EPFL

Thèse n° 8361

The organisation of science: topics, incentives and funding.

Présentée le 8 septembre 2020

au Collège du management de la technologie
Chaire en politiques d'innovation et de propriété intellectuelle
Programme doctoral en management de la technologie

pour l'obtention du grade de Docteur ès Sciences

par

Omar BALLESTER GONZALEZ

Acceptée sur proposition du jury

Prof. M. Finger, président du jury
Prof. G. J. A. de Rassenfosse, directeur de thèse
Prof. C. Sugimoto, rapporteuse
Prof. B. Van Looy, rapporteur
Prof. D. Foray, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

I deixaré d'estar,
de ser, no seré.
Res sóc.
En res quedaré.
— H. Mas

A Jorge i Mary-Paz...

Acknowledgements

I would like to take the opportunity to express my gratitude to Prof. Gaétan de Rassenfosse and Dr. Orion Penner for all the extensive work, ideas and comments that have greatly contributed to this thesis.

Over the last few years, I have also benefited from the discussions, remarks, questions, support to my frustrations and great deal of knowledge of Fabrizio, my Gerzensee cohort and colleagues at the IIPP and CEMI chairs at EPFL. They have often been an inspiration that kept me pushing in the hardest times.

I am too indebted to Fatine, Hèctor, Étoimoi and my bikes, who have been the pillars of my well being. Their endless caring and tolerance for my mistakes deserve my uttermost appreciation.

And of course my biggest recognition to my parents, who have always supported me in every spoilt wish that I've had. Their unconditional love is truly what has allowed me to get to this point.

Lausanne, August 19, 2020 Omar Ballester

Abstract

Ever since the links between the development of new technologies and economic growth became evident, researchers have attempted to study how the creation of knowledge fosters progress. If pushing the frontier of knowledge has an impact on progress and well-being, it is essential to pursue some form of science policy. Policymakers rely on the scholarly work of researchers in order to understand the likely impact of new policies and investments, and evaluate the state of the art in science and innovation policy. Therefore, the work of social scientists, economists of science and information scientists, among others, is vital to the characterisation, understanding and management of science. In recent years, the availability and quality of science data (including bibliographic data and metadata, funding, relational databases, ontologies and classifications) has boosted the empirical work in the depiction of the organisational structure of science. In turn, policy analysis has been able to accurately identify many unsuspected effects of past investments and policy decisions both at the macro and micro level.

Using topic models, we develop a novel method for evaluating the robustness of different text-to-text similarity models. Employing that procedure, we find that the neural-network-based paragraph embeddings approach seems capable of providing statistically robust estimates of document–document similarities. Finding methods to estimate the similarity between individual publications is an area of long-standing interest in the information science and scientometrics communities. These techniques enable researchers to build indicators and classification methods based on the analysis of large text corpora. We show that the most widely used techniques suffer from inconsistencies upon retraining, and provide a procedure to evaluate and compare the quality of different methods, regardless of the data.

Next, we present a game-theoretic model of rewards to scientific contributions. Our model of science may help explain the resulting social organisation of science from a simple social dilemma model. We model a researcher's payoff as a *common-pool resource* game, intrinsically connecting the appropriability of scientific output to a scientist's optimal strategy. This simple model of reward allocation sheds new light on a variety of behaviours that have been observed amongst researchers.

Finally, we propose an empirical analysis of the relationship between basic knowledge generation and spillovers to innovation. Using the United States' 2001 ban on federal funding of human embryonic stem cells (hESC), we disentangle the effect that policy had on downstream innovation. We employ recently developed data on patent-to-scientific-article citations to measure the spillovers, and we characterise the causal impact of the policy on subsequent

innovation with a difference-in-differences estimator. Our estimates suggest that in the years following the policy, scholarly publications subject to the ban received 65 to 80 per-cent fewer patent citations than the control group. We then apply topic modelling techniques to examine changes in the direction of science. In particular we build a topic-variety metric. Our findings show that variety decreased in the aftermath of the policy. Our results suggest that even the most modest policy changes have a profound impact on downstream innovation and the advancement at the frontier.

Keywords: Economics of Science, Economics of Innovation, Science Policy, Topic Modelling, Scientometrics, Econometrics, Machine Learning

Résumé

Après avoir établi les liens entre le développement de nouvelles technologies et la croissance économique, de nombreux travaux de recherches visent à présent à évaluer le rôle de la création de connaissances dans l'amélioration du niveau de vie. Les découvertes scientifiques menées à la frontière de connaissance ont un impact significatif sur le progrès et le bien-être. Les politiques économiques menées par rapport à la recherche scientifiques revêtent donc une importance particulière. En effet, les décideurs politiques ont besoin des résultats obtenus par les chercheurs afin d'estimer l'impact des nouvelles mesures implémentées et des nouveaux investissements fournis. Ils évaluent en conséquence les dispositions à prendre en matière de politiques scientifiques et d'innovation. Par conséquent, les travaux en sciences sociales, menés par des économistes des sciences et des informaticiens, entre autres, est vital pour la caractérisation, la compréhension et la gestion de monde de la recherche scientifique. Ces dernières années, la disponibilité et la qualité des données scientifiques (y compris les données bibliographiques et les métadonnées, le financement, les bases de données relationnelles, les ontologies et les classifications) ont favorisé la mise en place de travaux empiriques intégrant la représentation de la structure organisationnelle de la science. En parallèle, l'analyse détaillée des politiques a permis d'identifier avec précision de nombreux effets insoupçonnés des investissements passés.

En utilisant la technique de *topic models*, nous développons une nouvelle méthode permettant d'évaluer la stabilité de différents modèles de similitude de texte. Nous constatons que l'approche de vectorisation de paragraphes basée sur un réseau neuronal semble capable de fournir des estimations statistiquement robustes des similitudes entre deux documents. La recherche de méthodes pour estimer la similitude entre les publications individuelles est un domaine d'intérêt pour les communautés des sciences de l'information et de la scientométrie. Ces techniques permettent aux chercheurs de construire des indicateurs et des méthodes de classification basés sur l'analyse de grands corpus de texte. Nous montrons enfin que les techniques les plus fréquemment utilisées souffrent d'incohérences lors du ré-entrainement, et nous fournissons une procédure pour évaluer et comparer la qualité des différentes méthodes, pour tout type de données.

Dans un second temps, nous présentons un modèle théorique évaluant les récompenses attribuées aux contributions scientifiques. Notre modèle aide à expliquer l'organisation sociale de la science qui en résulte à partir d'un simple dilemme social. Nous modélisons le gain d'un chercheur comme un *jeu de ressources communes*, reliant intrinsèquement la pertinence de la production scientifique à la stratégie optimale d'un chercheur. Ce modèle simple d'attri-

bution de récompenses permet d'illustrer une variété de comportements observés chez les chercheurs.
Enfin, nous proposons une analyse empirique de la relation entre la génération de connaissances et son impact sur l'innovation. En exploitant l'interdiction imposée aux États-Unis en 2001 sur le financement fédéral des cellules souches embryonnaires humaines (CSEh), nous isolons l'effet de la politique sur l'innovation en aval. En utilisant des données récemment développées sur les citations du type brevet à article scientifique, nous mesurons l'impact causal de la politique sur l'innovation ultérieure avec un estimateur de différence dans les différences. Nos estimations suggèrent que dans les années qui ont suivi cette politique, les publications soumises à l'interdiction ont reçu de 65 à 80% de citations de brevets en moins par rapport au groupe de contrôle. Nous appliquons ensuite des techniques de *topic models* pour examiner les changements dans la direction de la science. En particulier, nous construisons une métrique de variété de sujet. Nous constatons que la variété a diminué après la mise en place. Ces résultats suggèrent que même les changements de politique les plus modestes ont un impact profond sur l'innovation en aval et l'avancement à la frontière.

Mots-clés : Économie des Sciences, Économie de l'Innovation, Politique Scientifique, Topic Modelling, Scientométrie, Économétrie, Machine Learning

Contents

List of Figures

List of Tables

Introduction

It is difficult not to sound pompous when professing the full extent of science's contribution to economic growth. The truth is that, today, science is undeniably recognised as having played a decisive role in humankind's development. There is an almost tautological consensus in acknowledging the impact that both basic and applied research have had on innovation, growth and social well-being in general. Indeed, nearly all modern governments and leading corporations strategically devote a considerable part of their budgets to research and development. Following efforts focused on defining decision-making accountability, the systematic study of science policy has seen an uptick in recent times. These efforts have, in turn, been motivated by widespread public sphere enquiries on the appropriateness of large expenditures in R&D, and on the real returns in competitiveness and economic security.

The present-day pursuit of a comprehensive science policy has its roots in Vannevar Bush's (1945) post-WWII report commissioned by U.S. President Franklin D. Roosevelt. This document established science as a central concern for governments, and it set in motion decisive lobbying for the creation of a national policy for science. The ensuing institutionalisation of science brought together an amalgam of actors working in the context of science policy, notably policymakers and researchers from a variety of fields (Marburger et al., 2011).

The work of this multidisciplinary community entails, broadly speaking, three areas of analysis. First, the understanding of the organisational structure of science —i.e., the institutional and sociological systems and networks in place. Second, the study of knowledge as a source of productivity growth —linkages between inputs and outputs— and the measurable impacts of policy interventions. Third, the *usability* of the information produced by researchers — analysis of impact, classification, manipulation, retrieval, and dissemination of information. The multiple dimensions that conform science policy imply that a plethora of academics from different specialities have targeted the most pressing questions in their respective fields.

Science is, by definition, a social system. Regarding science policy, sociologists, philosophers, historians and other social scientists alike have historically focused their research on the relationships that connect individuals and institutions. This work has allowed practitioners a better understanding of the organisation of science through the study of the formal (and informal) rules that drive researchers and institutions. The motivations, rewards, competitive nature, and social arrangements that emerge from the science system, as well as the power

forces —including ethical questions, autonomy and social responsibility— in place are under scrutiny in this area of research. It is, however, at the intersection between economics, sociology and informetrics that these questions have advanced the most in recent years. Starting with Dasgupta and David (1994), behavioural and applied economists have turned their minds towards explaining the governing rules of science too, quantifying —and modelling— the prevalent social norms.

That knowledge is a primary driver of economic growth goes mostly unquestioned by economists. In a quest for understanding the basic economics of science, the work of Nelson (1959) and Arrow (1962) constitutes the seed of an extensive line of economic writing on allocation of resources and productivity of science. Economists' work traditionally resides in theoretical —structural— analysis of the sources of technology, scientific labour supply and demand, and the empirical analysis of innovation. Economics of science —the analysis of basic science production systems—, on the contrary, has only developed in the last decades. Much of today's work in economics lies at the intersection with political science, with economists focusing their efforts on the most pressing policy questions: science funding, policy intervention effects and mechanisms of effective science-system management.

Managing science investment portfolios requires more than just productivity or impact analysis. Bibliometricians and information scientists have developed a myriad of knowledge management tools for the administration of the information generated by science. From technical solutions devoted to information curation and stewardship, to the development of indicators and characterisation of portfolios. Following Eugene Garfield's lead (1955), this area of work has enabled both policymakers and practitioners to acquire a deeper understanding of what science is being produced and how it is related to previous (and future) work. Today, the availability of electronic data has prompted the use of scientific text in the field for tasks as disparate as information retrieval or science mapping.

This thesis constitutes an effort that may speak to the three areas of science policy research. The aim of this work is no other than to advance the underlying fundamentals of how science operates by filling some of the gaps that the community has identified. It is my hope that, in covering such distinct topics, my contributions will be useful to a broad spectrum of policymakers and practitioners.

Contributions and Structure

This thesis is structured in three chapters. The first chapter develops a technical contribution to the use of topic models in the social sciences in general. These topic models are used in the following two chapters, where we apply them to target different questions. The second chapter proposes a simple game-theoretic model that helps explain some of the phenomena observed in research communities. The third chapter provides an empirical evaluation of a science policy shift.

Chapter 1 develops a fresh new look at the characterisation of text data in the social sciences in general, and in scientometrics in particular. I study the statistical robustness of different families of topic modelling techniques, namely NMF, LDA and neural paragraph embeddings. Using pairwise similarity metrics, I develop an original method to estimate the variability introduced in the models by stochastic processes, extrinsic to the practitioner's control (the parameters that are not available for tuning upon training). I also provide a simple validation test to account for the descriptive power of topic models as the researcher pushes the latent space size. Finally, by training the models in a number of different representative datasets typically used in the characterisation of science (i.e. subject headings, titles and abstracts), I provide relevant application examples of the most recent neural techniques.

This first chapter constitutes a significant technical contribution to the use of topic models in the social sciences. Topic models often face a number of roadblocks in their application to the social sciences. Notably, topic models have been under scrutiny due to their instability (i.e., difficulty to replicate). I provide evidence that the most recent methods —based on neural embeddings— are capable of providing more robust estimates than probabilistic techniques, at different degrees of granularity. I present an approach to evaluation that can be easily applied to any of the existing topic models, thus enabling practitioners to address robustness issues to their particular problems. Even though our central results emerge from clear-cut scholarly bibliographic data, the models and methods described in Chapter 1 apply to other sources of text analysis such as patents, discourse or news articles.

Chapter 2 presents a thought experiment about the strategic behaviour of scientists. Based on the extensive literature in the sociology of science, I model science as a common-pool resource (CPR) game. I argue that introducing appropriability in a public-good game of rewards results in a CPR game. In my model, I enable a researcher to appropriate the advances of her scientific field proportionally to her contribution. This distributional factor results in an optimal strategy that is supported by the weighted average of two components: the marginal and the average contribution. I argue that the CPR game results in an optimal strategy that represents the Kuhnian *essential tension* between new developments and consolidation efforts.

The exercise presented in chapter 2 is an effort to represent individual incentives in the simplest possible manner. I derive some implications that correlate well with observations from the present and the past. Namely, the trade-off between tradition and subversion and the existence of "competition" in science. Based on this model, I outline explanatory hypotheses about the links between the "old" sociology of science and the "new" economics of science. The introduction of appropriability in the payoffs to individuals helps us understand observed phenomena, and sheds new light on the mechanisms behind the incentives that shape science.

Chapter 3 analyses the impact of science funding policy on innovations and technology stemming from basic science. In this chapter, I exploit an exogenous shock —the 2001 U.S. human embryonic stem cell (hESC) policy— that impeded researchers from using certain materials (namely, newly-derived stem cells). I use a citation-based estimator to capture the

knowledge spillovers from frontier research to innovations. I estimate the causal effect of the policy with a difference-in-differences approach. Our estimates suggest that scientific articles subject to the policy restrictions received 65 to 80 per-cent fewer patent citations than unrestricted citations. The analysis constitutes one of the first to employ non-patent-literature citations and in-text references from patents to scholarly articles to establish the links between the two.

Additionally, in chapter 3, I explore the mechanisms behind the decrease in patent citations. I find that publications bound by the policy were placed in journals of comparatively lower rating and had fewer forward citations. Using publication-text data and the methods developed in Chapter 1, I characterise the yearly *diversity* of hESC publications. I observe a significant drop in variety in the aftermath of the policy, suggesting a concentration and clustering of topics in research.

Chapter 3 is a contribution to the study of science policy in itself, by inspecting how small changes to science funding policy have a profound effect on downstream innovation. We conclude that beyond funding, limitations to the materials available to researchers have negative consequences on the outlook of a field. The 2001 hESC ban affected the participation of researchers, the subsequent involvement of private sector actors and the capability to share and advance technology.

1 Robust similarity measures from topic modelling

"Remember that the primary purpose of this chapter, and the two which follow is not to develop the physical properties of ideal gases as such. Instead we shall use the ideal gas as an example to introduce some important thermodynamic ideas."
— G. Carrington (Basic Thermodynamics)

Finding methods to estimate the similarity between individual publications is an area of long standing interest in the scientometrics community. Traditional techniques have generally relied on references and other metadata, while text mining approaches based on title and abstract text have appeared more frequently in recent years. In principle, Topic Models have great potential in this domain. But in practice, they are often difficult to employ successfully and, in particular, they are notoriously inconsistent as latent space dimension grows. That is, running the same model, with the same parameters, on the same data, but with a different random seed, produces radically different similarity estimates as the number of topics increase. In this chapter, we develop a simple, but novel, method for evaluating the robustness of topic models. Employing that procedure, we find that the neural-network-based paragraph embeddings approach seems capable of providing statistically robust estimates of document–document similarities, even for topic spaces far larger than what is usually considered prudent for the most common topic model approaches.

1.1 Introduction

Methods for understanding the topics and concepts of individual documents —such as patents or scientific publications— are a matter of long-standing interest within the scientometric and informetric communities. Indeed, going back to some of Garfield's earliest thinking on citation

indexes (1955), he identified a goal of an "association-of-ideas" index. In those thoughts, he further developed the role such an index would play in the literature-search process, and highlighted the value of a "sub-micro" or "molecular" level approach over one focused on "classification".

Today, document similarity and clustering is a vibrant area of research within the scientometric and informetric community. Applications include information retrieval, the mapping of science, and metrics to enrich studies of the individuals and institutions engaged in the research production process. Much of today's work, in line with Garfield's early vision, find citations and co-citations at the centre of their formulation of contextual similarity, even though that relationship may be more tenuous than generally accepted (see Borner et al. (2003) for an in-depth exploration).

Upon the digitisation of bibliographic data, researchers began to use text data to study and characterise scientific literature. Co-word analysis, pioneered by Callon et al. (1991), laid the foundations of today's full-text usage for document interpretation. In parallel, the first hybrid approaches began combining co-citation methods with word analysis to generate speciality clusters (Braam et al., 1991). As processing power increased, different sources of text data were incorporated to the tool-set of quantitative science studies. The work by Noyons and van Raan (1994) framed the science-technology link by using keywords from both patents and scientific articles. Soon thereafter, research employing a combination of classification terms, subject headings and keywords emerged. Later, Glenisson et al. (2005) showed the potential of full-text analysis to map scientific disciplines.

Increases in computational capacity and the availability of electronic data have opened many new avenues for estimating document similarity and have enabled clustering. While the range of options and ideas is vast, in this manuscript we focus on "Topic Models" — a group of techniques arising mainly from the computer science literature. As the input to these techniques is textual data (specifically, a collection of text documents), they offer an exciting twist on traditional approaches for understanding the topics and concepts that make up individual publications and, in turn, estimating document similarities and clustering. As discussed below, these techniques are certainly not without their flaws (Velden et al., 2017), but they are also well positioned to exploit the rapidly growing body of textual, and perhaps even full text, data.[1]

In this manuscript, we develop a robust approach for calculating pair-wise similarities between documents based on state-of-the-art topic modelling techniques. We compute the similarity between researchers which, in turn, allows us to obtain the topical overlap (or proximity) between them. With this text-only approach, we obtain a continuous knowledge domain space from which we can cluster and delineate topics as narrowly as desired, estimate interdisciplinarity, and observe the evolution and direction of research.[2]

[1] In addition to the issues discussed throughout this chapter, current topic modelling techniques also fail to exploit citation data. A gap we are working to fill with further work.

[2] Code for all the experiments in this chapter is available in https://github.com/oballegon/Thesis

This chapter is structured as follows. In the Section 1.2, we introduce the most common topic modelling techniques and their applications to information science. In Section 1.3, we describe the data and methods of our analysis. In Section 1.4, we describe the method of evaluation of topic models, including the training set-up and the metrics of interest. Section 1.5 discusses the results extensively and their limitations. Finally, in Section 1.6, we discuss potential applications of neural-network-based topic models, and provide examples of successful applications.

1.2 Background

Topic models are statistical models designed to extract from a set of documents the relevant "topics", and in turn, provide a representation of each document within that "topic" or latent space. More pragmatically, topic modelling consists on inferring a set of document-topic vectors (i.e., establishing the extent to which each topic pertains to each document) and a set of topic-term vectors (i.e., establishing the extent to which each topic is associated with each term) from a set of document-term vectors. In this task, a topic model will exploit hidden semantic structures within and across the documents. Because each document is treated as a bag-of-words, topic models cannot exploit local structure (i.e., grammar or the specific order of words within a sentence). Instead, they exploit the structure that emerges at the document level. For example, that the word "table" in the context of a document also containing the words "wood" and "legs" conveys a different meaning than the word "table" in a document containing "row" and "column". It is ultimately through the exploitation of high-level correlations in the co-occurrence of individual terms (as well as groups of terms) that the topic model produces its document-term and document-topic vectors.

In this manuscript, we will test topic models in terms of their ability to estimate pairwise document similarities robustly. Specifically we have chosen Non-negative Matrix Factorisation (NMF) (Lee and Seung, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) and paragraph embeddings (Le and Mikolov, 2014). NMF decomposes the document-term matrix into a product of two matrices, which by design may have only non-negative entries. LDA is based on a probabilistic model of language in which decomposition produces two matrices stochastically. Generating paragraph embeddings, and in particular Doc2Vec, is a relatively novel neural-network-based approach built upon the similarly new Word2Vec word-embedding algorithm by Mikolov et al. (2013). Word2Vec formulates the problem as one of predicting an omitted word within a short (3 to 15) contiguous sequence of text.[3] Treating the neural network's hidden layer as the latent space, one can infer document-topic couplings from the model's parameters. Although it should be noted that, strictly speaking, Doc2Vec may not be considered a "true" topic model as the topic-term couplings are not easily inferred. However, this feature (or lack thereof) is acceptable as the fundamental elements

[3] Thus this is not, strictly speaking, a bag-of-words approach. However, we have reproduced each analysis within this manuscript shuffling the order of the terms, and all results are similar. Although this *is* curious and begs further consideration at an appropriate moment in the future.

of the statistical analysis presented herein are document-document similarity scores, which require only the document-topic vectors.

A specific feature of most, if not all, topic models is that the user must define the size of the "topic" (or latent) space. Indeed the question of what is the "best" or optimal size of the topic space comes up often in the literature, and no clear criteria exist (Glaser et al., 2017). However, we use the capability of models to be trained for different topic sizes as a feature rather than a bug. Increasing the size of the latent space increases the granularity in which the model represents ideas. Considering all journal publications as the corpus, a topic space of size five would presumably decompose documents into the highest level disciplines one may think of (e.g. biomedical science, the physical sciences, social sciences or humanities). A latent space dimension of one or two hundred may decompose only well-established fields (e.g. medicine, molecular biology, physics, economics, sociology or history). However, allowing a topic space up into the thousands, or even tens of thousands of dimensions, allows one to identify particular groups of documents — for example, those focusing on a specific form of cancer within a specific model organism. Thus the question of what is the correct number of topics should never be asked, but rather, one should ask, what is the proper number of topics to tackle a specific question.

Despite the many new lines of research that could potentially be attacked by pushing topic models to high dimensional topic spaces, topic models are rarely employed with a latent dimension greater than, perhaps, a few dozen. The reluctance of researchers to use large topic spaces does not arise, however, from a lack of vision. Rather, it originates from a technical limitation. As one increases the number of latent dimensions, the model, eventually, becomes unstable. To be more specific, at some point the exact same algorithm, with the exact same parameters, on exactly the same data, but with a different random seed will produce a quantitatively and qualitatively different set of document-topic and topic-term vectors (Belford et al., 2017). In the topic modelling literature, a variety of information-theoretic measures have been proposed for estimating the extent to which topic-term vectors vary from run to run. However, it is indeed the case that changes in the topic-term vectors may *not* preclude stability when considering only document-document similarities. That is, even if the topics themselves are inconsistent from one training to the next, the measure of pairwise similarity may not change.

In this chapter, we test and compare the performance of NMF, LDA and Doc2Vec regarding their scalability to high dimensions and their consistency across estimations. We illustrate the analysis on scientific bibliographic data but the concepts, methods and implementation are extensive to any social-science research involving or using text data (e.g. patents, news articles, social media or discourse). In the following section, we provide a detailed description of the data and methods we use.

1.3 Data and Methods

Before getting into the analysis, we will define the specific data and context in which we are working. This section briefly introduces topic models, in particular LDA, NMF and a particular case of paragraph embeddings (derived from Word2Vec) as well as the text corpus of the subsequent analysis.

1.3.1 Data

Topic models are a subclass of dimensionality-reduction techniques which map a high-dimensional space of document-terms into a lower-dimensional latent space of document-topics. In the analysis below, a *document* is the career output of a researcher and the *terms* are Medical Subject Headings (MeSH). For each researcher, we extract from their publications all assigned MeSH. We rely on the 2014 version of PUBMED, which provides the individual publication metadata, from which we use Journal, Year, Document Type and Medical Subject Headings. We identify the output of each researcher from the Author-ity disambiguation of PubMed carried out by Torvik and Smalheiser (2009).

To be explicit, the document-term vector resulting from this procedure is one in which each vector entry corresponds to the concatenated list of MeSH terms assigned to the given researcher's publications across the entirety of her career. Disambiguation is thus crucial for the reliability of the models. We deal with careers starting 1974 or later, noting that our data terminate in 2009 as that is when the disambiguation ends.

Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the National Library of Medicine and used for indexing, cataloguing, and searching for biomedical and health-related information and documents.[4] MeSH terms are, in a way, keywords that point in the direction of the content of the full article. While the use of MeSH terms does not fully leverage the power of the most recent natural language processing (NLP) techniques, it provides a controlled and curated vocabulary, which largely simplifies the pre-processing stage. Given the nature of the experiment presented in this chapter, which aims at comparing the performance of different topic modelling techniques, the more restrictive vocabulary set plays to our advantage. First, it reduces the document-term matrix dimension, and second, the analysis is technically simplified by eliminating the burden of text tokenisation.

MeSH terms have several advantages for the application to topic modelling: (i) They are standardised (both in spelling and in scientific terms — i.e., use of the prevalent terms for neoplasms, or cell nomenclature, or technique names. (ii) MesH terms are harmonised. They suppress the common problems of free-text including synonyms, term permutations and

[4] https://www.nlm.nih.gov/mesh/introduction.html

acronyms all referring to the same subject. (iii) Unlike author-keywords, they are assigned by a centralised agency, reducing the self-selection bias. (iv) They are content-descriptors. (v) MeSH terms include multi-word tokens, increasing the specificity while providing a different token for distinct or narrower concepts — e.g.: "Cell" or "Stem Cell" or "Embryonic Stem Cell" or "Murine Stem Cell".

Therefore, MeSH terms represent a meticulously curated description produced by a third party (alien to the authors of the manuscript or journal). The result is a unified set of keywords that encapsulate the content of the article they epitomise, with enough precision to identify the precise knowledge that the publication covers.

On the other hand, one should not forget that the use of reference words only might hide nuances in the topic structure of a text. Full-text approaches include relevant phrases for interpretation and categorisation, thus bringing greater potential benefits than the "noise" introduced (Glenisson et al., 2005). State-of-the-art topic models, specially neural-network-based models, are particularly well suited for full-text analysis, given that they incorporate contextual information upon training (as opposed to bag-of-words approaches) and their ability to scale the data processing step Mikolov et al. (2013) efficiently. Hence, we acknowledge that, by limiting our experiment to MeSH terms, we are not fully exploiting the capabilities of modern tools. Our approach, however, reduces the complexity of the problem, and we can compare the different families of topic models. In Section 1.4.3, we provide a working example using full text from titles and abstracts.

Authors/Documents

In order to work with comparable, large enough documents, we filter out researchers with fewer than 50 research publications.[5] Our full filtered corpus comprises about 147,000 researchers. We then construct a heuristic rule, based on journal classification, in order to assign an *a priori* topic to each document. We do not use the rule-based classification in this chapter for any purpose other than sub-sampling the database, or establishing comparison groups. We do not use the labels as ground truth for topic-generation, nor we incorporate them in any classification pipeline. To classify journals, we use Eigenfactor's journal classification labels (Bergstrom et al., 2008) identifying the subject family of all our indexed journals. We then associate a subject family to a researcher whenever she satisfies one of the following: (1) if 50% of her publications appear in journals of the same subject matter, we assign one major topic label to her; (2) if the top N categories (by the count of publications) account for at least 20% of all the publications each, we assign the N categories; (3) if none of the previous holds and the top class has at least 15 publications, we assign the top class to the researcher.

In many analyses below, we focus on a subset of 13,936 researchers in the Neurosciences.

[5] Amongst the publications, we exclude non-research manuscripts. For this, we filter out the following Document Types as provided by PUBMED: "News","Review","Letter","Comment","Editorial","Historical Article","Biography","Portraits","Interview","Newspaper Article","Bibliography".

This choice was based purely on a desire to reduce the scale of the analysis. With this corpus, all pairwise similarity scores can be calculated in a manageable amount of time and stored within the hardware at our disposal. Thus, a researcher belongs to the Neurosciences if she satisfies one of the three rules above. The full list of 371 Neuroscience Journals can be found in Appendix A.1.

Corpus

The dataset for the central analysis performed below comprises 13,936 "documents" (researchers). Each document contains, on average, 940 non-unique tokens that come from an average of 120 distinct publications. Therefore, our corpus of publications contains MeSH from over 1.6 million scholarly articles. In turn, each publication contributes about eight MeSH terms randomly sorted upon introduction in the term list. Table 1.1 contains a summary of the corpus data. To prevent any undesired ordering effect at training, the array of

Table 1.1 – Corpus Statistics – Neuroscientists

	Mean	StD	Min	Max
Publications/Document	120.5	79.1	50	1111
MeSH/Document	942.7	624.8	225	9985
MeSH Incidence	594.5	2877.5	1	114306
	Count			
Documents	13936			
Unique MeSH	22909			

document-terms is stored for reuse in each of the subsequent model retrainings. We then perform all the training and analysis on the same static corpus, using the models described below.

1.3.2 NMF

Non-negative matrix factorisation (NMF) (Lee and Seung, 1999) is an approach to matrix decompositions that assumes positive features (components) in the data. NMF finds a decomposition of a matrix into two matrices, all of which have non-negative components. Since the problem is not exactly solvable in general, it is commonly approximated numerically. Due to the positive nature of its components, it is particularly well suited for representations of text and image, and the resulting matrices are easily interpretable.

Although NMF is not technically a class of probabilistic topic models, when it is obtained minimising the Kullback-Leibler divergence, the optimisation function is equivalent to probabilistic latent semantic indexing (Ding et al., 2008). Furthermore, standard applications rely on stochastic elements in their initialisation phase (Belford et al., 2017).

In our analysis, we use Python's implementation of NMF from the sci-kit learn library (Pedregosa et al., 2011).

1.3.3 LDA

Generative models for documents are an attempt at describing how words in documents can be generated from a set of "latent" (arbitrary) variables. Upon training, the models find the best set of latent variables that can explain the observed words present in the documents (Steyvers and Griffiths, 2013). The only information relevant to statistical models of documents is the number of times words appear. This data input is known as the bag-of-words assumption.

For many years, Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) has been the algorithm of choice for modelling latent topics. LDA is a class of probabilistic topic model using a Dirichlet prior distribution to the generative model. The most widespread interpretation of LDA is equivalent to a dimensionality reduction (matrix factorisation) interpretation. LDA allows to decompose the bag-of-words of a set of documents into a low-dimensional representation of topics. The output are two sparse matrices with probabilistic interpretation. The first is a topic-word matrix which associates each latent dimension to a distribution of words. The second is a document-topic matrix that provides the adherence of each document to the latent dimensions.

In our analysis, we use Python's implementation of LDA from the sci-kit learn library (Pedregosa et al., 2011).

1.3.4 Neural Network Embeddings: Word2Vec and Doc2Vec

Distributed representation of words as vectors have become a standard in natural language processing. Amongst them, neural network-based representations have gained a reputation of encoding many linguistic attributes. Amongst these techniques, perhaps the most widely known is Word2Vec (Mikolov et al., 2013). Word2Vec is an algorithm based on neural networks that generates word embeddings (word vectors) based on the word context. It is substantially different from the previous dimensionality reduction approaches, in that it does not try to reduce the space based on document co-occurrence (bag of words), but rather tries to extract the semantics (meaning) associated to each word based on its surrounding text. In essence, neural word embeddings are a class of semi-supervised algorithms. The training objective consists of learning vector representations that correctly predict nearby words. The output is a non-sparse vector that represents the word's position in context.

In an attempt to generalise the success of word embeddings to larger corpora of variable length, Le and Mikolov (2014) developed a paragraph embeddings model, based on the same architecture as Word2Vec. The paragraph token can be thought of as another word (it is trained at the same level of words), acting as a sort of "memory" of the missing context in the prediction of another word. Therefore, each paragraph vector is updated in the many "next

word" predictions that take place in each sentence of the paragraph.[6]

In our analysis, we use Doc2Vec. Doc2Vec is Python's implementation of paragraph embeddings by Řehůřek and Sojka (2010), which follows the architecture laid out by Le and Mikolov (2014). In our implementation, we use the Distributed Bag of Words (DBOW) approach with negative sampling, as suggested in Levy et al. (2015) and Dai et al. (2015).

1.4 Analysis

Different applications of topic models have been extensively used in the scientometrics literature to greater or lesser success (Boyack et al., 2011). Such models are useful for a variety of information needs, including analysis, information retrieval or data management. The most recent developments in topic delineation take advantage of stochastic topic modelling methods. That family of models includes NMF and LDA as the most use widespread applications (Glaser et al., 2017). Albeit general purpose in their conception, probabilistic topic models are not newcomers in science characterisation at different levels of aggregation. Griffiths and Steyvers (2004) were the first to apply LDA to article abstracts in order to find scientific topics and illustrate the contextual relationships between different disciplines. The work by Rosen-Zvi et al. (2010) and Lu and Wolfram (2012) applied stochastic topic modelling to infer the author-research relatedness, incorporating information from several documents from the same author. The Author-Conference-Model (ACT) by Tang et al. (2008), uses probabilistic topic models to infer author's and conference's subjects of interest simultaneously. Similarly, Hall et al. (2008) use LDA to study the dynamics of conferences topics.

In the particular case of Latent Dirichlet Allocation (LDA), Blei et al. (2010) found that, in practice, most practitioners directly assume that the latent spaces (topic space) generated by the model are semantically meaningful without an in-depth quantitative evaluation. Often, researchers interpret the topics generated as "themes" (Hall et al., 2008), and use a manifold of techniques to label the generated categories, ranging from human interpretation (Chang et al., 2009) to automatic classification based on topic top-words (Mei et al., 2007; Newman et al., 2010). In a review of the applications of LDA, Chang et al. (2009) argue that researchers use model fit metrics to account for the validity of the models, completely disregarding measurements of the internal representation. In practice, topic models face a trade-off between human interpretability and improved fitness metrics (such as likelihoods or predictive probabilities).

In reality, topic models are seldom used as an all-in-one out-of-the-box topic identification tool, but rather as an intermediate step to facilitate the human readability and interpretability of large text corpora. One particular task that has been facilitated by topic models is similarity analysis. The relative proximity between pairs of documents becomes relevant when we describe knowledge as a "space" or "landscape" of ideas. In this framework, first proposed by

[6] Sentence in the algorithm sense: the context window of words surrounding each word that are used for prediction. Paragraph embeddings can be trained for text of different lengths, ranging from a *real* sentence, a *real* paragraph or a full document.

Garfield et al. (1978), similarity is a distance metric of the spatial composition of knowledge. As a metric, pairwise distances (or similarities) can be used for multiple purposes in the fields of scientometrics, economics and information sciences: (i) clustering documents (Boyack et al., 2011; Racherla and Hu, 2010); (ii) measuring interdisciplinarity (Wagner et al., 2011); (iii) identifying emerging topics and novelty (Wang et al., 2017); (iv) characterising research sub-specialities (Azoulay et al., 2015a; Fontana et al., 2019); (v) mapping science (Suominen and Toivanen, 2016); (vi) or finding overlaps between groups of documents Yan et al. (2012). Besides, information retrieval and search tasks employ text-inferred similarities (Hjaltason and Samet, 2003; Castells et al., 2007)

Despite their widespread adoption and multiplicity of applications, stochastic topic models suffer from systemic errors due to topic instability (Belford et al., 2017). Belford et al. (2017) argue that variation is due to different local solutions to the optimisation problem under different stochastic initialisations. Recent work by Hecking and Leydesdorff (2018) has tested the validity and reproducibility of out-of-the-box LDA by fixing the stochastic parametrisation of the data. Their work concludes that, while topics (defined by their top words) are coherent, they are not robust to small data variations. Furthermore, Agrawal et al. (2018) show that LDA suffers from ordering effects. Different models are generated if the data are shuffled upon training highlighting the unreliability of such models. Our contribution goes beyond prior art in testing stochastic topic models in that we study the stability across runs. We propose an evaluation that is not data-dependent nor requires from "human" interpretation. Additionally, our assessment of stability can extend to neural-network-based models such as Word2Vec/Doc2Vec (discussed below) or others (GloVe, FastText, BERT).

To be suitable for application in the social sciences, it is our view that it must be demonstrated that topic models possess three properties:

1. Statistical robustness. That is, running the same model on the same data with the same parameters should produce the same, or at least highly similar results.

2. Descriptive power increases with the size of the latent dimension. That is, changing the number of topics should alter the results both qualitatively and quantitatively.

3. Reflect reality. That is, the results produced by topic modelling, be they document-document similarities or clustering or otherwise, must be consistent with patterns and relations are known to exist within and across research domains.

Below we propose and execute specific statistical tests concerning the first two, while for the third, we provide preliminary evidence and highlight paths for further work. We will discuss the results in Section 1.5.

1.4.1 Statistical Robustness

The first property that we study is the statistical robustness of the models. Here, we suggest an approach to testing model stability that relies on the spatial arrangement of the document vectors.

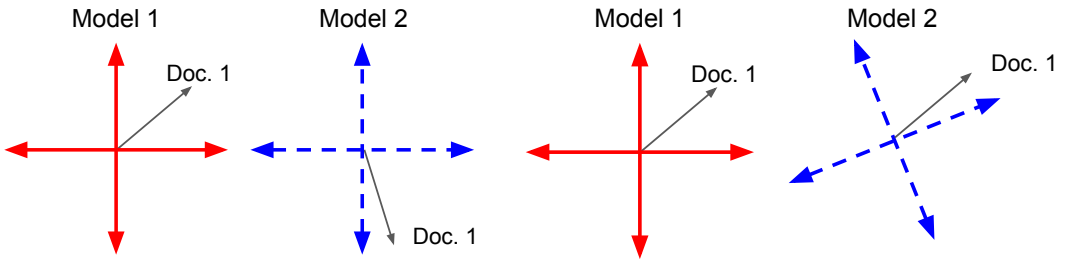
Given the non-linear nature of probabilistic topic models, there is no *a priori* ordering that makes the topics identifiable within runs of the algorithm (Steyvers and Griffiths, 2013). That means that, under each run, the space-vector is not necessarily the same, and the document representation is, therefore, difficult to compare. Figure 1.1 illustrates this effect by showing a 2D axis disposition of two different topic models. In each model, a different vector represents Document 1 due to the difference in the vector base that defines the latent space. For many applications, researchers have circumvented this problem by focusing on a single topic solution (see, e.g. Agrawal et al. (2018) and Steyvers and Griffiths (2013)).

However, it is necessary to know which topics are stable and not idiosyncratic to a particular solution. The majority of efforts to evaluate coherence in probabilistic topic models have concentrated around *expert evaluation* (Blei et al., 2010) or *entropy* metrics — i.e., information theory metrics closely linked to entropy, such as Pointwise Mutual Information (Velden et al., 2017), symmetrised Kullback-Leibler distance (Steyvers and Griffiths, 2013), Jensen-Shannon divergence (Boyack et al., 2011; Wagner et al., 2011) or Jaccard similarity (Agrawal et al., 2018). It is not uncommon that practitioners use a hybrid approach between the two methods such as comparing distributions over words manually (or automatically) selected from the most relevant for each topic (Boyack et al., 2011; Yan et al., 2012; Chang et al., 2009; Greene et al., 2014).

Cosine similarity as a metric

In light of the fact we can have these arbitrary transformations of the coordinate space, we require a different way of measuring agreement between models. We propose to evaluate the statistical robustness of a topic model via the extent to which it produces consistent estimates of pairwise document-document similarities. Being more specific, after retraining a model using the same parameters and data, one can track the mean and standard deviation of the cosine similarity of each pair of documents. If perfectly robust, a model would produce the same similarity each time, as conceptually represented in Figure 1.2. An imperfect, yet useful, model will produce slight variations in each pairwise similarity score, but over many retrainings, converge to a specific similarity value for each pair. Therefore, the use of inner pairwise comparisons eliminates the need for a *ground truth* or an *ad-hoc yardstick*, which is an important concern in the literature (Velden et al., 2017).

The use of cosine similarity (or, by extension, any Euclidean distance metric) enables the practitioner to incorporate to the comparison techniques beyond probabilistic topic models. In contrast, entropy metrics require the distributions to be probabilistic by nature. Despite



(a) Topic Model as-is (b) Rotatation of the latent in Model 2

Figure 1.1 – **2D spatial representation of topic models:** In theory these should agree, but the fact they do not, is not necessarily a problem. Because the axis is not necessarily the same. There can be rotations.

being commonly used when constructing distance metrics when applying topic models, cosine similarity has not, to the best of our knowledge, been used to study the robustness of topic models. Furthermore, cosine similarity is a more intuitive interpretation of distance in the context of a spatial representation of knowledge (Garfield et al., 1978) than entropy-based metrics.

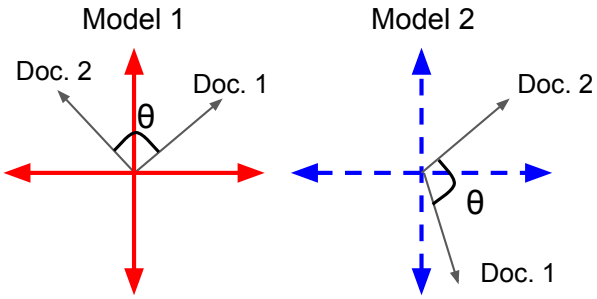


Figure 1.2 – **Measuring model robustness:** in light of the fact we can have these, essentially arbitrary transformations of the coordinate space, we require a different way of measuring agreement between models. We propose a cosine similarity.

The Experiment

Using the document-term corpus described in Section 1.3, we test the stability of LDA, NMF and paragraph embeddings generated by Doc2Vec. In order to better capture the potential systematic errors caused by the non-deterministic nature of the algorithms, we limit variability in the input to the maximum. First, the corpus is the same across all runs to enable comparison (Velden et al., 2017; Glaser et al., 2017; Klavans and Boyack, 2017), avoiding any deviation due to slightly different samples (Hecking and Leydesdorff, 2018). Second, not only data are the same, but the order in which they are fed to the algorithm in the training stage is the same (Agrawal et al., 2018). Finally, we use a fixed model tuning: all hyper-parameters are unchanged across runs with a fixed number of topics reducing the analysis of robustness to the

stochastic initialisation (Belford et al., 2017). Videlicet, two different runs of the same model with the same latent space size only differ in the random seed. In doing so, we expressly test for the variability that cannot be controlled for by the practitioner and examine the solutions for the idiosyncratic effects of random initialisation which, *a priori*, should not affect internal coherence of topic models.

As depicted in Figure 1.2, we expect that comparable models provide compatible representations of pairwise comparisons. Therefore, each document (researcher) i for each retraining k is represented by the vector d_{ik} where $k = 1, ..., K$ and K the total number of retrainings. We then compute the pairwise cosine similarity $s_{ijK} = \cos(d_{ik}, d_{jk})$ for each of the K retrainings (the only difference being the random initialisation). Subsequently, we compute the average similarity $\bar{s}_{ijK}$ and the standard deviation σ_{ijK} associated. For each pair of documents we have:

$$\bar{s}_{ijK} = \frac{1}{K}\sum_{k}^{K} s_{ijK} \tag{1.1}$$

and

$$\sigma_{ijK} = \left[\frac{1}{K}\sum_{k}^{K}\left[s_{ijK} - \bar{s}_{ijK}\right]^2\right]^{\frac{1}{2}} \tag{1.2}$$

Figure 1.3 shows the behaviour of a specific researcher-researcher similarity score produced by NMF, LDA and Doc2Vec under $K = 1, 2, 3...50$ retrainings. The X-axis represents the number of retrainings (k) included in the calculation of the average cosine similarity. The Y-axis is the average cosine similarity ($\bar{s}$) across the k models with the standard deviation as a confidence interval. The confidence interval for each calculation corresponds to the standard deviation. First note that each of the three topic models produces a different similarity score, despite having the same number of topics (50 in (b) and 100 in (a)). Second, and most importantly, note that the average similarity score for NMF and LDA has a far larger standard deviation of results than Doc2Vec. That is, they display weaker convergence than Doc2Vec. It is indeed the case that this figure is representative of the behaviour of the three models across all researcher-researcher pairs as well as a wide range of latent space sizes.[7] While LDA and NMF show a lack of convergence to a *true value*, Doc2Vec converges rapidly and accurately to a single value, independent of differing random seed. While this convergence is no guarantee of a better classification or closer-to-reality distance measure, it is a replicable measurement that warrants further analysis. Similar figures that compare the average similarity across the three models for many different number of topics (10, 25, 250 and 400) can be found in Appendix A.2. For consistency, all the figures represent the same pair of documents.

Next, we extend the analysis to measure all the document pairs and propose a general robustness metric based on the dispersion of similarities.

[7] Although at smaller latent space sizes, around 10, NMF and LDA will also converge perfectly well

Generalising the robustness metric

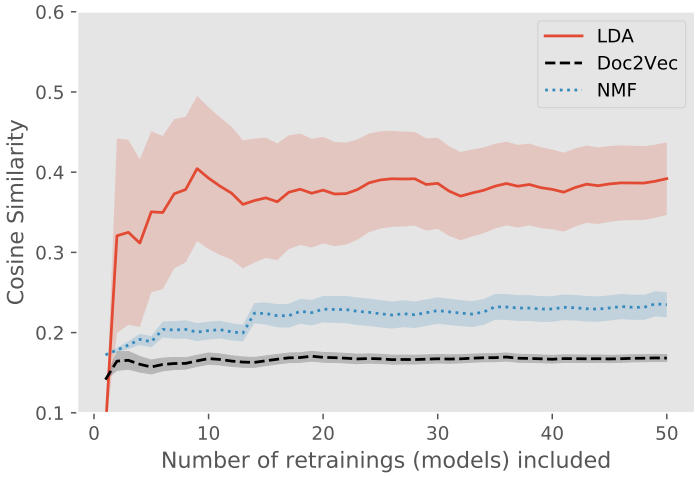
In order to measure the relative convergence of different retrainings to a pairwise similarity value, we need to take into account the entire corpus. The marginal contribution to the average similarity $\bar{s}_{ijK}$ decreases with K (i.e., with each additional model incorporated), smoothing out the variation. The smoothing might generate a false perception of convergence to a similarity value between a pair of documents. This effect is particularly well illustrated in Figure 1.3 (b). Thus, in order to compare the variation within each set of pairwise similarities, we resort to the standard deviation σ_{ijK}. In other words, in order to study the stability of multiple retrainings, we measure how broad the distribution of similarities produced for the same pair of documents is.

For each training k, we compute the similarity matrix M_k of all the pairwise similarities s_{ijk}. It is then possible to calculate the average similarity $\bar{s}_{ijK}$ and the standard deviations σ_{ijK}. We finally obtain the average standard deviation across all the pairs. A schematic representation of the experiment is represented in Figure 1.4. Our generalised robustness metric is the asymptotic (with K) average standard deviation, calculated from all the unique ($i \neq j$ and $i < j$) pairwise similarities:

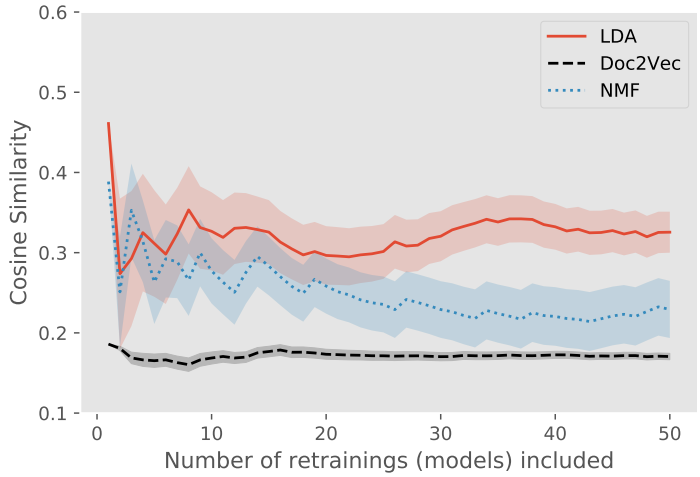
$$\Phi_K = \frac{1}{C_{(N,2)}} \sum_{i}^{N} \sum_{j>i}^{N} \sigma_{ijK} \tag{1.3}$$

where $C_{(N,2)} = \binom{N}{2}$ is the binomial coefficient of all the possible unique pairs with N documents.

