as to make these choices. We note, however, that we are operating within the context of using citation data to create models. Models created using textual characteristics or controlled vocabularies (such as MeSH) will have different properties. Such models may be appealing to different communities. For instance, disease categories using MeSH terms have been used by policy makers interested in comparing funding with disease burden for a defined set of diseases (Evans et al., 2014).

2.1. Topic coverage

Our goal is to identify all topics in science for the use of multiple stakeholders. Thus, the ideal would be to have access to all literature on all topics, and to then have a way to partition that literature into topics. This ideal, however, may not currently be reachable since no single database covers all of the scientific literature. In addition, there are differences of opinion over whether a full database is necessary to accurately identify topics or if smaller datasets using journals or keyword searches are adequate.

Although not definitive, there is a literature that addresses these issues and provides some guidance. Regarding database coverage, there are two large citation databases (Scopus and the Web of Science (WoS)) that cover a significant fraction of the scientific literature. However, it is also well known that coverage within these databases varies by field. Each is known