The asymptotic standard deviation, Φ_K, is a comparison metric, which helps us determine the variation between retrainings of the pairwise similarities. Large values indicate more dispersion of the average similarity for each pair of documents. The more robust a topic model is to retrainings, the lower the value of Φ_K. There is not an absolute value that denotes goodness of fit (robustness) nor a general benchmark to compare to. Rather, it is a relative metric, which will depend on the data, but that can be used to determine which model introduces fewer spurious similarities.



(a) Latent space size (number of topics) 50



(b) Latent space size (number of topics) 100

Figure 1.3 – **Pairwise Cosine Similarity across 50 model runs**: This figure shows the average cosine similarities between a randomly selected pair of documents across multiple retrainings of the same model. The only difference between two retrainings is the random seed. The X-axis represents the number of retrainings (k) included in the calculation. The Y-axis is the average cosine similarity ($\bar{s}$) across the k models with the standard deviation as a confidence interval.

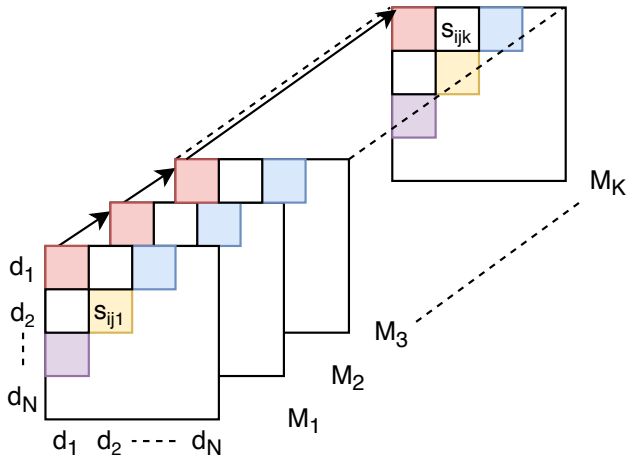


Figure 1.4 – **Schematic representation of the generalised experiment** We compute K similarity matrices M_k for each document pair. We then average the similarity for a given pair across the K retrainings and calculate the variation (standard deviation) for the pairwise similarity. Finally, we obtain the average standard deviation as an indicator of model stability.

1.4.2 Descriptive Power

One key parameter that practitioners should take into consideration when applying topic models is the size of the latent dimension. That is the target number of topics. Albeit subject to other interpretations, it usually varies to the likely use-case. Small latent spaces tend to give very broad generalisations, and larger sizes provide a greater level of detail. Often, if the granularity is pushed too high, the topics start to degrade into incoherent nonsense (Hecking and Leydesdorff, 2018; Greene et al., 2014; Steyvers and Griffiths, 2013). These parameters are also highly dependent on the size of the data and the variety inherently present in the text.

Determining the optimal number of topics has been the locus of research for multiple articles. In computational linguistics, Blei et al. (2003b) and Rosen-Zvi et al. (2010) propose to use perplexity (goodness of fit) to measure the generalisability across the different number of topics. Blei et al. (2003a) and Teh et al. (2006) resort to Bayesian statistics to automatically select the number of topics. In information science, the work to optimise the selection of topics has included topic-term stability metrics (Greene et al., 2014), and human-machine judgement tests (Chang et al., 2009). There is congruity amongst scholars from different knowledge areas that there is, however, a trade-off between interpretability and predictive power. As a result, as topics become increasingly fine-grained, they improve their predictive likelihood but become less useful for human interpretation (Chang et al., 2009).

In light of these results, we advocate for basing the latent space size selection depending on the objectives and subsequent tasks. As discussed above, a prevalent real-world task for topic models is to determine pairwise distances. In turn, pairwise distances can too be a good metric to evaluate topic size (Hecking and Leydesdorff, 2018). This comparison is, however, highly dependent on both the data characteristics and the research objectives and should vary on a case-by-case basis. Here, we take a different path. Rather than selecting the number of topics, we suggest a way of comparing the descriptive power as a function of the number of topics. Therefore, we can confront retrainings of the same model (i.e., changing hyperparameters), different models (e.g. NMF vs LDA), and assess the information gained by including additional dimensions.

To get a handle on the explanatory power of Doc2Vec, LDA or NMF (or any topic modelling approach), we propose a straightforward procedure based on principal component analysis (PCA). In this approach, we carry out factor decomposition (PCA) on the document-topic vectors (researcher-topic vectors in this instance). The algorithm then reorders the principal components explained variance, and we plot their cumulative explained variance.

The PCA explained variance plot allows us to understand the extent to which each dimension allows differentiation among documents vis-à-vis the latent space. For example, a perfectly straight line running from the lower left to the upper right would indicate that each dimension contributes equally to explaining the variation among researchers, and hence, allow for the differentiation between researchers. On the other hand, a curve that quickly reaches 1.0, perhaps after only $k < K$ dimensions, indicates that only those first k dimensions are contributing to

explaining the variance. This result can, indeed, be obtained for NMF, LDA, and especially Doc2Vec for various specific parameters. In other words, all dimensions beyond the first k do not add useful information. Thus, the explanatory power of a given topic model can be measured by the area over the curve (AOC) in such an explained variance plot.

1.4.3 Concordance with reality

Properly reflecting reality is, of course, the most important criterion for generating an abstract representation of data. It is often also the most difficult one, however. Since the literature is filled with use-cases for probabilistic matrix factorisation models, we focus on validating the neural network approach that we are using in this chapter. Here we provide three small examples as evidence that, at the very least, the results do not directly oppose expectations. For this, we train document embedding vectors for different data and compare the output in different scenarios.

Real-world communities

First, we demonstrate the ability of document embeddings to generate communities that reflect their real-world associations. To this end, we train a single Doc2Vec model on an extended corpus. We construct a researcher-term document following the same procedure as for the rest of the chapter — delineated in Section 1.3. This time, instead of limiting the number of documents to the 13,936 neuroscientists, we include the 147,000 researchers from all biomedical sub-specialities with at least 50 relevant publications in the database. We generate the document-term structure following the same steps as before, the only difference being the size of the corpus. Upon training, the sub-field speciality labels are never fed into the models in any way.[8] In particular, we train for a latent space size of 400, and we project the document embeddings into two dimensions using t-SNE van der Maaten and Hinton (2008).[9] In Figure 1.5 a (2D) t-SNE projection of the researcher embeddings is shown colour-coded by the sub-field speciality. From visual inspection, it is clear that, by and large, researchers of the same field cluster together, even if a 2D representation is limiting even though there are clearly some outliers and room for further investigation and/or refinement.

Going forward, we are pursuing two main avenues of analysis for evaluating the extent to which document similarities produced by Doc2Vec reflect reality. For these tests, we will use different levels of data aggregation to represent the documents.

[8] The speciality labels include, amongst others: Medicine, Molecular and Cell Biology, Dermatology, Radiology, Orthopedics, Dentistry or Obstetrics)

[9] t-SNE is a tool to visualise high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimise the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

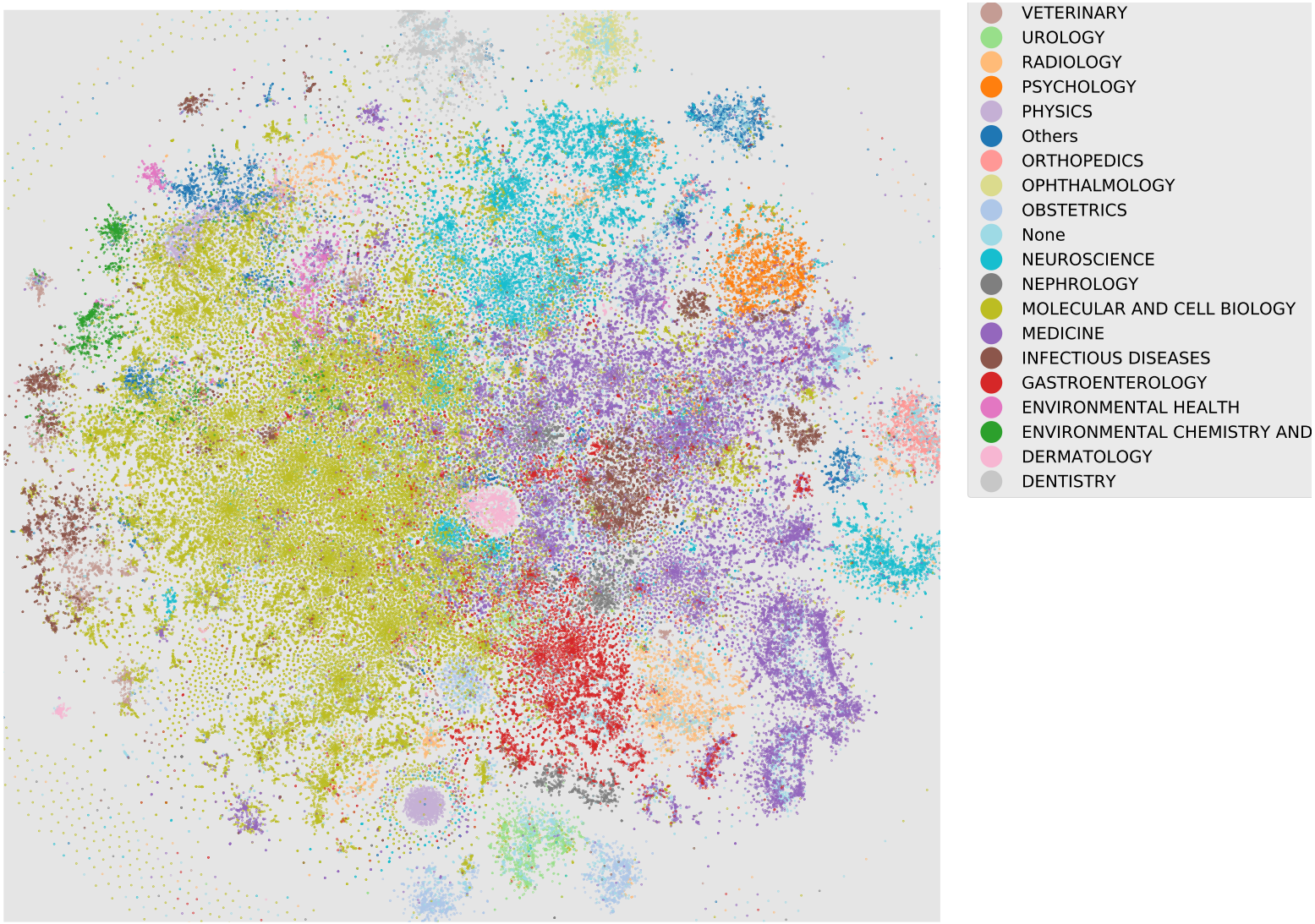


Figure 1.5 – **t-SNE projection in 2D of researcher embeddings**: Each point is a single researcher and colour indicates the researcher's field (the field in which the majority of his or her papers appeared).

Most similar to:		Score
Mamm Genome 2009	Mamm Genome 2008	0.794
	Mamm Genome 2007	0.789
	Mamm Genome 2006	0.787
	Mamm Genome 2005	0.728
	J Anim Breed Genet 2008	0.663
	PLoS Genet 2005	0.644
Curr Microbiol 1998	Curr Microbiol 2000	0.871
	Curr Microbiol 1997	0.858
	Curr Microbiol 1999	0.83
	Arch Microbiol 1997	0.787
	Arch Microbiol 1999	0.783
	FEMS Microbiol Lett 2001	0.781

Most similar to:		Score
J Hand Surg Br 1991	J Hand Surg Am 1990	0.88
	J Hand Surg Br 1993	0.874
	J Hand Surg Br 1989	0.872
	J Hand Surg Br 1990	0.87
	J Hand Surg Am 1991	0.862
	Handchir Mikrochir Plast Chir 1990	0.852
Adv Neurol 2002	Curr Opin Neurol 1999	0.744
	Neurol Clin 2002	0.743
	Curr Opin Neurol 2000	0.714
	Curr Opin Neurol 2003	0.7
	Curr Opin Neurol 2004	0.699
	Rev Neurol (Paris) 2004	0.676

Table 1.2 – Top-6 documents by similarity and scores to a representative random sample. Trained with Doc2Vec using journal-year as documents.

Journal Pairwise Similarity

The first avenue involves using external information to identify pairs of documents that should be highly similar and valid on the Doc2Vec based similarity measures. Therefore, we construct a data set of Journal-Year documents using MeSH Terms as the document content. We characterise a Journal-Year document as the compilation of Medical Subject Headings (MeSH) published in a given periodical throughout a year. That is, we compile the MeSH terms for each article that appeared in the same journal during a year grouping them in one single "document". That list of MeSH terms represents that Journal-Year.

We subsequently train a Doc2Vec topic model with Journal-Year documents ranging from 1985 to 2010 and generate inferred document embeddings (vectors) from 1985 to 2014. The corpus comprises almost the entirety of our in-house PUBMED database, with over 55,000 Journal-Year documents. Following the same visualisation methodology presented above, we plot the resulting embeddings. A 2D projection of these embeddings can be found in the Appendix B in Figure B.7. The figure shows how multiple clusters emerge "naturally" from the data, but more importantly, we observe how the same journals in consecutive years show high similarity. In Table 1.2, we display the most similar journals-years and the similarity score to four randomly selected documents.

Document Retrieval

For the second similarity test, we propose a document retrieval exercise. We train a Doc2Vec model on free text (Title and Abstract) of 2 million publications in the biomedical sciences. Each document is constituted by the concatenation of Title and Abstract words of a single publication. We extract data from Web Of Science and PUBMED. The 2 million publications constitute the majority of indexed articles (simultaneously in WoS and PUBMED) between

1997 and 2005 categorised as Journal Articles (not Editorials, News nor Reviews). In the pre-processing stage, we remove stopwords, homogenise the text to lowercase-only, stem the words and construct bigrams, trigrams and 4-grams using a pairwise mutual information statistic. For training, we use a latent space of size 200 and learn Word2Vec embeddings simultaneously.

In parallel, we extract 1885 articles from the reference list of 69 hand-picked review articles about human embryonic stem cells (hESC). We classify the sample of articles into two groups: hESC-related and non-hESC related. To do so, we implement a rigorous keyword rule. In particular, we include in the hESC list any document containing at least one of the following stemmed tokens:[10] "hESC", "human embryonic stem cell", "human ES", "human ES cell", "he cell". Additionally, we include any of the following in combination with the presence of "Humans" amongst the associated MeSH terms or the presence of "human" in the same paragraph (but not right next to them): "blastocyst", "embryon stem cell", "embryon (ES) cell", "ES cell", "embryon stem (ES) cell". This process yields 808 articles out of the 1885 that we will consider as hESC-related.

Using the trained model, we infer document embeddings for the hand-picked 1885 articles. It is worth noting that some escape the 1997–2005 period, and have not been part of the training process. We then compute cosine similarities between each of the article's embeddings and the trained word embedding for the 4-gram "human embryonic stem cell". Figure 1.6 shows the density of similarities between the 1885 articles and the selected embedding, split by keyword classification.

Despite a quasi-out-of-the-box approach to training, including over 2 million documents from multiple disciplines in the biomedical sciences, the training provides a reasonable separation between the two sub-sets of documents. Notably, two samples that had been originally selected from a list of review articles on a particular subject. This example provides compelling evidence that similarity scores reflect reality to a good extent.

1.5 Results

In this section, we discuss the results of the analysis presented above. First, we provide an overview of the robustness metric. We compare the performance of the three analysed models and motivate a second round of analysis that excludes the least-similar pairs of documents. Second, we comment on the scalability of the three models.

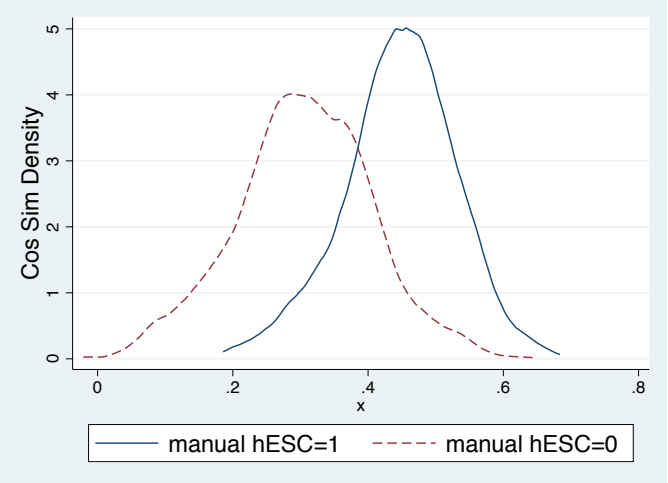


Figure 1.6 – **Kernel Density Plot of Cosine Similarity** between root articles and word embedding of "human embryonic stem cell". The blue line represents the keyword-rule-assigned hESC articles, with which we perform the core analysis of this chapter. The red line represents the non-hESC articles according to the keyword classification.

1.5.1 Robustness

Doc2Vec provides robust and multi-purpose topic models that overcome the main difficulties encountered by stochastic matrix factorisation (LDA or NMF). Figures 1.7,1.8 & 1.9 display the asymptotic convergence of Φ_K for $K = 75, 100$ trained on LDA, Doc2Vec and NMF respectively on our corpus. Consisting of $N = 13,936$ documents, the number of observations on each cosine similarity matrix $C_{(N,2)} = 97,099,080$. The results show that LDA is significantly less robust than NMF and Doc2Vec for smaller sizes of the latent space — Φ_K for LDA for 25 topics is three times larger than Doc2Vec and four times larger than NMF. NMF behaves opposite with a good robustness performance for lower dimensions until its internal coherence "breaks"— Φ_K stays constant for 25 to 50 topics, but doubles from 100 to 250 topics. Doc2vec seems to perform consistently across different topic sizes, with a decrease in Φ_K as the number of topics increases. This decrease is also observed for LDA, suggesting that dispersion occurs gradually. As the number of dimensions grows, the degrees of freedom increase and the vectors representing each document in the latent space disperse. Therefore, there should be a larger concentration of close-to-zero similarities which would, in turn, decrease the average standard deviation Φ_K.

These results are in accordance with our expectation for two reasons. First, as granularity increases, the changes in the retrainings will be absorbed along more dimensions, smoothing the dispersion in the pairwise similarities. Second, following each increment of the latent space, the distribution of pairwise similarities for any given document should systematically

[10] We Stem the words to their roots. We capture words such as *embryonal* or *embryonic* under the same token *embryo* will be captured by embryo. The bigram, trigram and 4-gram pre-processing allow us to have tokens such as "embryon stem cell" as a single word

approach zero, thereby reducing the room for variation. Besides, probabilistic models generate document-topic sparse matrices, whose sparsity will increase with the number of dimensions. This effect is not necessarily true for neural networks, which follows from the *density* of the output matrices.

It is true for many applications of topic models that the practitioner will be interested in the *most similar* documents. The degree of relatedness in information retrieval or the distances for clustering applications, to name two examples, require that the higher similarity values be well defined. In order to reduce the effect of increased granularity and test for larger cosine similarity values, we perform a similar analysis to that of Figures 1.7,1.8 & 1.9 filtering out low similarities. That is, Equation 1.3 takes now the following shape:

$$\Phi_K = \frac{1}{C^*} \sum_{i}^{N} \sum_{j>i}^{N} \sigma_{ijK} \cdot \delta_{ij} \tag{1.4}$$

where C^* is the number of observations left after filtering and:

$$\delta_{ij} = \begin{cases} 0 & \text{if} \quad s_{ijk} < \epsilon \quad \forall k \\ 1 & \text{if} \quad \exists s_{ijk} \geq \epsilon \end{cases} \tag{1.5}$$

where ϵ is the filtering value. That is, we calculate the robustness Φ_K taking into account only the pairwise similarities above ϵ it at least one of the $k = 1,...K$ similarity matrices (each for one retraining of the same model). The results are summarised in Figure 1.10. The top left figure, corresponding to an $\epsilon = 0$ summarises in one single plot the findings presented in Figures 1.7,1.8 & 1.9. Figure 1.10 allows us to grasp better the asymptotic behaviour of the two stochastic topic models and the neural-network approach. Without any filtering, we find evidence of a "breaking" point in both LDA and NMF, after which the model becomes significantly less robust. For the corpus in hand, for LDA, this happens for small-sized latent spaces. As the number of topics grows, the performance resembles that of Doc2Vec. For NMF, it evolves in the opposite direction. Yet, as we calculate the dispersion including only the pairwise similarities that, under at least one retraining, are larger than ϵ, the stochastic topic models' robustness degenerates. Consistently, as ϵ increases, LDA and NMF show higher average dispersion of the pairwise similarities. Doc2vec, on the other hand, performs at a similar level independent of the filter. Figures with model-by-model asymptotic behaviour after filtering can be found in Appendix A.3

In light of these results, one should be very careful at the tasks performed subsequent to topic modelling with stochastic dimensionality reduction methods. In the particular case of retrieving ranks or lists of "most similar" documents, LDA and NMF display considerable variation. Therefore, it is plausible that resulting documents are idiosyncratic to a specific training.

Comparison with other metrics

It seems evident that probabilistic topic models are more unstable than document embeddings. However, for our analysis, we have expressly omitted the discussion over other model parameters. Optimising any of the tested models to a given task escapes the purpose of this analysis. Nonetheless, we now provide succinct evidence that the models from our training would be considered robust under "traditional" metrics. In the benefit of conciseness, we display results for LDA only (as discussed, the most extensively used topic modelling technique).

Following standard practice laid out by Steyvers and Griffiths (2013), we study dissimilarity between pairs of topics in different runs. We take two estimated topic-word distributions — from two different retrainings — and compute the dissimilarity between topic i form the first model (t_i^1) and topic j from the second model (t_j^2). We measure dissimilarity as the Jensen-Shannon distance between the two probability distributions, t_i^1 and t_j^2. The topics of the second model are then re-ordered to correspond as best as possible (lowest Jensen-Shannon distance) with the topics of the first run using a greedy (by brute force) algorithm. The similarity matrix in Figure 1.11 (a) suggests that a large percentage of topics contain similar distributions over words. Additionally, for the pair of models, we display the top words of the two most similar and most dissimilar topics after pairing. That is: after matching topics from one model and the other, the pair with the lowest and highest JS distance respectively. The top-ten list of words is displayed in Figure 1.11. In all cases, the solutions from different models give slightly different results, but in practice, the two models would be considered as equivalent and stable across runs.

For consistency, we replicate this analysis for 25, 50 and 250 topics. The results are displayed in Appendix A.4.

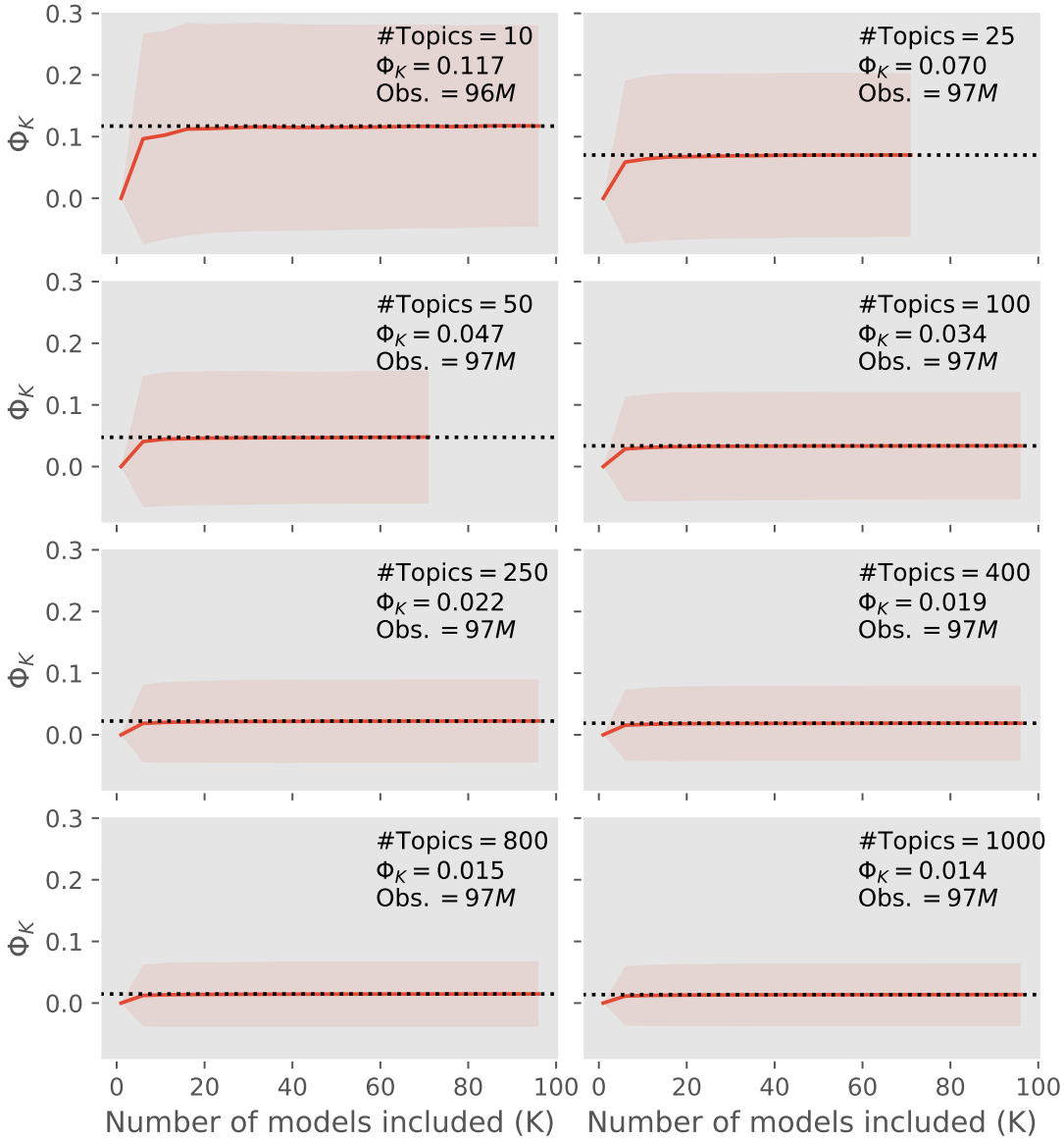


Figure 1.7 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions

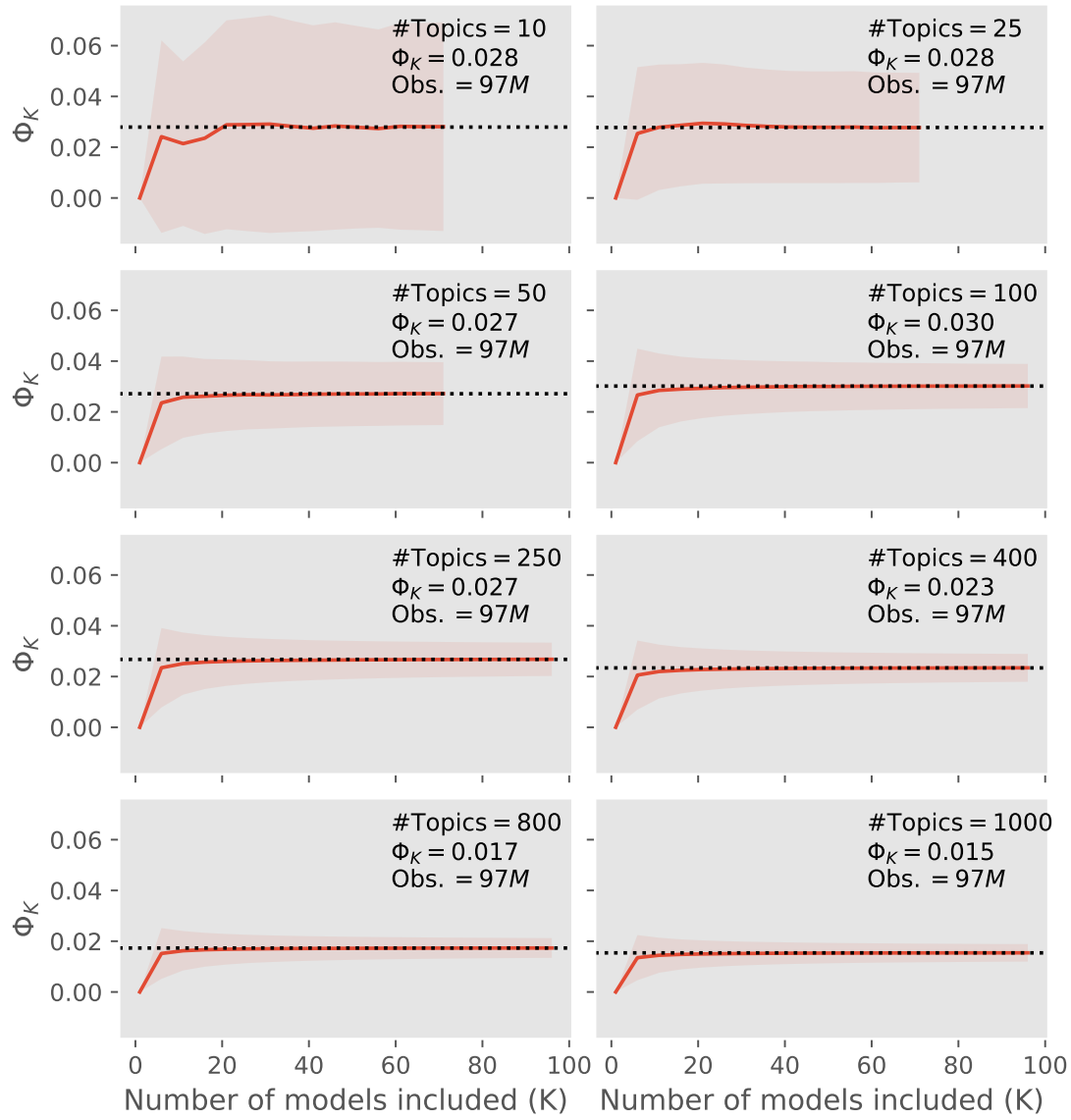


Figure 1.8 – **Average Standard Deviation for Doc2Vec:** Asymptotic value over multiple retrainings (75 or 100) of Doc2Vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions

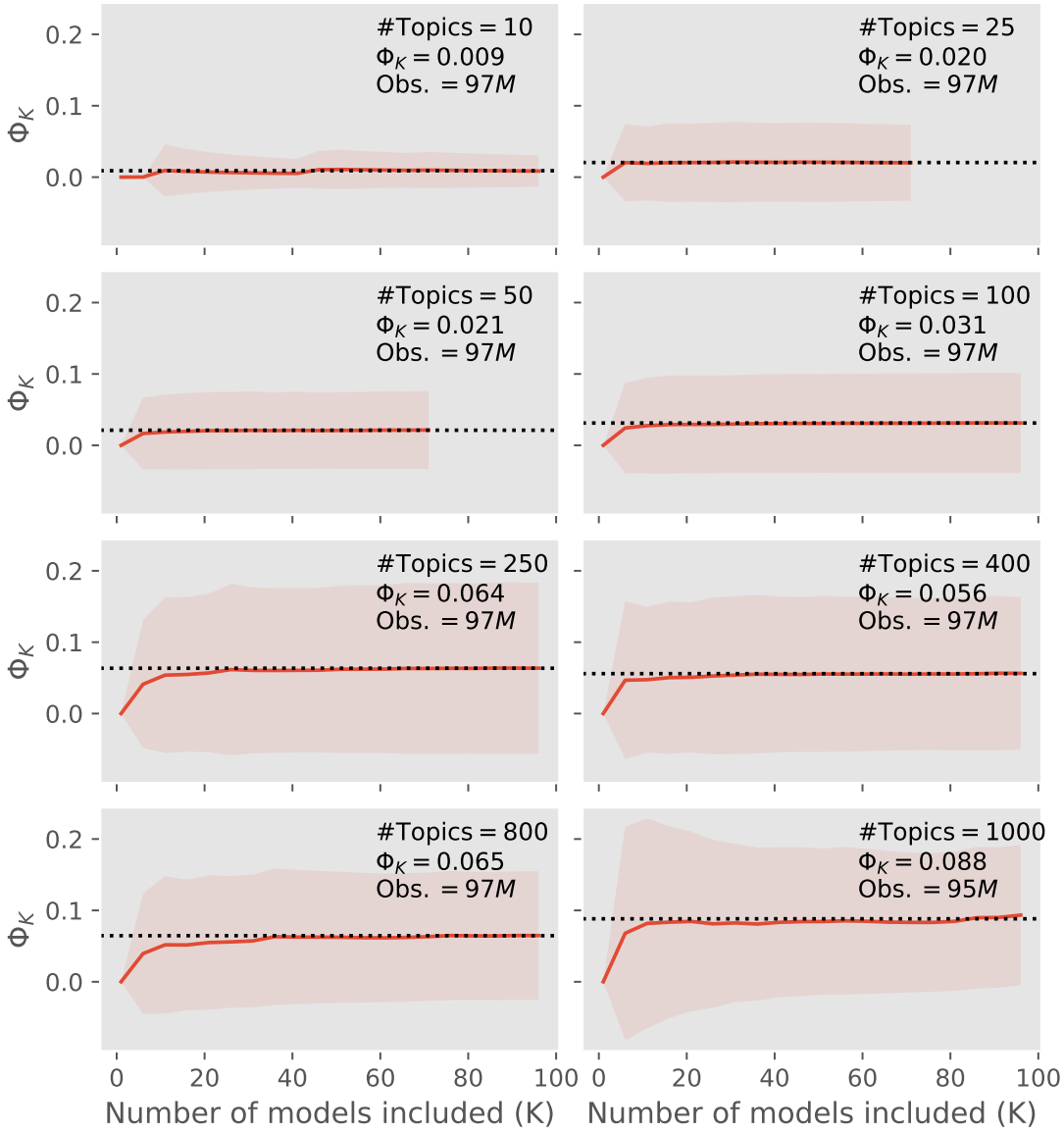


Figure 1.9 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions

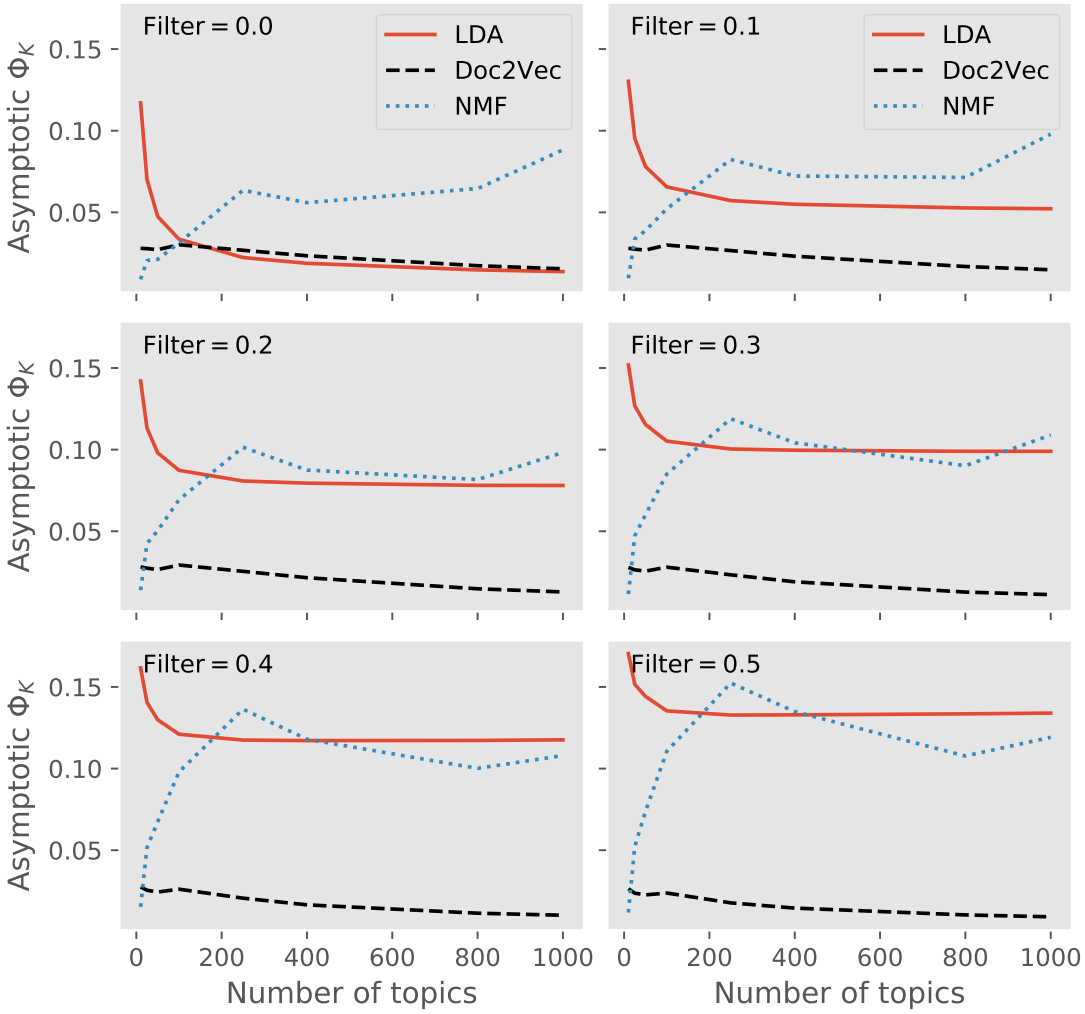
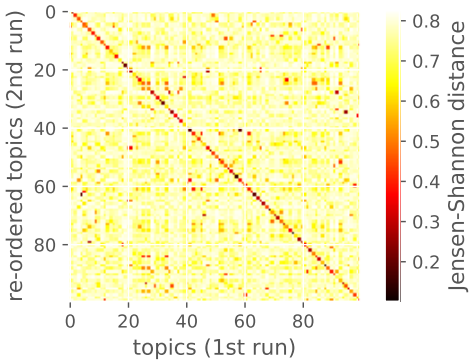


Figure 1.10 – **Asymptotic Average Standard Deviation comparison:** Asymptotic value over multiple retrainings as a function of the number of topics. 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions. Left-to-right and top-to-bottom, each figure displays Φ_K calculated after filtering out cosine similarities lower than $\epsilon = 0, 0.1, 0.2, 0.3, 0.4$ and 0.5 respectively.



(a) Similarity of topic-terms 1 and 2 (reordered)

	Model 1	Model 2
Most sim topic	Genotype	Genotype
	Polymorphism, Genetic	Polymorphism, Genetic
	Genetic Predisposition to Disease	Genetic Predisposition to Disease
	Gene Frequency	Gene Frequency
	Alleles	Alleles
	Polymorphism, Single Nucleotide	Polymorphism, Single Nucleotide
	Case-Control Studies	Case-Control Studies
	Haplotypes	Haplotypes
	Genetic Linkage	Schizophrenia
	Linkage Disequilibrium	Genetic Linkage
Least sim topic	Alcoholism	Ethanol
	Alcohol Drinking	Alcoholism
	Ethanol	Conditioning, Operant
	India	Alcohol Drinking
	Chiroptera	Cocaine
	Liver	Substance Withdrawal Syndrome
	Aldehyde Dehydrogenase	Self Administration
	Temperance	Behavior, Animal
	Acetaldehyde	Reinforcement Schedule
	Digoxin	Central Nervous System Depressants

(b) Top 10 words between most and least similar pair

Figure 1.11 – **"Traditional" stability of topics between different runs of LDA with 100 topics**

1.5.2 Scalability

Another important aspect is the scalability. Matrix factorisation approaches do not handle well large latent spaces (into the hundreds of topics) nor big increases in data (Ai et al., 2016). As we observe, changes in the latent space size also reconfigure the topic-word relationships, introducing noise in the similarities between researchers. Doc2Vec provides consistent results across different sizes of the latent space, with only marginally decreasing values of similarity with an increasing number of dimensions, as one would expect. Different latent space sizes allow for varying levels of granularity in the topic models, and thus, a more fine-grained clustering of topics or words.

Figure 1.12 shows the explained variance of each dimension in a PCA-transformed space for multiple different topic sizes and two different ways of text input. The speed at which the different models converge to 1 (the slope of the lines) explains the incremental gain of information of each additional dimension. For our particular corpus, albeit unevenly, both LDA and Doc2Vec carry descriptive power across the different dimensions under analysis. NMF, on the contrary, rapidly approaches 100% of the explained variance when trained with a large number of topics (250, 400, 800 and 100 in our analysis). Similar to the results presented in Figure 1.9, NMF models lose descriptive power relative to LDA and Doc2Vec as the latent space size increases, eventually "breaking" in the models above 100 topics.[11]

To discuss one particular example, for 250 topics we see that Doc2Vec is the closest to the diagonal, albeit far from overlapping. Still, a good 25% of the variance does reside in the last 100 topics after the transformation. On the other hand, we see that NMF models reach 100% of variance explained within 100 dimensions, which suggests there is no real advantage of training larger models.

Visualising the PCA transformation does not single-handedly provide evidence of a *better* model. Rather, it provides a test for the marginal increase in descriptive power by expanding the latent space. In other words, we abstain from declaring Doc2Vec superior to LDA solely from a smaller area under the curve in the explained-variance-ratio plots. We do, however, argue that NMF models trained under this particular corpus, provide no significant granularity gains for topic spaces larger than 150 topics. In general, fine-tuning the model parameters allows to decrease the area beneath the curves, attaining a significant gain in all the dimensions of the model. In other words, the models can scale in the number of dimensions carrying information at all times, allowing for a different optimisation depending on the objective of the training.

[11] For our data

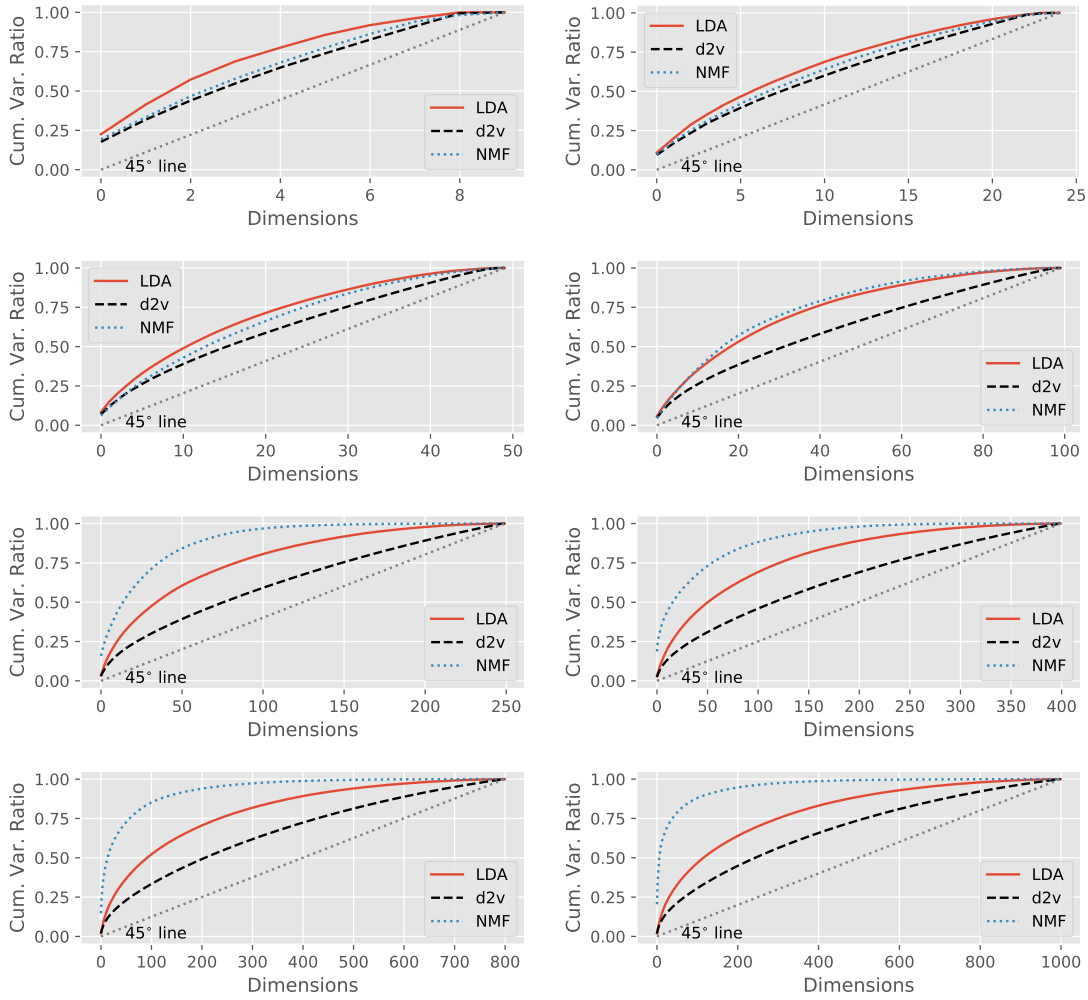


Figure 1.12 – **PCA-transformation of document-transformed vectors** Cumulative explained variance ratio of the transformed document (researcher) vectors, sorted by decreasing explained variance. The 45° line represents a model in which each dimension has the same descriptive power. The greater the Area Over the Curve (AOC), the greater the average descriptive power of each dimension.

1.6 Discussion and Limitations

In this chapter, we have focused on comparing the performance of topic models from two different classes: dimensionality reduction and neural networks. As a representation of the first family, we have analysed one of the simplest, NMF, and the most used LDA. From the second family, we have provided evidence of document characterisation using Doc2Vec, a particular case of paragraph embeddings that arises from Word2Vec (the most extensively used word-embedding model). On a general level, there is an overlap in the representation of documents delivered by either approach (probabilistic or neural-network-based). A vector represents every document in a "knowledge space", rather than in a discrete classification of topics. However, due to the sparse and positive components of the topic-term and document-topic matrices in topic models, LDA and NMF provide a more humanly-interpretable output, that enables classification tasks (topic labelling). On the other hand, neural network embeddings form a continuous space that is less readily explainable.

The comparison of models provided in this chapter is, to the best of our knowledge, the first effort to account for the idiosyncratic variability of solutions that escape the practitioner's analysis. We fix the data and hyper-parameters for several retrainings, allowing only the stochastic initialisation to vary across runs. In order to compare their performance, we devised a robustness metric based on pairwise similarity. Furthermore, we do this for different levels of granularity. This system allows us to measure the extent to which each model would produce arbitrary relatedness (false positives) in the association (or disassociation) of documents.

Up until this point, we have provided evidence that neural-network based topic models, in particular Doc2Vec, provide a more reliable — replicable — characterisation of documents. In turn, Doc2Vec is capable of scaling (increase the number of topics) without losing descriptive power at any point. Additionally, we provide evidence of different levels of support of the three models for different ranges of similarities — i.e., we show that for LDA and NMF robustness decreases as the pairwise similarity increases. Our results suggest that solutions obtained with stochastic dimensionality reduction methods are, on many occasions, contingent to a particular training. Even when traditional coherence metrics and ground-truth tests support the models, the optimality is frequently incidental. However, this analysis precludes us from claiming the superiority of one particular solution over the others. One would ideally study the complete set hyper-parameters and data-sample combinations to determine the resolution of a solution. Instead, our analysis provides practitioners with additional tools to evaluate their particular applications. Additionally, we provide a quantitative measure to the well-known issue of "instability" of probabilistic topic models.

These observations raise one question: should we then turn our backs on probabilistic models in favour of document embeddings? Not necessarily. First, neural-network methods are rather data-thirsty. That is, for a good embedding characterisation, one needs large amounts of data (Mikolov et al., 2013). Under these circumstances, they outperform probabilistic methods both in computational speed and robustness. However, for small document-term matrices,

convergence to a workable solution is rarely achieved. In this case, dimensionality reduction techniques (either of a probabilistic nature, such as NMF, LDA or pLSI) or factor analysis (such as PCA) might be more suitable and accurate. Second, neural-network methods do not provide humanly-readable topics. In contrast to "traditional" topic models, there is no *top-words* output that makes neural network embeddings intelligible. Thus, for classification or labelling tasks, additional work is required, over-complicating a task that simpler methods may still return. Proper understanding of the data and task requirements, along with the properties of each model, is more likely to provide the best results.

The numerous drawbacks of LDA (or NMF) should not automatically thwart previous efforts in their applications. This chapter raises concerns related to the reproducibility and sensibleness of topic modelling techniques relying on stochastic initialisation. However, we cannot generalise the findings to conclude that all topic representations be untoward. Depending on the data characteristics (corpus size, variability, document size), traditional topic modelling techniques might suffer from fewer issues than we highlight here. In turn, adequate analysis of the topic space can ensure that the local optima reached by the algorithm truthfully represents reality, which ultimately is the objective. As long as prior art has been carefully designed and tested, one should not distrust the outcome, but rather question the applicability to similar scenarios and aim for additional tests that prove that idiosyncratic errors are averted.

More work is needed to prove concordance with reality of neural network embeddings. Many of these solutions are still very recent and need more testing before they gain momentum in the scientometrics community and the social sciences in general. We have provided three rough examples that should provide a sanity check for the conclusions extracted in the rest of the analysis. We show how they can work in different contexts, at various aggregation levels and for miscellaneous tasks. Ultimately, the quality and performance of these tasks improve with parameter tuning, which we have overlooked in this chapter.

The explicit omission of parameter tuning is not the only limitation of the analysis above. Next, we discuss some of these limitations, such as data dependence and left-out approaches.

1.6.1 Limitations

As with any empirical study, we cannot discard any biases in the analysis. The conclusions made from this chapter must be considered with the following issues in mind: potential misalignment with a ground truth, sampling bias errors and model omission biases. We explain each of them in turn.

Ground Truth

A first and essential limitation to the analysis proposed above lies on the lack of a benchmark data set that provides a comparison of topic extraction results. Nevertheless, the evolution of research involving prior topic modelling techniques suggests that reflection of reality tasks

will be amongst the first to be tested by the community. While still a nascent area of interest, the work by Banerjee et al. (2018) provides a first successful (contrasted) effort in using word embeddings for information categorisation tasks; Ai et al. (2016) employ paragraph embeddings for information retrieval in short texts; and Thijs (2019) provides evidence of Doc2Vec-generated similarity between scholarly article sections, following *a priori* rational expectations. In a document retrieval exercise, Dai et al. (2015) show evidence of the superiority of paragraph embeddings over LDA and a static bag-of-words (no topic modelling) approach on a set of arXiv publications.

Sampling Bias

Second, we tested our results with only one curated dataset. Using a compilation of MeSH terms has multiple advantages for our testing procedure: (i) we avoid any pre-processing bias, (ii) we work with a curated set of words, (iii) the corpus size is restricted. Additionally, MeSH lists are freely available for any other researcher to replicate this work. The MeSH term list approach to the corpus construction could parallel an intensive keyword-extraction and stop-word removal that one would typically perform in pre-processing stages. However, these advantages are offset by a lack of comparison with other samples. Furthermore, it raises questions concerning to what extent the results are particular to the characteristics of the dataset.

We train the model on relatively large documents (943 terms on average). Paragraph embeddings tend to overfit when documents are too short (Ai et al., 2016). Our corpus is at a sweet spot in which probabilistic topic models can still be applied, and neural-network document embeddings have enough data to converge in the training stage. For a smaller corpus, training neural networks becomes more challenging, which compromises the conclusions of this work. We could see a potentially better performance from traditional topic modelling methods. For shorter documents, however, it has been shown that word embedding aggregations (Word2Vec) can represent short paragraphs adequately (Boom et al., 2016). That leads us to the next limitation, concerning the limited set of models we have put to the test.

Model Omission Bias

This chapter has only compared NMF, LDA and paragraph embeddings generated with Doc2Vec. The choice of models illustrates the most accessible (present in most software packages) and widely used topic modelling techniques. Nevertheless, we have left out other models that are commonly used and have demonstrated the ability to overcome some of the problems discussed above. Amongst the "traditional" topic models, there have been numerous developments which we have not tested here, such as Pairwise-Mutual Information (PMI), probabilistic Latent Semantic Indexing (pLSI), Latent Semantic Analysis (LSA). Many hybrid methods have been developed to complement probabilistic topic models in search of stability (Agrawal et al., 2018; Belford et al., 2017; Velden et al., 2017). On the other hand,

there are other models amongst the "new breed" of neural-network-based models. FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014) and Word2Vec (from where Doc2Vec stems) are other word (and sub-word) embedding models of the same generation (log-bilinear prediction-based semi-supervised models that generate static embeddings). Again, the choice of reporting Doc2Vec results lies in two factors: accessibility to training Doc2Vec and the length of the documents in our corpus.

Recent work has leveraged word embeddings to generate representations of sentences and small paragraphs. Most notably, researchers have used the algebraic properties of the latent spaces to represent documents. Socher et al. (2013) show how averaging the constituting word embeddings improves the performance on pairwise relationships, Garten et al. (2015) compare averaging and concatenating embeddings and Boom et al. (2016) suggest aggregating embeddings using their inverse document frequency (idf) weights.

For completeness, we tested two of these methods on our central corpus. We trained Word2Vec embeddings on the set of 13,936 researchers (documents) from the neurosciences. We then generated the document embeddings as the unweighted mean of all the embeddings present in a document (reported) and as the idf-weighted mean of the embeddings (unreported). In both cases, the results were very similar and consistent with our expectations. Although more testing would be required to generalise the results adequately, the robustness metrics (Φ_K) provide additional support for the use of neural-network-based models (as opposed to probabilistic) as shown in Figure 1.13. The pairwise-similarity variation across retrainings is much lower than stochastic models. However, neither aggregation procedure (mean and weighted-average) supported scaling. Hence, averaged word embeddings subsequently result in a loss of granularity which impedes segmentation tasks between documents. For these data, the distribution of similarities using word embeddings aggregations is dense and concentrated around very high values. Although further testing is required, we believe this loss of descriptive power is due to the length of the document combined with the aggregation procedure. Without retraining — i.e. without training the model in shorter text — we generated document embeddings to represent each researcher using only MeSH terms from one publication. This short text scales better than longer samples, as shown in Figure 1.14.

Previous results have provided evidence in the same direction. In a (long) document retrieval exercise, Dai et al. (2015) show that averaging word embeddings did not provide substantial accuracy gains, increasing the number of dimensions. On the contrary, LDA and Paragraph Embeddings did, before reaching a breaking point. Overall, paragraph embeddings displayed the best accuracy.

Together, these results provide suggestive evidence that, if adequately trained, neural network-based embeddings may provide a good document representation under different data. Given the relative novelty of neural network topic models, in the next section, we conclude by discussing some of the applications of document embeddings generated with neural networks.

1.6.2 Recommendations for testing topic models

In this subsection, we provide some guidelines to help practitioners test their particular applications of topic models. This framework builds upon the findings of this chapter and the works cited throughout.

First, researchers should allocate their efforts to the choice of the topic model. This involves taking into account the needs (interpretation and use of the outcome) as well as the data in hand for training. Large data corpora allow for neural network approaches, while smaller datasets will be at odds with such techniques. In this case, traditional topic modelling techniques might be more efficient. Second, we recommend that practitioners study which latent space size best fits their objectives and train numerous instances of the chosen model to determine the range of hyperparameters that produce models that adequately reflect reality. At this point, ad-hoc reflection-of-reality tests are sufficient to determine whether one is in the right direction — e.g. distribution of similarities for a set of documents, similarity ranks, top-word analysis, meta-data comparison, or any heuristic rule that may apply to the data in hand. These tests should be sufficiently quick to enable rapid pivoting and re-testing other implementations of the model. This stage should also allow the practitioner to put in perspective the choice of model, and assess the validity of the output — e.g. Paragraph Embeddings work for large corpora, but relatively short documents tend to produce highly concentrated similarity distributions for all documents, making classification tasks more difficult while word embeddings with vector averaging work better. In addition, we recommend the PCA-transformation test presented in this chapter to examine the usefulness of the chosen latent space size.

Once the model and preliminary hyperparameters have been adjusted, we recommend that practitioners test how small variations affect the outcome of the tasks (clustering, high/low similarity scores, rankings, topic relatedness) to perform by the topic models. As we show in this chapter, the level of variation between two retrainings is largely affected by the similarity threshold. Therefore, one should pay particular attention at this validation stage, since the replication of results is largely dependent on the idiosyncratic nature of the model's output. Models that display small variation in the tasks they are expected to perform should be preferred.

It is worth pointing out that there is nothing inherently wrong with a particular solution (topic model) that suffers from large variation upon retraining if practitioners acknowledge this feature and are exceptionally cautious in their analysis of the reflection of reality. That is, for a particular dataset, the model can lead (by accident or intent) to the optimal solution (best reflection of reality) and the outcome be valid for any subsequent application. However, given highly unstable models, one must be decidedly circumspect in the validation of the representation of reality, to eliminate any doubts.

Finally, researchers should perform the final tuning and more in-depth reality checks, to ensure that the output is consistent with what we know to be true from the documents in the

corpus.

1.6.3 Some Applications of Neural Embeddings

In this chapter, we have quantitatively assessed the robustness and scalability of recent developments in topic modelling and compared them to the prior art. Document embeddings have not yet been fully incorporated in the information science researcher tool-set. We hope that our work will shed new light into the capabilities, strengths and weaknesses of neural-network-based models. To conclude, we discuss some potential applications of non-probabilistic word and document embeddings for science characterisation and evaluation.

In Section 1.4, we show some straightforward tasks performed by document embeddings. The first consists of visualisation of knowledge domains. The vectors representing each document equip the practitioner with a high dimensional spatial representation of a knowledge domain. Traditional visualisation techniques (Borner et al., 2003) can be applied directly to dense vector representations. In our example, we reduce the space to two dimensions using t-SNE, a method that preserves some of the multi-dimensional clusters through the projection.

Our second example, built around journal-year embeddings, suggests that community detection is possible through word embeddings. The vector representation caters euclidean distance metrics between pairs. Adding a layer of analysis, clustering methods that rely on distance metrics can be applied to automatically detect communities of documents (researchers, journals, publications) in the data. In Chapter 2, we apply Agglomerative Clustering on journal-year embeddings to generate communities of journals.

In our last reality-check, we suggest validating the model through retrieved and labelled documents. By reversing the experiment, document embeddings can be used for document retrieval. Using distance metrics as a relatedness measure (to a subject), it becomes possible to determine similarity rankings and relevance.

Beyond these simple use cases, the embeddings can accomplish more intricate tasks. For example, it is possible to study the dynamic properties of the knowledge space. Lenz and Winker (2020) show how it is possible to measure the diffusion of innovations through paragraph embeddings. Following a similar analysis, it is possible to characterise *ex post* novelty, hot topics and disappearing topics through topic models. In Figure 1.15, we display the t-SNE projection of Neuroscientists (our corpus of 13936 researchers) by 5-year window. We see knowledge areas dissipating and new concentration points emerging. Additionally, neural network models are not fixed entities that need complete retraining for every application. It is possible to update, or, in other words, to sequentially train the models. Successive training would allow for a dynamic study of vocabulary, or even to study the convergence/divergence of disciplines through time.

Another form of analysis of document embeddings is to measure the coherence or variety within a group of documents. In Chapter 3 we use document embeddings trained on free text to

measure the narrowing focus of a field of study. Similarly, Ayoubi et al. (2020) use embeddings to measure the disparity between past and present grant applications of researchers. Using these methods, it would be possible, for instance, to characterise the distancing process between a researcher and her mentor. Or the knowledge fit between peers in mobility events.

Finally, we are confident that future lines of work in scientometrics will exploit other geometric properties of embeddings. The geometric properties of embeddings that transform semantic meaning through addition and subtraction of vectors present an exciting avenue of research. With adequate data and careful training, it might be possible to subtract (or add) the contribution of, for example, a coauthor to a publication. Similarly, if researchers are capable of disentangling the direction of methodologies, one could potentially divide empirical and theoretical contributions in a publication. Bolukbasi et al. (2016) are capable of removing gender bias from word embedding representation by finding the *direction* in which gender is expressed in the latent space. Similar approaches could yield significant advances in the characterisation of scientific contributions.

1.7 Conclusion

Topic models are powerful statistical techniques with great potential to contribute to scientometrics, especially as textual data become more available going forward. However, they also suffer from specific flaws that must be carefully weighed against the benefits. In particular, establishing statistical robustness is challenging, evaluating their explanatory power is critical, and ultimately verifying they reflect reality is necessary.

In this manuscript, we have proposed a simple approach for estimating the statistical robustness of topic models that is based on pairwise similarity scores between documents. Applying that method, we found that Non-negative Matrix Factorisation does not appear to be exceptionally robust for large latent spaces (dimension far greater than 10). Latent Dirichlet Allocation is comparatively robust amongst the distant pairs. However, its instability rapidly increases as one evaluates more closely related pairs. Doc2Vec, a neural network-based approach does, on the other hand, appear to produce relatively stable estimates of pairwise similarity across all scenarios.

We further proposed a principal component analysis based approach for assessing the descriptive power of topic models. Applying that method to researcher-topic vectors, we find that, while they do not produce perfect results, LDA and Doc2Vec explanatory power does persist into the highest dimensions of the latent space. On the contrary, NMF maintains only meaningful descriptive power in the lower dimensions.

In terms of the extent to which Doc2Vec results reflect reality, many questions remain. We provided three small pieces of evidence that what Doc2Vec produces is not entirely out of bounds. However, careful quantitative validation is still required.

The analysis presented in this chapter, thus, provides ground for the application of neural embeddings approaches in the social sciences. The remainder of this thesis provides a vivid example of paragraph and word embeddings in practice. In Chapter 2 we use journal-year embeddings to delineate scientific fields. In Chapter 3, we construct an alternative treatment-control sample for an econometric analysis using a combination of word and paragraph embeddings. In addition, we use publication similarity metrics, generated from paragraph embeddings, to characterise the disparity in topics within a set of publications.

Acknowledgements

Preliminary versions of this chapter have been presented at the Global Tech Mining Conference (Atlanta, 2019), the IV Summer School in Science and Technology Indicators (KUL-EPO, Vienna, 2019), International Conference on Scientometrics and Informetrics (Rome, 2019) and EPFL Workshop on Computational Methods in Social Science (2019). I wish to thank the anonymous comments and remarks received from referees and attendants, who have played a role in the improvement of the analysis presented above. A working-paper version of this chapter was published as part of the ISSI 2019 proceedings (Ballester and Penner, 2019).

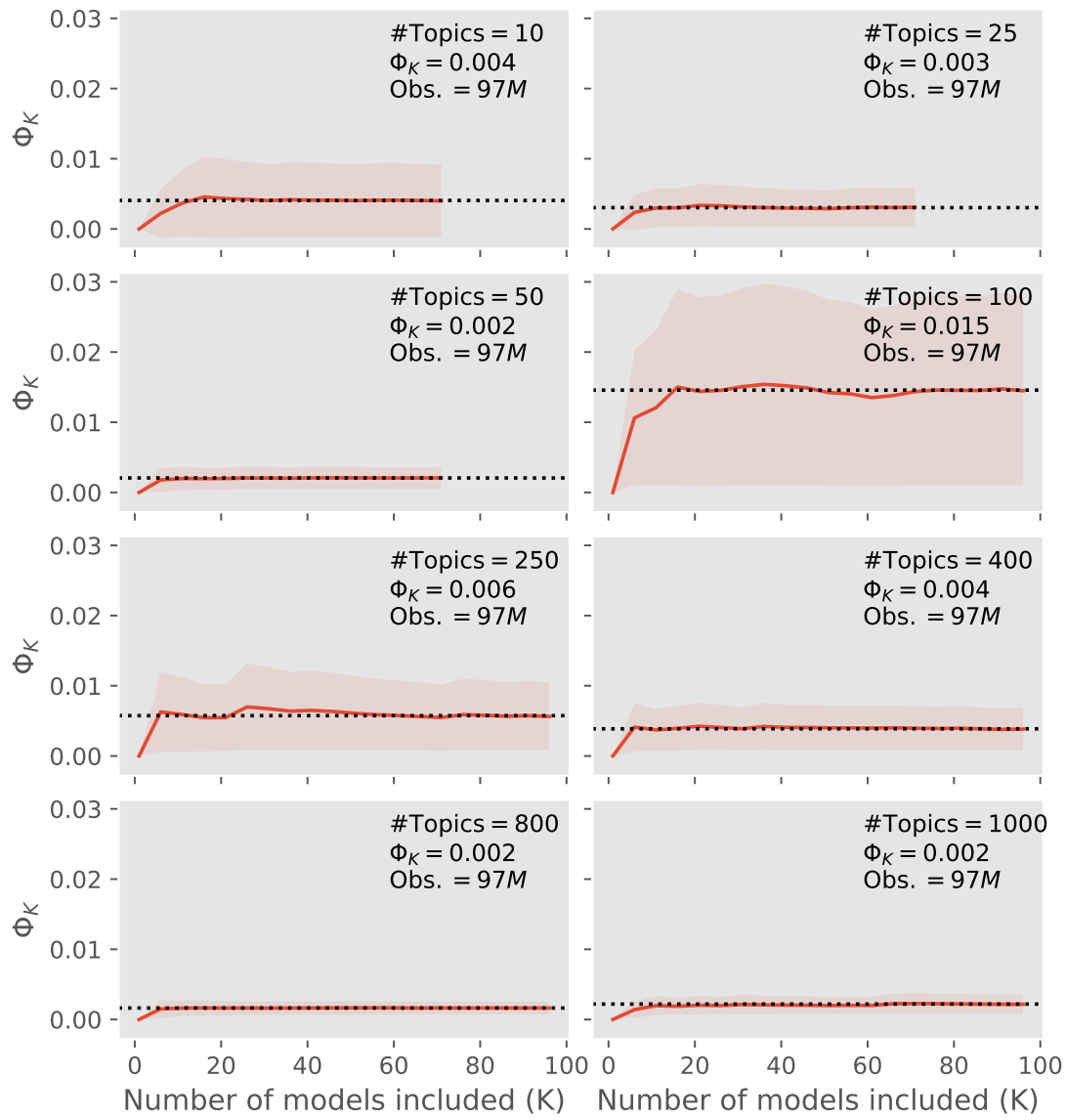


Figure 1.13 – **Average Standard Deviation for Word2Vec (Averaged word embeddings):** Asymptotic value over multiple retrainings (75 or 100) of w2v for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions

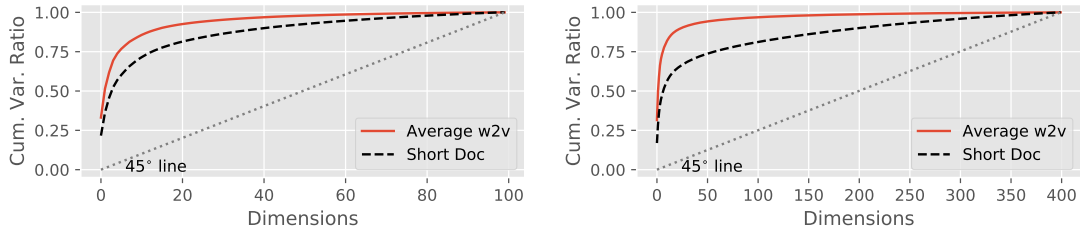


Figure 1.14 – **PCA-transformation of document-transformed vectors** Cumulative explained variance ratio of the transformed document (researcher) vectors, sorted by decreasing explained variance. The 45° line represents a model in which each dimension has the same descriptive power. The greater the Area Over the Curve (AOC), the greater the average descriptive power of each dimension.

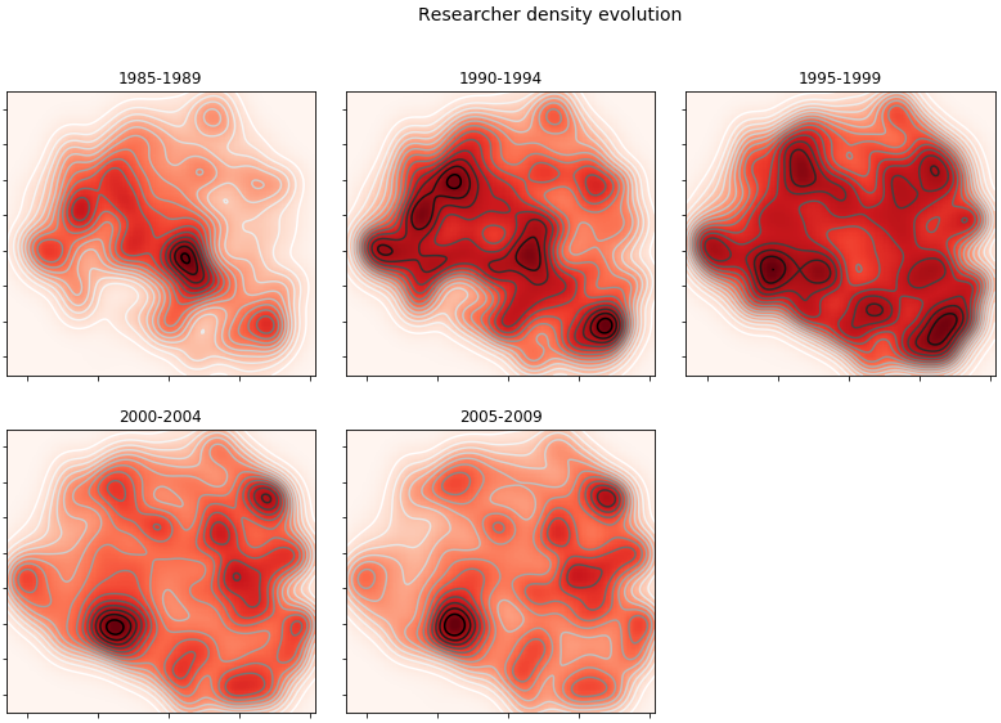


Figure 1.15 – **Dynamic t-sne representation of Neuroscientists**: Concentration of researchers in the 2D knowledge space by year (2D Kernel Density).

2 Rivalry in science: Modelling science as a CPR game.

"There is an enormous difference between the creative work of the genius and the monograph of a specialist. Yet in the field of empirical research it is possible to cling to this fiction. The great innovator and the simple routinist resort in their investigations to the same technical methods of research. (...) The outward appearance of their work is the same. Their publications refer to the same subjects and problems. They are commensurable."

— Ludwig von Mises (Human Action)

Reputation plays a key role in determining the allocation of rewards among scientists. Alongside individual ability, reputation creates a strong path dependency in the trajectory of individual careers. We present a theoretical model in which the rewards of scientific production captured by an individual are proportional to the magnitude of that individual's contribution. Our model deviates from the classical approach in which science is treated as a pure *public good*. We argue that this point of view may help explain for the heterogeneity in scientific production we observe among peers arising from their career choices. Specifically we model a researcher's payoff as a *common-pool resource* game, intrinsically connecting the appropriability of scientific output to a scientist's optimal strategy.

This simple model of reward allocation sheds new light on a variety of behaviours that have been observed amongst researchers. In particular, actions often attributed to social networks and community effects, but that have been measured only approximately.

2.1 Introduction

Economic growth arises from productivity growth, which is, in turn, driven by technology. The link between economic growth and technological advance has been long understood in the literature (Romer, 1986; Aghion and Howitt, 1992). However, recent macroeconomic data suggest a slowdown in the productivity of ideas. The reason behind this seems to be that ideas are harder to find than ever before (Bloom et al., 2020; Jones, 2009). The global trend of *production* of scientists has been steadily increasing for decades (Stephan, 2012), so the observed dynamic becomes both a question of rate and direction of scientific inquiry.

The economics literature on individual incentives and rewards for scientific production is vast. In recent years a steady stream of applied work has studied a variety of reward schemes intercepting reputation and status. These often study the trade-offs researchers face: explore *vs* exploit (Azoulay et al., 2011), specialist *vs* generalist (Teodoridis et al., 2018) or optimal strategies in scientific collaboration (Bikard et al., 2015). The discovery race sets a competitive environment that has some benefits, such as the efficient allocation of scientific effort amongst problems (Hagstrom, 1974) and also some important downsides, e.g. when researchers stray into fraud (Azoulay et al., 2015b; Jin et al., 2013). On the other hand, economists have somewhat overlooked the dynamics and *direction* of scientific ventures, a quest targeted by philosophers, historians and sociologists (Azoulay et al., 2015a).[1] In their seminal paper *The new economics of science*, Dasgupta and David (1994) suggest that interactions in science could be modelled as a game, an idea that Kealey and Ricketts (2014) and Kiri et al. (2018) exploit.

In this chapter, we model science as a local quasi-public good introducing appropriability, partially challenging the assumption of non-rivalry in basic science. The beneficiaries of science production are the actors at the heart of that process — i.e. the contributing scientists themselves. We assume there are no free-riders in *common-pool resource* games. Within this framework, we find a plausible explanation for the direction of science given the level of congestion of a field, as well as the incentives to work on a new field. We also provide a framework under which there exist barriers to entry and competition amongst researchers in the space of ideas.

The remainder of this chapter is structured as follows. First, we set up the ground for the model, laying out the prior literature that precedes this chapter. We then provide some macro and micro empirical evidence reinforcing the idea that researchers' strategic behaviour plays an active role in the organisation of science. In Section 2.3 we present the model. First, we characterise the problem as a CPR game and derive the fundamental implications. Next, we introduce heterogeneity in the players with two illustrative cases. Finally, we discuss how the model links to a longstanding literature in the sociology of science and discuss the model's many limitations.

[1] Except for, perhaps, the theoretical work of Jones (2009) and Bramoulle and Saint-Paul (2010).

2.2 Background

The organisational characteristics of science

The rewards scientists receive for producing science arise both from creating a particular new piece of knowledge *and* from the public recognition that they were the person to do so (Stephan, 2012). Similar to innovations, scientific discoveries are a first-come-first-serve notoriety monopoly. In order to increase reputation, one must first *win* a discovery race, as priority is the basis by which one can legitimately claim her contribution (Dasgupta and David, 1994). This fact is well established within the sociology of science, where disputes over priority, and the incentives arising thereof, are understood to play a central role in the organisation of science (Merton and Storer, 1973).

The study of the direction of science has been part of the sociology literature for a long time. Upon studying research communities, Crane (1969) identified leaders within a community, who had a powerful influence in the direction of the field. These highly visible individuals were considered the intellectual leaders by their peers. Crane's findings helped explain why knowledge in a field follows a Pareto distribution, with the majority of contributions corresponding to a well-knit circle of resources, literature and academics in orbit (Swanson, 1966). While old, these ideas are "en vogue" in economics. Numerous recent publications point in the direction of few influential researchers guiding (or signalling) prevalent research directions (Azoulay et al., 2010, 2015a; Agrawal et al., 2017; Higgins et al., 2011; Oettl, 2012). Others have suggested full independence in the choice of a research agenda as a key characteristic of academic research (Aghion et al., 2008). In either case, what really drives scientific developments are creative ideas. Creativity leads to exploration, and the exploration of new avenues of research is paramount for the advancement at the frontier (Azoulay et al., 2011).

Along these lines, Azoulay et al. (2011) tested empirically how the existence of the right incentives plays a crucial role in the creativity of researchers. They show how a funding mechanism (HHMI) organised around the principal investigator is more effective at promoting riskier behaviour than a project-oriented program like NIH.[2] The work of Manso (2011) for the case of innovation and Bramoulle and Saint-Paul (2010) for the case of science provide a framework in which a scheme of incentives helps explain the trade-off between exploration and exploitation (March, 1991). The study of this dichotomy in basic sciences traces back to Kuhnian theory of paradigm shifts (Kuhn, 1962).

Appropriability

If we want to grasp the real economic significance of science we need to recognise it as a source of variety and to admit that it can be more or less rival or appropriable according to the strategic configurations into which it enters. The current notion of appropriability (in

[2] Howard Hughes Medical Institute (HHMI) is an American non-profit research organisation with a focus on the life sciences.

economics) developed from Arrow's classic 1962 paper (Arrow, 1962) — where he discussed a model of incentives to invent — and reached maturity upon David Teece's seminal work (Teece, 1986). Information (and knowledge) is understood to be appropriable under rights of exclusivity — i.e. intellectual property rights — that grant innovators an opportunity to protect their investments in research. Appropriable (and rival) science is, thus, generally discussed in applied developments, usually downstream in the research and innovation pipeline and commonly supported under private foundations that derive profit under "proprietary rules" (David, 2003; Callon and Bowker, 1994).

Basic science, however, relies on the full disclosure of findings and in a cooperative organisation of exploration and discovery. The basic science system relies on the quest for expanding the reliable stock of knowledge alien of personal or corporate interests. The new knowledge generated should only serve as a "public" contribution that benefits the entire community in its cooperative program of inquiry (David, 2003). David argues that the societal benefits of the advancement of science must be incorporated in the "incentive mechanism that induces individual effort". Thus, beyond the fixed monetary compensation, researchers are allowed to appropriate non-pecuniary rewards proportional to the size of their contribution to the stock of knowledge. These payoffs include the public recognition by peers as key contributors to subsequent research, and the right to "own" the finding.

The prospect of gaining the non-pecuniary rewards, which, in turn, enables a better positioning in the organisational scheme for obtaining more substantial monetary (or influential) payoffs are driving individuals in their choice of problem and the direction of their contributions. Since reputation is built upon the reactivity of peers in the scientific community, research agenda choices are biased toward "research spillovers" (David, 2003) that the individuals can appropriate. The basic science regime, thus, must be looked from the standpoint of the individual researcher.

Science from the standpoint of the researcher

Due to the complexity inherent in tracking researchers' careers, macroeconomic trends have never (to the best of our knowledge) looked into the headcount (input) correlation with the number of scientific publications (output). Similarly, very few studies have leveraged microdata on scientists' productivity and career choices. Disambiguation issues, as well as the difficulty in tracking all the affiliations, have led to empirical work focusing on specific grant programs or very limited samples. We believe that accounting for the returns to human capital provides new insights into the dynamics of science and its organisation. In Appendix B.1, we present descriptive evidence of the complex relationships between field growth evolution and the direction of research, which help motivate the model presented in Section 2.3.

In particular, we observe a correlation in the data that points towards diminishing marginal returns as the headcount increases, both in terms of publication counts and overall reach. Furthermore, the correlation seems to be amplified as specialities mature, suggesting that

breakthrough research is more likely to be published in smaller fields. In contrast, exploitative research appears in subfields that attract researchers at a higher rate. This effect seems to relate back to the notion of appropriability, as researchers are aware of the competitive nature of a field and respond accordingly.

A large part of the empirical work studies science at a publication or grant level. Rather than projects, we ascribe our analysis to researchers. Beyond monetary incentives, researchers are not necessarily altruistic in their work, as they engage in a race for recognition. This reputation (the achievements) is, in turn, a gateway for additional monetary resources (Packalen and Bhattacharya, 2018). Thus, we question whether we can explain the direction of science, and perhaps its organisation, from the viewpoint of scientists (be it alone or in a team). We propose a simple game-theoretic model that introduces appropriability of the non-pecuniary benefits from scientific output. Researchers earn higher payoffs the more significant their contributions, which results in the exploitation-exploration trade-off to arise naturally. We develop a model in which researchers make informed decisions based on their beliefs and the potential payoffs. With our model we try to explain trends that emerge from researchers' strategic behaviour.

2.3 A static game of a research speciality

In this section, we develop a game-theoretic model of knowledge proliferation, inspired by research in ecology and plant-species competition (Gersani et al., 2001). The model builds upon the notion of knowledge as a public good and introduces appropriability to the yield of the newly-generated knowledge. We understand appropriability as the factors that determine the researcher's ability to capture the returns generated by a contribution. This derivation results in a common-resource pool game. It predicts an evolutionary stable strategy, in the sense that individuals cannot improve its performance by unilaterally deviating from the optimal strategy. As expected, the optimal strategy does not coincide with the social optimum, but rather results in a sort of *tragedy of the commons*, in which researchers seek to maximise their own good over the population welfare. We argue that this model, albeit quite simple, accounts for the competitive nature of scientific research, and sets a theoretical foundation to the macroscopic behaviour of researchers as a group.

To understand the competitive ecosystem of researchers, it is convenient to split the ensemble of scientists into its organisational units. On a grand scale, science is divided into fields. In turn, each of these fields contain multiple subdivisions. The areas of knowledge covered within these subdivisions, regardless of their root, can be significantly distant. Typically, researchers specialise in much narrower topics, together with a community of similar individuals —a social circle— with whom they primarily interact. Networks generally arise either directly through collaboration or participation in the same forums or indirectly through the action/reaction to the contributions from other stakeholders. We will call this interacting group —which constitutes our organisational unit of analysis— a "speciality" or "sub-field". After all, most

competition in science will happen amongst members of the same speciality (Hagstrom, 1974).

A sub-field is a group of N researchers that is characterised by a stock of knowledge K, non-excludable to all incumbents. Every member can easily access and build upon that stock of knowledge as part of the social circle of researchers. As community members, researchers have free access and the capacity to reuse this knowledge. We define K as a cumulative function of the individual contributions v_i, which is information readily available for use by other individuals (typically papers). Thus, $K = \sum v_i$. We will assume each contribution is a function of a combination of elements —some chosen, some inherent to the individual characteristics— including, but not restricted to, the ability of the researchers, the effort devoted to the contribution, financial constraints, network and social skills. This stock of knowledge generates, in turn, a set of outputs to society y. The aggregated yield of this knowledge stock K is, therefore, $y(K)$. We do not consider the returns to individual contributions, but rather to the entire knowledge capital. It is eminently complicated to isolate the specific output of a given piece of codified knowledge. The construction of both K and $y(K)$ respects the notion of accrual of information and knowledge (its cumulative nature) in the generation of rewards. The stock of scientific work, in turn, generates a return on all active users, and not merely on the marginal contributor.

2.3.1 Public Good vs Common Resource Pool

A Public-Good Game of Science

In a standard public-good/social dilemma game, the rewards from the total product are evenly split amongst all players such that $r_i = y(K)/K$. In our researcher-benefit model, the individual captures the average return to knowledge accumulation. A public-good game requires non-excludability as well as non-rivalry in the profit function. It can generally be reduced to a prisoner's dilemma game. The individual payoff π for researcher $i = 1...N$ in such a game is represented as:

$$\pi_i = \frac{y(K)}{K(v_i, v_{-i}, N)} - C(v_i) \tag{2.1}$$

where $C(v_i)$ is the cost associated with producing contribution v_i, and v_{-i} are the contributions of the other researchers in the community.[3] K depends on the total number of researchers in the speciality N.

Studying science from a public-good perspective results in a social dilemma. On the one hand, public expenditure and Government intervention on basic science is economically justified by its intrinsic nature. Public goods generate economic inefficiencies, and the market incentives are not sufficient to support them (Callon and Bowker, 1994). Non-rivalry and non-exclusion imply industry and business will tend to under-invest, a market failure that has been empirically and theoretically shown (Romer, 1990). On the other hand, research

[3] The cost can be understood as the amount of labour that goes into the production of a contribution.

in experimental behavioural economics has shown that the framing of public-good games leads to an under-contribution of effort compared to the social optimum. Gintis (2009) claims that *"people measure movements from the status quo and hence tend to under-contribute in the public goods game and over-contribute in the common pool resource game, compared to the social optimum"*.

Another way to look at the economic inefficiencies is through the multiplicity of efforts. In science production mechanism, redundancy and over-contribution are crucial to ensure discoveries will take place. Duplication of efforts and replication activities are required for the validation of theories and experimental designs, as well as for arriving to a timely solution to a problem (Hagstrom, 1974; Merton and Storer, 1973).

The public good nature of science imposes limits on the outcomes of scientific production, and of scientific contribution to society. As such, recent research has argued in favour of an alternative framing. Kealey and Ricketts (2014) argue that treating science as a *contribution good* has implications on the incentives that explain the mechanisms behind the organisation of science which, in turn, changes the game from the *prisoner's dilemma* above to a *pure-coordination* game. For them, tacit knowledge constitutes an entry barrier to research in a sub-field. Hence, only contributors (actively involved members) will benefit directly from an accumulating pool of research results — hence the *contribution* good.

Competition lies in the organisational foundations of science. Researchers compete for the acquisition of financial resources in order to fund their projects, but also for a limited number of positions within academia, which are becoming more and more scarce (Stephan, 2012). However, even beyond the monetary rewards, science is ultimately (socially) organised as a meritocracy, where researchers capture higher rewards through priority or significance of their contributions, awarded and recognised within their community. The non-pecuniary benefits from science production are thus largely appropriable and sought after by the community members.

Thus, the desirable model should have two properties:

- Result in a prisoner's dilemma type of game that has natural inefficiencies.
- Incorporate rivalry in the appropriation of the non-pecuniary rewards, such that it captures the competitive element and eliminates the propensity to under-invest.

We propose common-pool resource games as a solution that meets these criteria.

Science as a Common-Pool Resource Game

Common-Pool resource games (CPR) are a type of game extensively studied in a branch of economics, the rent-dissipation literature (Gordon, 1954; Walker et al., 1990; Gardner et al., 1990). While both public-good and CPR games reduce to a prisoner's dilemma, there are

significant differences in the response behaviour of players (Apesteguia and Maier-Rigaud, 2006). Apesteguia and Maier-Rigaud argue that, when players capture individual payoffs weighted by an individual distributional factor, a degree of rivalry is introduced in the game, generating a distinct strategic environment.

We use the contribution to the stock of knowledge by each individual, v_i, as the distributional factor. Then, the payoff becomes: $r_i = v_i \cdot y(K)/K$. This seemingly small modification introduces rivalry through appropriability of the common resource. The assumption underlying this model is that researchers capture a larger payoff proportional to the weight of their contribution. K is still a public good from which everyone can source, but the total yield $y(K)$ is a limited resource, a common-pool, from which rewards are distributed amongst its members. v_i works as an individual distributional factor. Hence, from equation 2.1:

$$\pi_i = v_i \cdot \frac{y(K)}{K(v_i, v_{-i}, N)} - C(v_i) \tag{2.2}$$

This model assumes monotone, increasing functions of the structural components of a contribution ($v' > 0$) and a concave, monotone increasing function of returns to knowledge (y) where

$$K = v + \mathbb{U} \tag{2.3}$$

with

$$\mathbb{U} = \int^{N-1} v_{-i} dv$$

Notice that v changes with the focal individual, and π depends on the frequency distribution of the strategic choices of the different individuals v_{-i}. The choice of an individual is independent of the others. At the equilibrium, we maximise the profit function with respect to the individual's strategy, so we set $\frac{\partial \pi}{\partial v} = 0$ for an interior solution. From equation 2.2 and the first order condition:

$$\frac{\partial \pi}{\partial v} = -\frac{\partial C}{\partial v} + \frac{v}{K}\frac{\partial y}{\partial K}\frac{\partial K}{\partial v} + \frac{y}{K^2}\left[K - v\frac{\partial K}{\partial v}\right] = 0 \tag{2.4}$$

From equation 2.3, we know that $\frac{\partial K}{\partial v} = 1$, and introducing this into equation 2.4 we get:

$$\frac{\partial C}{\partial v} = \frac{v}{K}\frac{\partial y}{\partial K} + \left[\frac{\mathbb{U}}{K}\right]\frac{y}{K} \tag{2.5}$$

Equation 2.5 is our main result as it describes the cost support allocation of the optimal strategy. As we can see, there is no path dependence on the distribution of efforts since it only affects the lump sum $\mathbb{U}$ and K. Two results emerge from Equation 2.5:

(a) The marginal cost for supporting a research choice is a weighted sum of the average yield y/K and the marginal yield $\partial y/\partial K$.

(b) The individual researcher cost supported by the average return increases with the size of the group. An increase of size N implies an increase of the knowledge stock by others $\mathbb{U}$.

If there is no competition, $N = 1$, and all the weight is supported by the marginal value of the contribution. As N goes to infinity, researchers will enlarge their production until the average return equals the marginal cost of pursuing that strategy.

The second finding (b) results in a tragedy of the commons, where there is an over-exploitation of the common resource. The equilibrium (the optimal strategy for the individual player) differs from the social optimum (the one that maximises the productivity of scientific output). In this sense, the model is consistent with the institutionalisation of science policy and government intervention, specifically, in public funding of science. For instance, by providing the right incentives, governments can steer funding to breakthrough research. The model presented, however, tries to shed new light on the individual motivations of researchers and how these affect the dynamics of science. While it is convenient that the resulting dynamics align with policy design at a larger scale, we must not forego the signification for the individual researcher regarding optimal choice. Section 2.4 provides a full discussion of these two results.

2.3.2 Heterogeneity: Two illustrative (special) cases.

Weak heterogeneity amongst researchers

The solution obtained in the previous section for the general case assumes homogeneous players which conform to the optimal —uniform— strategy (the choice of contribution). It is a reasonable assumption that, within speciality, all established researchers are, to a good approximation, homogeneous in their abilities and allocation of efforts. While academic careers require of very particular skill-sets, they allow for a great degree of self-selection into the topics of interest. It is then credible that individuals with very similar characteristics conform the communities. Naturally, there exist differences between the researchers. In what follows, we extend the previous results for weakly heterogeneous players, whose abilities are concentrated around an average value.

Heterogeneous players in simple one-period simultaneous-move games have been studied extensively, with numerical solutions for complex scenarios or few numbers of players (Lockard and Tullock, 2001). The solution proposed by Lockard and Tullock falls outside the scope of this chapter. Given the characteristics of science, in the following, we develop a special case for weakly heterogeneous players, which builds upon the rent-seeking game solutions proposed by Pérez-Castrillo and Verdier (1992), Nitzan (1991) and particularly Ryvkin (2007). For an analytical derivation, we further assume the costs $C(v)$ are a linear function of the contributions $C(v) = c_i \cdot v_i$.

Let $\bar{c}$ denote the average cost of generating a contribution in a sub-field. Individual costs differ slightly, but will be concentrated around $\bar{c}$. To model these individual differences, we introduce the relative abilities of the players θ_i such that $c_i = \bar{c}(1-\theta_i)$ where $|\theta_i| \ll 1$. Similarly, following the results from Ryvkin (2007), small heterogeneity in abilities can only lead to small heterogeneity in the contributions made by researchers, in turn concentrated around $\bar{v}$. Hence

$v_i = \bar{v}(1+x_i)$ where $|x_i| \ll 1$ are the relative contributions. Notice the different sign in the construction of c_i and v_i. A larger relative ability results in a lower effort, while a larger relative contribution results in a larger contribution. Equation 2.5 then takes the form:

$$c_i = \frac{v_i}{K}\frac{\partial y}{\partial K} + \frac{(K-v_i)y}{K^2} \tag{2.6}$$

If we introduce the expressions above, and rewrite K as a sum of its components, we get:

$$\begin{aligned}\bar{c}(1-\theta_i) =& \frac{\bar{v}(1+x_i)}{\sum \bar{v}(1+x_j)}\frac{\partial y}{\partial K} + \frac{\sum \bar{v}(1+x_j) - \bar{v}(1+x_i)}{\left[\sum \bar{v}(1+x_j)\right]^2} y \\ =& \frac{1+x_i}{\sum(1+x_j)}\frac{\partial y}{\partial K} + \frac{\sum(1+x_j)-(1+x_i)}{\bar{v}\left[\sum(1+x_j)\right]^2} y\end{aligned}$$

Using $X = \sum_j^N x_j$, $K = N\bar{v}$, $\bar{y} = \frac{y}{k}$ and rearranging leads to:

$$\begin{aligned}\bar{c}(1-\theta_i) =& \frac{1+x_i}{N+X}\frac{\partial y}{\partial K} + \frac{N+X-(1+x_i)}{\bar{v}\,[N+X]^2} y \\ =& \frac{1}{N+X}\left[(1+x_i)\frac{\partial y}{\partial K} + N\bar{y} - \frac{1+x_i}{N+X}N\bar{y}\right]\end{aligned}$$

And finally:

$$\theta_i = 1 - \frac{1}{\bar{c}}\frac{1}{N+X}\left[(1+x_i)\frac{\partial y}{\partial K} + \left(1 - \frac{1+x_i}{N+X}\right)N\bar{y}\right] \tag{2.7}$$

Equation 2.7 is an expression of the relative ability as a function of the size of the group N and the relative contribution x_i. From here we derive the elasticity of contribution with respect to ability, that is, the cost:

$$\epsilon_{ij} = -\frac{c_j}{v_i}\frac{\partial v_i}{\partial c_j} = \frac{\partial x_i}{\partial \theta_j}$$

For sufficiently large N, the last term of equation 2.7 becomes $N\bar{y}$, and ignoring second order and cross derivatives,

$$\epsilon_{ij} = \frac{\partial x_i}{\partial \theta_j} = \frac{1}{\frac{\partial \theta_j}{\partial x_i}} \simeq \bar{c}(N+X)^2\left(\frac{1}{(1+x_i)\partial_y + N\bar{y}}\right) \quad \forall\, i \neq j \tag{2.8}$$

Hence, $\epsilon_{ij} > 1 \quad \forall\, i \neq j$ for as long as the equilibrium holds.

Intuitively, the cross-elasticity ϵ_{ij} measures the reaction of player-i's strategy —i.e. the relative contribution of player i— to changes in the relative ability of other members of the community. Researchers will adjust their contributions increasingly with the ability of other members.[4]

On the one hand, this result motivates the belief that the most productive researchers drive scientific fields. They can have a significant weight on the marginal returns to the stock of

[4]The results for the self-elasticity ϵ_{ii} are ambiguous, and depend on the absolute values: the response of a researcher to her ability depends on the environment and size of the group.

knowledge. The presence of such above-average individuals drives production-crowded fields, while the less-able still reap the average return with their production, as equation 2.5 suggests. On the other hand, it is compatible with the notion of rivalry and competition in science. Researchers will respond to their perception of their peer's ability through more impactful contributions.

In the next section, we take a different simplifying assumption in order to observe the implications for the respective payoffs.

Fixed cost and contributions as a function of ability. Competition and the invisible college

Let us now fix the costs of a contribution in order to study the effect of ability in the rewards of scientists. We assume that all contributions require a constant level of effort, which all incumbents perform in order to be in the community. This premise is in line with the *contribution good* approach by Kealey and Ricketts (2014), where the investment in tacit knowledge acquisition is required to take part in the community. Under this scenario, we impose no restriction in the magnitude of the relative abilities, but rather that they are bounded and follow a given distribution. Then, equation 2.2 now becomes a profit function where $v_i = v(\theta_i)$. Note that now θ_i does not refer to the relative ability, but rather to the absolute ability of researcher i. The ensemble of abilities of researchers in a group are bounded $\theta \in [\underline{\theta}, \bar{\theta}]$ and follows the distribution $\theta \sim f(\theta)$. The total stock of knowledge K is, then:

$$K = \int_{\underline{\theta}}^{\bar{\theta}} v(\theta) f(\theta) d\theta \tag{2.9}$$

where

$$\frac{\partial K}{\partial \underline{\theta}} = -v(\underline{\theta}) f(\underline{\theta}) < 0 \tag{2.10a}$$

and

$$\frac{\partial K}{\partial \bar{\theta}} = v(\bar{\theta}) f(\bar{\theta}) > 0 \tag{2.10b}$$

From equation 2.9, K is a function of the bounds $K = K(\underline{\theta}, \bar{\theta})$. We will use the notation $\hat{\theta}$ to generically refer to either bound, depending on which one we fix, so that $K = K(\underline{\theta}, \bar{\theta}) = K(\hat{\theta})$. Equation 2.2 becomes then a function of ability and the bound $\pi_i = \pi_i(\theta_i, \hat{\theta})$, so we can derive the following result:[5]

$$\frac{\partial \pi(\theta, \hat{\theta})}{\partial \theta} = v'(\theta) \frac{y(K)}{K(\hat{\theta})} > 0 \tag{2.11}$$

Equation 2.11 shows that, assuming that the marginal contribution increases with ability, $v' > 0$, a marginal increase in the ability of a researcher yields an increase in his or her profit function. This straightforward result implies that, with a fixed level of effort, researchers with a higher innate ability (higher θ) will have a higher payoff π. As in the previous section, the

[5] Dropping the subscript i in what follows, for simplicity of notation.

cross derivative (individual payoff with respect to the bound) yields less trivial results:

$$\frac{\partial \pi(\theta,\hat{\theta})}{\partial \hat{\theta}} = \frac{v(\theta)}{K}\left[\frac{\partial y}{\partial K} - \frac{y}{K}\right]\frac{\partial K}{\partial \hat{\theta}} \tag{2.12}$$

Equation 2.12 shows the effect of a change in the bounds of the researcher's group ability on a researcher's profit function.

Assuming $y(K)$ is a strictly concave function, $\frac{\partial y}{\partial K} < \frac{y}{K}$, combined with equations 2.10a, 2.10b, 2.12 yields:

$$\frac{\partial \pi}{\partial \underline{\theta}} > 0 \tag{2.13a}$$

$$\frac{\partial \pi}{\partial \bar{\theta}} < 0 \tag{2.13b}$$

Individual payoffs of members of a community increase with the lower bound of abilities, while they decrease with an increase in the upper bound of abilities. Given the fixed distribution of abilities $\theta \sim f(\theta)$, this means incumbents will receive larger payoffs by directly or indirectly limiting the size of their communities. The marginal increase of the lower-bound ability, within a collective, increases the profit function of the researchers within the group.

In the following section, we rationalise the previous results and present a framework for their compatibility with the existing literature. We then proceed to discuss the limitations of the model.

2.4 Discussion and Limitations

2.4.1 Discussion

The model presented in Section 2.3.1 is a simple and illustrative challenge to the notion of science as a pure public good in economic terms. We introduce the notion of rivalry through an individual distributional factor —namely, the weight of the individual contribution— which affects the allocation of non-pecuniary payoffs from contributing in a scientific speciality. The need for such a model arises from a rather naive association of ideas: science is socially organised around priority, and rewards are attained by individuals — spurring competitive behaviour— but knowledge (especially codified, published, accessible publications) has the characteristics of a public good.

Much like a contribution good game, the introduction of an individual distributional factor generates a scenario where there is no free-riding. Researchers must be individually engaged and proactive in order to aspire to the gains —just by being a community member does not provide any gains. From a sociological point of view, the community judges and defines the reputation of individual researchers based on the specific merits of their work.

The derivation of the optimal strategy for each individual from the central, represented by equation 2.2 suggests that:

i. Assuming y is a concave monotone increasing function of K, the individual's perceived benefit from contributing to a large field is larger than to a small field.

ii. An area of knowledge grows optimally in its early stages as long as the marginal contribution has more weight than the average.

The first finding (i) results in a tragedy of the commons. In a larger field, the strategy is supported by the average — rather than the marginal— contribution, which represents a deviation from the social optimum. We interpret the individual payoff pursuit (instead of the social optimum) as the mechanism behind the economic inefficiency of science, i.e. the multiplicity of efforts for a given *problem*. As discussed above, this is, in fact, a desirable property of the science ecosystem, which *ensures* discoveries.

The second finding (ii) is a consequence of the weighted average represented in equation 2.2. The socially-optimal (the one that maximises welfare) allocation of effort is that which is entirely supported by the marginal contribution. Hence, the *socially optimal* growth happens for low N. We believe this result to be consistent with the first observations in the sociology of science. Crane (1969) observes that all the *high producers* enter the field in the early stages (or during exponential growth of the sub-field). She argues that this might suggest *"a sensitivity to potentialities of growth in a field in making their selection of research problems"*, in line with our suggestion that sub-fields stagnate towards average returns in time, making them less attractive to researchers with high potential.

When average returns support the cost structure, there is an excess of production (of research). Incumbent researchers are "foraging" the common-pool of resources at the cost of influencing the total knowledge stock. In the branch of behavioural economics that studies these games —the rent-seeking literature— this behaviour is known as active *exploitation* of the resource. As the stock of knowledge K grows, it provides strategic support to the average player, not just the marginal contributor, sacrificing collective socially-optimal gains.

Invention and Consolidation (explore vs. exploit)

In his highly cited public lecture at Virginia Polytechnic Institute, Callon (1994) questions whether science is a public good. He identifies two types of knowledge being generated, one that transcends the boundaries by taking risks and pushing the frontier limits, and one that frames the application and consolidates the established knowledge. Within these trade-offs, he conjectures that science is not a public good in pure economic terms.

The model we present in the previous section is the reverse exercise: introducing appropriability to the payoffs of the researchers, we theorise about the dynamics of science. Equation

2.2 shows a weighted average and suggests that the optimal strategic behaviour for the actors weighs on the shoulders of two components. Reflecting on the work of sociologists, we suggest interpreting the marginal and average yield as the exploration and exploitation of research, respectively.[6] Or, in words of Callon, invention and consolidation. Given the clarity of his arguments, it is better to cite his work directly. Callon summarises the contributions of sociology and anthropology of science up to that point as follows:

> *The first outcome [that of a well-defined community] is associated with routine work, consolidation and continued and stubborn improvement. The second outcome corresponds to what is generally called invention: an unexpected association of several preexisting networks that up to that point were strangers to each other.*

In this parallelism (that of our model), the researcher bears the cost of supporting her optimal strategy by the marginal contribution and the average contribution. The first (the marginal contribution) represents a subversive contribution and the latter, a traditional approach, a reliable strategy to accumulate recognition. The efficient allocation between the two tasks depends on the size of the community to which the researcher belongs. Therefore, we suggest that breakthrough research that is highly innovative or that defies the "accepted knowledge" will happen in smaller (unconsolidated) fields. The second (the average contribution), consolidation, work will follow. As fields become large, most published literature will be, to a certain degree unsurprising and continuist (Foster et al., 2015).

Modelling science as a common pool resource game played by researchers conduces to an optimal strategy that balances exploration and exploitation of knowledge. It is an indirect consequence of the optimal allocation of efforts that maximise the payoffs for the individual player, rather than an active choice. The model, in a way, frames the *essential tension* of strategic choices in academia (Bourdieu, 1975). Further back in time, Kuhn (1977) had framed fields in science as a competitive endeavour in which researchers faced a strategic choice between succession and subversion. Recent work by Foster et al. (2015) tests Bourdieu's model empirically and finds supporting evidence for it.

The model presented in this chapter derives the essential tension as a result of the optimal strategy, rather than an active choice (to explore or exploit) of academics. In Appendix B.5, we address this limitation by turning the tables around, suggesting how a simple two-period model in which researchers have an active choice on whether to innovate or consolidate could

[6] The concepts of exploration and exploitation have largely been used in the literature of learning and innovation, especially since the seminal work of March (1991). He introduces exploration as variation-seeking, risk-taking and experimentation oriented. Conversely, exploitation is variety-reducing and efficiency-oriented. The two concepts have been used in a wide range of ways, making it hard to unify a definition in the context of Knowledge and Learning. The work by Li et al. (2008) collates the main uses and suggests a consolidated definition based on the *"knowledge distance domain."* For them, one exploits by searching for knowledge within the boundaries of an organisation, local to the existing stock. On the other hand, one explores by searching distant knowledge that is unfamiliar

be characterised.

Heterogeneity

The introduction of weak heterogeneity between the players —i.e. small differences in relative abilities or contributions— allows the computation of the cross-elasticity of contribution with respect to ability. For established specialities, one can derive Equation 2.8 where the cross-elasticity measures the percentage change of a researcher i relative contribution following a change in researcher j relative ability. $\epsilon_{ij} > 1$ implies a positive and strong reaction to the other player's abilities for the members of the community. Therefore, researchers take positive, assertive strategic reactions to the ability of their peers, *competing* for the gains. The distributional factor introduced in the public-good game, we argued, introduced rivalry in the appropriation of the payoffs of science. With the derivation of the cross-elasticity of effort-ability, we show a degree of competition in science derived from the model.

The reactivity of researchers proposed by the model to other's ability hints towards the most productive researchers as *guiding* members of research specialities. The contributions of the incumbent colleagues will be adjusted to respond to the existence of such leaders.

In the second particular case, using researcher's abilities drawn from a bounded distribution, Equation 2.13 highlights a preference to be part of a group with researchers of uniform and static characteristics.[7] We suggest this introduces *entry barriers* — a bound in the distribution of abilities can be understood as an exclusion term below the threshold. That is, researchers within a community perceive as detrimental the inclusion, in the same community, of researchers below the lower-bound. On the other hand, the marginal increase of the upper-bound ability is detrimental to the researchers within the group. The former could potentially explain impediments to new-entrants of high ability to well-formed groups which translates into the reluctance to adopt new high-impact ideas. These results in Equation 2.13 highlight, once again, the competitive environment of science, suggesting elitist motivation —with the tendency to prefer a higher lower-bound for one's group— and fear to new-entrants that challenge the incumbents' status-quo. They could help explain the creation of an *invisible college* (Crane, 1969; Price and Beaver, 1966) within a speciality, which comprises the closest social circles to a researcher. In turn, it theorises the existence of social entry barriers to such a close-knit group. In other words, the model suggests a form of social-pressure effect through the loss of gains by individual researchers amidst changes in status-quo.

Anew, sociologists of science have extensively studied the social characteristics of research communities, and it is worth looking back at their findings to illustrate the model's derivations. Crane (1969) studied the social organisation of researchers. In her work, she sought to find which type of sociological structure best described the interactions and whether there existed leadership in the fields. In her typification of actors, she found *high producers* to have a

[7] Consistent with the assumption of *weak heterogeneity* used for the first special case.

higher degree of visibility in the problem area.[8] Price and Beaver (1966) argue that the most productive researchers were the nexus that held communities together, while ties amongst other incumbents (of similar lower productivity) were weaker. Wagner (2008) explains that the science system evolves, splits and merges with other subfields where "nascent fields" are led by groups that have exercised control over the direction of research in the area. Recent work in Economics by Azoulay et al. (2015a) finds similar evidence in that the peer-pressure exerted by colleagues of "superstar" researchers discourages "revolutionary" new entrants.

2.4.2 Limitations

Motivated by the observed phenomena and by the sociologists' view on field dynamics, we attempt to provide a simplified framework that uses a new type of social dilemma to explain science. The optimal strategies for researchers result in a tragedy of the commons, as in a public-good game. However, introducing rivalry uncovers a set of features that are missing from the classical description of science. Modelling scientific fields as a common-pool resource game in the way presented in Section 2.3 has, still, many limitations.

First, the omission of a temporal dimension. The model provides a simplified version of a static world — as if science was defined by static independent communities— in which researchers have no active choice on which community to engage, they belong to one unmovable category. Second, the oversimplification of contributions as produced by single researchers. The reputation and contribution game has a much more nuanced story with the inclusion of collective effects, which shape the outcome of the non-pecuniary payoffs through the mechanics of communication channels, and sociological constructions (e.g. the Matthew effect (Merton, 1968)). Team science is essential in current research (Bikard et al., 2015; Jones, 2009), so perhaps this model is instead one of PIs. Third, we disregard the multidisciplinary researcher and the many interactions that happen in the real world connecting communities. Scientists often contribute to more than one stream of literature and are involved in several communities at the same time. In each, they interact at different levels of participation, in some they might be leaders, in some followers. Fourth, we overlook the financial constraints and the *burden of knowledge* of different specialities (Jones, 2009). This omission comes as a consequence of the previous limitations since the assumption is small within-group heterogeneity. Fifth, we assume all knowledge present in a subfield contributes to the stock of knowledge. There is, of course, knowledge that leads to a wrong track — i.e., contributions that open up unfruitful avenues of research— or knowledge that *destroys* prior art by, for instance, finding opposing evidence. Finally, we assume scientists have no active voice in the decision to engage in more innovative or cumulative research. It comes as given with the optimal strategy (which would be costly to deviate from).

Team science and barriers to entry present in fields —such as cost of laboratory set-up, or

[8] High Producers are the highest contributors in the research area that Crane studied, and were advisors for several of the Moderate Producers at some stage.

access to funding, rather than imposed by the social norms— are necessary for a full-fledged empirical analysis of the model. They must be included in testing the validity of this or similar models, not to incur any omitted variable bias. Furthermore, multiple examples study the reputation effects of team construction or failure (Jin et al., 2013; Bikard et al., 2015). In order to address time-invariance and choice, a different class of models needs to be used, such as growth models in the likes of Jones (2009), infinite-horizon overlapping generation models such as Bramoulle and Saint-Paul (2010), or ultimately complex dynamic discrete games (Aguirregabiria and Mira, 2010) which are a branch of economic theory themselves, and fall outside the scope of this chapter.

One way to simply incorporate an active choice between exploration and exploitation (and time dependence) is through a simple two-period dynamic discrete choice model. A simple two-period model helps illustrate how infinite-horizon games might help model science. For the sake of completeness, we derive and discuss this model in Appendix B.5.

2.5 Conclusion

This chapter presents a simple model that tries to explain organisational characteristics of science from a pure optimisation of strategies by the researcher. From the maximisation of her payoff, we derive a set of observations consistent with the literature.

First, present a model of scientific rewards drawing from a common-pool resource game. The simple configuration of the model sheds new light on several observed sociological effects with a straightforward model of rewards. Allowing researchers their choice of a contribution, we incorporate appropriability in the context of scientific production. This way, we account for the incentives of subsequent choices in a researcher's career. The most basic setting results in optimal support being defined by a weighted average of the marginal and average contribution. We link these results to the extensive literature on the *essential tension* of researchers.[9] Depending on the number of players N, contributing to a stream of literature, we derive implications on the incentives to publish extensively (exploiting) or innovate (explore).

Introducing two particular cases with players' heterogeneity allows us to make several claims regarding the effect of competition on science production. Researchers respond to increases in their peers' abilities positively, enhancing their production. Similarly, our results suggest that researchers with potentially above-average abilities drive new sub-fields, which leads to them having a higher impact. Setting upper and lower bounds to the abilities within a social circle provides a framework compatible with the existence of an *invisible college*. More specifically, our model provides a setting in which that incumbent researchers of high ability will benefit from a higher entry barrier within a given invisible college. In contrast, new entrants at the upper bound of ability decrease the payoffs for its members. The intuition behind these barriers to entry could be explained by resilience to challenges to status-quo.

[9] The dichotomy in knowledge creation that researchers face: explore vs exploit; innovate vs consolidate; tradition vs subversion.

The model excluded from the analysis several variables and factors that influence the organisation of research circles and their dynamics. We present the dichotomy of researchers between succession and subversion as a consequence of a strategy-maximisation game. This framework explicitly omits researcher conscious choice in strategy and timing.

The model presented in this chapter extends beyond science to communities with similar characteristics such as the open-source coding. (Open) software developers build upon significant initial contributions with marginal improvements. They are rewarded by either being active in many marginal contributions that consolidate the great leaps or by proposing original first-mover algorithms (or languages, solutions, bugs, etc.). The incentives of community reputation are similar, peer review works systematically, and the underlying structure seems to follow a social circle model.

Policy implications

We argue that researchers will naturally tend to prioritise exploitation as fields mature. Funding policy design is then crucial to ensure that innovation is not discouraged. There is evidence that risk-supporting schemes yield greater creativity (Azoulay et al., 2011). Career pressure in the current system drags the creative behaviour of scientists. Hence, decoupling job security from productivity could result in more original work. Understanding that researchers will prioritise their individual payoffs suggests that grant design (i.e., how grants are implemented) is vital to ensure that the objectives of the grant scheme are met (Jacob and Lefgren, 2011; McKnight, 2009; Kaplan, 2005).

The model may also help illustrate the institutional trade-off. Academics increasingly required to reach out and disseminate their findings and are evaluated accordingly. If dissemination work or technology transfer practices that do not always push the frontier of knowledge report benefits to the researcher, they might represent a disincentive towards breakthrough science. While these activities undoubtedly bring value to society, researchers may face yet another trade-off in their effort allocation. In this sense, the model might help explain how the use of altmetrics and technology-transfer outputs to measure productivity (and evaluate performance) might engulf the pursuit of breakthrough research (Larédo, 2015; Philpott et al., 2011).

Extensions

The natural extension to the work presented in this chapter is a sequential discrete game with heterogeneous players. Such a model would overcome the temporal limitation of the work presented above. A first-step approximation is provided in Appendix B.5 where we present a two-period model of choice based on the *essential tension*.

Furthermore, evidence presented in Section B.1.2 can be explored further with a fully empirical exercise. First, there is unobserved field-level heterogeneity that can be correlated with the size

and growth of the field. Second, there are researcher level covariates that might influence the decisions. Third, one should track the "landing" field of a researcher and how mentors' private information influences their strategic action. Fourth, exit events are as important as entry events to model strategic behaviour. Finally, we do not track the "mode" of entry, whether exploration or exploitation. Respecting the game structure, the work could be extended using a model of entry-exit from a field as a dynamic discrete game of incomplete information. These set of models are often not present in the mainstream literature due to the difficulty of estimation. However, the sequential estimation method (Aguirregabiria and Mira, 2010) provides a framework that works for both heterogeneous players and permanent unobserved field heterogeneity.

External shocks might also influence the strategic best replies. Bhattacharya and Packalen (2011) find evidence that researchers respond to demand pulls from societies fundamental needs. Nevertheless, what if market demand happens in the exploration phase of a field? In other words, how do researchers strategically respond to a "hot" topic when exploitation is still not likely? The work presented in this chapter could then be extended with an empirical setting that studies the rate and direction of science a high-growth highly-novel field where exploitation is discouraged.

Chapter 3 provides such an environment. Using a natural experiment, we study the effect of a funding ban on a highly novel (and with great potential) field in the life sciences.

Acknowledgements

Preliminary versions of this chapter have been presented at the Atlanta Conference on Science Policy (Atlanta, 2019); Barcelona Summer Forum in Economics of Science and Innovation (Barcelona, 2019); Competition and Innovation Summer School (CISS) (Ulcinj, 2019); and 6th PhD Workshop in Economics of Innovation, Complexity and Knowledge (WICK) (Torino, 2019). I wish to thank the anonymous comments and remarks received from referees and attendants, who have played a role in the improvement of the analysis presented above, and in particular to Prof. Aldo Geuna, discussant of the presentation at WICK.

3 Innovation Stems from Science: The Impact of Funding Policy on Innovation

"We do not invent it ourselves and nor do we ask for it,
yet it is our job to find the hour when needs might erupt,
and salmons defiantly and insanely jump against the tide for...
who knows what reason?"
— Morrissey (List of the Lost)

This chapter explores the relationship between advances at the frontier of science and downstream innovations using the United States' 2001 policy regarding the federal funding of human embryonic stem cells (hESC). We employ patent-to-scientific-article citations to evaluate the impact of the policy on the innovations stemming from basic science. We characterise the causal impact of the policy on subsequent innovation with a difference-in-differences estimator. Our estimates suggest that in the years following the policy, scholarly publications subject to limitations received 65 to 80 per-cent fewer patent citations than the control group. We show that lesser quality publications lead, at least partially, to this relative decline. Using topic modelling techniques applied to publication abstracts, we construct a topic-variety metric. Our findings show that diversity decreased in the aftermath of the policy, suggesting publications covered narrower topics than before. The results suggest that modest policy changes —such as restrictions to materials— have a profound impact on downstream innovation and the advancement at the frontier.

3.1 Introduction

This chapter analyses the impact of the 2001 U.S. human embryonic stem cell (hESC) policy on downstream innovations and technology. We exploit an exogenous shock to science funding that restrained researchers from using certain materials (namely, newly-derived stem cells). Using a citation-based approach to capture the knowledge spillovers, we estimate the causal effect of the policy on innovation with a difference-in-differences approach. With these techniques, we are able to quantify the differences in innovative output between the private sector and research institutes, as well as the scientific quality of publications in the aftermath of the ban. Therefore, our work provides evidence to answer two important policy questions: First, what was the impact of the 2001 hESC policy on innovation? And second, how did limitations in research materials affect the underlying quality of scientific research?

Recent research has provided concrete evidence that public funding of research plays a vital role in enabling innovation, particularly in the biomedical sciences (Li et al., 2017). Our analysis is thus motivated by a central question in science policy: how do basic science funding conditions affect the rate of innovation? Measuring the spillovers generated from science —i.e., the dollar return on frontier research expenditure— is of particular interest for policymakers. In the U.S., the belief that public-research spillovers directly fuel innovations has accelerated the federal investment in basic science, but there is little evidence of these spillovers (Jaffe, 2002; Azoulay et al., 2019).

Our estimates suggest that scientific articles subject to the policy restrictions received 65 to 80 per-cent fewer patent citations than unrestricted citations. The analysis constitutes one of the first to employ non-patent-literature citations and in-text references from patents to scholarly articles to establish the links between the two.

Additionally, we quantify the differences in innovative output between the private sector and research institutes. We further examine the mechanisms behind the decrease in patent citations, and we find that publications bound by the policy were placed in journals of comparatively lower rating and had fewer forward citations. Using publication-text data and the methods developed in Chapter 1, we characterise the yearly *diversity* of hESC publications. We observe a significant drop in *diversity* in the aftermath of the policy, suggesting a concentration of topics in research.

The remainder of this chapter is structured as follows. In the next section, we introduce the research setting of the hESC funding ban. We detail the characteristics of biomedical research in the late 1990s and early 2000s, as well as the particularities of the regulatory environment. We then proceed to describe the data and methods employed for our research question, including a sample validity check and the construction of an alternative sample for robustness checks. In Section 3.5, we discuss the main effect of the ban and the mechanisms of response that we believe triggered the decrease in innovative output. Finally, we comment on the implications of our findings.

3.2 Background

Innovation increasingly relies on scientific knowledge (Ahmadpoor and Jones, 2017). Following a linear model of knowledge generation, basic research provides the foundations for later applications (Bush, 1945), paving the path of innovation. This model assumes knowledge generated from fundamental questions ultimately contributes to technological progress. At the same time, modern growth theory embraces the idea that the productivity leaps extensively support long-run economic growth (and welfare) that new technologies bring (Romer, 1990). However, the free flow of ideas from basic research into innovations, i.e. the knowledge spillovers, result in private firms that under-invest in the production of the most basic science. As broadly discussed in Chapter 2 the market failure is due to the public good nature of knowledge, that is non-rival and difficult to appropriate (Griliches, 1992), and therefore public policy must intervene in order to overcome the inefficiencies.

These ideas date back to Vannevar Bush's (1945) report. However, the views in support of direct spillovers have been continuously challenged, and substantial evidence has only been recently developed. The works by Ahmadpoor and Jones (2017); Fleming et al. (2019) and Poege et al. (2019) have advanced the understanding of the linear model of spillovers between science and technology. Using patent citation approaches, these three publications show that: first, most scientific works links forward to future patents; second, the measured quality of science is a good predictor of patent impact; and third, and perhaps more importantly, more and more innovations come from government-funded research.

The impact of publicly-funded research has been thoroughly documented. Most work has concentrated around the study of particular funding programs, measuring the direct returns to science funding (Azoulay et al., 2018). The works of Jacob and Lefgren (2011) and Bhattacharya and Packalen (2011) study the impact of public funding on productivity, particularly at the frontier. Their analyses provide a deep understanding of the (scientific) returns of the NIH funding scheme. Beyond productivity, Azoulay et al. (2011) examine how different incentives shape the career-path decisions taken by researchers, depending on their allocated funding schemes. The work by Furman et al. (2012) establishes a causal relationship between an exogenous shock in funding and the advancement of knowledge at the frontier. Furman et al. (2012) observe a decline in research output following the intervention, which differently affects institutions of multiple status as well as shapes the international collaborations.

Today, we understand that there are numerous ways in which public funding of science generates spillovers, including training, creating instruments or methodologies, networks and even firms (Salter and Martin, 2001). These (sometimes overlapping) channels of spillovers between public and private R&D are not easy to study. One particular way in which economists have measured the knowledge flows is by examining the patenting and licensing activity at universities. Since the 1980 Bayh-Dole Act, research institutes have actively engaged in the entrepreneurship and innovation ecosystem. Extensive studies have turned their focus to IP-based academic contributions, such as Henderson et al. (1998). Following the citation analysis

proposed by Garfield (1955) in the bibliometrics literature, forward patent citation counts are now used as a proxy for patent quality and impact (Jaffe and de Rassenfosse, 2017). IP-based analysis provides deep insights into the technological prowess of a particular technology, but it fails to trace the direct flows from scientific contributions. Recent work by Azoulay et al. (2019) uses patent-to-article citations in order to study the impact of NIH Funding, linking the innovation efforts to the sources of knowledge upon which they are built. Conversely, Hegde and Sampat (2015) study the inverse effect by tracing money flows. In their work, they show how private money (and lobbying) affect the funding ecosystem, and ultimately the research output (through distributional alterations of grants).

Using patent-to-article citations, our contribution goes beyond previous work, trying to assess the causal impact of federal-policy funding policy on innovation. In this manuscript, we use a quasi-natural experiment (an exogenous shock) to science funding, the 2001 human Embryonic Stem Cell (hESC) ban. In particular, we study how limitations in the methods and materials available for research — which, as we argue below, limited the autonomy of scientists, shaping the direction of scientific inquiry — impacted the innovation ecosystem. To the best of our knowledge, the closest prior art used exogenous sources of variations in funding (rather than methods) to study causal impact (Azoulay et al., 2019; Moretti et al., 2019). Both of their contributions thoroughly document the response rate to funding shocks. Our findings suggest a decline in follow-on innovations as a consequence of the methodological limitations. Following the literature that exploits external shocks to the direction of science (independent of funding policy) (Azoulay et al., 2010; Teodoridis et al., 2018), we argue that the root cause for the negative impact of the ban was due to a decreased quality and variety at the frontier, which resulted in lower potential applicability.

Before we dig into the analysis, we will describe the scientific ecosystem at the time of the ban, and the particularities of the research setting.

3.2.1 Timeline

In 1997, sheep Dolly unanticipated disclosure left the public (and the scientific community) in awe. After years of development in embryonic research and DNA implantation, the implications of the cloning of Dolly (Wilmut et al., 1997) were huge. For medicine, it meant that cell development might be more malleable than once thought, and most importantly, that through the appropriate techniques, differentiated cells (adult cells) could be reprogrammed into undifferentiated cells. Beyond the scientific community, it ignited a ferocious (public) debate on the ethics, suitability and risks of such development.

During the late 1990s and early 2000s, extraordinary advances in biomedical sciences signified a leap forward in realising new therapeutic approaches (Holland et al., 2001). The propitious scientific landscape was fuelled by one of the favourable funding environments for medical research in history. The NIH budget (the single largest funding agency worldwide) roughly doubled between 1995-2005 (Huang and Jong, 2019). Under these circumstances, and aided

by rapid technological progress, scientists completed the expression of the human genome sequence; simultaneously measured the expression of thousands of genes; and widely advanced recombinant-DNA techniques and cell reprogramming. On the other hand, the late 1990s and early 2000s were also characterised by a strong public debate on the ethical issues surrounding overall progress in biotechnology. The secrecy around Dolly's project and the potential implications for humans added fuel to the flames, in particular surrounding embryo and in-vitro-fertilisation research. Amidst this controversy, in 1998, the first isolation of human Embryonic Stem Cells (hESC) was unveiled by Thomson et al. (1998).

A stem cell is an undifferentiated cell (that is, without a specific task) present in the majority of tissues in adult mammals. Stem cells are capable of transforming themselves into the type of cell present in their originating tissue (e.g. blood cells, muscular cells, etc.). Human *adult* stem cells were isolated long before hESC, in the 1960s. Typically, the number of adult stem cells present in a tissue is rather small, but they are fundamental to the repair regeneration of that specific tissue (Chen et al., 2014). By contrast, embryonic stem cells are *pluripotent*, a term used to designate cells that can differentiate into any other cell-type of the individual, hence the relevance of the scientific breakthrough. As their name suggests, hESC are primarily extracted from the inner cell mass of a *blastocyst* — a 5-day-old human embryo. The embryos are sourced from donated in-vitro-fertilised eggs. These cells are then cultured indefinitely with the help of what is known as feeder layers. Each cell cultured from a single blastocyst is part of what is called a *cell line* (Russo, 2005).

Thomson's breakthrough was not as controversial as Dolly, however. It came 16 years after similar developments in mice were revealed, and only two years after Thomson's lab had reported the first isolation of monkey ESC. Nonetheless, it was the first on human cells. The relevance and impact of this discovery were rapidly acknowledged by the community, highlighting the importance of this tool to both basic research and applications to novel therapies in medicine (Murray, 2007). In Thomson's own words *"hES cells capture the imagination because they are immortal and have an almost unlimited developmental potential"*. hESC had the potential to help understand medical concepts both in the fetal and adult stages of human development. Through hESC, it was anticipated that one could potentially test drug toxicity in the lab; propose regenerative medicine therapies; create tissues (and perhaps organs) artificially in the lab; target some of the most prevalent chronic diseases such as diabetes and heart disease; or understand neurodegenerative diseases such as Parkinson's or Alzheimer's. The list was long, leading to human embryonic stem cells to be proclaimed *Science's* "Breakthrough of the Year" in 1999 (Vogel, 1999).

Despite this pathway to promise (NIH, 2001), in August 2001, the Bush Administration introduced a controversial ban on hESC research. Following months of speculation in the concurrent public debate, the ban came as a surprise to most. It neither prohibited hESC research nor encouraged it actively. Instead, the policy became a moratorium on federal funding (and only federal) effective only to *new* stem cell lines. Little and yet so much changed.

Up until 2001, the NIH was not openly supportive of the development of new stem cell lines. Research on embryos was amidst the 1990s life science's ethical controversies. After multiple amendments, since 1995 (Dickey-Wicker Amendment), federal funding was prohibited for research that either created or destroyed human embryos. hESC were, therefore in a legal grey area. hESC are not technically embryos, but the generation of a new cell line involved the destruction of one embryo. Therefore, even before the ban, NIH supported hESC research but not the creation of cell lines (Vogel, 1999). In the case of the first-ever development, Thomson and his team used support from Geron (a biotechnology company) to derive the first cells (Furman et al., 2012). So, in technical terms, there was no difference to the research that could be performed under federal funding before and after the ban. However, the 2001 policy impeded the use of any new cell lines that had not yet been developed. The moratorium constrained researchers in the materials — cell lines — and methods —feeders—, limiting their academic freedom.

In the aftermath of the ban, hESC research continued to gather support from both the public and academics, while the research community profoundly criticised the ban. A report from the National Academy of Sciences, in 2002, highlighted, once again, the relevance of hESC for developing new therapeutic methods, while pointing out two critical problems under the prevailing situation. The report highlighted how cultured cell lines accumulate genetic mutations over time, which make them non-viable for human implantation (or even lab experiments). Moreover, the majority of the (then) existing cell lines had been cultured in the presence of non-human cells or dangerous feeders, which could lead to potential human health risks (Council and of Medicine, 2002). In total, 71 lines from 14 different laboratories met the eligibility criteria, but only 21 of them proved to be of any use to investigators (Murugan, 2009).

The situation even forced some labs to duplicate their structures, based on federal vs other funding for staff, equipment and labs (Murugan, 2009; Cyranoski, 2018). The Harvard Stem Cell Center notoriously introduced physical access (keycard) barriers to hESC labs in order to ensure complete separation of funds and staff (Dreifus, 2006). The additional efforts meant that, even for elite institutions, it took time to find other sources of funding and generate the infrastructure necessary to comply with federal policies. Therefore, as opportunities arose, many chose to take their research elsewhere (nationally or internationally) (Russo, 2005).

By 2005, the uncertainty had started to dissipate. The funding outlook for research on hESC significantly improved as state-funded research on stem cells was promoted. While the moratorium was still in place, California approved (not without controversy) *Proposition 71*, promising an impressive US $3 billion (Vakili et al., 2015) in funding for Stem Cell research over ten years (Russo, 2005). Other states followed, and between 2005 and 2006 Connecticut, Illinois, Massachusetts and New Jersey had also adopted state-level funding programs. At the federal level, a reversal started to seem plausible. The Stem Cell Research Enhancement Act passed in 2005 by a bipartisan majority at the U.S. House of Representatives, only to receive presidential veto. However, a general belief that the veto would be lifted sooner than later

prevailed amongst politicians and scientists (Huang and Jong, 2019). All the efforts from 2001–2006 alongside the active limitations contributed to one of the greatest innovations in regenerative medicine in the 2000s: the discovery of induced pluripotent stem cells (iPS) (Takahashi and Yamanaka, 2006). The process offered, in theory, the same potential as hESC without the ethical conundrum. The following year, Takahashi et al. (2007) achieved the same feat in human cells.

Finally, in March 2009, President Obama fulfilled his campaign promise to end the ban. Figure 3.1 shows the summary of the timeline (1996–2006) for hESC.

3.2.2 Research setting

Following the comprehensive timeline review, we proceed to examine the setting in which we will implement our study. We analyse the impact that changes in (basic) science funding policy had in the innovation landscape over the period 1996-2006. In particular, we focus on the rather unexpected nature of the US moratorium on hESC research and follow-on patents that build on the scientific knowledge of the time.

In August 2001, an executive order by President George W Bush addressed the changes in public funding for hESC. The policy differed greatly from the expectations of scientists - and was received with substantial disappointment from all sides of the political spectrum (Wertz, 2002). Rather than imposing a complete ban on human embryo research, the policy prohibited only the development of new cell lines under federal funds, and imposed no restrictions on other sources of funding (be it private, state or local), Wertz (2002) states. In many ways, the actions taken by the government seem to have opted for a compromise between cultural facets in the United States.[1]

The policy settled the opposing currents with the following particularities:

i hESC research was enabled nationwide. Federal funding, however, would only support research on the set of hESC lines that had already been developed before the policy.

ii Development of new lines or subsequent work with unapproved lines was prohibited under federal funds

iii Any other source of public (state or local) or private funds was allowed.

Therefore, the ban, if we may call it so, was enacted in a particular point of time, with very particular delineations. It applied to a singular area of research, without directly influencing work on other areas and it was localised to the United States. Albeit the uncertainty period that preceded the moratorium, the shock appears to have been exogenous, due to its unexpected

[1] On the one hand, the free agency of the private sector. Free enterprise is amongst the most treasured values by U.S. conservatives. On the other hand, the American is a rather religious culture, and embryo-derived research has faced hurdles for many years.

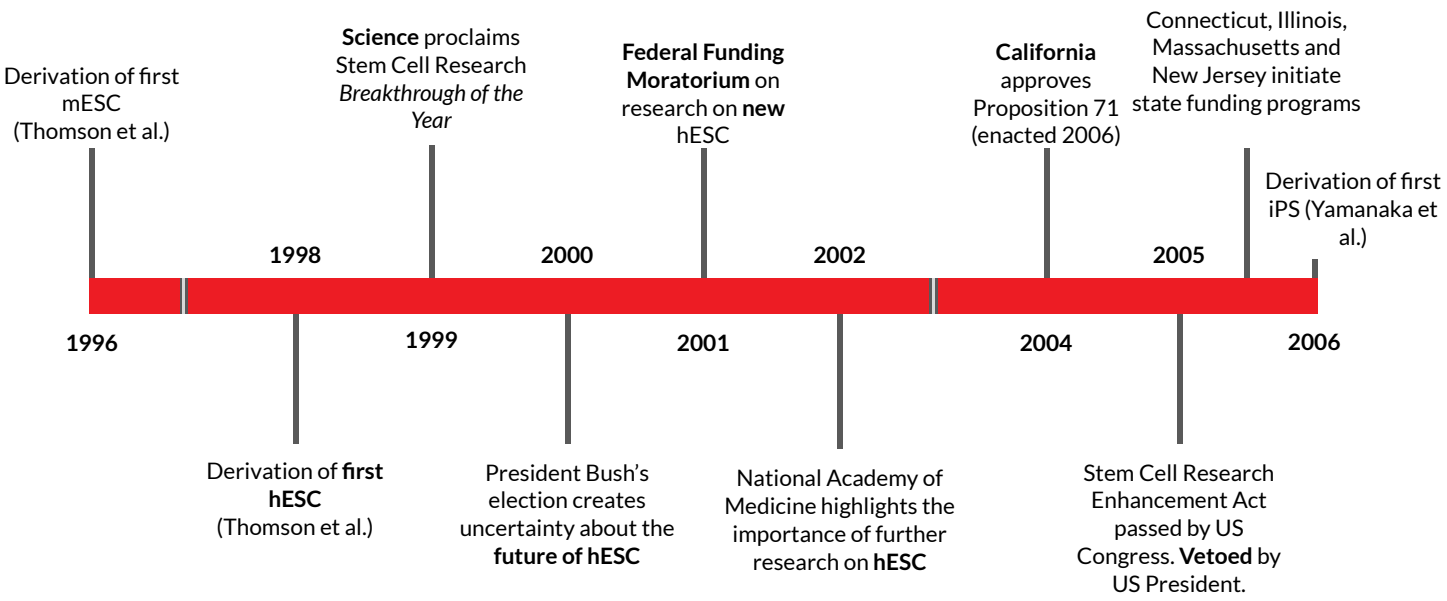


Figure 3.1 – **hESC Timeline** Schematic timeline of the period of analysis

nature and nuanced limitations, as it has also been argued by Furman et al. (2012) and Huang and Jong (2019). The policy did not place restrictions on the topic, but rather in the materials (namely the cell lines) that researchers could use. Thus, the years following the ban until 2006 —when state-level funding was largely introduced and iPSC were first developed— provides a quasi-natural experiment through which to study the impact of the policy shift.

Numerous publications have explored the 2001-policy effects on science production patterns and geographical distribution (McCormick et al., 2009; Owen-Smith and McCormick, 2006; Scott et al., 2009; Löser et al., 2010; Levine, 2004; Furman et al., 2012). Recent work has mapped the uncertainty to a decrease in R&D investments in the private sector for Stem Cell-related projects (Huang and Jong, 2019). The previous literature has captured rates of production of science as well as industry investment in the aftermath of the ban as a result of the shock using different techniques and models. While this work has paved the path to understanding the phenomenon, it is yet to exploit the specificity of the intervention. We, therefore, extend the previous literature by capturing the context in which research direction (academic freedom) was limited. A central element of our research design is the identification of the funding source of research publications. Thus, we can accurately delineate those publications that were limited in their materials by the ban. We subsequently use the exogenous shock to examine the effects of such limitations on the innovation pipeline.

3.3 Setting, Data and Descriptive Statistics

In this section, we provide a detailed summary of the identification strategy, the process through which we compiled data and the Machine Learning (topic modelling) tools we used to characterise scientific articles and hESC research. Finally, we provide some descriptive statistics of the sample.

Measuring the impact of basic science

The spillovers generated by basic science have been a subject of study by economists for a long time. Mostly supported by public expenditure, the actual return to investments in science are, however, hard to measure. Innovative activity and private sector knowledge spillovers have, however, been largely accounted for. A common approach to measuring the innovative activity and such spillovers is through patent citations to other patents (Jaffe et al., 1993). While the idea of patent-to-article citations is not new (Trajtenberg et al., 1997), it has not been thoroughly exploited to capture the impact of science in downstream innovation until recently (Ahmadpoor and Jones, 2017; Fleming et al., 2019; Poege et al., 2019). These publications show how a large portion of patent activity traces back to scientific advancements, and their quality signals the patent impact. Hence, by linking patents back to the articles that they cite, we can measure the influence of basic science on innovations.

Particularly in the life sciences, innovations are thought to be fundamental drivers of economic

and social welfare (Azoulay et al., 2019). Due to the enormous development costs and potential profitability, intellectual property protection is widespread in the life sciences. For the case in hand, in 2001 hESC were seen as critical to the development of gene and cell therapy (Murray, 2007; Cyranoski, 2018). hESC showed potential to address novel therapies in yet-to-be tamed diseases such as Parkinson's or diabetes. It is not uncommon for biopharma companies to collaborate with academic researchers, and gradually build up their advancements to shorten product development times. Patents are the primary way that biotechnology and pharmaceutical companies have to appropriate the returns to R&D (Cohen et al., 2000). Furthermore, life science patents cite scientific literature more intensively than other patent categories do (Narin and Olivastro, 1998), allowing us to capture the spillovers better. Therefore, given that the NIH (federal funds) is the largest basic life science funding agency, it is natural to assume that the uncertainty in the policy landscape affected the subsequent rate of innovation. However, we argue that, since the 2001 ban did not prohibit hESC research nor its applications, if substantial developments were still taking place, firms would commit their resources and protect the new developments.

Suppose NIH-funded research was pushing the frontier of knowledge despite the limitations in place. In that case, we should not observe any difference in the follow-on innovation compared to unrestricted research (i.e., non-NIH-funded research). With this in mind, this chapter goes beyond previous work by trying to assess the causal effect on subsequent innovation of public policy hurdles to scientific autonomy. We use the 2001 exogenous shock in a DiD configuration. In order to do so, we combine several data sources as detailed below.

3.3.1 Data

Treatment and Control groups. First, we construct a corpus of hESC-related research spanning the years before and after the ban. Finding publications dealing with a particular subject is no easy task. Traditional approaches involve journal classification, keyword approaches or, in the case of the life sciences, Medical Subject Headings (MeSH) terms. However, being only a nascent sub-field of Stem Cell research, neither approach provides a fine-enough subset of articles. We, therefore, rely upon a more canonical approach to scientific discovery. We use review articles as a starting point, to capture the underlying (original) contributions which constitute a significant (impactful) subset of the publications on the topic.

We are particularly interested in articles between 1996–2006. To collect them, we gather all PUBMED articles between 2000 and 2012 related to hESC through a search engine (The Lens).[2] We match the results with PUBMED, we filter them by document type, references and citations, limiting ourselves to *"Reviews"* with at least 50 references and cited at least 60 times. Consequently, we filter out concise review publications and work of lesser impact. Finally, we restrict the sample to those publications with (explicitly) "human embryonic stem cells" (or hESC) in their titles and abstracts. This procedure yields 69 Review articles with over

[2] www.lens.org

12,000 references, of which 3930 are unique. Filtering out news, opinions, reviews and other non-relevant references and limiting ourselves to the period between 1996–2006, we finally obtain 1885 core articles that constitute our unit of analysis.

Once the core group of articles has been defined, the next step is to characterise the treatment and counterfactual groups that allow us to estimate the trajectory of follow-on innovations. To analyse the causal effect on innovation of the ban, we need to unambiguously identify the recipients of public funds from the federal government. In order to do so, we incorporate Funding and Acknowledgement Data from Clarivate's Web of Science, enriched with PUBMED's API. Despite the efforts in recent years to populate these databases backwards, there are plenty of missing data points. About a 65% of the articles are missing some, or all, of the acknowledgement data. Fortunately, the majority of omitted funding (or acknowledgements) data is for non-U.S. publications. In either case, we manually annotate 931 full-text PDF publications to discern whether they were a recipient of US federal funding.

Despite the careful identification above, the manual labelling allowed us to find a large proportion of articles questionably related to hESC. We use topic modelling techniques, in particular, Doc2Vec and Word2Vec (see Chapter 1 for a detailed description of the methods) to help us differentiate hESC from other neighbouring topics. Using free text data from titles and abstracts, we simultaneously train a Doc2Vec and a Word2Vec model for all PUBMED-indexed journal articles between 1997 and 2005. To do so, we first group pairs, triplets and quadruplets of words that statistically appear often together. Common combinations of words such as "stem" and "cell" will be grouped as a single token (hence a single vector) whenever they are found next to each other, while still being trained as individual words when other terms surround them. We group words using "normalised mutual information" as a decision tool.[3] As a result, we obtain a list of the most similar terms to "hESC", comprised, amongst others of tokens like: "embryoid body", "murine stem cells" or "pluripotent cells". By strictly selecting the tokens that refer to embryonic stem cells amongst the most similar to hESC, we construct a set of keyword rules to delineate hESC-related and not-hESC-related articles amongst the 1885 present in the set. In particular, we include any document containing at least one of the following stemmed words:[4] "hESC", "human embryonic stem cell", "human ES", "human ES cell", "he cell". Additionally — in combination with the presence of "Humans" amongst the associated MeSH terms or the presence of "human" in the same paragraph (but not right next to them) — we include any of the following: "blastocyst", "embryon stem cell", "embryon (ES) cell", "ES cell", "embryon stem (ES) cell". This process yields 808 articles out of the 1885 that we will consider as hESC-related.[5]

Patents citing scientific literature. Once the treatment and control groups are properly delineated, we link them with a set of patents that reflect the knowledge spillovers into innovation. Patents include a section on Non-Patent-Literature (NPL) citations. While a high proportion

[3] Example source code on https://github.com/oballegon/Text-Similarity-Basics

[4] We stem the words to their roots so that suffixes are eliminated. We capture words such as *embryonal* or *embryonic* under the same token *embryo*

[5] We later provide an alternative identification method which we use as a robustness check

of patent-to-patent citations come from examiners (Sampat, 2010), non-patent-literature is more likely to come from the inventors themselves. In addition, in-text patent citations (in many occasions different from the header citations) provide additional linkages between patents and articles. As a result, we are almost certain to be capturing spillovers from the underlying basic science when we observe direct citations.

The process certainly has its limitations too. We fail to observe interrelated advancements of second degree. That is, citations to academic articles that cite the root article that we observe. The omission of these may lead us to estimate the effect falsely, but we believe that the identification strategy helps to smooth the problem. Having compiled the root science from references in review articles, we expect both the treatment and control samples to be of similar characteristics, in terms of exposure and relevance. The review articles synthesise the knowledge in the field by taking the most eminent (and significant) contributions. Hence, we expect them to be central to patent knowledge too.

Recent work by de Rassenfosse and Verluise (2020) has extracted and parsed NPL citations both from in-text and header sources. We combine their database along with *The Lens* (Jefferson et al., 2018) in order to obtain 76,878 patent citations of which 26,055 are unique.

For the analysis, we further restrict the sample to USPTO granted patents with priority-year dates no older than seven years after the root article publication.[6] As a second step, we extract patent citations to the article-citing patents. For this, we allow a 5-year window since the patent publication. This count that we call "2nd-degree Patent Citations" aims to capture the longer-run effects on innovation, namely the R&D developed from the initial efforts (first patent). Finally, we use PatentsView harmonised organisation data to infer the assignee origin by institution and country (U.S. vs non-U.S.): (1.) Research institutes include all non-profit research-intensive centres such as universities, hospitals, institutes, national institutes; (2.) Private Sector includes all for-profit corporations, laboratories and companies.[7]

Authors. Finally, we enrich the article-level data combining Author-ity (Torvik and Smalheiser, 2009), PUBMED, WoS as shown in Figure 3.2. This way we capture the origin of the coauthors, international collaborations (two or more countries amongst the affiliations), coauthor based in hESC-favorable countries, reprint and last author "ages" (a measure of experience, proxied

[6] For robustness, we also gather patents within 5 and 10 years, which provide no significant differences in the count distribution. USPTO patent kind codes changed precisely in 2001. Therefore, in order to capture granted patents, we compile patents with Kind Code "A" until 2001 and kind codes "B1" and "B2" afterwards.

[7] Despite restrictive, limiting the patent counts to USPTO granted patents is both convenient and does not necessarily diminish the potential impact of the ban. On the one hand, it is fair to assume that highly relevant medical innovations will seek US patent protection, regardless of origin and will, therefore, be captured by USPTO patents. On the other hand, more than two-thirds of our citations are by USPTO patents. These figures are not a particularity of our dataset, but rather a general trend in the time of analysis: the USPTO documents cite 3.5 times more Non-Patent-Literature than the European or Japanese counterparts (Michel and Bettels, 2001). Therefore, we argue that despite having filtered, we are still capable of capturing the innovation spillovers at stake. Additionally, USPTO patents are analysed in Patentsview, incorporating much relevant information on the assignees and inventors. Thus, we can efficiently collect harmonised organisation (patent holder) data by origin, without the burden of additional disambiguation.

by years since their first publication indexed by PUBMED), Scholarly Citations (5-year window), coauthors affiliated to the private sector, journals, journal quality (Scimago Journal Rank) and the number of coauthors.

The sample contains 1170 unique principal investigators (PIs), identified as the last author in each publication. The hESC-related article sub-sample, with 806 articles, has 6.5 authors per paper on average. Before the ban (up to 2001), 977 researchers are cited as authors of hESC publications, of which 769 unique. After the ban, the 4269 researchers, of which 2776 unique, are listed as coauthors of a publication in the sample. Only 210 researchers are present and active both before and after the ban in the sample.

3.3.2 Descriptive Statistics

Table 3.1 displays the summary statistics of the full article sample. Amongst the 1885 articles, 70% were published in 2001 or later. Overall, 35% received federal funding, representing a large proportion of articles with at least one coauthor based in the US. Reprint authors for 44% of the articles are based in the United States, so we capture a large proportion of research that is led from the US, and is potentially affected by the moratorium. Scientific articles receive almost ten patent citations on average within the first seven years after publication. Moreover, a large proportion, 63%, are being cited by patents. Follow on innovation is rather consistent across the sample, with 47% of the articles having 2nd-degree patent citations (within five years of the original patent-to-article citation).

Prevalence of international collaborations is lower on the sample. Only 26% have coauthors with affiliations pertaining to (at least) two different countries. However, about one-fourth of the articles have collaborators from one of the hESC-affine countries (Israel, Singapore, Denmark, UK, Taiwan).

Table 3.2 shows the summary statistics of the data grouped by hESC and Federal Funding. The prevalence of patent citations to articles is significantly larger for publications with some sort of federal funds. However, the average number of citations received inverts the trend. Federally funded hESC research receives, on average, fewer citations than non-federally funded, while the contrary is true for non-hESC articles. The same trend follows for second-degree patent citations, suggesting the innovation spillovers from these articles might be lower. The table also documents how the presence of authors from hESC-favorable countries is larger amongst non-federally funded publications and significantly larger for hESC-related articles in general. Remarkably, the average number of patent citations follows the same trend as in the full sample described above for all the organisation origins except for private-sector U.S. patents, which cite, on average, more Federally Funded hESC patents than non-hESC.

Table 3.1 – Summary Statistics Full Sample

	Mean	*StD*	*Min*	*Max*
Year	2002.05	2.99	1996	2006
Year ≥ 2001	0.69	–	0	1
Years 2001-2003	0.29	–	0	1
Federal Funding	0.35	–	0	1
hESC-related	0.43	–	0	1
At least one Author in USA	0.54	–	0	1
Reprint Author in USA	0.44	–	0	1
Number of Authors	6.24	6.26	1	223
Cited by Patent	0.63	–	0	1
Cited by 2nd degree Patent	0.47	–	0	1
Scholar Citations 5y	84.65	149.36	0	3211
Patent Citations 5y	7.82	18.18	0	347
Patent Citations 7y	9.49	22.21	0	413
Patent Citations 10y	11.05	26.11	0	447
2nd degree Pat. Cit.	30.61	105.34	0	2152
Research Institute No-US Pat Cit	0.65	2.37	0	64
Research Institute US Pat Cit	1.81	5.21	0	109
Private Sector No-US Pat Cit	1.35	3.97	0	55
Private Sector US Pat Cit	4.38	11.15	0	136
IProduct 2nd Degree	1.10	9.59	0	228
Coauthor from Priv Sector	0.15	–	0	1
International collab.	0.26	–	0	1
CoAuthor from fav. country	0.23	–	0	1
US Contract Patents	0.04	0.41	0	6
US Grant Patents	1.65	5.63	0	77
Similarity to hESC	0.37	0.12	−0.02	0.71
Articles	1885			
Journals	332			
P.I.	1170			

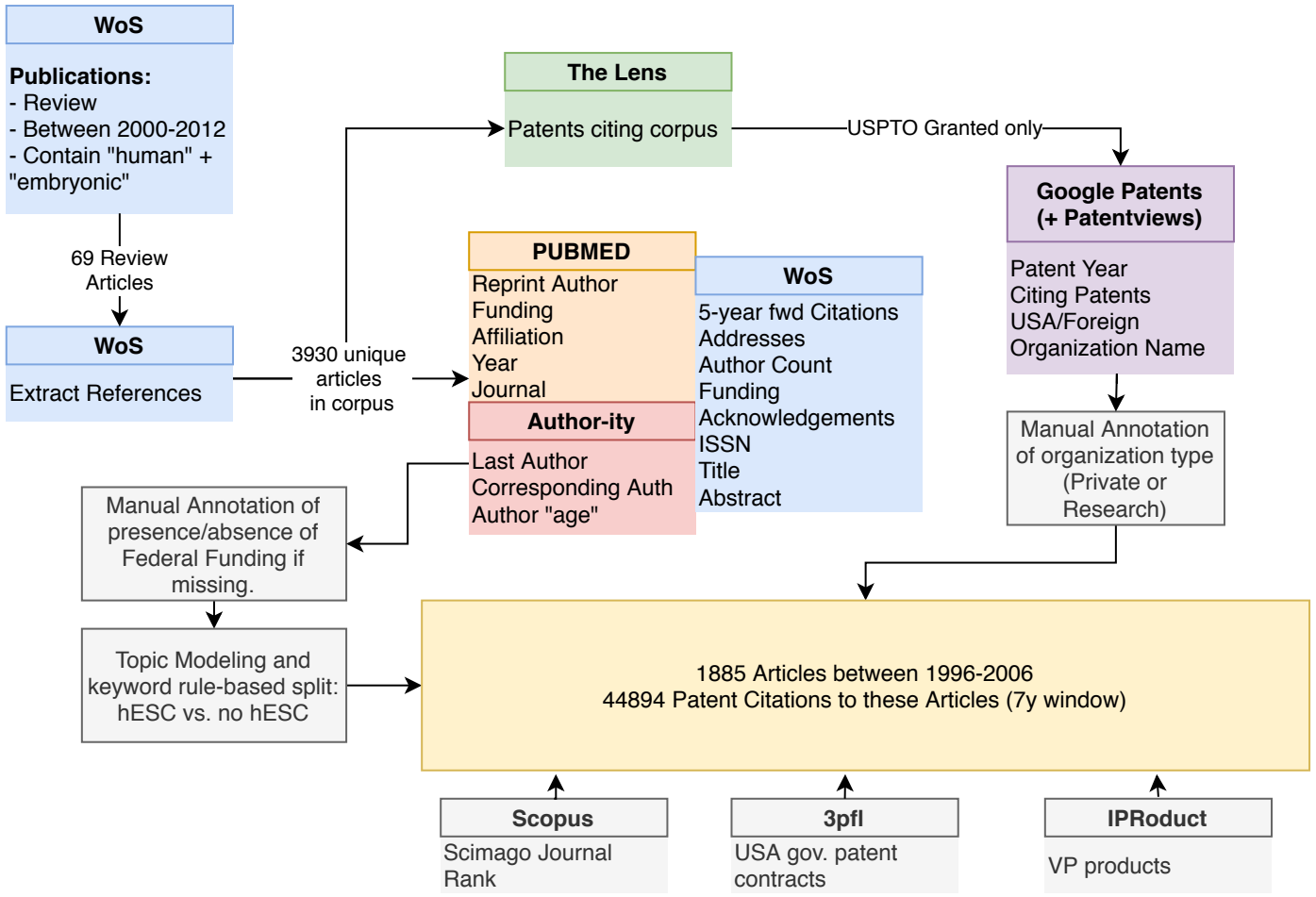


Figure 3.2 – **Data Construction** Schematic summary of data construction

Table 3.2 – Summary Statistics by group

	No Federal Funding (n = 1216)				Federal Funding (n = 669)			
	no hESC		hESC		no hESC		hESC	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Year	2001.28	2.95	2003.02	2.64	2001.17	2.95	2003.40	2.74
Year ≥ 2001	0.59	0.49	0.82	0.39	0.60	0.49	0.85	0.36
Years 2001-2003	0.31	0.46	0.27	0.45	0.34	0.47	0.20	0.40
At least one Author in USA	0.30	0.46	0.27	0.44	1.00	0.00	1.00	0.00
Reprint Author in USA	0.22	0.41	0.19	0.39	0.90	0.30	0.83	0.37
Number of Authors	6.03	9.44	6.33	3.31	6.04	3.53	6.93	4.04
Cited by Patent	0.52	0.50	0.71	0.45	0.61	0.49	0.77	0.42
Cited by 2nd degree Patent	0.38	0.49	0.54	0.50	0.47	0.50	0.59	0.49
Scholar Citations 5y	62.25	104.44	75.64	161.50	121.20	185.03	101.79	139.73
Patent Citations 7y	5.48	12.74	13.23	30.18	9.18	21.47	12.19	19.98
2nd degree Pat. Cit.	19.66	84.92	39.92	136.16	32.84	102.35	34.63	70.83
Research Institute No-US Pat Cit	0.30	1.07	1.10	3.62	0.46	1.59	0.88	2.20
Research Institute US Pat Cit	1.12	3.25	2.00	6.21	2.57	6.68	1.87	3.59
Private Sector No-US Pat Cit	0.97	2.77	1.89	5.09	1.30	4.34	1.22	2.82
Private Sector US Pat Cit	2.33	7.01	6.49	15.13	3.75	8.92	6.11	11.84
IProduct 2nd Degree	0.75	8.12	1.13	10.41	1.88	12.50	0.59	3.46
CoAuthor from Priv. Sec.	0.12	0.33	0.20	0.40	0.12	0.33	0.20	0.40
International collab.	0.25	0.43	0.23	0.42	0.30	0.46	0.28	0.45
CoAuthor from fav. country	0.24	0.43	0.34	0.48	0.10	0.30	0.16	0.36
SimHesc	0.31	0.10	0.45	0.09	0.29	0.09	0.44	0.08
Observations	654		562		425		244	

Summary Statistics by group

3.3.3 Validity

In this section, we put to the test the validity of our treatment and counterfactual articles. In order to do so, we replicate the main results from Furman et al. (2012). In their article, the hESC-related articles are extracted from the NIH report 2001 on stem cell research. They identify 17 articles published before 2001 which would have been subject to the ban had they been published later. As a counterfactual, they suggest three candidates: (1) RNAi publications (a separate, equally novel and promising field of research); (2) "nearest neighbour articles" that appeared in the same journal issue as the core hESC articles; (3) "other stem cell articles", i.e. adult/animal stem cells. If anything, we would expect our counterfactual to behave similarly to the nearest neighbour control group. Nearest neighbours will be composed of articles describing similar methods or concrete diseases that hESC might target, and they would, therefore, be captured trough the review-reference identification we suggest.

Without entering in too much detail here, since it escapes the purpose of this chapter, the authors explained the effect of the ban on U.S. science and firms using a conditional fixed effects (Hall et al., 1984) regression in the lines of:

$$\begin{aligned} \text{CITES}^r_{it} = \quad & \epsilon_{it} + \gamma_i + \beta_t + \text{age}^r_{it} + \\ & \alpha_0(\text{hESC}_i \cdot 2001_{it}) + \alpha_1(\text{hESC}_i \cdot (t > 2001_{it})) + \\ & \phi_0(\text{US}^r_t \cdot \text{hESC}_i \cdot 2001_{it}) + \phi_1(\text{US}^r_t \cdot \text{hESC}_i \cdot (t > 2001_{it})) \end{aligned} \tag{3.1}$$

where *CITES* is a per-year count of citations (or projects) to focal publications from two different stacks: U.S. and non-U.S.; hESC is a dummy for hESC-related articles; 2001 and > 2001 are time dummies that cover the specified periods; age represents the years since the publication of the focal article; and γ are publication fixed effects. Therefore, ϕ_0 and ϕ_1 are the coefficients of interest, indicating the marginal impact of the policy intervention on US citations or projects.

We present the results of regressing Equation 3.1 on our data in Table 3.3. The coefficients (and the Incidence Rate Ratios are consistent (and very similar in value) to the estimated coefficients by Furman et al. (2012) (Table 4, column 4-2). We therefore conclude that, despite potential pitfalls and selection biases, our core sample and counterfactual are correctly identified so as to perform a quasi-replication of contrasted results.

Table 3.3 – Data validation: replication of Furman et al. (2012)

	Conditional fixed effects negative binomial, stacked DV = Cites with reprint Author from USA (or non USA Reprint Author)	
	(1)	(2)
hESC×2001	0.411***	0.413***
	(0.103)	(0.103)
hESC×(2002-2005)	0.674***	
	(0.103)	
hESC×(2002-2003)		0.485***
		(0.103)
hESC×(2004-2005)		0.690***
		(0.103)
USA×hESC×2001	−0.388***	−0.388***
	(0.067)	(0.067)
USA×hESC×(2002-2005)	−0.496***	
	(0.031)	
USA×hESC×(2002-2003)		−0.451***
		(.046)
USA×hESC×(2004-2005)		−0.533***
		(0.041)
N	6977	6977
Number of articles	577	577
Log-likelihood	−15652.191	−15653.077

Standard errors in parentheses

Models include constat, hESC*Year FEs, article age FEs and article FEs.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.4 Methods

Our research approach involves an econometric estimation of the causal impact of the 2001 hESC ban on the innovation pipeline. For this, we estimate the treatment effect of federally-funded research on pooled hESC-related publications:

$$E\left[y_{it}|X_{it}\right] = f\left[\epsilon_{it}; \quad \beta_0 + \beta_1 \text{BAN}_t + \beta_2 \text{FedF}_i + \beta_3 \text{FedF}_i \times \text{BAN}_t + \delta_t\right] \tag{3.2}$$

where the dependent variable y_{it} is a measure of downstream innovation impact, namely Patent Citation counts or second-degree Patent Citation counts; BAN is an indicator variable that switches to 1 in 2001; FedF is an indicator variable that equals one when federal funds financially support at least one of the coauthors; δ_t are publication-year fixed effects; and X_{it} is a vector of article characteristics (controls).

The coefficient of the interaction term β_3 identifies the difference in follow-on innovation experienced by federally funded research on hESC relative to the control group. It is the central focus of our analysis. In other words, it indicates the additional increment (or decrement) to innovative output that hESC Federally-Funded articles published during the ban receive, relative to the rest of the sample.

This model specification allows us to causally interpret whether the limitation in the cell lines *allowed* by the 2001 ban had an effect on downstream applicability of advancements at the frontier. The particular characteristics of the policy shock, which did not completely ban research but rather put limitations on new cell lines, provides a unique framework that allows us to study the impact of hurdles to the frontier of research. By regressing Patent Citations, we can causally infer the effect of limiting the applicable methods in the innovation pipeline.

Matching

As an attempt to overcome the selection bias present in the data, we estimate the effects of the ban using treatment weights for the estimation. While the Stem-Cell ban provides a quasi-natural experiment scenario, it only affects Federally Funded Research. NIH accounts for the vast majority of (federal) funding in the life sciences (Azoulay et al., 2019), and funding works through grants. Financial support is allocated through a set of future expectations and past performance of the researchers. Therefore, we must confront a selection bias in NIH-supported research and address a potential imbalance between the treated and untreated groups. On the other hand, the confronting effect between uncertainty, limited access to funds and field relevance (trendiness) introduces additional heterogeneity in the publications. There are notorious differences between state-funded and corporate science (Matheson, 2008), which we might see in the data.

To overcome these challenges, we use coarsened-exact matching (CEM) (Iacus et al., 2012). The difference with other matching methods happens in the coarsening stage, where indis-

tinguishable values are grouped and then matched. The imbalances are eliminated within strata, which results in a minimal model dependence (Ho et al., 2007) (due to the remaining differences being constrained by the coarsening). Figure C.1 in Appendix C.2 illustrates the results from CEM matching for Patent Citations.

To satisfy the assumption of ignorable treatment assignment, it is important to include in the matching procedure variables known to be related to both treatment assignment and the outcome (Rubin and Thomas, 1996; Heckman et al., 1998). We use Year, Reprint Author Origin and Number of Coauthors. For robustness, we also tested the inclusion of journal identifiers, with no substantial difference in the results. However, journal placement may have been influenced by the treatment, and we finally decided to exclude the variable from the matching procedure. In order to evaluate the performance of the matching, we use the multivariate distance L_1, as suggested by Iacus et al. (2012). L_1 is effectively a distance metric between the multivariate distribution of all the possible binings of the raw data, with values ranging from 0 to 1, increasing with the level of separation (complete overlap corresponds to $L_1 = 0$. In our data, CEM-matching improves the L_1 imbalance metric from 0.6828 down to 0.3354. There are just 22 unmatched samples out of 244 treated elements. For the pre-ban period, the L_1 imbalance metric from 0.7056 down to 0.3871.

Model Specification

The dependent variables of interest, including citations from patents, citations from 2nd-degree patents and scientific publications are skewed and non-negative count data. We will use, as is standard, models for count variable outcomes, namely from the *Negative Binomial* family (Hall et al., 1984). We use a logarithmic link function so that the coefficient estimates remain consistent, and robust (sandwich) standard errors that account for heteroskedasticity. Because the Poisson assumes equidispersion, the descriptive statistics shown on Table 3.1 suggest that the estimators will be biased (ML-based estimator). Figure C.2 in the appendix shows the kernel density estimates of the scientific article and patent counts.

The Negative Binomial Regression Model (NBRM) adds a parameter to the Poisson Regression Model (PRM), which allows the conditional variance of y_{it} to exceed the conditional mean. This term introduces variation due to *unobserved heterogeneity*. There is, however, an alternative explanation to the presence of this term, which is more suitable for the current exercise. It is based on the idea of *contagion*. Contagion occurs when the probability of an event occurring changes as events occur. Given the nature of citations and the visibility they bring, it is highly likely than receiving a citation changes the probability of receiving future citations. Hence, the probability is likely to change with the arrival of citations. Hence, contagion violates the independence assumption of the Poisson distribution, but not for NBRM.

We provide additional evidence in appendix C.3. Table C.1 provides a comparison of the model statistics for PRM, NBRM and, additionally, Zero-Inflated Poisson Regression Model. For the three main dependent variables for which we display results, both Akaike's and Bayesian's

Information Criterion (AIC, BIC) point towards NBRM as the model that best describes the data. Additionally, the one-tailed test of H0: $\alpha = 0$ for NBRM is rejected at $p < 0.01$, where alpha is the free parameter distinguishing NBRM from PRM.[8]

Alternative sample

Using the topic modelling methods described in Chapter 1 as specified in Section 3.3, we use the trained model to generate an alternative treatment and control groups. We compute the word embedding for the 4-gram "human embryonic stem cell", and we calculate its cosine similarity with each of the document embeddings of the articles in the sample. The density plots of the article-word similarities are represented in Figure 3.3 for each of the two groups: hESC and no-hESC (manually labelled using the keyword rules previously described).

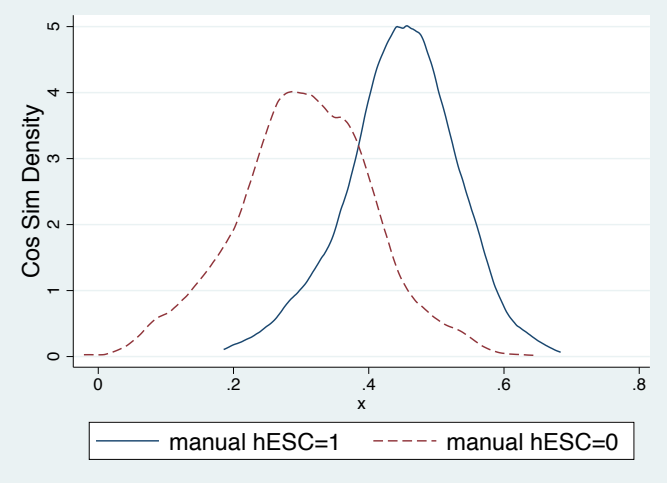


Figure 3.3 – **Kernel Density Plot of Cosine Similarity** between root articles and word embedding of "human embryonic stem cell". The blue line represents the keyword-rule-assigned hESC articles, with which we perform the core analysis of this chapter. The dotted red line represents the non-hESC articles according to the keyword classification.

For the alternative sample, we then proceed to split the sample between hESC-related and not-hESC related publications at the similarity value of CosSim = 0.38. This value is chosen arbitrarily while still fulfilling the following: it is the rounded value median similarity for all the sample (1885 articles); it is the median and mean value for the articles published after 2001; it corresponds to approximately the 25% percentile of the keyword-indexed hESC-related articles from the entire sample. We, therefore, capture a higher number of articles, while only discarding a small portion of the original sample. This method corrects potential false positives in the keyword rule approach by relaxing the imposed keyword rules. Having imposed a strict set of rules as described in Section 3.3, the false positives (i.e. the articles identified as hESC that in reality are not) are few and still likely related to hESC. However, false negatives are, potentially many (i.e. the articles not identified as hESC that, in reality, are hESC-related). The

[8] As α approaches 0, NBRM converges into PRM.

exact combination of words might not appear in the title or abstract, but the article still be highly related to hESC research, describing a method, or other nomenclature that we may miss with the keyword approach. Therefore, the article-to-keyword similarity cut-off may help incorporate these closely-related articles that were otherwise left out.

3.5 Results

In our regression analysis, we start by presenting the results that pertain to the main effect of the Federal Funding Ban to hESC on the rate of innovation. For this, we use a difference-in-difference approach, as shown in Equation 3.2, where the treatment is having at least one researcher with acknowledged Federal Funds. The dependent variables of interest are Patent Citation Counts, and 2nd-degree Patent Citation Counts to the focal publications. We subsequently attempt to find the drivers of the effect and potential mechanisms by slicing the dependent variables. In particular, we regress citations by origin (U.S./Non-U.S.) and Institutional Origin (research-centred institution/corporation). Finally, we dwell into the quality of the publications as a potential mechanism that explains the effect. For clarity, we present only summarised tables in the body of the chapter, and full tables in Appendix C.

3.5.1 Main effect of the ban

We present our regression results, starting in Table 3.4. We include publication-year fixed effects in order to account for the between-year citation differences. The coefficient of the variable *Federal Funding*× *Ban* describes the average difference in the number of citations received by focal hESC (scientific) publications from patents and 2nd Degree Patents. This coefficient is negative and significant under the different specifications, suggesting that, relative to non-federally-funded research, these publications led to fewer follow-on inventions after 2001. The effect holds in sign and significance after introducing controls for the quality of the publication (JIF), Reprint author location (U.S./non-U.S.), presence of private-sector-affiliated researchers amongst the authors, International Collaborations, presence of coauthors from hESC-favorable countries and researcher age (columns (2) and (4)). Transforming the coefficient to an incidence-rate ratio, which refers to the percentage change compared to the reference group, suggests that federally-funded article citations from patents fell up to 65 to 85 per-cent during the ban. Unsurprisingly, the coefficients measuring the quality of the publication (proxied by the Journal Rank Score) and the presence of a private-affiliated author are both positive and significant (not reported).

Table 3.5 examines the alternative sample, constructed as explained in Section 3.4. Columns (1) and (2) show the results of a negative binomial regression, while columns (3) and (4) show the unit-offset logarithmic counts linearly regressed. The variable of interest is, once again, negative and significant, albeit the effect is smaller.[9] Similarly, although we believe

[9]The pattern, however, is different for the non-interacted variables. While it allows us to confirm the direction of the results illustrated in Table 3.4, we believe these results are inherent to the construction of the sample. Through

Table 3.4 – NB Regression; Treatment=Federal Funding, hESC=1

	Patent Citations		2nd Degree Pat Cit	
	(1)	(2)	(3)	(4)
Federal Funding	1.579***	2.805***	3.859***	7.253***
	(0.483)	(0.674)	(0.766)	(1.295)
	[4.84]	[16.5]	[47.4]	[1412]
Ban	2.462***	3.746***	4.075***	6.984***
	(0.379)	(0.584)	(0.604)	(1.122)
	[11.73]	[42.3]	[58.9]	[1079]
Federal Funding × Ban	−2.070***	−3.229***	−4.368***	−7.588***
	(0.524)	(0.688)	(0.801)	(1.283)
	[0.13]	[0.039]	[0.013]	[.0005]
Year FE	*Yes*	*Yes*	*Yes*	*Yes*
Article Controls	*No*	*Yes*	*No*	*Yes*
$log(\alpha)$	0.751***	0.636***	1.504***	1.397***
	(0.0950)	(0.0939)	(0.103)	(0.107)
Observations	655	655	655	655
Log-likelihod	−2369.8	−2331.1	−2545.2	−2510.2

Incidence-rate ratios in brackets

Standard errors in parentheses adjusted for heteroskedasticity

Including unreported constant, controls for Reprint Author location, International colab.,JIF,
CoAuthor from hESC-favorable country, Number of Authors, Researcher "age",
and Presence of Private-Sector-Affiliated Authors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.5 – Alternative Sample; Treatment=Federal Funding, Sim-hESC>0.38

	NB Reg		Unit-offset-log (OLS)	
	Pat Cit	2nd Pat. Cit.	Pat Cit	log(2nd Pat. Cit.)
	(1)	(2)	(3)	(4)
Federal Funding	0.537	0.860	0.210	0.456
	(0.483)	(0.578)	(0.444)	(0.692)
Ban	1.131**	0.0751	0.776	0.0438
	(0.568)	(0.628)	(0.661)	(1.181)
Federal Funding × Ban	−1.058**	−1.312**	−0.818*	−1.293*
	(0.497)	(0.580)	(0.473)	(0.726)
Year FE and Controls	*Yes*	*Yes*	*Yes*	*Yes*
$log(\alpha)$	0.711***	1.376***		
	(0.106)	(0.105)		
Observations	677	677	677	677
Adjusted R^2			0.177	0.196
Log-likelihod	−2486.2	−2753.5	−1153.4	−1367.2

Standard errors in parentheses adjusted for heteroskedasticity

Including unreported constant, controls for Reprint Author location, International colab.,JIF,
CoAuthor from hESC-favorable country, Number of Authors, Researcher "age",
and Presence of Private-Sector-Affiliated Authors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the comparison is illustrative to demonstrate robustness with a different regression model (OLS on log-counts), it is well known that it leads to biased coefficients. Log-linear results with the original sample are reported in Appendix B (Table C.4). The coefficients of interest are similarly significant under the OLS specification.

Because of the heterogeneity in regulatory environments and the particularities of the ban itself, we extend the analysis, including the patent origin for first-degree patent citations. Using the harmonised organisation data from *PatentsView*, we split the patent citation counts by U.S./Other Country and by type of institution. Furthermore, in light of the results obtained by Furman et al. (2012) and Huang and Jong (2019), we interact the treatment variable with the period 2001–2003 immediately following the ban.[10] Table 3.6 shows the results. Like in the previous tables, the interaction of the ban with the treatment is the coefficient of interest. Columns (1)–(4), then, show that the average treatment effect during the ban by patent origin. As in Tables 3.4 and 3.5, the effect is negative and significant except for non-us research institutes where it is not significant. Columns (5)–(6) present an entirely different picture. The strong negative effect disappears for the private sector. The average treatment in the aftermath of the ban (2001–2003) is small and not significant. Thus, it highlights a delayed reduction in the effect post-ban for the private sector, in line with the results of Huang and Jong (2019), albeit from a completely different analysis. The existence of a larger and weakly significant effect for U.S. research institute Patents suggests a different mechanism in response to the ban between the for-profit sector and research-oriented organisations.

The results presented so far are consistent with the claim that the moratorium enacted by the Bush administration had a significant effect on the rate of innovation. Federally funded research resulted in significantly less innovative developments than its counterpart.[11] So, to what extent was this decrease due to the quality of publications? And, more importantly, how did the moratorium affect the direction of federally-funded research? To answer this questions, we examine three new measures for the root publications, namely: (i.) the count of citations from scientific publications as a proxy for quality;(ii.) the Journal Rank Score as another proxy for publication-quality;(iii.) the *topic variety* amongst publications as a way of measuring direction.

similarity metrics, we introduce articles in the analysis that are very close substitutes to the core hESC. Hence, they are a control sample that is within reach of hESC researchers. The topic-based similarity implies these articles discuss either similar techniques or concepts, albeit probably non-hESC. Researchers can transfer relatively easily, hence resulting in not the most suitable control sample.

[10] Furman et al. (2012) show how the rate of publications decelerates in the U.S. in the immediate aftermath of the ban, while researchers figure out ways of overcoming the limitations of the ban. Between 2004–2007 the rate of arrival of citations slowly restores to the pre-ban levels.

[11] Marginal effects unreported due to collinearity between Ban and Year fixed effects. However, we "forced" the calculation for the intuition of the reader, and display the Federal Funding/Ban Marginal effects in Figure C.5 the Appendix

Table 3.6 – NB Reg Patent Citations by Origin; Treat=Fed Fund., hESC

	Research Institute		Private Sector		Research Institute		Private Sector	
	(1) Non-US	(2) US	(3) Non-US	(4) US	(5) Non-US	(6) US	(7) Non-US	(8) US
Federal Funding	0.141	1.834***	3.947***	2.373***	−0.395	0.0160	−0.275	−0.131
	(1.606)	(0.646)	(1.223)	(0.802)	(0.270)	(0.244)	(0.349)	(0.271)
Ban	7.564***	3.593***	3.213***	3.510***				
	(2.492)	(0.675)	(1.047)	(0.836)				
Fed Fund × Ban	−0.734	−2.142***	−4.507***	−2.837***				
	(1.609)	(0.654)	(1.212)	(0.822)				
(2001-2003)					8.042***	3.639***	0.978	1.970***
					(2.698)	(0.877)	(0.960)	(0.751)
Fed Fund ×(2001-03)					−0.854*	−0.884*	−0.394	−0.527
					(0.513)	(0.465)	(0.498)	(0.448)
$log(\alpha)$	0.765***	0.901***	1.226***	1.160***	0.744***	0.904***	1.296***	1.192***
	(0.220)	(0.166)	(0.149)	(0.126)	(0.210)	(0.167)	(0.139)	(0.125)
Observations	655	655	655	655	655	655	655	655
Log-Likelihood	−894.6	−1218.3	−1040.4	−1777.2	−893.1	−1222.3	−1055.4	−1787.0

Standard errors in parentheses adjusted for heteroskedasticity

Including unreported constant, controls for Reprint Author location, International colab.,JIF,CoAuthor from hESC-favorable country, Number of Authors, Researcher "age", and Presence of Private-Sector-Affiliated Authors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.5.2 Mechanisms of Response

The analysis heretofore has focused on identifying to what extent the ban had an impact on the innovation landscape. In Table 3.7, we examine changes in the scientific output that may, to some extent, explain the relative decrease in innovative output from federally-funded research. Columns (1)–(3) present a negative and (weakly) significant effect in the average scholarly citation counts received by *treated* publications during the ban. This result suggests that lower-quality output might be related to the average treatment effect observed on Table 3.4. Similarly, columns (4)–(5) show a negative and significant effect on the interaction term as well. This time, however, the dependent variable is the Journal Rank Score. Scientific articles with at least one acknowledged federal source of funds landed in lower-ranked periodicals.

Overall, the coefficients of *Federal Funding* × *Ban* suggest that lower-impact work might be the root of the observed decrease in innovative output. Nevertheless, withal, the policy did not place restrictions on expenditure, which may explain this deceleration. The ban only limited the materials (namely new stem-cell lines) that researchers could use. We suspect that this constraint to the autonomy of researchers may have created a *niche effect* for federally funded research. Building up only on existing lines enacted an effective limitation on the direction of science, potentially reducing the variety and scope of experiments. The triple interaction term on Table 3.7 (column 3) puts this hypothesis to test. The average treatment effect of a collaboration with an author whose affiliation is based on an hESC-favorable country is positive and significant, suggesting that the lower impact might be due to the novelty and variety (or lack thereof) of the publications.

Using the Doc2Vec model trained, as explained in Section 3.3, we compute the pairwise cosine similarity matrix for different strata of the full sample (hESC and no-hESC) by year. That is, we infer the document embeddings for the $N(t)$ articles published in a given year, and then compute the similarity matrix $M_{N\times N}$, where each element M_{ij} represents the cosine similarity between articles i and j. Each row then represents a vector of similarities S_i of article i such that $S_i = (\mathbf{s}_j) \quad \forall \quad j = 1...N$. The dispersion of each row d_i, calculated as the standard deviation $d_i = std(S_i)$ allows us to measure of how far apart, as a whole, the articles are from article i. In other words, if an article is very close to some, and very far to others, we will observe a large dispersion, and the group of articles has a larger variety. However, suppose the article is very close (or very far) from the majority (concentration of similarities). In that case, the standard deviation will be lower, and we will conclude the N articles are less spread out (highly clustered topic). The average standard deviation $\bar{d}$ is represented in the upper row in Figure 3.4 for different groupings of articles. The standard deviation is independent of the mean, potentially misrepresenting spread for lower means. Essentially, a small standard deviation d_i in a vector S_i where the average $\mu_i = mean(S_i)$ is small, may mean that all articles are far apart, hence very varied. In order to overcome the magnitude problem, we introduce variation as a measure of weighted standard deviation $v_i = d_i/\mu_i$. The lower row in Figure 3.4 shows variation across years.

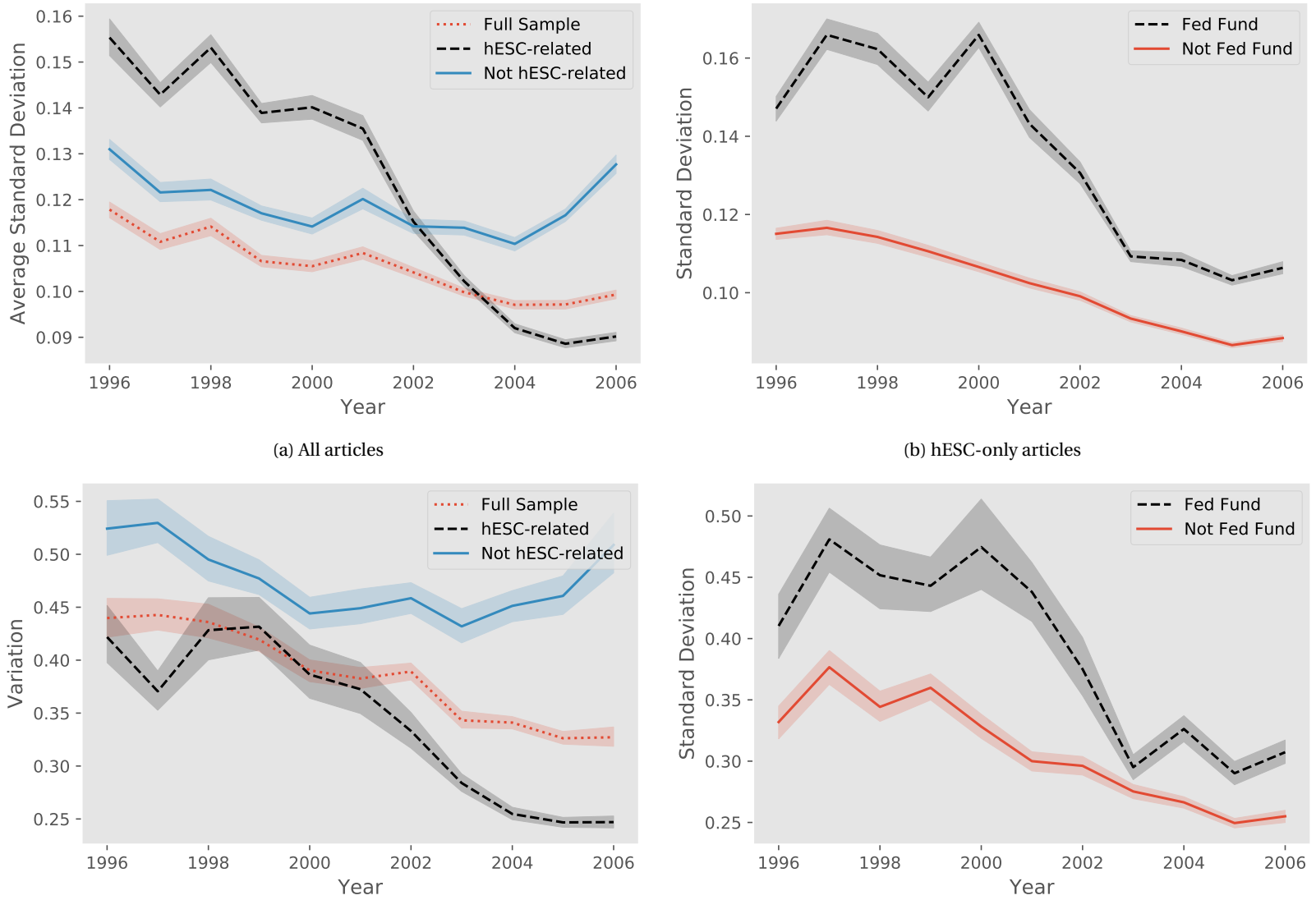
Table 3.7 – Scientific Output: Measures of Quality; Treat=Fed Fund., hESC=1

	Negative Binomial Regression D.V.: Scholar Citations			OLS D.V.: Journal Score	
	(1)	(2)	(3)	(4)	(5)
Federal Funding	0.995	1.191	1.155	1.340*	1.375*
	(0.606)	(0.728)	(0.714)	(0.781)	(0.808)
Ban	2.087***	2.142**	2.150**	−2.779***	−2.726***
	(0.753)	(0.887)	(0.854)	(1.064)	(1.050)
Federal Funding × Ban	−1.238**	−1.427*	−1.406*	−1.416*	−1.444*
	(0.625)	(0.741)	(0.734)	(0.793)	(0.782)
Fed Fund × CoAuthor fav. country			−1.794**		
			(0.812)		
Fed Fund × Ban × CoAuthor fav. country			2.113**		
			(0.870)		
Year FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Journal FE	*No*	*No*	*No*	*Yes*	*Yes*
Article Controls	*No*	*Yes*	*Yes*	*No*	*Yes*
$log(\alpha)$	0.320***	0.214	0.190		
	(0.119)	(0.132)	(0.134)		
Observations	655	655	655	806	806
Adjusted R^2				0.935	0.936
Log-Likelihood	−3646.4	−3608.6	−3600.5	−1358.7	−1348.1

Standard errors in parentheses adjusted for heteroskedasticity

Including unreported constant, controls for Reprint Author location, International colab., Number of Authors, Researcher "age", and Presence of Private-Sector-Affiliated Authors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) All articles

(b) hESC-only articles

Figure 3.4 – **Dispersion across the root articles**: Mean standard deviation (variation) by year with 95% confidence intervals by group. (left) (a) includes the 1885 root articles. (right) (b) includes the 806 hESC-related articles only

This experiment suggests that hESC-related research became narrower in scope in the aftermath of the ban. The effect seems to be mostly driven by federally-funded research, hinting towards another mechanism behind the decrease in innovative impact. The similarity measures illustrate that research after the ban became decreasingly disruptive, as the limitations in place impeded the exploration of new avenues. It seems that new work consisted increasingly of consolidation of previous research, and thus, innovations also tended to cite only the prior-art. If our interpretation is correct, unrestricted research should see its variation increased.

In order to study the spread of the publications in our sample beyond simple conditional means, we regress (OLS) variation. The results are presented in Table C.7 in the appendix. This table expands the information presented in Figure 3.4 accounting for additional covariates while interacting the variables of interest. The interaction coefficients between treatment and ban are all negative and significant, meaning that federally-funded research was, on average, less varied in scope. Column (5) shows a triple interaction that includes collaboration with a coauthor from an hESC-favourable country, which is positive and significant as one would expect from our interpretation.

In this section, we have presented the results from our regression. Using a DiD approach, we have shown that the moratorium lowered the potential for follow-on innovation of federally funded research. We have observed a decrease in the quality of publications with at least one author acknowledging federal funds. Besides, the publications subject to the limitations of the ban were placed in lower-quality journals. A striking pattern emerged post-ban: from a text modelling analysis, hESC publications (and notably federal-funded research) *approached* each other, reducing their apparent diversity. That is, the spread of the topics covered was more limited in scope.

3.6 Discussion

Our analysis indicates that limitations in the materials researchers can use have a negative impact on the innovation pipeline. We demonstrate that the ban on federal funding on certain stem cell lines significantly decreased the downstream applicability of the restricted research. hESC research performed under the federal regulations received up to 85% fewer patent citations, relative to hESC research performed under less restrictive grounds. The relative decline in innovative output resulting from basic science extends to the second-degree patent counts (downstream applications), suggesting a lower quality of innovations in the first place: not only the publications received fewer patent citations, but the citing patents also received fewer patent citations.

In addition to quantifying the impact of Federal Funding limitations on overall patenting, we also study which types of inventors it affects the most. The extent to which the uncertainties in the regulatory and political landscapes affected the willingness of the private sector to invest in hESC research and commercial applications certainly affected the outcome too (Huang

and Jong, 2019). Additionally, we observe that, on average, the ban affected innovations from private and research sectors in the U.S. to a similar extent. This sharp political decision lead to a new hostile environment for stem-cell research that, interestingly, did not affect innovation instantaneously, but with a lag. While scientific output decayed rapidly, the effect on innovation in the (immediate) aftermath of the ban is smaller for patents from research institutes and negligible for the private sector. These findings are in line with project initiations data (Huang and Jong, 2019). However, previous literature had only quantified the effects of stem cell advancements (and innovations) at large, regardless of funding source. Additional to the prior art, our contribution draws the line between federally funded and non-federally funded research.

It is open to interpretation whether the observed decline in patentability is a direct consequence of the targeted funding policy or other unobserved factors at play. However, by comparing hESC-only publications where the treatment is the source of funding, we are confident that we estimate the impact of the limitations of the ban. Our results show that not only research output (quantity) was lower after the ban (Furman et al., 2012), but the quality decreased overall for federally-funded research compared to other hESC publications. Our analysis of the mechanisms suggests that a lower quality at the frontier of hESC coming from NIH-funded (and other U.S. national agencies) publications might be behind the decreased returns to research efforts. On average, hESC publications after the ban received fewer citations when they were under policy limitations compared to the others. They were also published in journals of comparatively lower rating (and reach).

Discussions with then-members of the research community emphasised the poorer quality of publications. In these, they highlighted that the timing of the ban coincided with a series of high profile papers suggesting that adult-derived cells had pluripotency. One source, in particular, noted that "these studies tended to be less rigorously reviewed as they allowed a more politically-correct alternative to hESC-based cell therapies. The ban also promoted more of that lower-quality *adult stem cell* work until iPS were developed".

Considering the framing of the ban, research support for US hESC put bounds to the direction, more than the rate, of scientific advancement. In our last set of results, we show a decrease in diversity within the topic. Whether the decrease in variety may be the root of lower marginal advancements is certainly debatable. We suggest that federal funds supported research that inevitably became stalled by the lack of newer tools (new stem cells). Therefore, hESC research under the regulated funding scheme became more concentrated around a subset of possible topics, which subsequently hindered its development potential. One should not entirely discard the possibility that these results are data-driven. Using text modelling analysis only, it is possible that the lower variety merely captures a sequential homogenisation of language. It is possible that, as the field developed, phrases and common terms were standardised, hence decreasing the variety that we measure.

We interpret our results as consistent with a picture in which autonomy expands the scientific

competitiveness, and allows researchers to explore new lines of inquiry. In turn, this framing is consistent with Azoulay et al. (2011) who show that, under a less-restrictive grant environment, researchers produce publications with higher degrees of impact and novelty. Our results provide support for the idea that scientific autonomy is a driver of greater innovative output. The incentives provided by the competitive environment in science assure the *inevitability* of scientific findings (Merton and Storer, 1973). Furthermore, our work contributes to framing the theoretical model by Aghion et al. (2008), which attempts to explain the process of innovation based on the trade-off between academic freedom and private-sector focus.

At the time of the ban, hESC research was still a very recent and promising field. Thus, access to private funds, generally available for research that is close to the market, was minimal. Additionally, the first cell lines had been developed under public-private partnerships and had tight IP control over them (Murray, 2007). As a consequence, the quickest way to access funds that enabled the hESC research agenda was through accessible, yet outdated, cell lines. The ban effectively generated lower-quality science, which resulted in a slow-down in time-to-market, and the pursuit of marginally unproductive research strategies.

While it seems clear (taking together ours and previous results) that the ban had a spillover effect on both the corporate and research sectors, our measure of innovation only relies on patenting activity directly related to scientific publications. This shortcoming neglects the effects associated with other research which stemmed from the hESC publications. For instance, it has been argued that, without the experience acquired through trial-and-error with hESC, the derivation of iPS cells would have taken longer (Russo, 2005). We might be, therefore, underestimating the actual social value of research published during the ban. On the other hand, iPSC were first derived in Japan, outside of the ban's jurisdiction. One might argue that the finding was on its way, regardless of the ban (Hagstrom, 1974). A survey conducted in 2010 after a court ruling against Obama's reinstatement of hESC funding found that 41% of U.S. stem cell scientists not working with hESCs reported that the temporary ban impacted their research (Levine, 2011). Moreover, we observe lower quality publications. Be as it may, these results seem to suggest that regulators play a central role in the innovative landscape, and directly or indirectly affect scientific output (and therefore, innovation in the long run). In the next section, we extend our analysis to the policy implications of our results, which, as we see, caused unintended negative externalities.

One limitation and potential future extension of the work presented herein involves the study of the policy shock with the individual researcher as a focal point. The data construction process potentially limits the presence of many articles by the same researchers. Indeed, the publication sample studied incorporates a large proportion of unique authors, with only 210 being active both before and after the ban in hESC-related papers. Therefore, the existence of such a large pool of researchers raises questions surrounding the individual strategic response to the shock, as well as the topics covered in their other publications — and publications in their labs — around the time of the ban.

Policy implications

The causal assessment of the policy implications on innovation provides insights on the degree to which funding agencies (especially larger funding bodies) can shape the direction and rate of production of research. It provides a unique assessment of how limitations or the failure to promote novel research has downstream effects on later-stage (closer to market) innovations. The 2001 hESC ban ultimately affected participation, involvement of the private sector and research trajectory. In a sector highly dependent on basic research public funding, such as biotechnology and pharma, the analysis addresses questions regarding policy effects on research (and technology) competitiveness. The ban had distributional consequences on the areas that researchers could engage into, which signified a decrease in their applicability. Thus, our results add to the extensively discussed question of the relevance of public funding in the early stages of R&D.

More broadly, our findings contribute towards the understanding and development of *a science of science funding* (Azoulay and Li, 2020; Azoulay et al., 2018). Designing scientific funding schemes is crucial to provide the right incentives that encourage the best possible returns (Jacob and Lefgren, 2011; Li, 2017). We conclude that limitations on the methods and materials have negative consequences both in the long-term and in the short term. Higher science quality, unimpeded by any regulatory framework, yields greater inventive value. In the life sciences, where social returns to inventions far exceed the private costs of development, promoting good science is vital for a stronger innovation ecosystem. Funding should be directed at challenging and then polishing the status quo and advancing knowledge. Free agency, autonomy and creativity encouragement have proven to be amongst the most reliable predictors of scientific success, regardless of the funding mechanisms (Ayoubi et al., 2019), and our results align with these conclusions.

Our analysis opens up other relevant questions in the realm of knowledge appropriability — and subsequently, accessibility to scientific findings. The *forced* over-exposure to private funds encourages appropriability — through IP protection— of the findings. Another notable example being the *oncomouse* (Murray, 2010). For hESC, the ban encouraged the entry of other private stakeholders, which meant that the available cell lines came from private investments, and had very tight IP control (Murray, 2007), limiting their use (and development) by third parties. Furthermore, any downstream commercial applications derived from subsequent research (either federally or privately funded) would be subject to the same IP protection as the original cell line. The Open Science debate was not as active in 2001 as it is today, but it was widely acknowledged that openness played a fundamental role in the production of knowledge (Dasgupta and David, 1994; Stephan, 2012). The ban played a role against openness, the capability to share results, and more importantly, the market accessibility of downstream innovations publicly funded. Beyond simple characterisation of the causal effect of the ban on innovation, our results provide a deeper understanding of how subtle or even naive policy interventions affect best-practices that ultimately dampen scientific progress.

3.7 Concluding remarks

This chapter analyses the impact of science funding policy on innovations and technology stemming from basic science. To do so, we exploit an exogenous shock (the 2001 hESC ban) that impeded researchers from using certain materials (namely, newly-derived stem cells) and its subsequent effect on patents. Using a citation-based approach, we find that innovative output substantially decreased in the aftermath of the ban compared to unrestricted science. We argue that lower quality science emerged in the aftermath of the ban, and decreased variety in publications emerged as a result of the limitations. Therefore, we argue, the constraints imposed on scientific output generate contributions of lesser value and limited downstream applicability.

Our work provides insight into the degree to which science funding policy and policy uncertainty shape the direction and impact of, not only research, but also R&D. The complicated hESC situation, however, poses limitations in the generalisation of the effect's magnitude. By restricting our analysis to hESC-related publications, we are capable of providing an integral acumen of the differences in methods and materials to the direction of science. However, failing to account for the specific IP control over particular cell lines overshadows the accounted effects. We tracked citations to hESC-research only, disregarding the displacement effect on surrounding fields that the ban may have caused.

In addition, this chapter provides a hands-on application of the topic-modelling methods described in Chapter 1. We use similarity metrics between the core articles to study their topic convergence over time, to find a narrowing scope of research output. These type of analysis provides a basis for other lines of work comprising the study of novelty, spread and scope for many applications: from scientometric field characterisation to econometric evaluation of specific funding programs.

The results and their limitations altogether open alluring avenues for future research. In particular, the research trajectories of individual researchers directly affected by such a ban would shed new light in the strategic responses of individuals to certain limitations. The broad heterogeneity in the individual's affiliations, prior experience and personal objectives (within research) provide a compelling case for study for the responses to policy shocks. Understanding the career-path dependencies that funding policy and uncertainty create, would shed new light on the mechanisms that drive the production of science and shape its direction, perhaps to a greater extent than the analysis of individual publications.

Acknowledgements

In the course of preparing this manuscript, I had the opportunity to exchange some questions with hESC field experts who were active at the time of the policy intervention. I wish to thank Prof. George Honig, Prof. Loren J. Field, Prof. Mark H. Tuszynski and Prof. Robert Langer for their time and effort to clarify the situation. Their input greatly improved my understanding of the turmoil surrounding the ban.

Conclusion

The body of work presented in this dissertation sets new boundaries for the consolidation of a science of science. Each of the three chapters tackled one of the main open problems in the discipline. Namely: (i) how to improve the instability of topic models applied to scientific text; (ii) why it is relevant to study science from the viewpoint of researchers and (iii) how feeble policy design hinders innovation. Because science does not spontaneously occur in a vacuum, and its outcomes and triggers are deeply intertwined with all aspects of society, multiple research specialities are involved in pushing the field. The analysis conducted within the three chapters of this dissertation approaches an identical number of scientific disciplines and methods which all speak to the methodological infrastructure required to enable more a scientific (science) policymaking.

In this work, first, I target the automatic representation of textual data, as an enabler for richer scientific output analysis. Second, I turn the discussion to the debate of individual incentives and motivation in the organisation of scientific social circles. Finally, I make use of novel data to link research to innovation and impact to measure the effect of a policy intervention. These contributions expand the horizon for a unified, scientific and collaborative science of science policy, as discussed at length along the dissertation. In the following, I summarise the main takeaways from each chapter.

Chapter 1 revisited questions regarding the validity of topic modelling for the characterisation of scientific text. I proposed a simple approach to estimate the statistical robustness of topic models based on pairwise similarity scores between documents. I found that the most extensively used generative model, Latent Dirichlet Allocation (LDA), does not appear to be exceptionally robust as similarity increases. A typical matrix factorisation approach to topic modelling, Non-Negative Matrix Factorisation (NMF), suffers from instability as the latent space size increases, and is comparatively unstable for dimensions far higher than 10. In contrast, Doc2Vec, a neural-network-based approach to paragraph embeddings, performs consistently across retrainings. I further propose a principal component analysis (PCA) based approach to assess the descriptive power of marginal increases of latent dimensions in topic models. I find that, while not perfect, LDA and Doc2Vec produce models that maintain explanatory power into the highest dimensions of the topic space. NMF and neural word embedding aggregation fail to scale with the tested data.

I conclude that neural embeddings appear to produce relatively stable estimates of pairwise similarities, compared to other methods, that allow for different levels of granularity and reflect reality relatively well. Neural network topic models are, thus, a promising avenue of further research and application in the social sciences in general, and in scientometrics in particular. Beyond providing proof of performance superiority in the quest for robustness, Chapter 1 provides the practitioner with a tool-set to self-evaluate the performance of her training of topic models, regardless of the method or data chosen, hoping to improve the credibility concerns that some researchers have raised upon the application of topic models for analytical purposes.

Chapter 2 modelled science from the viewpoint of researchers in a payoff-maximisation game. Starting from a public-good game of equitable allocation of rewards, I introduced appropriability of non-pecuniary payoffs as a distributional factor. This simple modification to a simple game, equivalent to accounting for rivalry, implies that the support for the optimal strategy is unevenly split between the marginal and the average contribution. I argued that it is possible to interpret these components as a tension between exploration (invention) and exploitation (consolidation) strategies. This dichotomy, which emerges from the maximisation of individual payoffs in our model, has been primarily documented in the sociology of science. I then introduced heterogeneity in the model through two particular examples. In the first example, I showed that incorporating weak heterogeneity in the ability of the researchers, the model describes the existence of competition in science. In the second example, I showed that by incorporating upper and lower bounds to the abilities of a group, the model helps account for the existence of well-knit social circles. These circles would explain why there exist schools of thought and new ideas are sometimes challenged.

Science policy practitioners must understand the mechanisms that drive researchers. Chapter 2 provides a new look at how individual incentives might shape science. I argue that these individual incentives are central to the organisation of scientific fields. This understanding ultimately helps design better funding schemes and steer science programs in the desired direction. The model lays out the groundwork for further research in modelling the interactions between researchers, largely unexplored in theoretical economics.

Chapter 3 proposed an evaluation of the impact of science funding policy on innovations and technology. I analysed the 2001 U.S. policy on human embryonic stem cell (hESC) research which limited the materials available for research under federal funding (namely, newly-derived stem cells). Taking advantage of the recent identifications of non-patent-literature and in-text scholarly references from patents, I used a citation-based approach to measure the knowledge spillovers from basic research into R&D. In order to determine the causal effect of the exogenous policy shock, I employed a difference-in-differences setting. I found that innovative output substantially decreased in the aftermath of the ban compared to unrestricted science. I showed that lower quality science emerged after the ban, and the limitations imposed by the new policy affected the variety of topics in published material negatively. Therefore, the constraints imposed on scientific output generate contributions

of lesser value and limited downstream applicability. I employed novel topic modelling techniques to characterise the variety of topics within the core publications. I showed that the years of the policy led to decreasing levels of diversity at the frontier or research.

Chapter 3 contributes to the understanding and development of an empirical science of science policy. Thanks to a careful identification of the publications subject to the policy intervention, I conclude that limitations on the materials available for research have negative consequences in the long-term innovation landscape. This finding supports the growing literature on science funding (Azoulay and Li, 2020), and provides evidence that even nuanced policy modifications have a significant impact in the long-term spillovers. It will be interesting to see, as work being carried out by other researchers, how the partial reintroduction of (state) public funds for stem cell research affected the societal value of the investments. Alternatively, more currently, how the top-down adoption of Open Data and Open Access practices affects researchers' work.

Outlook

The topic models presented in Chapter 1 are techniques with great application potential to the study of science. The technical contribution presented in this thesis should provide practitioners with a better understanding of the advantages and limitations of such models. More importantly, I aim at democratising the use of topic models by suggesting an evaluation method that should ultimately lead to better modelling. Improving the granularity and robustness of topic models, not only increases the credibility and reproducibility of the work developed, but it opens new avenues of research. Highly granular and robust models will enable researchers to study dynamics and changes in the direction of science to a greater extent than ever before. Furthermore, while I demonstrate these techniques using scientific literature, the analysis can be extended to any other source of textual data: patents or trademarks (direction of innovation), news articles (societal challenges)or even regulations (policy or discourse analysis). Following the work presented above, I would like to implement neural network techniques to characterise the mentor-mentee intellectual pathways using full-text analysis of their publications.

Chapter 2 opens promising areas for future research. Its more descriptive and exploratory nature delves into research questions that deserve further attention. However, more importantly, it emphasizes the study of the individual as a driver of science organisation evolution. Following the call for richer data sources in order to study researcher mobility (Fernández-Zubieta et al., 2015), we draw the attention to tracking individual researcher's careers in the knowledge domain. In the parallelism between physical and epistemological "mobility", it would be instructive to study the determinants of *entry* or *exit* of a knowledge area as a competitive strategic response. The tools to empirically analyse sequential discrete games are already used in theoretical economics, and present an opportunity to study social interactions in the public research domain as a competitive market. Likewise, the model presents a succinct

revision of incentives in science, which is useful for policymaking. Understanding the rewards that motivate certain behaviours is crucial for funding schemes to be successful, and for the effective management of scientific portfolios.

Finally, the empirical analysis developed in Chapter 3 has implications on science policy and mechanism design. On the one hand, our findings contribute to the development of a science of science funding, and open questions in the realm of knowledge appropriability. If the limitation to use certain materials in basic science discourages innovation, it is reasonable to ask whether IP-protected basic science is affecting innovation similarly. Similarly, it raises ethical questions regarding the market accessibility of fundamental science under protection, which are currently under scrutiny as Open Science is slowly becoming mainstream. On the other hand, the question of how policy affects the careers of individual researchers remains unsolved. The 2001 hESC ban provides an excellent framework to study the career-path of individual researchers and reflect on how different scenarios affect the direction of their research. For this, the application of topic models such as those described in Chapter 1 presents a promising line of future work. To be more precise, it is possible to study the intellectual-space (space of ideas) mobility of researchers by characterising their contributions dynamically, observing how they approach and distance themselves from given topics, particularly when affected by an exogenous shock.

Final word

The negative effect of policy on innovation observed in Chapter 3 does not call for leaner government intervention, but rather for a redesign of political decision-making in the scientific arena. Overall, my vision is that the science of science policy (and governance) should emerge from a scientific process and a carefully designed analysis. An increasingly scientific policy-making process must rely on the availability and accessibility of data, to test and propose policy measures that are evidence-based and, at the same time, inform future decision-makers.

A number of structural changes — including the policy shift toward Open Access and FAIR data practices (Hodson et al., 2018)— in data availability and interoperability will facilitate the analysis in coming years. Besides, incremental researcher efforts that facilitate the link between different contributions to the knowledge stock — e.g. patent-to-article citations (Marx and Fuegi, 2020; de Rassenfosse and Verluise, 2020) or product-to-patent links (de Rassenfosse and Higham, 2020)— are already showing their potential to address questions related to the societal impact and value of science. Full-text disclosure of text paired with the state-of-the-art computing power can only be expected to improve the analytical capabilities of research.

This dissertation modestly contributes to the development of such an environment, providing a broad discussion that should appeal to practitioners at multiple levels and disciplines. Nevertheless, more work is needed. The hunt for an evidence-based governance must begin by recognising that the generation of further knowledge is among the most important uses of new knowledge (David, 2003). Appropriate disclosure of breakthroughs not only promotes

the creation of incremental knowledge (that increases the knowledge stock) (David and Foray, 1995) but it provides the tools to analyse the production system in place. A clear policy implication arises in the development of a scientific science policy formulation, in that however difficult to predict the future value of discoveries, a full Open Science ecosystem is needed to push the frontiers of knowledge adequately.

A brighter future for science, science policy, and science of science policy awaits.

A Chapter 1: Appendix

A.1 Neuroscience Journals

ISSN WoS Short Name WoS Long Name
0194-2638 PHYS OCCUP THER PEDI PHYS OCCUP THER PEDI
1047-9651 PHYS MED REH CLIN N PHYSICAL MEDICINE AND REHABILITATION CLINICS OF NORTH AMERICA
1571-0645 PHYS LIFE REV Physics of Life Reviews
0091-3057 PHARMACOL BIOCHEM BE PHARMACOLOGY BIOCHEMISTRY AND BEHAVIOR
0898-5669 PEDIATR PHYS THER PEDIATR PHYS THER
1016-2291 PEDIATR NEUROSURG PEDIATRIC NEUROSURGERY
0887-8994 PEDIATR NEUROL PEDIATRIC NEUROLOGY
1353-8020 PARKINSONISM RELAT D PARKINSONISM & RELATED DISORDERS
0304-3959 PAIN PAIN
1330-1403 PAEDIATR CROAT Paediatria Croatica
1539-4492 OTJR-OCCUP PART HEAL OTJR-OCCUPATION PARTICIPATION AND HEALTH
0966-7903 OCCUP THER INT OCCUP THER INT
1028-415X NUTR NEUROSCI NUTRITIONAL NEUROSCIENCE
1300-0667 NOROPSIKIYATRI ARS NOROPSIKIYATRI ARS
1029-8428 NEUROTOX RES NEUROTOXICITY RESEARCH
1933-7213 NEUROTHERAPEUTICS Neurotherapeutics
0148-396X NEUROSURGERY NEUROSURGERY
0344-5607 NEUROSURG REV NEUROSURGICAL REVIEW
1092-0684 NEUROSURG FOCUS NEUROSURGICAL FOCUS
1050-6438 NEUROSURG QUART NEUROSURGERY QUARTERLY
1042-3680 NEUROSURG CLIN N AM NEUROSURGERY CLINICS OF NORTH AMERICA
1424-862X NEUROSIGNALS NEUROSIGNALS
1073-8584 NEUROSCIENTIST NEUROSCIENTIST
1319-6138 NEUROSCIENCES Neurosciences
0306-4522 NEUROSCIENCE NEUROSCIENCE
0168-0102 NEUROSCI RES NEUROSCIENCE RESEARCH
1673-7067 NEUROSCI BULL NEUROSCI BULL
0304-3940 NEUROSCI LETT NEUROSCIENCE LETTERS
0149-7634 NEUROSCI BIOBEHAV R NEUROSCIENCE AND BIOBEHAVIORAL REVIEWS
0959-4965 NEUROREPORT NEUROREPORT
1545-9683 NEUROREHAB NEURAL RE NEUROREHABILITATION AND NEURAL REPAIR
1053-8135 NEUROREHABILITATION NEUROREHABILITATION
0028-3940 NEURORADIOLOGY NEURORADIOLOGY
1303-5150 NEUROQUANTOLOGY NEUROQUANTOLOGY
0893-133X NEUROPSYCHOPHARMACOL NEUROPSYCHOPHARMACOLOGY
0894-4105 NEUROPSYCHOLOGY NEUROPSYCHOLOGY
1040-7308 NEUROPSYCHOL REV NEUROPSYCHOLOGY REVIEW
0028-3932 NEUROPSYCHOLOGIA NEUROPSYCHOLOGIA
0960-2011 NEUROPSYCHOL REHABIL NEUROPSYCHOLOGICAL REHABILITATION
0302-282X NEUROPSYCHOBIOLOGY NEUROPSYCHOBIOLOGY
1176-6328 NEUROPSYCH DIS TREAT NEUROPSYCH DIS TREAT
0090-2977 NEUROPHYSIOLOGY+ NEUROPHYSIOLOGY
0987-7053 NEUROPHYSIOL CLIN NEUROPHYSIOLOGIE CLINIQUE-CLINICAL NEUROPHYSIOLOGY
0028-3908 NEUROPHARMACOLOGY NEUROPHARMACOLOGY
0143-4179 NEUROPEPTIDES NEUROPEPTIDES
0174-304X NEUROPEDIATRICS NEUROPEDIATRICS
0305-1846 NEUROPATH APPL NEURO NEUROPATHOLOGY AND APPLIED NEUROBIOLOGY
0919-6544 NEUROPATHOLOGY NEUROPATHOLOGY
0896-6273 NEURON NEURON
1740-925X NEURON GLIA BIOL Neuron Glia Biology
0960-8966 NEUROMUSCULAR DISORD NEUROMUSCULAR DISORDERS
1535-1084 NEUROMOL MED NEUROMOLECULAR MEDICINE
0028-3878 NEUROLOGY NEUROLOGY
1094-7159 NEUROMODULATION NEUROMODULATION
1074-7931 NEUROLOGIST NEUROLOGIST
0213-4853 NEUROLOGIA NEUROLOGIA
0161-6412 NEUROL RES NEUROLOGICAL RESEARCH
1590-1874 NEUROL SCI NEUROLOGICAL SCIENCES
0028-3843 NEUROL NEUROCHIR POL Neurologia i Neurochirurgia Polska
0470-8105 NEUROL MED-CHIR NEUROLOGIA MEDICO-CHIRURGICA
0028-3886 NEUROL INDIA NEUROLOGY INDIA
0353-8842 NEUROL CROATICA NEUROLOGIA CROATICA
0733-8619 NEUROL CLIN NEUROLOGIC CLINICS
1823-6138 NEUROL ASIA NEUROLOGY ASIA
1539-2791 NEUROINFORMATICS NEUROINFORMATICS
1053-8119 NEUROIMAGE NEUROIMAGE
1364-6745 NEUROGENETICS NEUROGENETICS
1052-5149 NEUROIMAG CLIN N AM NEUROIMAGING CLINICS OF NORTH AMERICA
0947-0875 NEUROFORUM NEUROFORUM
1874-5490 NEUROETHICS-NETH Neuroethics
0251-5350 NEUROEPIDEMIOLOGY NEUROEPIDEMIOLOGY
0172-780X NEUROENDOCRINOL LETT NEUROENDOCRINOLOGY LETTERS
1660-2854 NEURODEGENER DIS Neurodegenerative Diseases
1541-6933 NEUROCRIT CARE Neurocritical Care
1130-1473 NEUROCIRUGIA NEUROCIRUGIA
0028-3770 NEUROCHIRURGIE NEUROCHIRURGIE
0364-3190 NEUROCHEM RES NEUROCHEMICAL RESEARCH
1819-7124 NEUROCHEM J+ Neurochemical Journal
0197-0186 NEUROCHEM INT NEUROCHEMISTRY INTERNATIONAL
1355-4794 NEUROCASE NEUROCASE
1074-7427 NEUROBIOL LEARN MEM NEUROBIOLOGY OF LEARNING AND MEMORY
0969-9961 NEUROBIOL DIS NEUROBIOLOGY OF DISEASE
0197-4580 NEUROBIOL AGING NEUROBIOLOGY OF AGING
1673-5374 NEURAL REGEN RES NEURAL REGENERATION RESEARCH
0893-6080 NEURAL NETWORKS NEURAL NETWORKS
0792-8483 NEURAL PLAST NEURAL PLAST
1749-8104 NEURAL DEV Neural Development
0899-7667 NEURAL COMPUT NEURAL COMPUTATION
0954-898X NETWORK-COMP NEURAL NETWORK-COMPUTATION IN

Appendix A. Chapter 1: Appendix

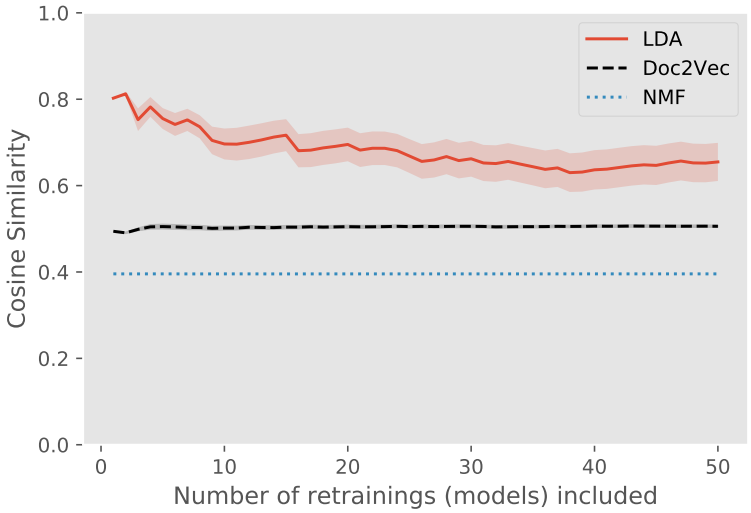
NEURAL SYSTEMS
0722-1541 NERVENHEILKUNDE NERVENHEILKUNDE
0028-2804 NERVENARZT NERVENARZT
1471-0048 NAT REV NEUROSCI NATURE REVIEWS NEUROSCIENCE
1759-4758 NAT REV NEUROL Nature Reviews Neurology
1097-6256 NAT NEUROSCI NATURE NEUROSCIENCE
1745-834X NAT CLIN PRACT NEURO Nature Clinical Practice Neurology
0730-7829 MUSIC PERCEPT MUSIC PERCEPTION
0148-639X MUSCLE NERVE MUSCLE & NERVE
1352-4585 MULT SCLER J Multiple Sclerosis Journal
0885-3185 MOVEMENT DISORD MOVEMENT DISORDERS
1087-1640 MOTOR CONTROL MOTOR CONTROL
1359-4184 MOL PSYCHIATR MOLECULAR PSYCHIATRY
1744-8069 MOL PAIN Molecular Pain
1750-1326 MOL NEURODEGENER Molecular Neurodegeneration
0893-7648 MOL NEUROBIOL MOLECULAR NEUROBIOLOGY
1044-7431 MOL CELL NEUROSCI MOLECULAR AND CELLULAR NEUROSCIENCE
0946-7211 MINIM INVAS NEUROSUR MINIMALLY INVASIVE NEUROSURGERY
0885-7490 METAB BRAIN DIS METABOLIC BRAIN DISEASE
0885-1158 MED PROBL PERFORM AR MEDICAL PROBLEMS OF PERFORMING ARTISTS
0306-9877 MED HYPOTHESES MEDICAL HYPOTHESES
1072-0502 LEARN MEMORY LEARNING & MEMORY
1357-650X LATERALITY LATERALITY
1474-4422 LANCET NEUROL LANCET NEUROLOGY
1434-0275 KLIN NEUROPHYSIOL KLINISCHE NEUROPHYSIOLOGIE
1534-7362 J VISION JOURNAL OF VISION
0957-4271 J VESTIBUL RES-EQUIL JOURNAL OF VESTIBULAR RESEARCH-EQUILIBRIUM & ORIENTATION
1052-3057 J STROKE CEREBROVASC J STROKE CEREBROVASC
1079-0268 J SPINAL CORD MED JOURNAL OF SPINAL CORD MEDICINE
0748-7711 J REHABIL RES DEV JOURNAL OF REHABILITATION RESEARCH AND DEVELOPMENT
1650-1977 J REHABIL MED JOURNAL OF REHABILITATION MEDICINE
0269-8803 J PSYCHOPHYSIOL JOURNAL OF PSYCHOPHYSIOLOGY
0269-8811 J PSYCHOPHARMACOL JOURNAL OF PSYCHOPHARMACOLOGY
1180-4882 J PSYCHIATR NEUROSCI JOURNAL OF PSYCHIATRY & NEUROSCIENCE
0022-3751 J PHYSIOL-LONDON JOURNAL OF PHYSIOLOGY-LONDON
0928-4257 J PHYSIOL-PARIS JOURNAL OF PHYSIOLOGY-PARIS
1085-9489 J PERIPHER NERV SYST JOURNAL OF THE PERIPHERAL NERVOUS SYSTEM
1526-5900 J PAIN JOURNAL OF PAIN
0897-7151 J NEUROTRAUM JOURNAL OF NEUROTRAUMA
1355-0284 J NEUROVIROL JOURNAL OF NEUROVIROLOGY
1933-0707 J NEUROSURG-PEDIATR Journal of Neurosurgery-Pediatrics
0390-5616 J NEUROSURG SCI JOURNAL OF NEUROSURGICAL SCIENCES
0022-3085 J NEUROSURG JOURNAL OF NEUROSURGERY
0888-0395 J NEUROSCI NURS JOURNAL OF NEUROSCIENCE NURSING
0360-4012 J NEUROSCI RES JOURNAL OF NEUROSCIENCE RESEARCH
0165-0270 J NEUROSCI METH JOURNAL OF NEUROSCIENCE METHODS
0270-6474 J NEUROSCI JOURNAL OF NEUROSCIENCE
1748-6645 J NEUROPSYCHOL Journal of Neuropsychology
0150-9861 J NEURORADIOLOGY JOURNAL OF NEURORADIOLOGY
0895-0172 J NEUROPSYCH CLIN N JOURNAL OF NEUROPSYCHIATRY AND CLINICAL NEUROSCIENCES
0022-3077 J NEUROPHYSIOL JOURNAL OF NEUROPHYSIOLOGY
0022-3069 J NEUROPATH EXP NEUR JOURNAL OF NEUROPATHOLOGY AND EXPERIMENTAL NEUROLOGY
0911-6044 J NEUROLINGUIST JOURNAL OF NEUROLINGUISTICS
1302-1664 J NEUROL SCI-TURK JOURNAL OF NEUROLOGICAL SCIENCES-TURKISH
1557-0576 J NEUROL PHYS THER J NEUROL PHYS THER
0022-510X J NEUROL SCI JOURNAL OF THE NEUROLOGICAL SCIENCES
0022-3050 J NEUROL NEUROSUR PS JOURNAL OF NEUROLOGY NEUROSURGERY AND PSYCHIATRY
0340-5354 J NEUROL JOURNAL OF NEUROLOGY
1742-2094 J NEUROINFLAMM Journal of Neuroinflammation
1759-8478 J NEUROINTERV SURG JOURNAL OF NEUROINTERVENTIONAL SURGERY
0165-5728 J NEUROIMMUNOL JOURNAL OF NEUROIMMUNOLOGY
1051-2284 J NEUROIMAGING JOURNAL OF NEUROIMAGING
1743-0003 J NEUROENG REHABIL Journal of NeuroEngineering and Rehabilitation
0167-7063 J NEUROGENET JOURNAL OF NEUROGENETICS
1866-1947 J NEURODEV DISORD Journal of Neurodevelopmental Disorders
0953-8194 J NEUROENDOCRINOL JOURNAL OF NEUROENDOCRINOLOGY
0022-3042 J NEUROCHEM JOURNAL OF NEUROCHEMISTRY
1070-8022 J NEURO-OPHTHALMOL JOURNAL OF NEURO-OPHTHALMOLOGY
0303-6995 J NEURAL TRANSM-SUPP JOURNAL OF NEURAL TRANSMISSION-SUPPLEMENT
1741-2560 J NEURAL ENG Journal of Neural Engineering
0300-9564 J NEURAL TRANSM JOURNAL OF NEURAL TRANSMISSION
0022-2895 J MOTOR BEHAV JOURNAL OF MOTOR BEHAVIOR
0895-8696 J MOL NEUROSCI JOURNAL OF MOLECULAR NEUROSCIENCE
0271-0137 J MIND BEHAV JOURNAL OF MIND AND BEHAVIOR
2005-3711 J KOREAN NEUROSURG S Journal of Korean Neurosurgical Society
0219-6352 J INTEGR NEUROSCI Journal of Integrative Neuroscience
1355-6177 J INT NEUROPSYCH SOC JOURNAL OF THE INTERNATIONAL NEUROPSYCHOLOGICAL SOCIETY
0964-704X J HIST NEUROSCI Journal of the History of the Neurosciences
0885-9701 J HEAD TRAUMA REHAB JOURNAL OF HEAD TRAUMA REHABILITATION
1129-2369 J HEADACHE PAIN JOURNAL OF HEADACHE AND PAIN
0891-9887 J GERIATR PSYCH NEUR JOURNAL OF GERIATRIC PSYCHIATRY AND NEUROLOGY
0022-0930 J EVOL BIOCHEM PHYS+ JOURNAL OF EVOLUTIONARY BIOCHEMISTRY AND PHYSIOLOGY
0929-5313 J COMPUT NEUROSCI JOURNAL OF COMPUTATIONAL NEUROSCIENCE
0021-9967 J COMP NEUROL JOURNAL OF COMPARATIVE NEUROLOGY
0898-929X J COGNITIVE NEUROSCI JOURNAL OF COGNITIVE NEUROSCIENCE
0967-5868 J CLIN NEUROSCI JOURNAL OF CLINICAL NEUROSCIENCE
0736-0258 J CLIN NEUROPHYSIOL JOURNAL OF CLINICAL NEUROPHYSIOLOGY
1738-6586 J CLIN NEUROL Journal of Clinical Neurology
1380-3395 J CLIN EXP NEUROPSYC JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY
0883-0738 J CHILD NEUROL JOURNAL OF CHILD NEUROLOGY
0891-0618 J CHEM NEUROANAT JOURNAL OF CHEMICAL NEUROANATOMY
0271-678X J CEREBR BLOOD F MET JOURNAL OF CEREBRAL BLOOD FLOW AND METABOLISM
0092-0606 J BIOL PHYS JOURNAL OF BIOLOGICAL PHYSICS
1387-2877 J ALZHEIMERS DIS JOURNAL OF ALZHEIMERS DISEASE
1123-9344 INTERV NEURORADIOL INTERVENTIONAL NEURORADIOLOGY
0074-7742 INT REV NEUROBIOL INTERNATIONAL REVIEW OF NEUROBIOLOGY
1747-4930 INT J STROKE International Journal of Stroke
0342-5282 INT J REHABIL RES INTERNATIONAL JOURNAL OF REHABILITATION RESEARCH
0167-8760 INT J PSYCHOPHYSIOL INTERNATIONAL JOURNAL OF PSYCHOPHYSIOLOGY
0020-7454 INT J NEUROSCI INTERNATIONAL JOURNAL OF NEUROSCIENCE
1461-1457 INT J NEUROPSYCHOPH INTERNATIONAL JOURNAL OF NEUROPSYCHOPHARMACOLOGY
0129-0657 INT J NEURAL SYST INTERNATIONAL JOURNAL OF NEURAL SYSTEMS
0899-9457 INT J IMAG SYST TECH INTERNATIONAL JOURNAL OF IMAGING SYSTEMS AND TECHNOLOGY
0736-5748 INT J DEV NEUROSCI INTERNATIONAL JOURNAL OF DEVELOPMENTAL NEUROSCIENCE
1534-4320 IEEE T NEUR SYS REH IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING
0019-1442 IDEGGYOGY SZEMLE IDEGGYOGYASZATI SZEMLE-CLINICAL NEUROSCIENCE
0167-9457 HUM MOVEMENT SCI HUMAN MOVEMENT SCIENCE
1065-9471 HUM BRAIN MAPP HUMAN BRAIN MAPPING
0018-506X HORM BEHAV HORMONES AND BEHAVIOR
1569-1861 HONG KONG J OCCUP TH Hong Kong Journal of Occupational Therapy
1050-9631 HIPPOCAMPUS HIPPOCAMPUS
0017-8748 HEADACHE HEADACHE
0894-1491 GLIA GLIA
1601-1848 GENES BRAIN BEHAV GENES BRAIN AND BEHAVIOR
0393-5264 FUNCT NEUROL FUNCTIONAL NEUROLOGY
1662-5129 FRONT NEUROANAT FRONT NEUROANAT
0091-3022 FRONT NEUROENDOCRIN FRONTIERS IN NEUROENDOCRINOLOGY
1662-5110 FRONT NEURAL CIRCUIT FRONTIERS IN NEURAL CIRCUITS
1662-5161 FRONT HUM NEUROSCI Frontiers in Human Neuroscience
1662-5102 FRONT CELL NEUROSCI FRONTIERS IN CELLULAR NEURO-

SCIENCE
1662-5188 FRONT COMPUT NEUROSC FRONTIERS IN COMPUTATIONAL NEUROSCIENCE
1641-4640 FOLIA NEUROPATHOL FOLIA NEUROPATHOLOGICA
0014-4886 EXP NEUROL EXPERIMENTAL NEUROLOGY
0014-4819 EXP BRAIN RES EXPERIMENTAL BRAIN RESEARCH
0924-977X EUR NEUROPSYCHOPHARM EUROPEAN NEUROPSYCHOPHARMACOLOGY
0014-3022 EUR NEUROL EUROPEAN NEUROLOGY
1090-3801 EUR J PAIN EUROPEAN JOURNAL OF PAIN
1090-3798 EUR J PAEDIATR NEURO EUROPEAN JOURNAL OF PAEDIATRIC NEUROLOGY
0953-816X EUR J NEUROSCI EUROPEAN JOURNAL OF NEUROSCIENCE
1351-5101 EUR J NEUROL EUROPEAN JOURNAL OF NEUROLOGY
0920-1211 EPILEPSY RES EPILEPSY RESEARCH
1294-9361 EPILEPTIC DISORD EPILEPTIC DISORDERS
1535-7597 EPILEPSY CURR EPILEPSY CURR
1525-5050 EPILEPSY BEHAV EPILEPSY & BEHAVIOR
0013-9580 EPILEPSIA EPILEPSIA
1011-288X DOULEUR ANALG Douleur et Analgesie
0963-8288 DISABIL REHABIL DISABILITY AND REHABILITATION
0378-5866 DEV NEUROSCI-BASEL DEVELOPMENTAL NEUROSCIENCE
1932-8451 DEV NEUROBIOL Developmental Neurobiology
8756-5641 DEV NEUROPSYCHOL DEVELOPMENTAL NEUROPSYCHOLOGY
0012-1622 DEV MED CHILD NEUROL DEVELOPMENTAL MEDICINE AND CHILD NEUROLOGY
1940-5510 DEV DISABIL RES REV Developmental Disabilities Research Reviews
1420-8008 DEMENT GERIATR COGN DEMENTIA AND GERIATRIC COGNITIVE DISORDERS
1092-8480 CURR TREAT OPTION NE CURRENT TREATMENT OPTIONS IN NEUROLOGY
1531-3433 CURR PAIN HEADACHE R CURRENT PAIN AND HEADACHE REPORTS
1350-7540 CURR OPIN NEUROL CURRENT OPINION IN NEUROLOGY
0959-4388 CURR OPIN NEUROBIOL CURRENT OPINION IN NEUROBIOLOGY
1567-2026 CURR NEUROVASC RES CURRENT NEUROVASCULAR RESEARCH
1570-159X CURR NEUROPHARMACOL Current Neuropharmacology
1528-4042 CURR NEUROL NEUROSCI Current Neurology and Neuroscience Reports
1567-2050 CURR ALZHEIMER RES Current Alzheimer Research
0010-9452 CORTEX CORTEX
1758-8928 COGN NEUROSCI-UK COGNITIVE NEUROSCIENCE
0264-3294 COGN NEUROPSYCHOL COGNITIVE NEUROPSYCHOLOGY
1871-4080 COGN NEURODYNAMICS Cognitive Neurodynamics
1866-9956 COGN COMPUT Cognitive Computation
1543-3633 COGN BEHAV NEUROL Cognitive and Behavioral Neurology
1530-7026 COGN AFFECT BEHAV NE COGNITIVE AFFECTIVE & BEHAVIORAL NEUROSCIENCE
1755-5930 CNS NEUROSCI THER CNS Neuroscience & Therapeutics
1172-7047 CNS DRUGS CNS DRUGS
1871-5273 CNS NEUROL DISORD-DR CNS & Neurological Disorders-Drug Targets
0269-2155 CLIN REHABIL CLINICAL REHABILITATION
1869-1439 CLIN NEURORADIOL CLINICAL NEURORADIOLOGY
1385-4046 CLIN NEUROPSYCHOL CLINICAL NEUROPSYCHOLOGIST
0362-5664 CLIN NEUROPHARMACOL CLINICAL NEUROPHARMACOLOGY
1388-2457 CLIN NEUROPHYSIOL CLINICAL NEUROPHYSIOLOGY
0722-5091 CLIN NEUROPATHOL CLINICAL NEUROPATHOLOGY
0303-8467 CLIN NEUROL NEUROSUR CLINICAL NEUROLOGY AND NEUROSURGERY
0749-8047 CLIN J PAIN CLINICAL JOURNAL OF PAIN
1550-0594 CLIN EEG NEUROSCI CLINICAL EEG AND NEUROSCIENCE
0256-7040 CHILD NERV SYST CHILDS NERVOUS SYSTEM
1936-5802 CHEMOSENS PERCEPT Chemosensory Perception
0379-864X CHEM SENSES CHEMICAL SENSES
1210-7859 CESK SLOV NEUROL N CESKA A SLOVENSKA NEUROLOGIE A NEUROCHIRURGIE
1473-4222 CEREBELLUM CEREBELLUM
1015-9770 CEREBROVASC DIS CEREBROVASCULAR DISEASES
1047-3211 CEREB CORTEX CEREBRAL CORTEX
0333-1024 CEPHALALGIA CEPHALALGIA
0044-4251 CENT EUR NEUROSURG Central European Neurosurgery
0272-4340 CELL MOL NEUROBIOL CELLULAR AND MOLECULAR NEUROBIOLOGY
0008-4174 CAN J OCCUP THER CAN J OCCUP THER
0317-1671 CAN J NEUROL SCI CANADIAN JOURNAL OF NEUROLOGICAL SCIENCES
0268-8697 BRIT J NEUROSURG BRITISH JOURNAL OF NEUROSURGERY
0896-0267 BRAIN TOPOGR BRAIN TOPOGRAPHY
1935-861X BRAIN STIMUL Brain Stimulation
1863-2653 BRAIN STRUCT FUNCT Brain Structure & Function
0165-0173 BRAIN RES REV BRAIN RESEARCH REVIEWS
0361-9230 BRAIN RES BULL BRAIN RESEARCH BULLETIN
1015-6305 BRAIN PATHOL BRAIN PATHOLOGY
0006-8993 BRAIN RES BRAIN RESEARCH
0093-934X BRAIN LANG BRAIN AND LANGUAGE
1931-7557 BRAIN IMAGING BEHAV Brain Imaging and Behavior
1443-9646 BRAIN IMPAIR BRAIN IMPAIRMENT
0269-9052 BRAIN INJURY BRAIN INJURY
0387-7604 BRAIN DEV-JPN BRAIN & DEVELOPMENT
0278-2626 BRAIN COGNITION BRAIN AND COGNITION
0006-8977 BRAIN BEHAV EVOLUT BRAIN BEHAVIOR AND EVOLUTION
0889-1591 BRAIN BEHAV IMMUN BRAIN BEHAVIOR AND IMMUNITY
0006-8950 BRAIN BRAIN
1471-2202 BMC NEUROSCI BMC NEUROSCIENCE
1471-2377 BMC NEUROL BMC Neurology
0301-0511 BIOL PSYCHOL BIOLOGICAL PSYCHOLOGY
0006-3223 BIOL PSYCHIAT BIOLOGICAL PSYCHIATRY
0340-1200 BIOL CYBERN BIOLOGICAL CYBERNETICS
0208-5216 BIOCYBERN BIOMED ENG BIOCYBERNETICS AND BIOMEDICAL ENGINEERING
0735-7044 BEHAV NEUROSCI BEHAVIORAL NEUROSCIENCE
0955-8810 BEHAV PHARMACOL BEHAVIOURAL PHARMACOLOGY
0953-4180 BEHAV NEUROL BEHAVIOURAL NEUROLOGY
1744-9081 BEHAV BRAIN FUNCT Behavioral and Brain Functions
0166-4328 BEHAV BRAIN RES BEHAVIOURAL BRAIN RESEARCH
1566-0702 AUTON NEUROSCI-BASIC AUTONOMIC NEUROSCIENCE-BASIC & CLINICAL
0045-0766 AUST OCCUP THER J Australian Occupational Therapy Journal
1040-0435 ASSIST TECHNOL ASSISTIVE TECHNOLOGY
1759-0914 ASN NEURO ASN NEURO
0004-282X ARQ NEURO-PSIQUIAT ARQUIVOS DE NEURO-PSIQUIATRIA
0003-9993 ARCH PHYS MED REHAB ARCHIVES OF PHYSICAL MEDICINE AND REHABILITATION
0003-9942 ARCH NEUROL-CHICAGO ARCHIVES OF NEUROLOGY
0003-9829 ARCH ITAL BIOL ARCHIVES ITALIENNES DE BIOLOGIE
0914-9465 ARCH HISTOL CYTOL ARCHIVES OF HISTOLOGY AND CYTOLOGY
0887-6177 ARCH CLIN NEUROPSYCH ARCHIVES OF CLINICAL NEUROPSYCHOLOGY
1090-0586 APPL PSYCHOPHYS BIOF APPLIED PSYCHOPHYSIOLOGY AND BIOFEEDBACK
0908-4282 APPL NEUROPSYCHOL APPLIED NEUROPSYCHOLOGY
0268-7038 APHASIOLOGY APHASIOLOGY
0147-006X ANNU REV NEUROSCI ANNUAL REVIEW OF NEUROSCIENCE
0364-5134 ANN NEUROL ANNALS OF NEUROLOGY
0972-2327 ANN INDIAN ACAD NEUR ANNALS OF INDIAN ACADEMY OF NEUROLOGY
0940-9602 ANN ANAT ANNALS OF ANATOMY-ANATOMISCHER ANZEIGER
1748-2968 AMYOTROPH LATERAL SC Amyotrophic Lateral Sclerosis
0894-9115 AM J PHYS MED REHAB AMERICAN JOURNAL OF PHYSICAL MEDICINE & REHABILITATION
0272-9490 AM J OCCUP THER AMERICAN JOURNAL OF OCCUPATIONAL THERAPY
0195-6108 AM J NEURORADIOL AMERICAN JOURNAL OF NEURORADIOLOGY
1533-3175 AM J ALZHEIMERS DIS AMERICAN JOURNAL OF ALZHEIMERS DISEASE AND OTHER DEMENTIAS
1552-5260 ALZHEIMERS DEMENT Alzheimers & Dementia
0893-0341 ALZ DIS ASSOC DIS ALZHEIMER DISEASE & ASSOCIATED DISORDERS
0741-8329 ALCOHOL ALCOHOL
0302-4350 AKTUEL NEUROL AKTUELLE NEUROLOGIE
1355-6215 ADDICT BIOL ADDICTION BIOLOGY
0360-1293 ACUPUNCTURE ELECTRO ACUPUNCTURE & ELECTROTHERAPEUTICS RESEARCH
0001-6322 ACTA NEUROPATHOL ACTA NEUROPATHOLOGICA
0001-6314 ACTA NEUROL SCAND ACTA NEUROLOGICA SCANDINAVICA
0300-9009 ACTA NEUROL BELG ACTA NEUROLOGICA BELGICA
0001-6268 ACTA NEUROCHIR ACTA NEUROCHIRURGICA
0065-1400 ACTA NEUROBIOL EXP ACTA NEUROBIOLOGIAE EXPERIMENTALIS
0031-9023 PHYS THER PHYSICAL THERAPY
0031-9384 PHYSIOL BEHAV PHYSIOLOGY & BEHAVIOR
0079-6123 PROG BRAIN RES PROGRESS IN BRAIN RESEARCH
0278-5846 PROG NEURO-PSYCHOPH PROGRESS IN NEURO-

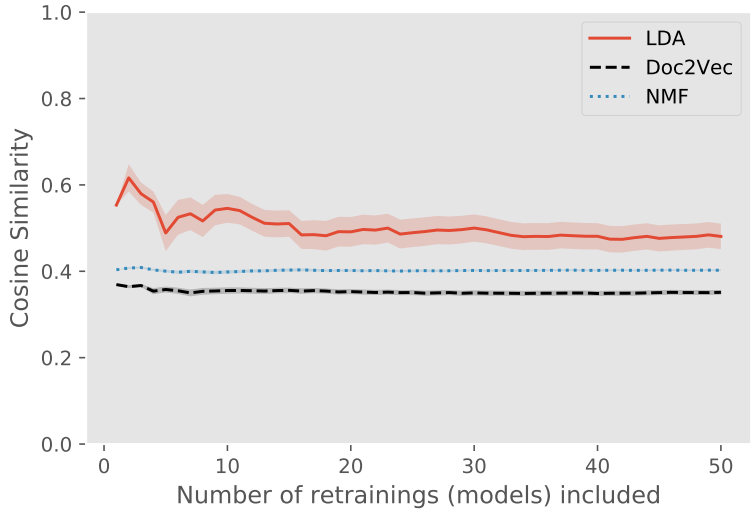
Appendix A. Chapter 1: Appendix

PSYCHOPHARMACOLOGY & BIOLOGICAL PSYCHIATRY
0301-0082 PROG NEUROBIOL PROGRESS IN NEUROBIOLOGY
0925-4927 PSYCHIAT RES-NEUROIM PSYCHIATRY RESEARCH-NEUROIMAGING
0306-4530 PSYCHONEUROENDOCRINO PSYCHONEUROENDOCRINOLOGY
0033-3158 PSYCHOPHARMACOLOGY PSYCHOPHARMACOLOGY
0048-5772 PSYCHOPHYSIOLOGY PSYCHOPHYSIOLOGY
0922-6028 RESTOR NEUROL NEUROS RESTORATIVE NEUROLOGY AND NEUROSCIENCE
0035-3787 REV NEUROL-FRANCE REVUE NEUROLOGIQUE
0334-1763 REV NEUROSCIENCE REVIEWS IN THE NEUROSCIENCES
0033-698X RLA-REV LINGUIST TEO RLA-REVISTA DE LINGUISTICA TEORICA Y APLICADA
1103-8128 SCAND J OCCUP THER SCANDINAVIAN JOURNAL OF OCCUPATIONAL THERAPY
0932-433X SCHMERZ SCHMERZ
1878-4755 SEEING PERCEIVING SEEING AND PERCEIVING
1059-1311 SEIZURE-EUR J EPILEP SEIZURE-EUROPEAN JOURNAL OF EPILEPSY
0271-8235 SEMIN NEUROL SEMINARS IN NEUROLOGY
1071-9091 SEMIN PEDIATR NEUROL SEMINARS IN PEDIATRIC NEUROLOGY
1749-5016 SOC COGN AFFECT NEUR Social Cognitive and Affective Neuroscience
1747-0919 SOC NEUROSCI-UK Social Neuroscience
0899-0220 SOMATOSENS MOT RES SOMATOSENSORY AND MOTOR RESEARCH
0169-1015 SPATIAL VISION SPATIAL VISION
1362-4393 SPINAL CORD SPINAL CORD
1011-6125 STEREOT FUNCT NEUROS STEREOTACTIC AND FUNCTIONAL NEUROSURGERY
1025-3890 STRESS STRESS-THE INTERNATIONAL JOURNAL ON THE BIOLOGY OF STRESS
0039-2499 STROKE STROKE
0090-3019 SURG NEUROL SURGICAL NEUROLOGY
0887-4476 SYNAPSE SYNAPSE
1074-9357 TOP STROKE REHABIL Topics in Stroke Rehabilitation
1364-6613 TRENDS COGN SCI TRENDS IN COGNITIVE SCIENCES
0166-2236 TRENDS NEUROSCI TRENDS IN NEUROSCIENCES
1019-5149 TURK NEUROSURG Turkish Neurosurgery
0042-6989 VISION RES VISION RESEARCH
0952-5238 VISUAL NEUROSCI VISUAL NEUROSCIENCE
WIRES COGN SCI WIRES COGN SCI
1878-8750 WORLD NEUROSURG World Neurosurgery
1016-264X Z NEUROPSYCHOL ZEITSCHRIFT FUR NEUROPSYCHOLOGIE
1997-7298 ZH NEVROL PSIKHIATR Zhurnal Nevrologii I Psikhiatrii imeni S S Korsakova
0044-4677 ZH VYSSH NERV DEYAT+ ZHURNAL VYSSHEI NERVNOI DEYATELNOSTI IMENI I P PAVLOVA

A.2 Pariwise Cosine Similarities

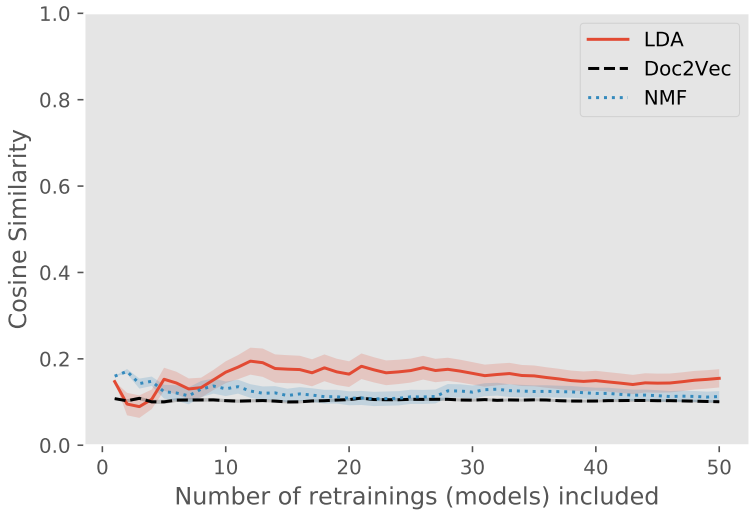


(a) Latent space size (number of topics) 10

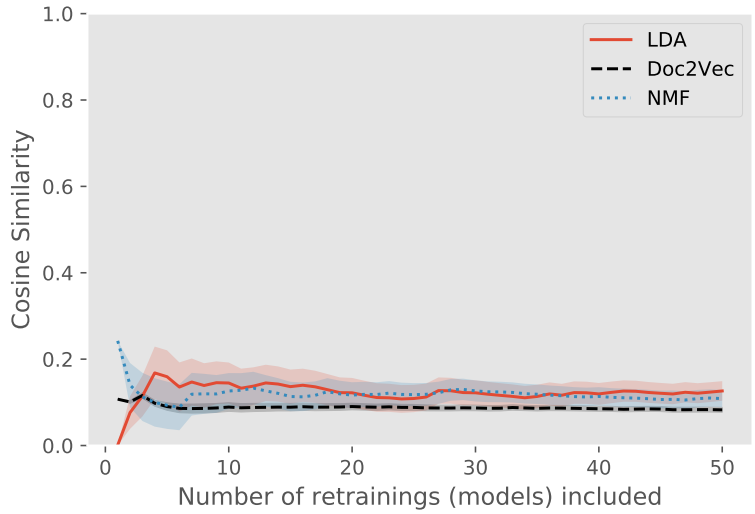


(b) Latent space size (number of topics) 25

Figure A.1 – **Pairwise Cosine Similarity across 50 model runs**



(a) Latent space size (number of topics) 250



(b) Latent space size (number of topics) 400

Figure A.2 – **Pairwise Cosine Similarity across 50 model runs**

A.3 Asymptotic Φ_K with filters

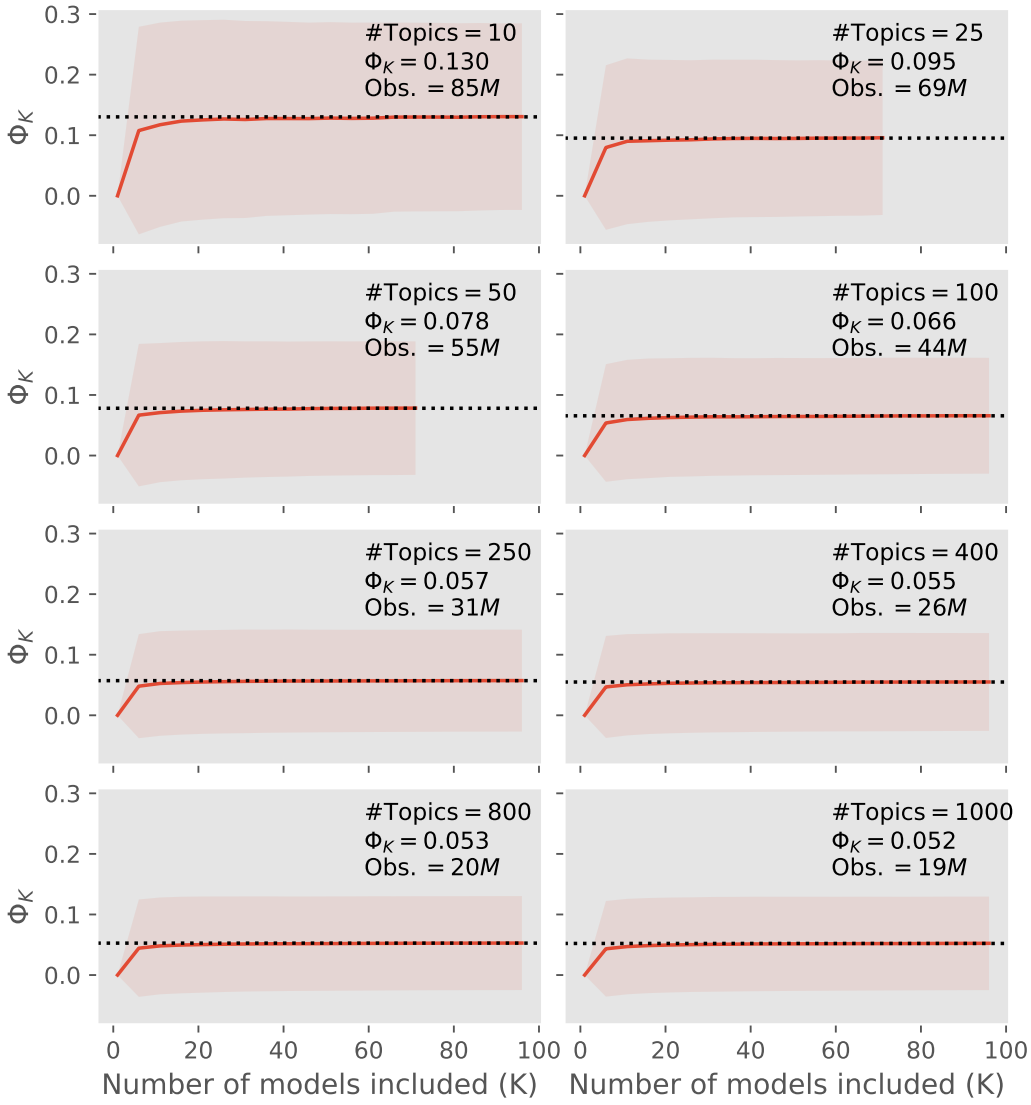


Figure A.3 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.1 \forall k$

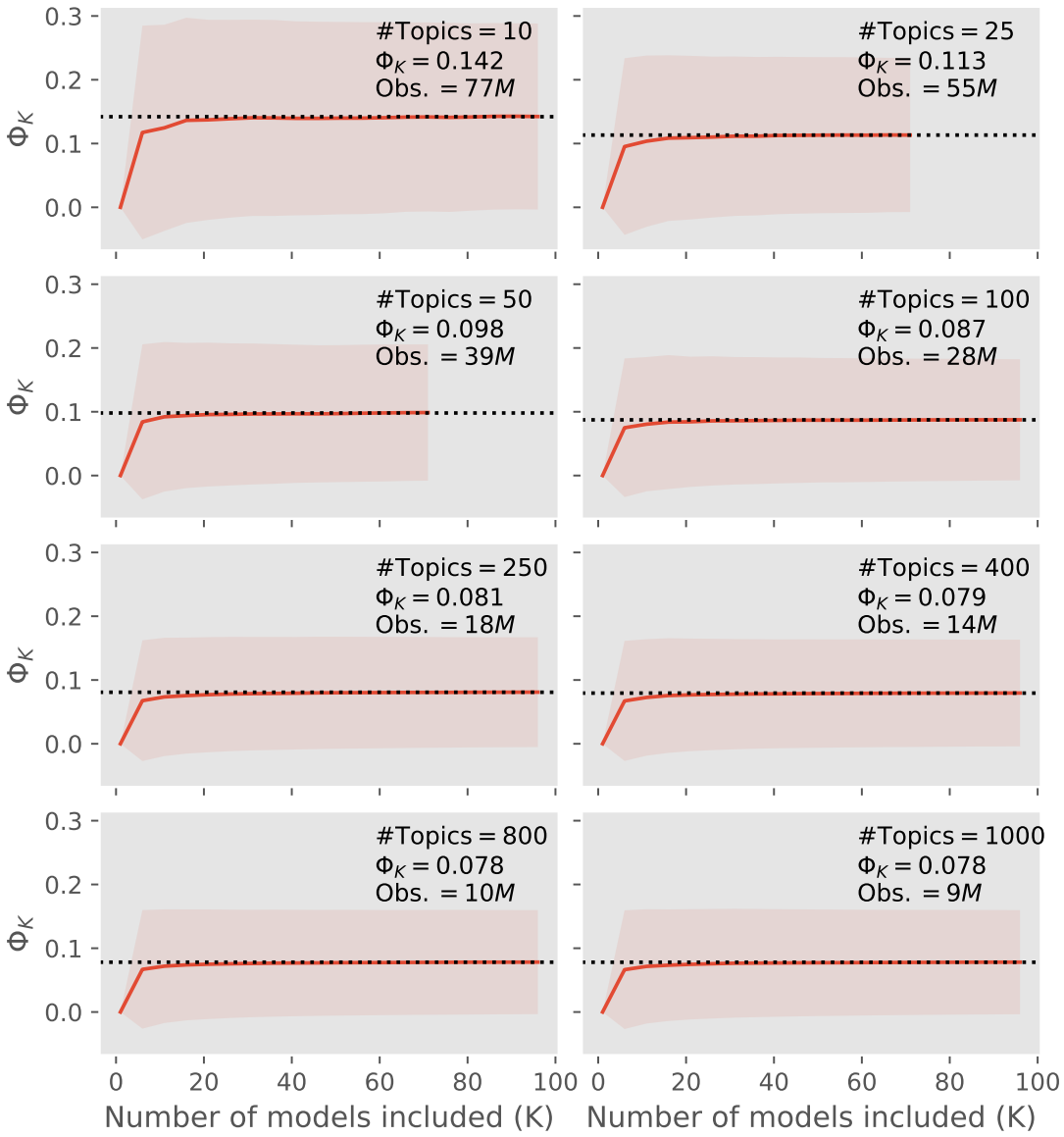


Figure A.4 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.2 \forall k$

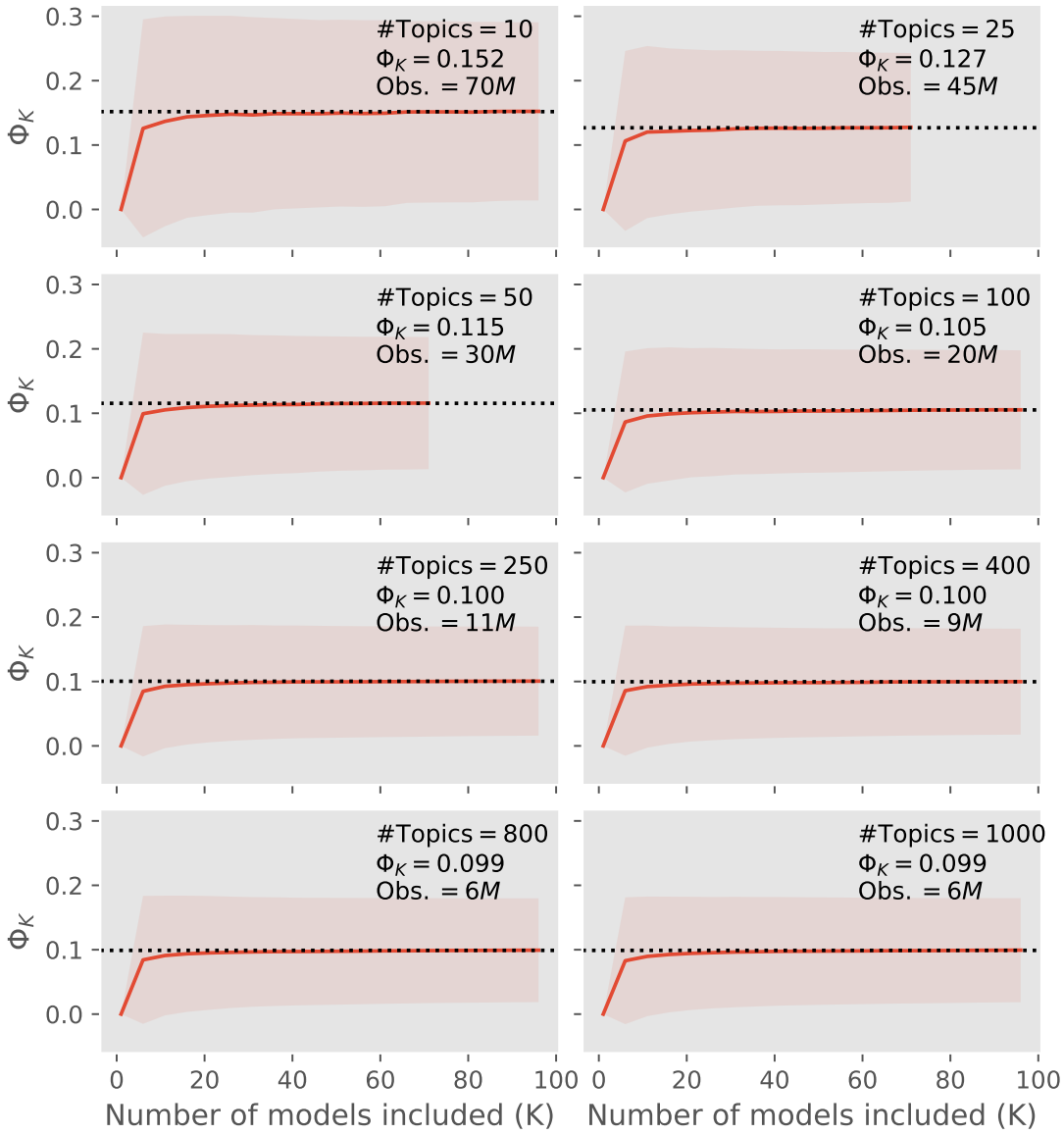


Figure A.5 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.3 \forall k$

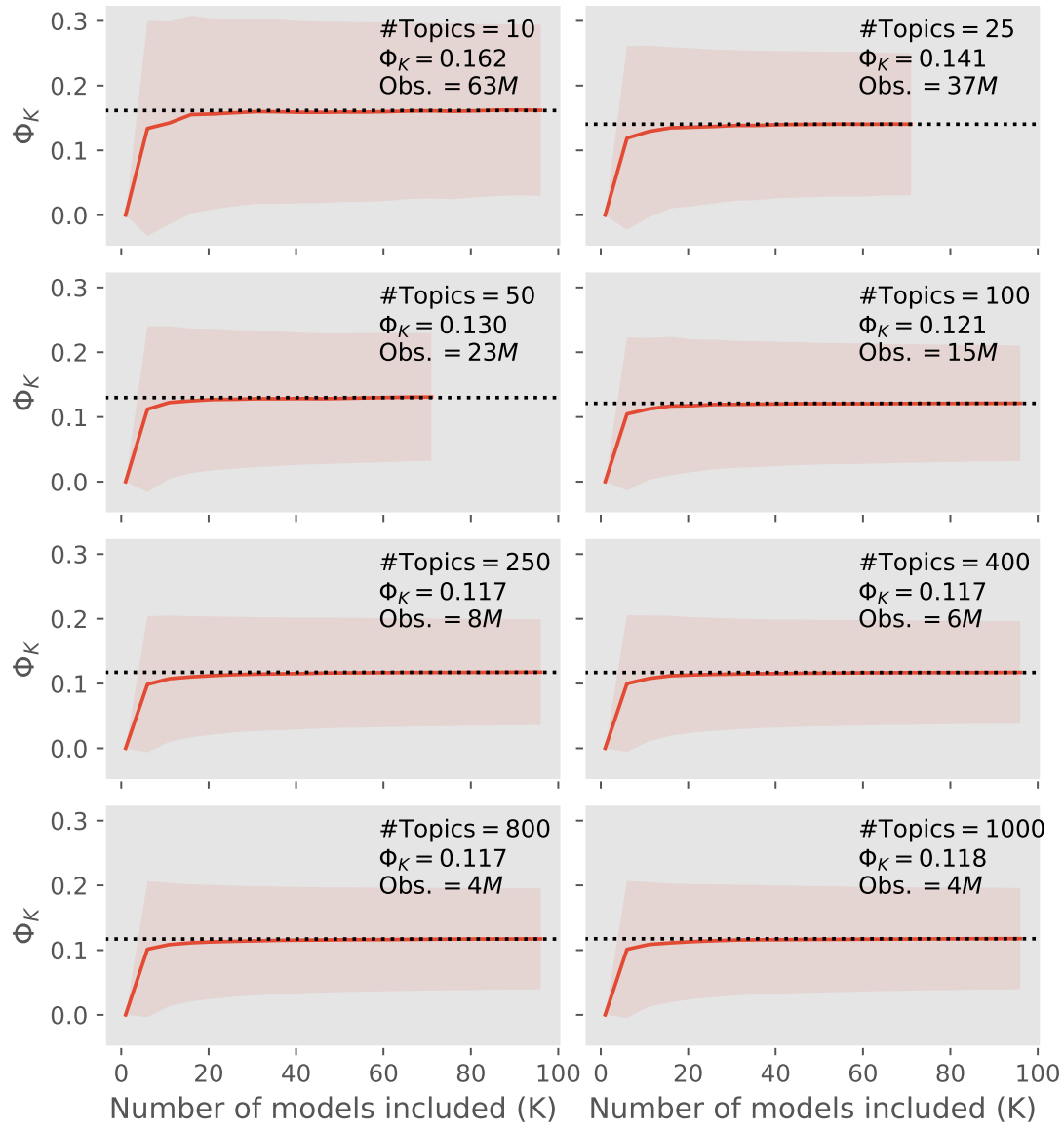


Figure A.6 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.4 \forall k$

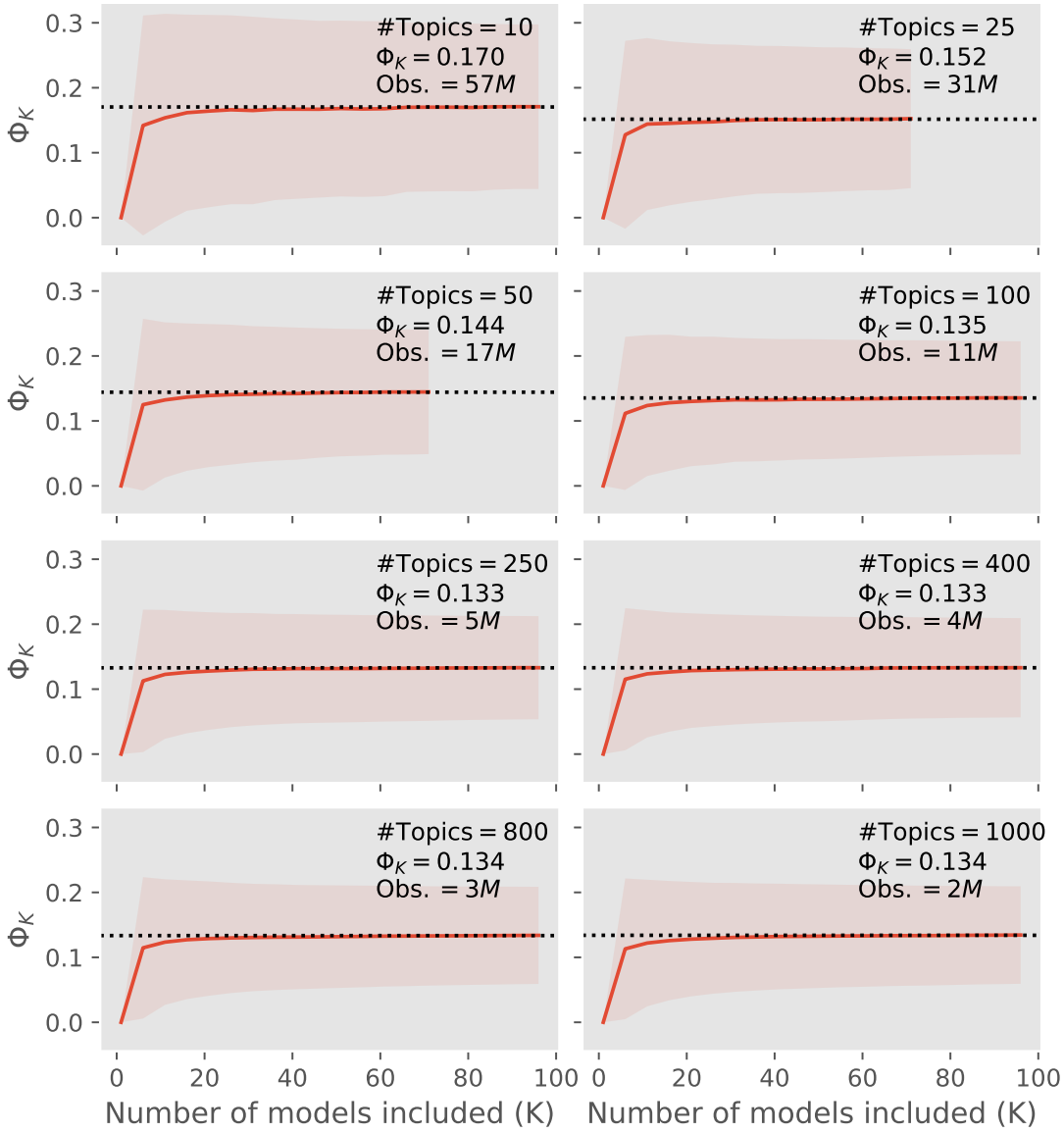


Figure A.7 – **Average Standard Deviation for LDA:** Asymptotic value over multiple retrainings (75 or 100) of LDA for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.5 \forall k$

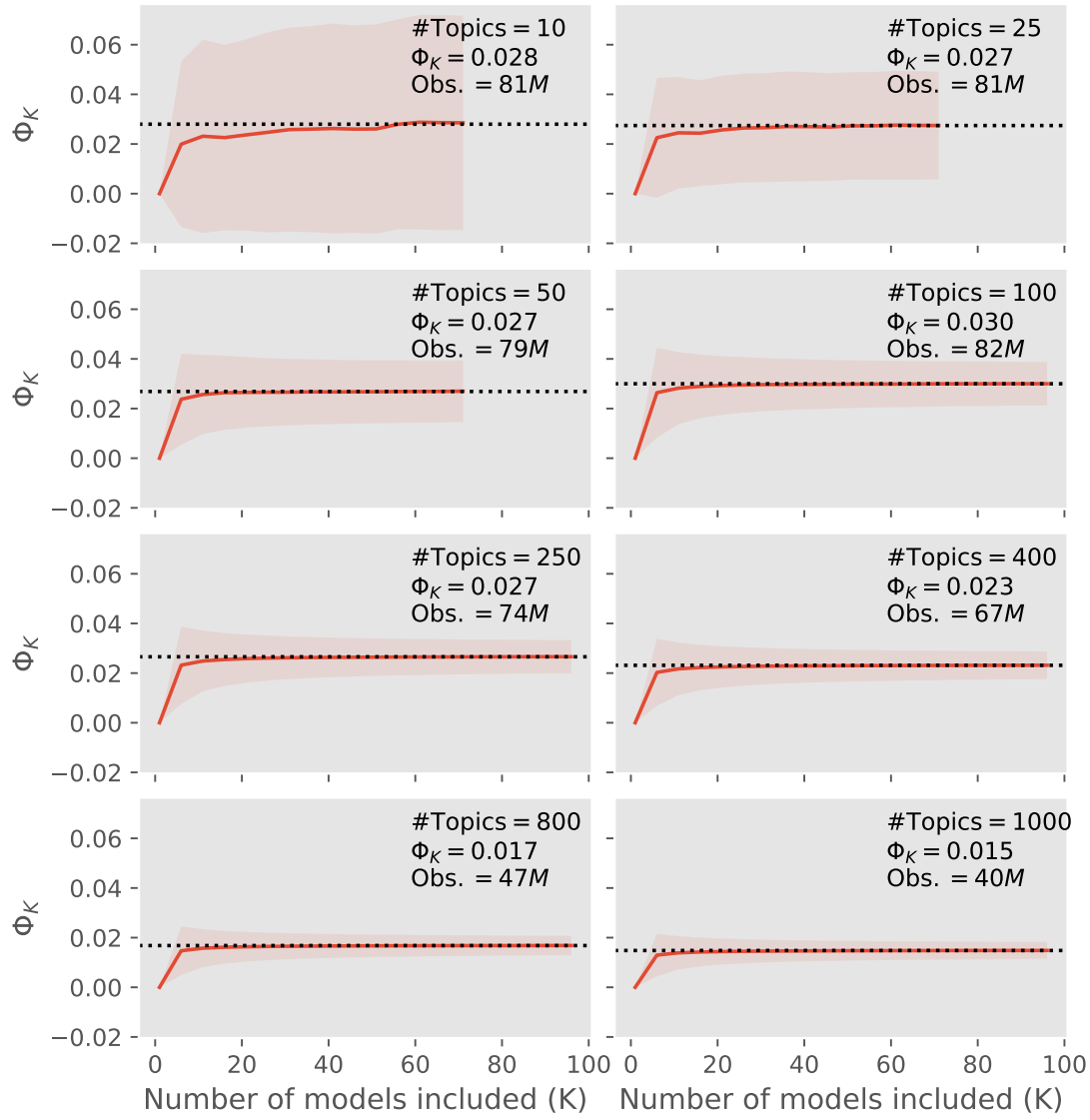


Figure A.8 – **Average Standard Deviation for doc2vec:** Asymptotic value over multiple retrainings (75 or 100) of doc2vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.1 \forall k$

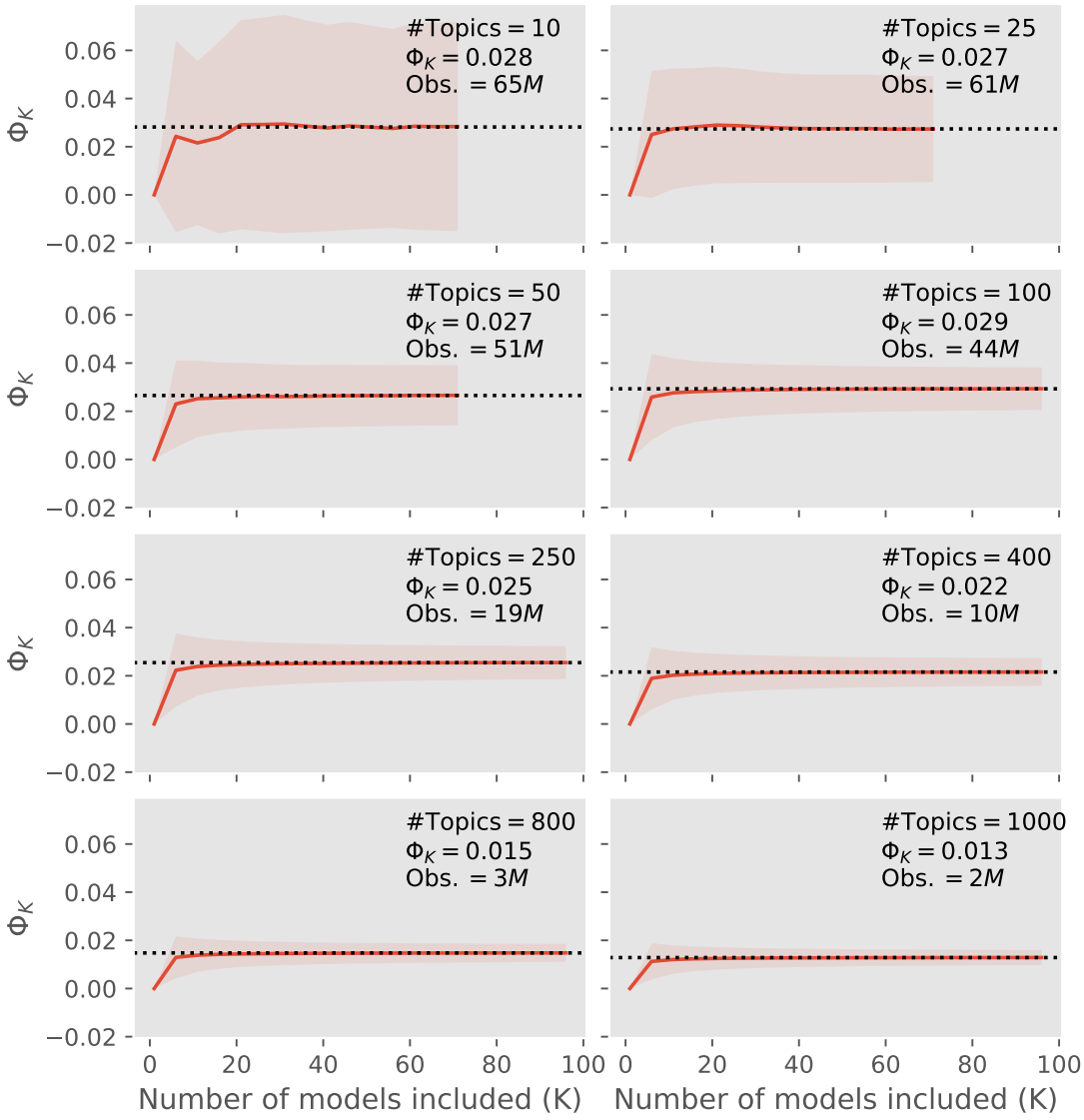


Figure A.9 – **Average Standard Deviation for doc2vec:** Asymptotic value over multiple retrainings (75 or 100) of doc2vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.2 \forall k$

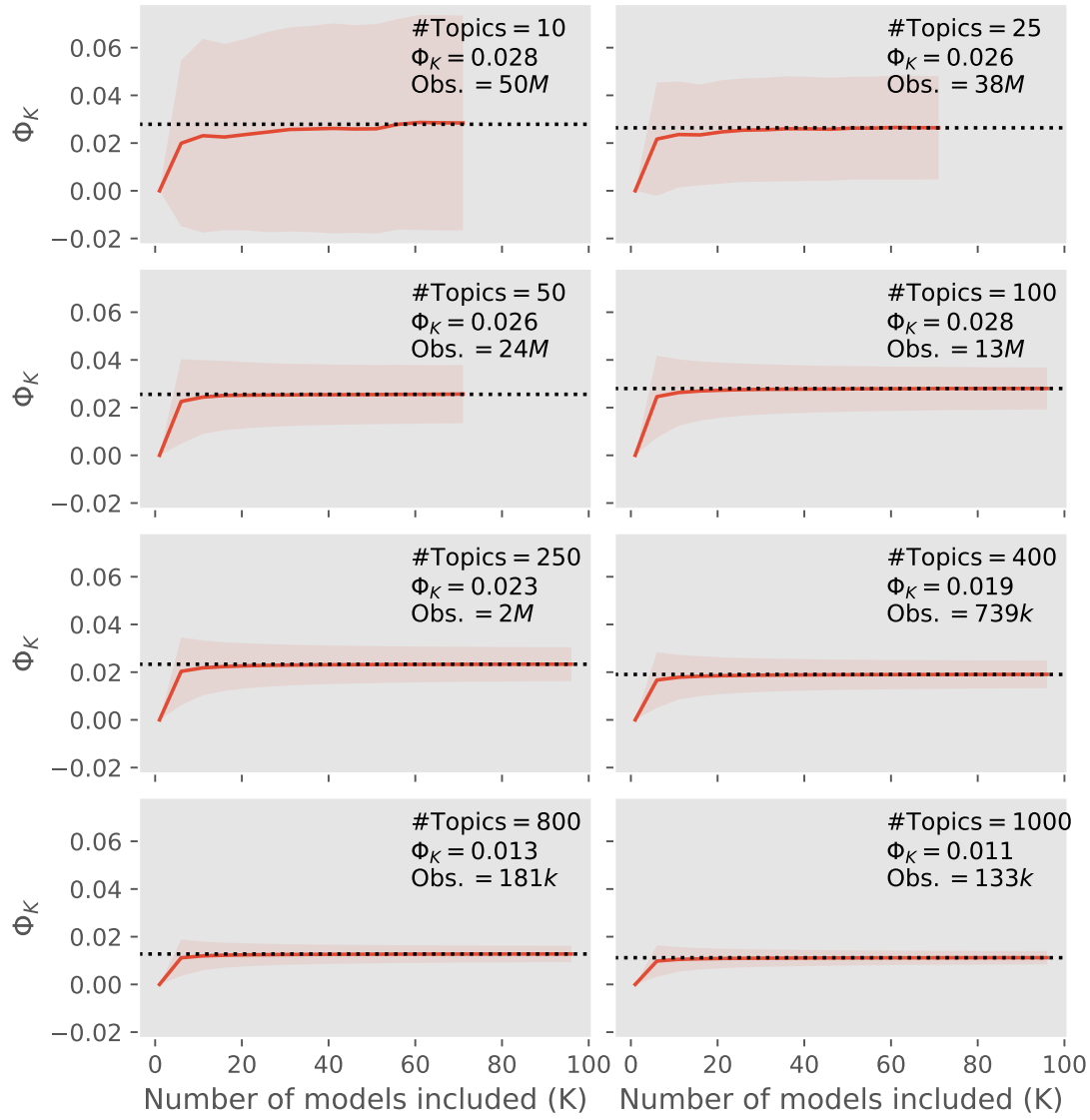


Figure A.10 – **Average Standard Deviation for doc2vec:** Asymptotic value over multiple re-trainings (75 or 100) of doc2vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.3 \forall k$

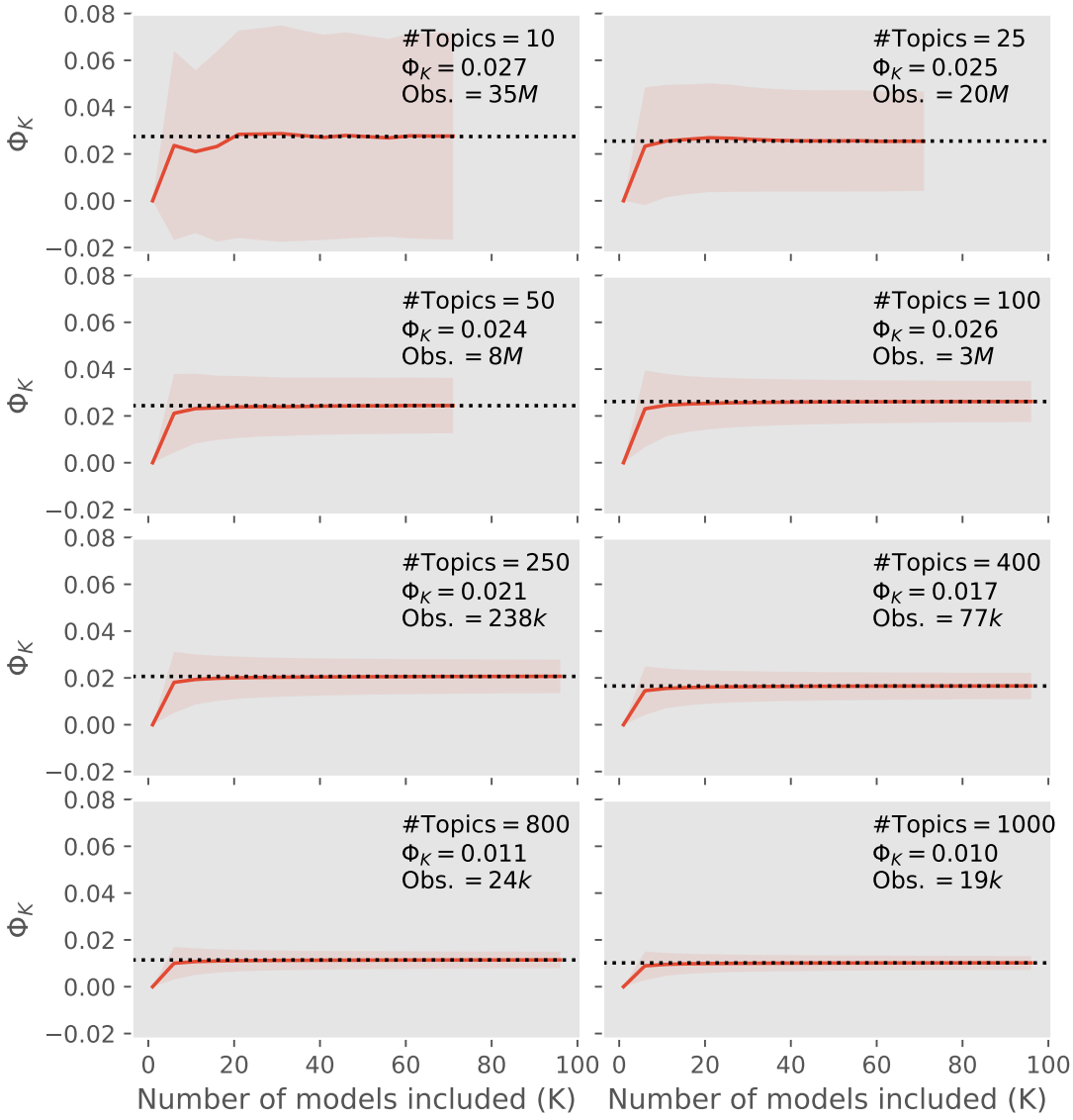


Figure A.11 – **Average Standard Deviation for doc2vec:** Asymptotic value over multiple re-trainings (75 or 100) of doc2vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.4 \forall k$

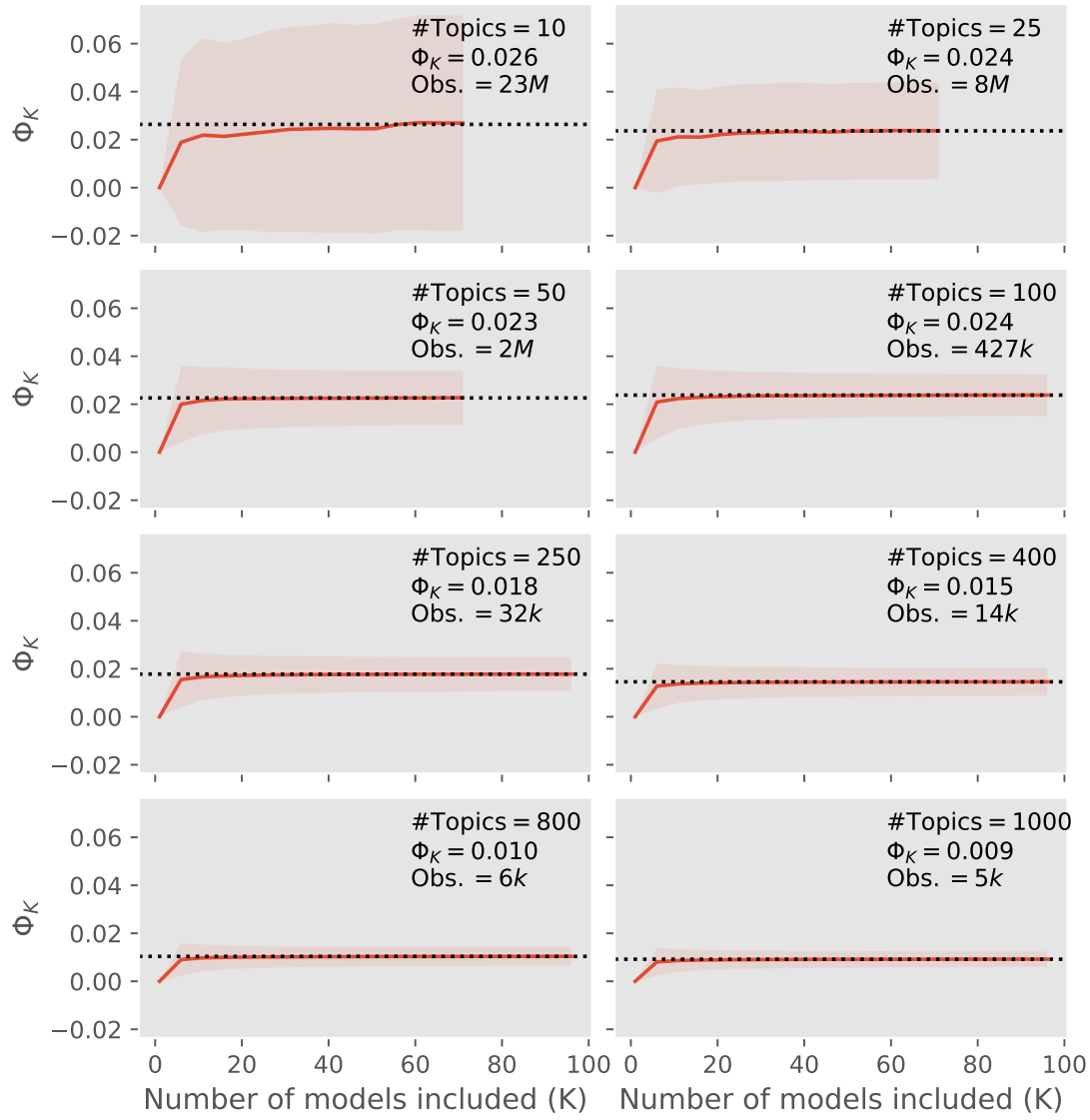


Figure A.12 – **Average Standard Deviation for doc2vec:** Asymptotic value over multiple re-trainings (75 or 100) of doc2vec for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.5 \forall k$

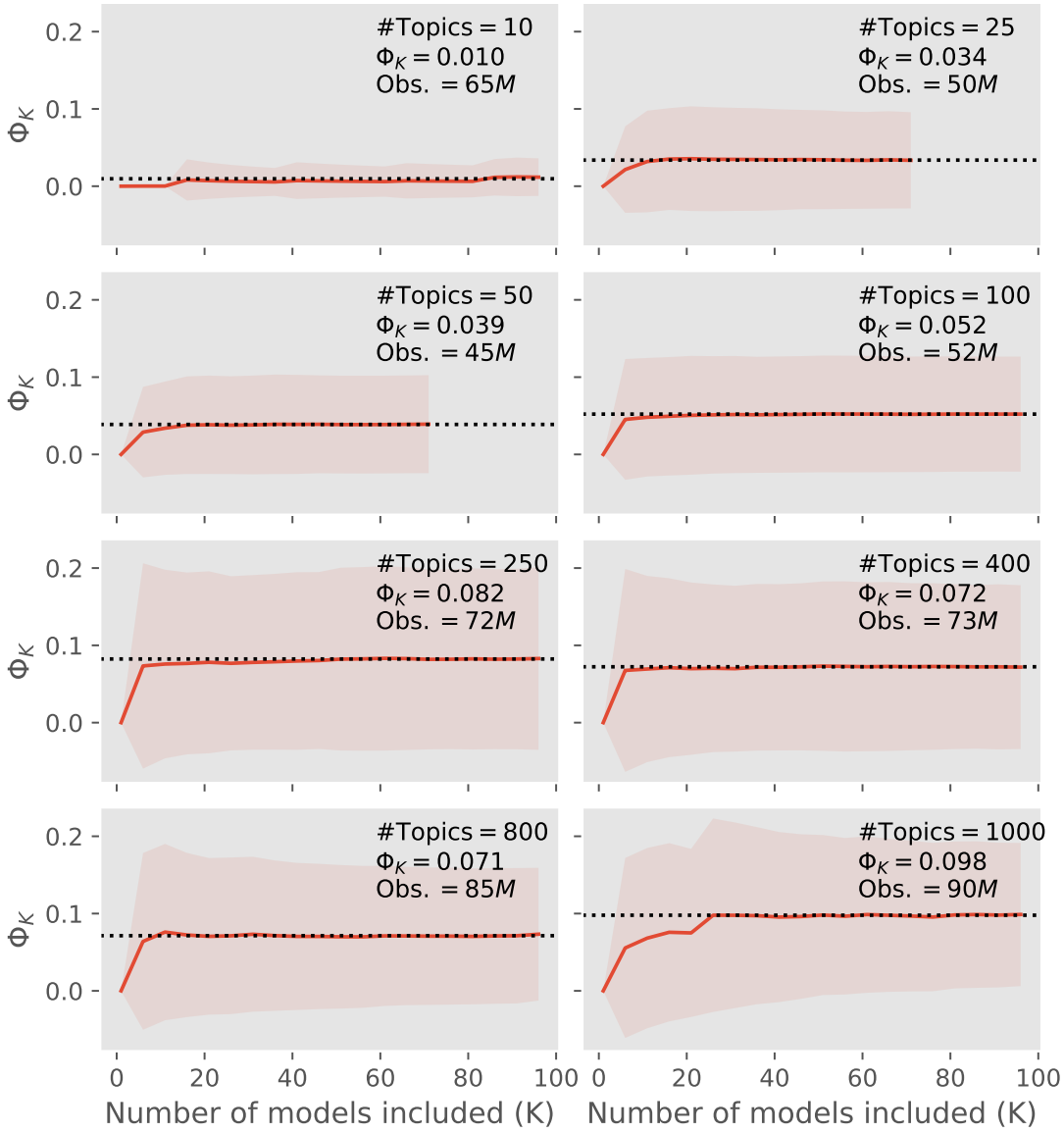


Figure A.13 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.1 \forall k$

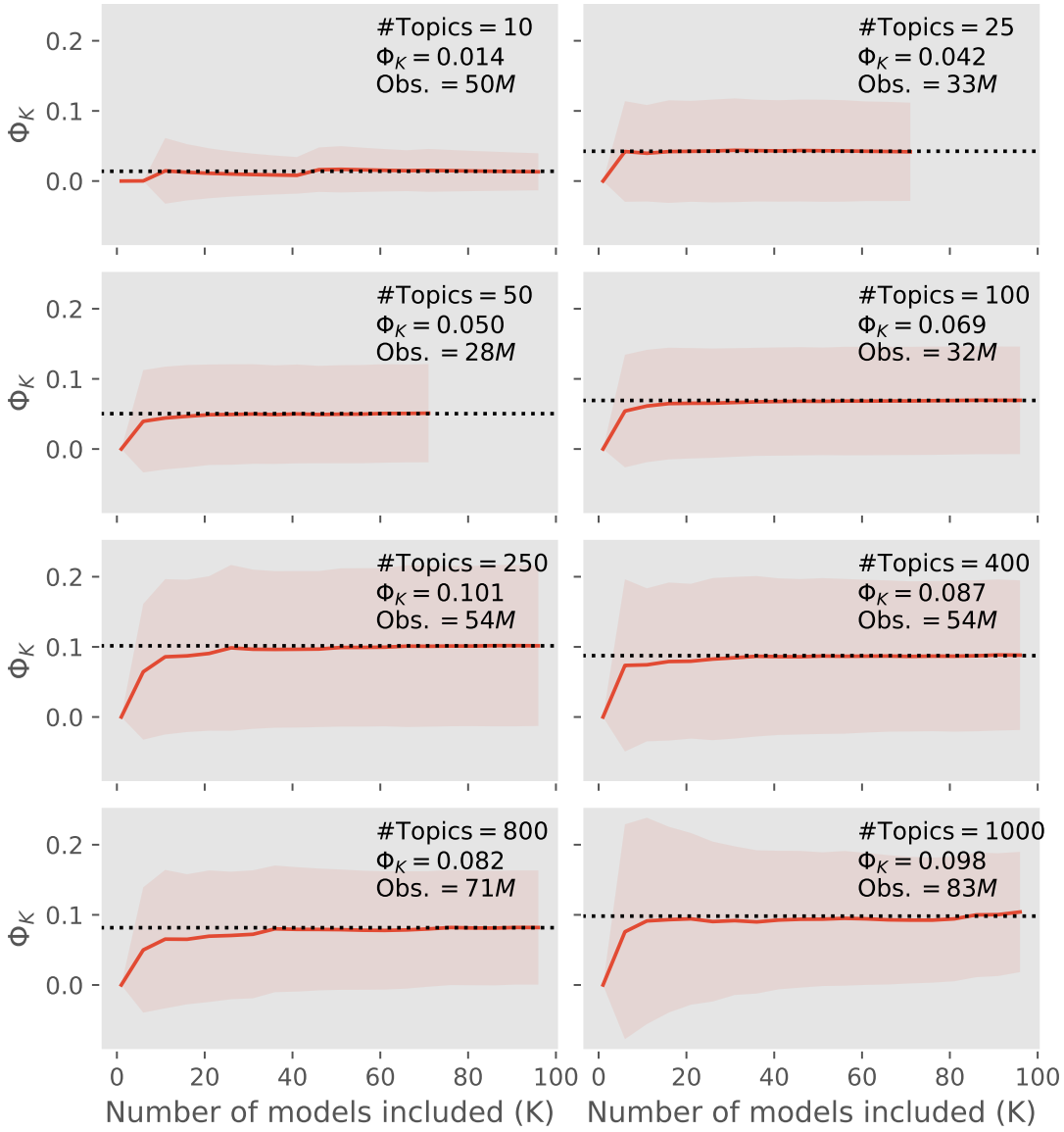


Figure A.14 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.2 \forall k$

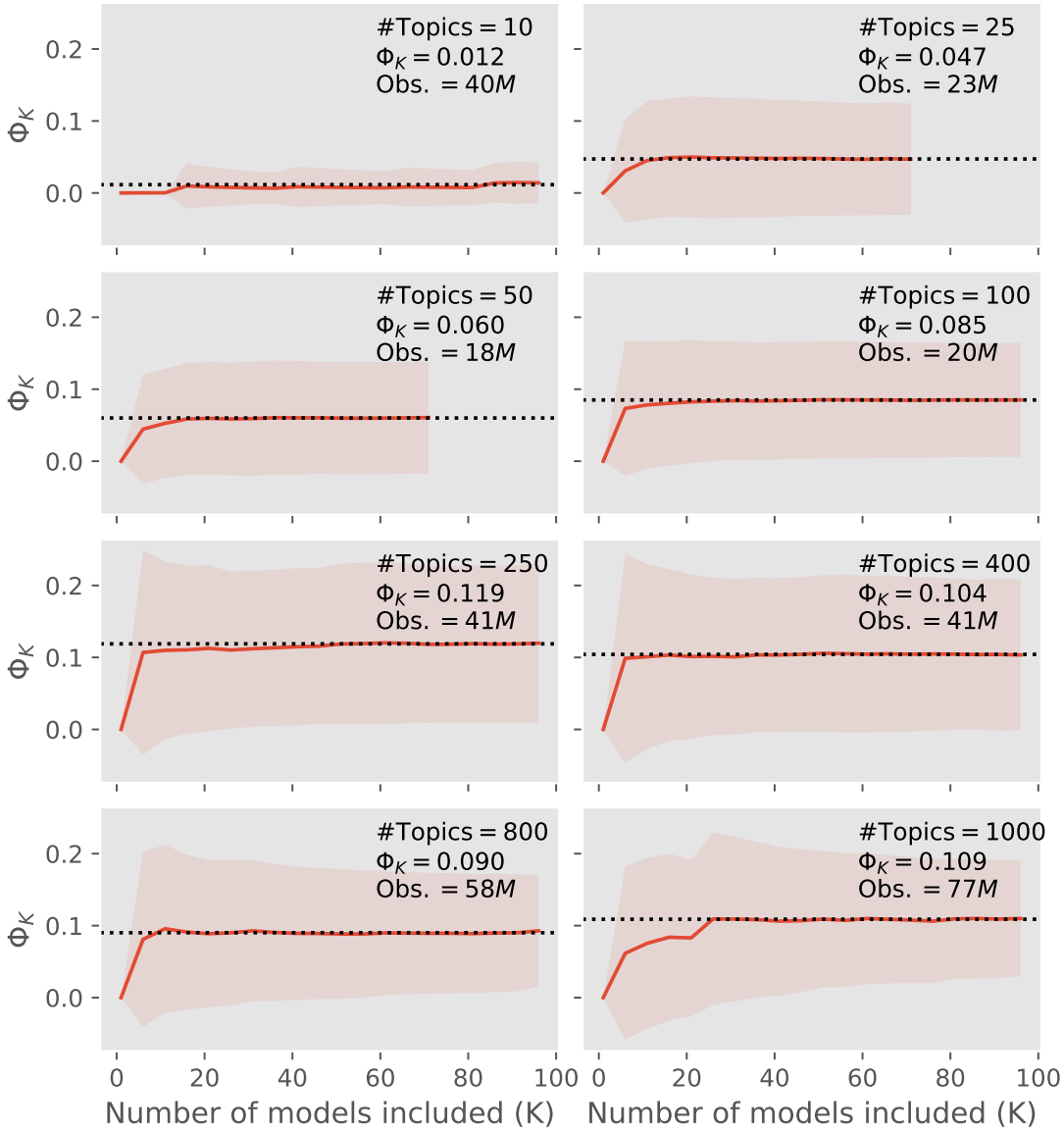


Figure A.15 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.3 \forall k$

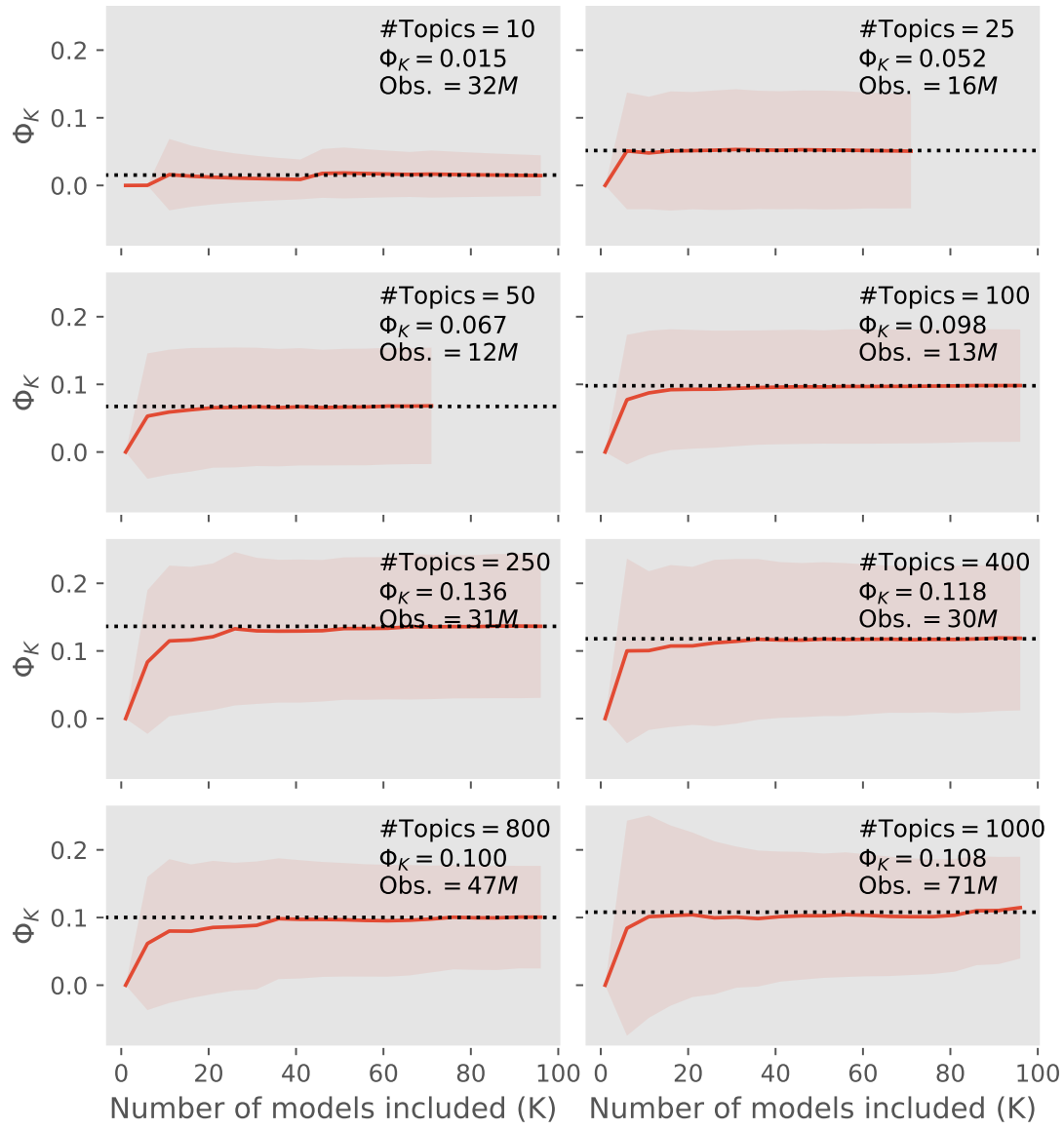


Figure A.16 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 0.4 \forall k$

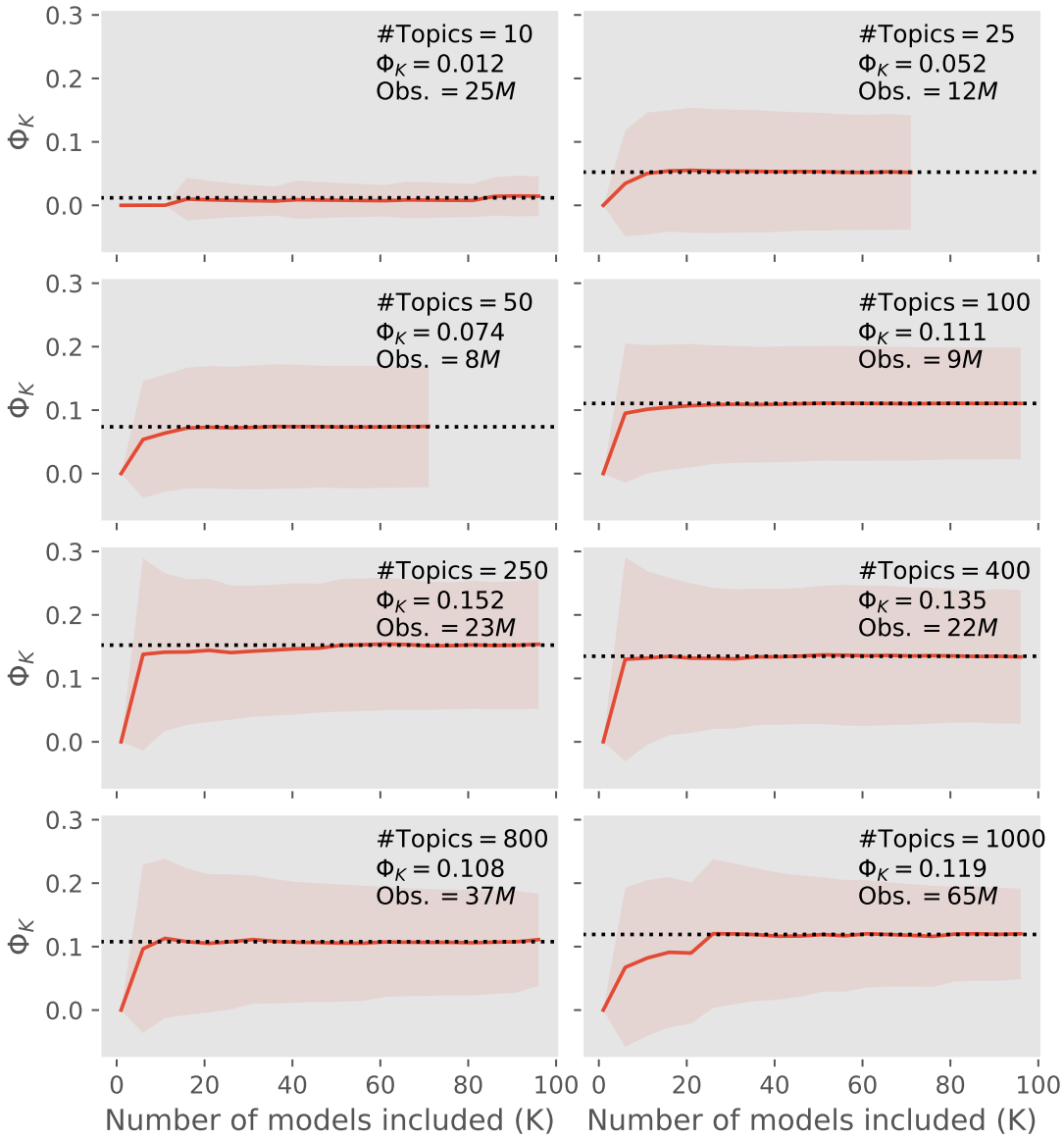
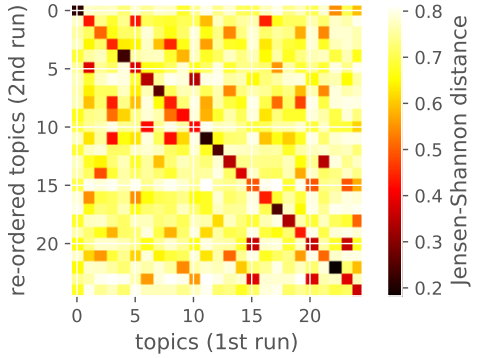


Figure A.17 – **Average Standard Deviation for NMF:** Asymptotic value over multiple retrainings (75 or 100) of NMF for 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions filtering out pairwise similarities that satisfy $\epsilon < 051 \forall k$

A.4 Jensen-Shannon and Top Words LDA

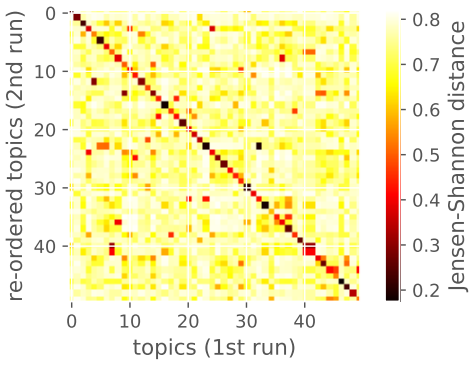


(a) Similarity of topic-terms 1 and 2 (reordered)

	Model 1	Model 2
Most sim topic	Parkinson Disease	Parkinson Disease
	Levodopa	Levodopa
	Antiparkinson Agents	Antiparkinson Agents
	Aged, 80 and over	Aged, 80 and over
	Huntington Diseas	Substantia Nigra
	Parkinsonian Disorder	Corpus Striatum
	Dystonia	Dystonia
	Glbus Pallidus	Parkinsonian Disorders
	Tremor	Dopamine
	Severity of Illness Index	Parkinson Disease, Secondary
Least sim topic	Rats, Inbred Strains	Rats, Inbred Strains
	Hypothalamus	Behavior, Animal
	Estradiol	Rats, Wistar
	RNA, Messenger	Rats, Sprague-Dawley
	Immunohistochemistry	Motor Activity
	Rats, Wistar	Corticosterone
	Radioimmunoassay	Conditioning, Operant
	Pregnancy	Estradiol
	Rats, Sprague-Dawley	Hypotalamus
	Sexual Behavior, Animal	Sexual Behavior, Animal

(b) Top 10 words between most and least similar pair

Figure A.18 – **"Traditional" stability of topics**: different runs of LDA with 25 topics

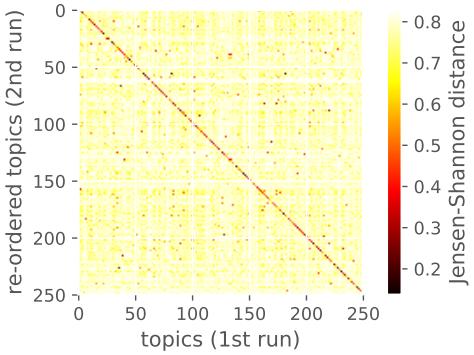


(a) Similarity of topic-terms 1 and 2 (reordered)

	Model 1	Model 2
Most sim topic	Parkinson Disease	Parkinson Disease
	Levodopa	Levodopa
	Antiparkinson Agents	Antiparkinson Agents
	Aged, 80 and over	Aged, 80 and over
	Huntington Diseas	Substantia Nigra
	Parkinsonian Disorder	Corpus Striatum
	Dystonia	Dystonia
	Glbus Pallidus	Parkinsonian Disorders
	Tremor	Dopamine
	Severity of Illness Index	Parkinson Disease, Secondary
Least sim topic	Immunohistochemistry	Estradiol
	Substance P	Testosterone
	Neuropeptides	Sexual Behaviour, Animal
	Smell	Hypotalamus
	Sexcual Behavior, Animal	Circadian Rhythm
	Odors	Progesterone
	Testosterone	Sex Characteristics
	Preoptic Area	Preoptic Area
	Taste	Ovariectomy
	Olfactory Pathways	Immunohistochemistry

(b) Top 10 words between most and least similar pair

Figure A.19 – **"Traditional" stability of topics**: different runs of LDA with 50 topics



(a) Similarity of topic-terms 1 and 2 (reordered)

	Model 1	Model 2
Most sim topic	Motor Cortex	Motor Cortex
	Electric Stimulation	Electric Stimulation
	Evoked Potentials, Motor	Electromyography
	Transcranial Magnetic Stimulation	Evoked Potentials,Motor
	Electromyography	Transcranial Magnetic Stimulation
	Magnetics	Reaction Time
	Reaction Time	Magnetics
	Functional Laterality	Movement
	Neural Inhibition	Hand
	Hand	Functional Laterality
Least sim topic	Amino Acid Transport System X-AG	Carrier Proteins
	ATP-Binding Cassette Transporters	Membrane Transport Proteins
	Symporters	Membrane Glycoproteins
	Excitatory Amino Acid Transporter 2	Nerve Tissue Proteins
	Glutamate Plasma Membrane Transport Proteins	Symporters
	Glutamic Acid	Dopamine Plasma Membrane Transport Proteins
	Excitatory Amino Acid Transporter 1	Serotonin Plasma Membrane Transport Proteins
	Excitatory Amino Acid Transporter 3	Vesicular Monoamine Transport Proteins
	Carrier Proteins	Vesicular Biogenic Amine Transport Proteins
	Biological Transport	Serotonin

(b) Top 10 words between most and least similar pair

Figure A.20 – **"Traditional" stability of topics**: different runs of LDA with 250 topics

B Chapter 2: Appendix

B.1 Motivation & Empirical Evidence

B.1.1 Macro Evidence

Because macroeconomic trends have never (to the best of our knowledge) looked into the headcount statistics, we suggest two examples from the life sciences as an attempt to illustrate better the rate and direction (choice of direction) of scientific effort. The social returns to these scientific efforts are often measured through publication counts or allocation of funding. However, due to the difficulty of working with researcher data (mainly due to disambiguation issues), the returns to human capital are rarely accounted for. We believe this perspective provides new insights into the dynamics of science and its organisation since ultimately, all projects are performed by active researchers. In order to do so, we analyse PUBMED data, for which we can leverage author data by using Author-ity, a high-quality disambiguated database of researchers by Torvik and Smalheiser (2009).

The Case for Cancer

Research productivity is usually measured through publication counts, and citations are often a proxy for scientific quality. Bloom et al. (2020) suggest an alternative approach. In their work, they use publication counts as an input and measure productivity with societal advancements (related to the scientific fields of the input literature). For the life sciences, they use life expectancy increases that link to research in different sub-fields. Following Bloom et al. (2020)'s lead, we analyse one particular example, cancer deaths (all types). However, rather than publications, we introduce workforce (count of researchers) as input. Furthermore, we also incorporate FDA-approved cancer-related drugs as an output.

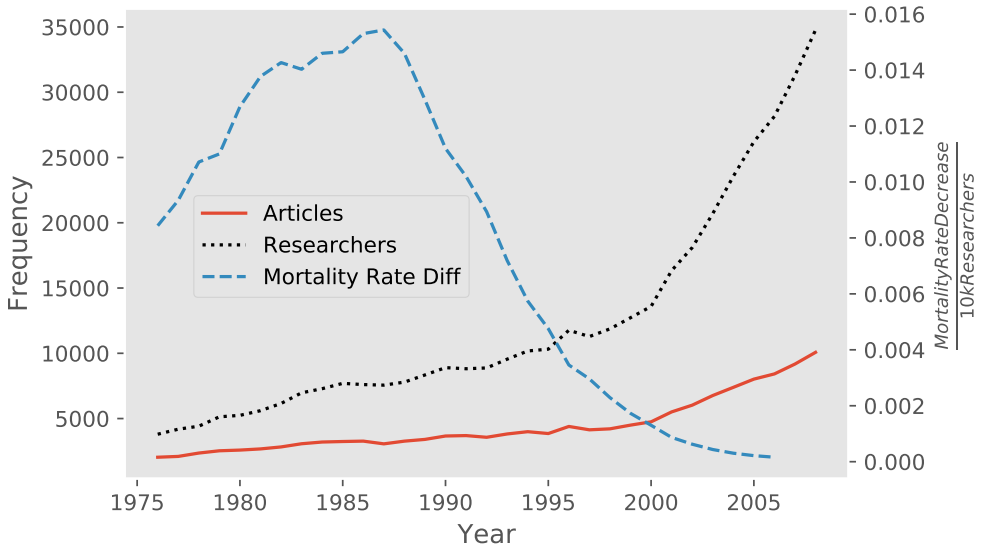


Figure B.1 – **Cancer Mortality Decrease per Researcher:** (left axis) Yearly count of Publications containing MeSH Term "Neoplasms" (solid red). Unique count of researchers contributing to the identified publications (dotted black). (right axis) Yearly mortality rate decrease, computed from survival rates five years after diagnostic for ages 50+ (dashed blue). Data extracted from PUBMED, Author-ity and https://seer.cancer.gov/

.

Figure B.1 compares the rate of arrival of publications and unique researchers to cancer research. For that, we count all the scientific publications indexed in PUBMED that include the Medical Subject Heading (MeSH) term *"Neoplasms"* (solid red). We then count the unique number of researchers involved in these publications (dotted black).[1] The exponential growth in researchers has not directly translated into the same rate of growth in publications. The same figure also shows the Mortality Rate Decrease (changes in mortality rate from the previous year) by researcher involved (dashed blue).[2]

There is not a single reason that can fully explain the divergence in these trends. First, as recent literature has pointed out, there could be an exhaustion of ideas (Bloom et al., 2020). Second, as science advances, larger teams need to work in a single problem (Jones, 2009). Alternatively, even, it is possible that through better disease understanding and prevention measures, marginal gains are have rarefied. The headcount growth coincides with a worldwide paradigm shift in cancer research, which moved to more applied work (Eckhouse et al., 2008). Moreover, with the subsequent increase in funds directly targeted at cancer research (von Eschenbach, 2003). Figure B.2 displays similar trends. This time, however, the publication counts are a subset of the previous, identified as "Clinical Trials" by either typology or MeSH term. Similarly, the researcher count takes only into account scientists involved in this article

[1] Unit count regardless of the number of contributions per year, as long as there is at least one
[2] See Appendix B

subset. Along with the frequencies, we show the number of FDA-Approved drugs per active researcher. [3]

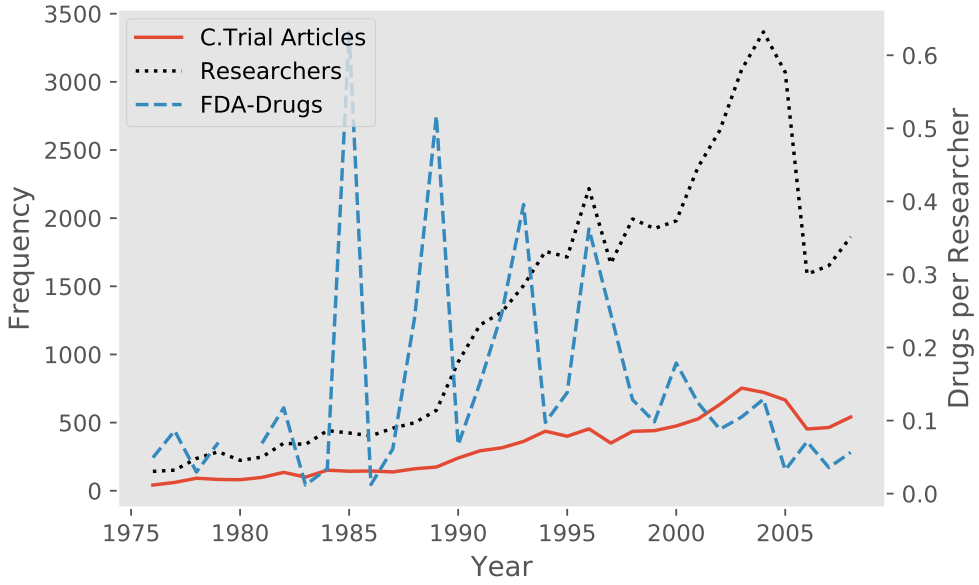


Figure B.2 – **FDA-Approved Drugs per researcher:** (left axis). Yearly count of publications containing MeSH Term "Neoplasms" and "Clinical Trial" or where the document type is "Clinical Trial" (solid red). Yearly (unique) count of researchers contributing to the identified publications (dotted black). Yearly count of FDA-Approved drugs (dashed blue). Data extracted from PUBMED, Author-ity and https://nctr-crs.fda.gov/fdalabel/ui/search

.

The cancer trends seem to suggest that scientists engaging in a mature and resourceful topic decrease their productivity both in terms of publication counts and overall reach (societal impact) of their research. The increase in the number of active researchers is unparalleled by neither the publication count nor the approval of new drugs.

We should not disregard the limitations to the figure. For instance, the latest breakthrough research in cancer, has focused on personalised treatments — such as immunotherapy or the underlying mechanisms of the genetic mutations that ultimately cause the disease (ACI, 2020). These techniques will certainly have an enormous impact on the mortality rate, but require time and the effort of many, while they probably do not show in the FDA statistics (yet, if at all). Personalised therapies are driven on a case-by-case basis, and many of the developments may also not be published. In addition, companies seem to be publishing *science* less often (Arora et al., 2018), which might lead to a lower count of researchers.

[3] FDA-Approved data extracted from FDA Databases keyword search https://nctr-crs.fda.gov/fdalabel/ui/search. Results containing "Cancer", "Tumour", "Metastatic" and "Chemotherapy" are aggregated together by Year of Initial USA Approval.

Next, we provide further evidence in support of our ideas from a different disease: malaria.

The Case for Malaria

By 1950, malaria was close to eradication in Europe and the United States. The discovery of DDT insecticides and the first antimalarials between the great wars led to a substantial decrease of outbreaks.[4] In 1955 the recently created WHO established a campaign for its full eradication. In subsequent years, however, resistance to the first treatments and insecticides —along with a generalised ban on DDT compounds— led to a resurgence of the disease. From 1970 onward, the scientific community engaged in a *race* towards the control of malaria diffusion (or eradication whenever possible). These research efforts led to the first vaccine trials in 1987 (SPf66-vaccine) and 1992 (RTS,S vaccine) and the development of several treatment compounds in the late 1980s.

Albeit only in developing countries, a drastic increase in the malaria burden drew international coordination efforts through the Global Malaria Control Strategy 1993–2000 (WHO). In the late 1990s international cooperation and private non-profits began targeting malaria incisively (The Global Fund, 2002; Malaria R&D Alliance, 2004; Bill and Melinda Gates Foundation, 2003-), thus boosting the availability of funds for research up to this day (Vaughan et al., 2012).

In order to gauge the productivity of research in malaria, we examine the input-output ratio of research. As an input measure, we employ the number of (academic) researchers who took part in at least one contribution in any given year. The output from the human capital is measured with either the count of scientific publications in PUBMED that have "Malaria" or "Malaria Vaccine" as a MeSH term, or a USPTO Patent count. We target two different counts of patents: those that cite at least one scientific publication from the sample (follow-on patents); and those that contain the term "malaria." Only granted patents are considered in our analysis. Data are extracted from The Lens (Jefferson et al., 2018), and researchers are uniquely identified through their Microsoft Academic Graph ID. [5]

Figure B.3 shows a rather suggestive trend in the field of malaria. One can only speculate about the sudden drop in workforce and scientific publications following the first *potential* vaccine trials in 1987 and 1992. The coetaneous increase in malaria-related patents hints towards a shift of focus.[6] During the 1980s, the race for a *big hit* on malaria was on, attracting an ever-

[4] DDT (dichlorodiphenyltrichloroethane) pesticides was first synthesised in 1874 by the Austrian chemist Othmar Zeidler. DDT's insecticidal action was discovered by the Swiss chemist Paul Hermann Mueller in 1939. DDT was used in the second half of World War II to control malaria and typhus among civilians and troops. After its widespread adoption and commercialisation, DDT was discovered to be extremely toxic, dangerous to the environment, likely carcinogenic, leading to its ban a few years later.

[5] The researcher count in the early years might be lower than the real numbers due to missing information in the retrieved publications. Researchers are indexed by a unique *Microsoft Academic Graph ID*

[6] Although we have tried to verify the data consistency or a source of literature that points towards the drop, we have not been successful. Therefore, we do not discard that the sudden and rather large drop in publications is a pure data artefact. In the Appendix – Figure B.5– we tested whether there was a displacement of researchers towards HIV research, which we do not find.

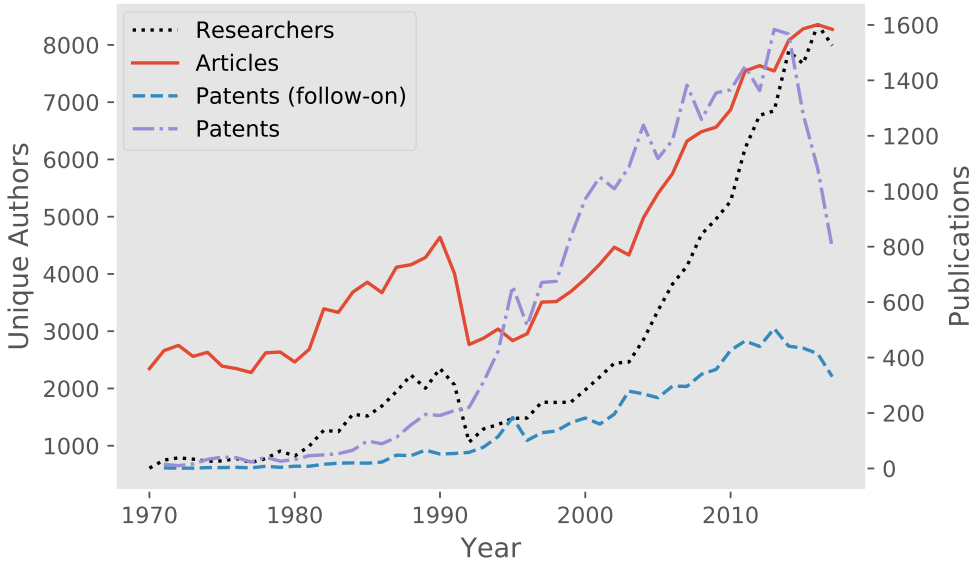


Figure B.3 – **Research input-output for malaria:** Left axis is the count of unique researchers with at least malaria-related scientific article (dotted black). Right axis is the count of: malaria-related scientific publications *(solid red)*; follow-on patents *(dashed blue)*; patents containing the word "malaria" *(dash-dot purple)*. Data extracted from *The Lens*

.

growing pool of researchers thriving for the reputation gains. The search for a suitable solution to malaria infections was, by and large, a competition for priority, that spurred exploration of ideas. With the advent of the first clinical trials, we conjecture, the focus shifted from exploratory (basic) research to more applied, close-to-market research. This change of approach, changed the nature of research in the field to a consolidatory (or exploitative) approach, increasing patentability and effectively slowing down the rate of publications in the field during the 1990s.[7] With the arrival of private capital in the 2000s, the number of researchers rapidly rose at a faster-than-ever rate, while the rate of arrival of publications increased at a much slower speed. The large stipends from private funds directed towards malaria research and the association gains attracted many researchers, who engaged in cumulative research rather than breakthrough advancements.

Unlike cancer, malaria is an infectious viral disease. The transmission mechanisms and potential solutions —i.e., a working vaccine— are well understood. Therefore, we wonder whether research steered from exploratory (and competitive) to exploitative as figure B.4 seems to indicate. It shows that the rate of patents per researcher directly citing academic work has remained constant, while the number of articles and patents per researcher has not.

The two examples provided above raise many intriguing questions, and they have motivated

[7] With the inclusion of follow-on patents, we hope to convince the reader that the Bayh–Dole Act is not entirely responsible for the rapid increase in patents.

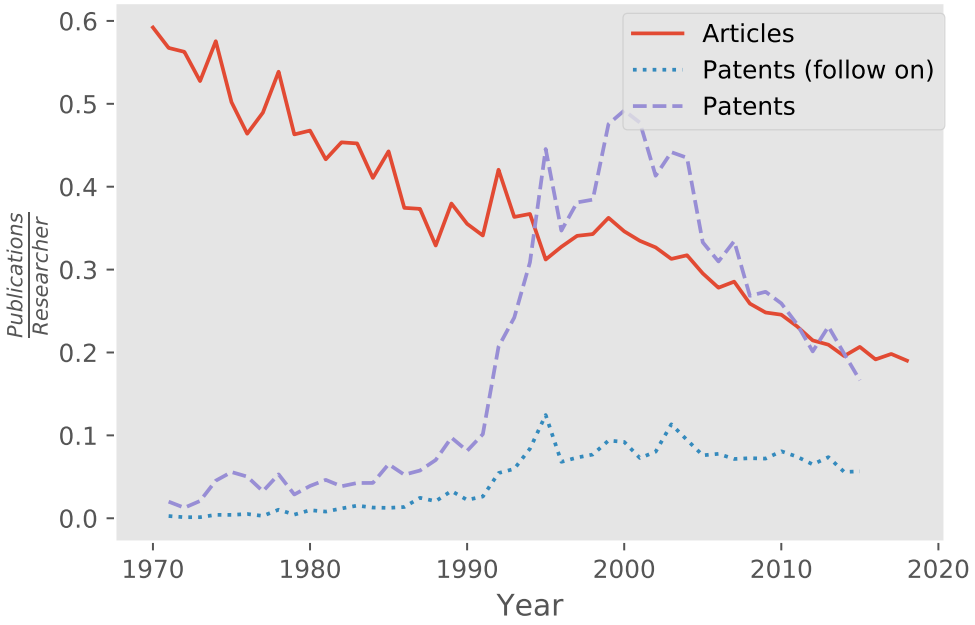


Figure B.4 – **Research output ratio for malaria:** Count ratio of publications per unique active researcher in a given year for: scientific publications *(solid red)*; follow-on patents *(dotted blue)*; patents containing the word "malaria" *(dashed purple)*. Data extracted from *The Lens* .

the ones we will address in this work. We are interested in studying the organisation of science from the researcher's viewpoint. However, before we do so, we attempt to find evidence that can help explain two particular problems:

- What are the characteristics of fields displaying more breakthrough research?
- To what extent are researchers aware of the competition? How do they respond?

In the next section, we provide tentative empirical evidence spanning other specialities of the life sciences, which will help motivate the formal model of science as a common-pool resource game presented in Section 2.3

B.1.2 Micro-Empirical Evidence

This section presents descriptive evidence of a complex relationship between field characteristics and the direction of research. Guided by the two questions raised in the previous section, we provide some suggestive results. It should be noted, however, that our analysis does not address endogeneity nor omitted variable bias. The objective of this section is merely to document these observations and motivate the model developed in Section 2.3. We use a data-driven approach that makes no use of ad-hoc field classifications nor heuristic groupings.

Using topic modelling, researchers are associated based on the contextual similarity of their work through techniques described in Chapter 1.

B.1.3 Where is breakthrough research published?

Wang et al. (2017) show how novel research within the boundaries of a subfield (that they call the *home field* to the publication) suffers from a lower impact in the short term. In contrast, highly novel papers are more likely to be highly cited in foreign fields (other than the field where the paper was published). They find that novel papers are more likely to become big hits, with follow-on research that generates a larger impact. Their results are not alien to the scientometrics literature (Yegros-Yegros et al., 2015) nor other social science's understanding of science (Kuhn, 1977; Bourdieu, 1975). Groundbreaking research frequently faces confrontation in the knowledge areas where it belongs. We use this distinguishing feature as a proxy for impact to understand the characteristics of the field where an article is published. For this, we regress citation counts from the *home* and *foreign* field on sub-field covariates.

Data: We collect PUBMED articles published between 2000–2002. The years chosen correspond to a period of almost constant growth in NIH funding, the major source of funds for the life sciences in the United States. Moreover, these years are well covered in Author-ity (Torvik and Smalheiser, 2009). All these characteristics during this time period allow us to have a buffer of high-quality data (a merge between the two databases) both before and after the period, in order to compile forward citations and determine the years a researcher has been active.

Journal Classification methods have largely been used in order to define specialities in science (Wang et al., 2017; Boyack et al., 2005). We use a combination of Machine Learning tools to automatically group journals that deal with similar topics. The groupings generated by topic similarity will be our basis to define sub-fields. We use the methods described in Chapter 1 applied to Journals for each year. Hence, we start by defining our "documents". We characterise a Journal-Year document as the compilation of Medical Subject Headings (MeSH) published in a given periodical throughout a year. That is, for each article that appeared in the same journal during a year, we compile the MeSH terms. We then group all of the terms as a single list of tokens and generate the document that represents any given Journal-Year.

We subsequently train a Doc2Vec model on Journal-Year documents ranging from 1985 to 2010 and generate inferred document embeddings (vectors) from 1985 to 2014. That comprises almost the entirety of our in-house PUBMED database. Following the same visualisation methodology presented in Chapter 1 (t-SNE), we plot the resulting embeddings. A 2D projection of these can be found in the Appendix B in Figure B.7.[8] The figure shows how multiple clusters emerge "naturally" from the data. We self-validate the topic model with a similarity-coherence test: excepting counted occasions, for each Journal-Year, the document

[8] Interactive figure online at: https://github.com/oballegon/Thesis/blob/master/doc2vec_journalYear_19852010_150_5year.html

embeddings corresponding to Journal-Year(-1) and Journal-Year(+1) are within the top-five most similar documents. Finally, we cluster the journal-year vectors using Agglomerative Clustering (a bottom-up approach to Hierarchical Clustering) in 100 groups that constitute sub-fields.[9]

Dependent variables: We use unnormalised citation counts as dependent variables.[10] There are three variables of interest. First, internal forward citations, which come from papers that belong to the same cluster as the focus article, the *home* field — i.e., are published in a Journal-Year inside the same cluster. Second, outer forward citations, which come from papers from other clusters, the *foreign* clusters. Third, for robustness, we also compute the three Nearest Neighbours to a cluster, and count the citations received from outer fields *excluding* the three nearest clusters. We also regress the percentage of outer citations. The raw counts are regressed as unit offset logs, as it is standard with count data. Given the size of the sample and the goal of finding simple descriptive relationships, other count models (Negative Binomial and Poisson) were not considered, and the log-linear approximation deemed sufficient for the level of analysis.

Independent variables: we collect basic article metrics including the number of coauthors and the principal investigator (which we identify as the last author, as it is standard in the life sciences). Using the article data, we characterise the clusters. Size of a cluster is computed for each cluster-year as the count of PIs with at least three (we also compute for two) publications in the cluster in the previous two years. The Growth of a cluster is also computed yearly as the slope of a linear regression of *Size* on a five-year window around the focus year (two years before and after). Size and Growth are then standardised $\sim N(0,1)$.

Algorithm 1 summarises the process to obtain the data for the cluster citations. The resulting data is a panel of Articles from all biomedical specialities between years 2000–2002. Summary statistics are displayed in Table B.1. We estimate Equation B.1.

$$\mathrm{CIT}_i = \quad \epsilon_i + \gamma_j + \delta_t + \beta_0 + \beta_1 \mathrm{FSize}_{ijt} + \beta_2 \mathrm{FGrow}_{ijt} + \mathrm{R}_i \tag{B.1}$$

where CIT represents the *home (IN)* or *foreign (OUT)* forward citations, γ_j are field fixed effects, δ_t are time fixed effects, FSize_{ijt} and FGrow_{ijt} are the *home*-field Size and Growth in the year of the publication of the main article and R_i is a vector of article characteristics including number of coauthors and impact (total number of forward citations). The results are displayed in Table B.2.

The *home* field size has a negative correlation with forward citations from *foreign* fields

[9] The choice of 100 sub-topics is entirely arbitrary, and chosen for simplicity. Agglomerative Clustering is a hierarchical method in which each element starts on its own individual cluster, and is paired up with the nearest in each step until the threshold is achieved. This method is equivalent to an inverse dendrogram approach, where pairs of elements are grouped together only when they are the closest among all the possibilities. It is therefore convenient for grouping journals together based on their relative positions in the latent (topic) space.

[10] The use of unnormalised data introduces aa bias that we overcome through Year and Field fixed effects in the regression model

Algorithm 1: Pseudo Code to generate In-Out citations

input: PUBMED Articles 1985-2014
foreach *Journal* **do**
  foreach *Year* **do**
    Compile all publication MeSH;
    Generate document embedding;
output: Journal-Year document vectors
↓
Clusters←`Cluster(`*Journal-Year*`)`
↓
input: Clusters; Author-ity DB; WoS DB
foreach *Article* **do**
  ArtClus←Assign cluster **from** clusters(J-Y);
  TotCit←Get Citations;
  InCit← Citations **from** ArtClus;
  OutCit← Citations **not from** ArtClus;
  PI← Get Principal Investigator;
  NumAuth ←Get Number of Authors;
output: ArticleData (ArtClus; TotCit; InCit; OutCit; PI; NumAuth)
↓
input: Clusters; ArticleData;
foreach *Article* **do**
  sizeCl-PI←`Count(`*PI* **in** *Cluster in 2-year window*`)`;
  sizeCl-J←`Count(`*Unique Journals* **in** *Cluster*`)`;
  growCl-PI←Slope of `Regress(`*size-PI*`)` in 5 year window;
  growCl-J←Slope of `Regress(`*size-J*`)` in 5 year window;
output: ArticleDataEnhanced(size-PI; size-J; grow-PI; grow-J)

Table B.1 – Summary Statistics Article Data

	Mean	*StD*	*Min*	*Max*
Year	2001.03	0.82	2000	2002
Tot Cit	27.87	73.68	0	27700
Citations from same Cluster	5.07	11.02	0	1497
Citations from other Cluster	10.50	29.38	0	5938
Cit other Clust excluding 3 NN	7.91	23.52	0	5493
Num Coauthors	4.25	3.44	1	551
Growth Field (min 3P)	52.29	160.74	−460	709
Growth Field (min 2P)	120.29	299.04	−729	1235
Cluster Size (min 3P)	1094.16	987.07	0	4169
Cluster Size (min 2P)	2475.46	2208.12	2	9047
N	1172761			
Clusters	100			
Unique PIs	461074			

Table B.2 – OLS - Citation from Inside-Outside the article cluster

	In Cit	Out Cit	Out Cit - 3NN	% Out Cit
	(1)	(2)	(3)	(4)
Cluster Growth	0.144***	−0.00757	0.00256	−0.00768***
	(0.00521)	(0.00585)	(0.00573)	(0.00129)
Cluster Size (min 3P)	0.000307	−0.0401***	−0.0526***	−0.0101***
	(0.00893)	(0.00984)	(0.00956)	(0.00196)
Coauthors	0.0573***	0.0730***	0.0666***	−0.000471***
	(0.00283)	(0.00356)	(0.00338)	(0.000106)
Total Citations	0.00496***	0.00621***	0.00594***	
	(0.000560)	(0.000697)	(0.000665)	
Observations	1.17M	1.17M	1.17M	0.97M

Standard errors in parentheses adjusted for heteroskedasticity

Including Year FE, Cluster FE, Principal Investigator FE, controls for number of coauthors, total citations and constant.

Dependent variables (1),(2),(3) are regressed using unit-offset natural logs.

Cluster Growth and Cluster Size enter the regression after standardisation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

controlling for total number of citations. This outcome may, of course, stem from a simple mechanical effect. The influence increases with field distance when we exclude the nearest fields from the citation counts, as columns (3) and (4) in Table B.2 show. On the contrary, Size is not significantly associated with citations within the field, while field Growth (arrival of more researchers) does. Again, this could be a mechanical effect. Altogether these results suggest that more innovative research (i.e. more citations from foreign specialities) are more likely to be published in smaller fields. In contrast, *exploitative* research is more likely to appear in fields attracting researchers at a higher rate. But, to what extent is this an active choice? Are researchers aware of field size? In the next section, we explore whether there is a correlation between the size and the placement of an article.

B.1.4 Are researchers aware of field size? How do they respond?

In this subsection, we show anecdotal evidence about the researcher's decision to enter a field. We would expect researchers to learn about field characteristics over time, and make an informed choice on where to publish. In order to do so, we make use of the fields constructed following Algorithm 1, and we generate a panel of entry into a field by a researcher. We then proceed to estimate the following probability:

$$\begin{aligned} \text{Entry}_{jit} = \quad & \epsilon_{ji} + \beta_1 \text{R}_i + \beta_2 \text{F}_{jt} + \gamma_t + c_{ji} \\ & \beta_3(\text{FSize}_{jt} \cdot R_i\text{exp}) + \beta_4(\text{FGrow}_{jt} \cdot \text{R}_i\text{exp}) + \\ & \beta_5(\text{FSize}_{jt} \cdot \text{R}_i\text{qual}) + \beta_6(\text{FGrow}_{jt} \cdot \text{R}_i\text{qual}) \end{aligned} \tag{B.2}$$

where Entry is a binary variable that takes value 1 when a PI publishes at least two articles in a given year in a cluster where she has never published before. R_i is a vector of researcher i. F_{jt} is a vector of field j characteristics in year t. $R_i$exp is the experience of a researcher, as the number of years between the entry event and the first indexed publication. $R_i$qual is a proxy of the researcher's ability through the log(Citations). Of particular interest are the signs of the interaction terms between field size and growth, and researcher years of experience and ability.

Data: We extract from Author-ity all PIs whose first indexed publication in PUBMED is between 1992 and 1996. Using the field classification from algorithm 1, we compute, for each PI, the fields where she is active yearly. Entry in a field takes value 1 whenever she publishes two or more articles in a field for the first time within a year. Entry takes value 0 for all the other fields where she has past or future publications. The zeroes, then, capture those fields susceptible of an entry event in a different year. We register entry events for five years, between 1999 and 2003. We enrich the data with the average number of citations of articles in a field, average number of coauthors of articles in a field, number of fields in which a researcher takes part (spread), the average number of coauthors in the researcher's publications and forward citations received by the researcher.

Table B.4 shows the probability regression delineated in equation B.2. The negative coefficients

Table B.3 – Summary Statistics Entry

	Mean	*StD*	*Min*	*Max*
Entry	0.01	−−	0	1
Year	2001.00	1.41	1999	2003
Researcher Experience	7.59	1.96	3	11
Rs Spread	9.44	4.79	1	41
Rs Avg Coauthors	5.18	1.71	1	23
Rs Avg Citations	31.65	31.31	0	903
Field Growth	34.66	130.37	−479	908
Field Size	781.49	767.45	0	3751
Field Avg Coauthors	4.68	0.82	3	7
Field Avg Citations	33.34	17.95	0	122
N	520500			
Researchers	15314			

of the first row suggest that researchers with more years of experience might be more aware of the size of the field. Therefore, should the number of incumbent researcher be a good proxy for competition in the field, there is a possible interpretation of the coefficients as awareness of competition. The probability of entering an above-average-size field decreases steadily with researcher experience.

In turn, researchers who ultimately receive more credit are less likely to enter fields while they are extensively growing. The interaction term —F Growth × Res logCit — provides tentative evidence that higher ability researchers might deter from publishing in high-growth fields. It suggests that more impactful scientists avoid entering directly into "exploitation" fields. Growth does not significantly correlate with the probability to enter of more experienced researchers in the same way that size effects do not vary across ability.

Reflecting upon the empirical descriptive tables presented in the section herein hints that researcher strategic behaviour may have an impact on the organisational characteristics of science. The evidence is consistent with researchers that are aware of the existence of competition and react accordingly. Scientists will then pursue the strategy that will yield the most significant benefit. Researchers with higher impact (with more citations) seem to avoid exploitation, while the more experienced avoid competition. At the same time, scientific advancements yielding a higher impact are published in below-average-size fields that are not (yet) rapidly attracting more investigators. Scientists are seemingly aware of their abilities and surrounding and place their work accordingly. As a consequence of these observations, we instinctively propose the introduction of appropriability in science modelling as developed in the following section.

Table B.4 – Field Entry Probability

	OLS (1)	OLS (2)	Logit (3)	Probit (4)
F Size × Res Exp		−0.000173**	−0.0250***	−0.00816**
		(0.0000681)	(0.00884)	(0.00320)
F Growth × Res Exp	−0.000166**		0.00450	0.000166
	(0.0000785)		(0.00741)	(0.00283)
F Size × Res logCit		−0.000165	−0.000161	−0.000361
		(0.000145)	(0.0220)	(0.00771)
F Growth × Res logCit	−0.000428**		−0.0329*	−0.0126*
	(0.000207)		(0.0194)	(0.00716)
Field Size		0.00307***	0.291***	0.105***
		(0.000747)	(0.0996)	(0.0355)
Field Growth	0.00514***		0.322***	0.136***
	(0.00100)		(0.0912)	(0.0342)
Res Exp	−0.000746***	−0.000769***	−0.150***	−0.0543***
	(0.0000593)	(0.0000598)	(0.0126)	(0.00450)
Res logCit	0.000257*	0.000247*	−0.0330	−0.0117
	(0.000135)	(0.000134)	(0.0228)	(0.00811)
Res Spread			0.0316***	0.0117***
			(0.00333)	(0.00123)
Res Coau			0.0104	0.00381
			(0.0107)	(0.00380)
Field Coau			0.0844***	0.0290***
			(0.0228)	(0.00810)
Field logCit			0.0355	0.00702
			(0.0459)	(0.0161)
Year FE			*Yes*	*Yes*
Constant	0.0115***	0.0117***	−4.879***	−2.407***
	(0.000651)	(0.000652)	(0.187)	(0.0658)
Observations		520500		

Standard errors in parentheses clustered at the Researcher Level
Cluster Growth and Cluster Size enter the regression after standardisation.
Researcher Fixed Effects were considered but the models failed to converge.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B.2 HIV-Malaria

The exponential increase of resources dedicated to HIV/AIDS research in the late 80s and early 90s meant it became the "hot topic". Figure B.5 shows that the decrease in Malaria research is not due to academics jumping fields. The researchers with publications in both fields within the decade never pile up to more than 1% of the total (unique) researchers per year in HIV-related publications.

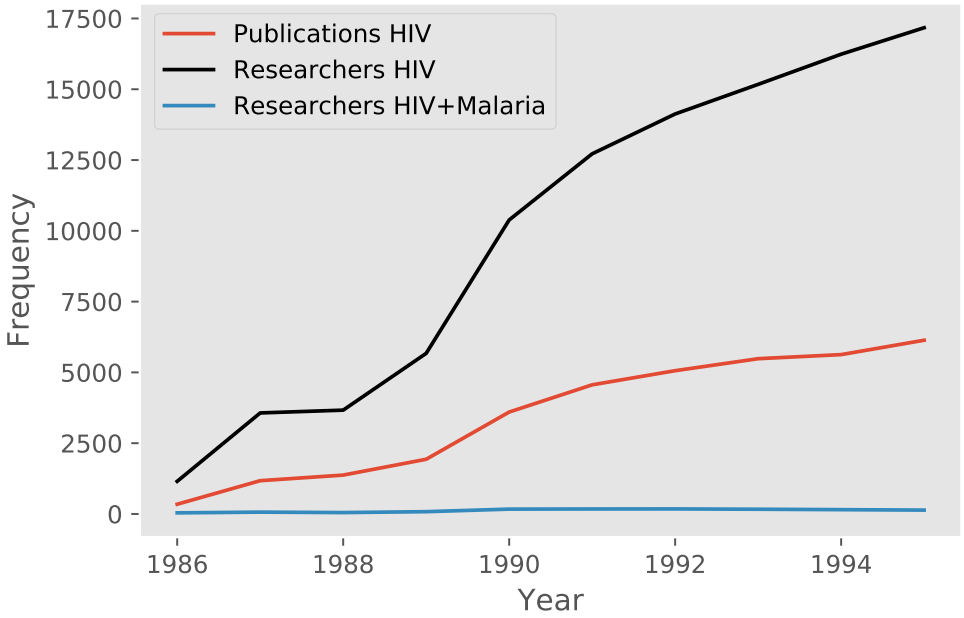


Figure B.5 – **Research input-output for HIV:** Count of: scientific publications with MeSH term "HIV" or "HIV-1" or "HIV Infections" *(red)*; Unique Researchers contributing at least to one publication in a given year *(black)*; Researchers with at least one Malaria-related publication between 1983-1992 contributing to an HIV-publication *(blue)*. Data extracted from *The Lens*

B.3 Cancer

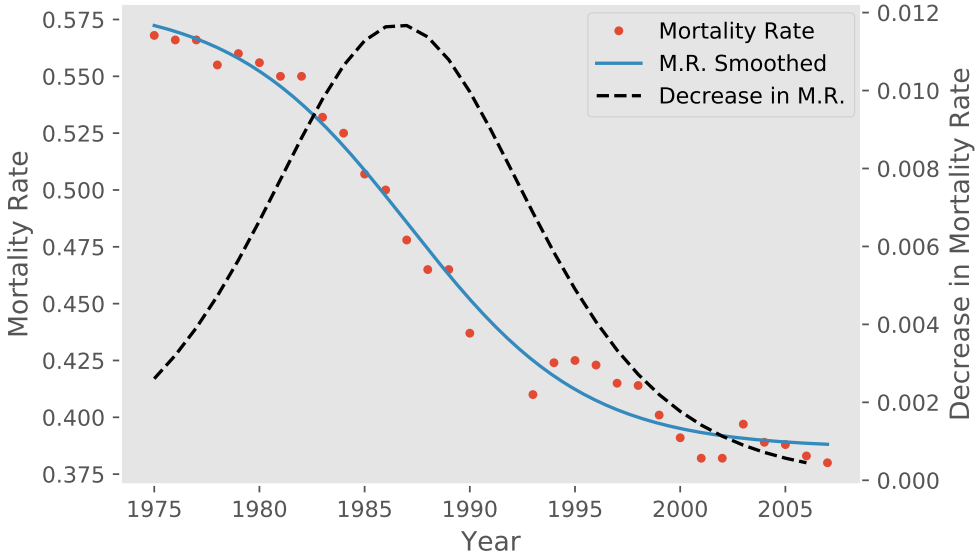


Figure B.6 – **Mortality Rate All Cancer Age 50+:** Mortality rate 5 years after diagnostic *(red)*; Smoothed mortality rate *(blue)*; Decrease in Mortality rate (difference from $t-1$) *(black)*. Data from https://seer.cancer.gov/

B.4 Clustering Fields

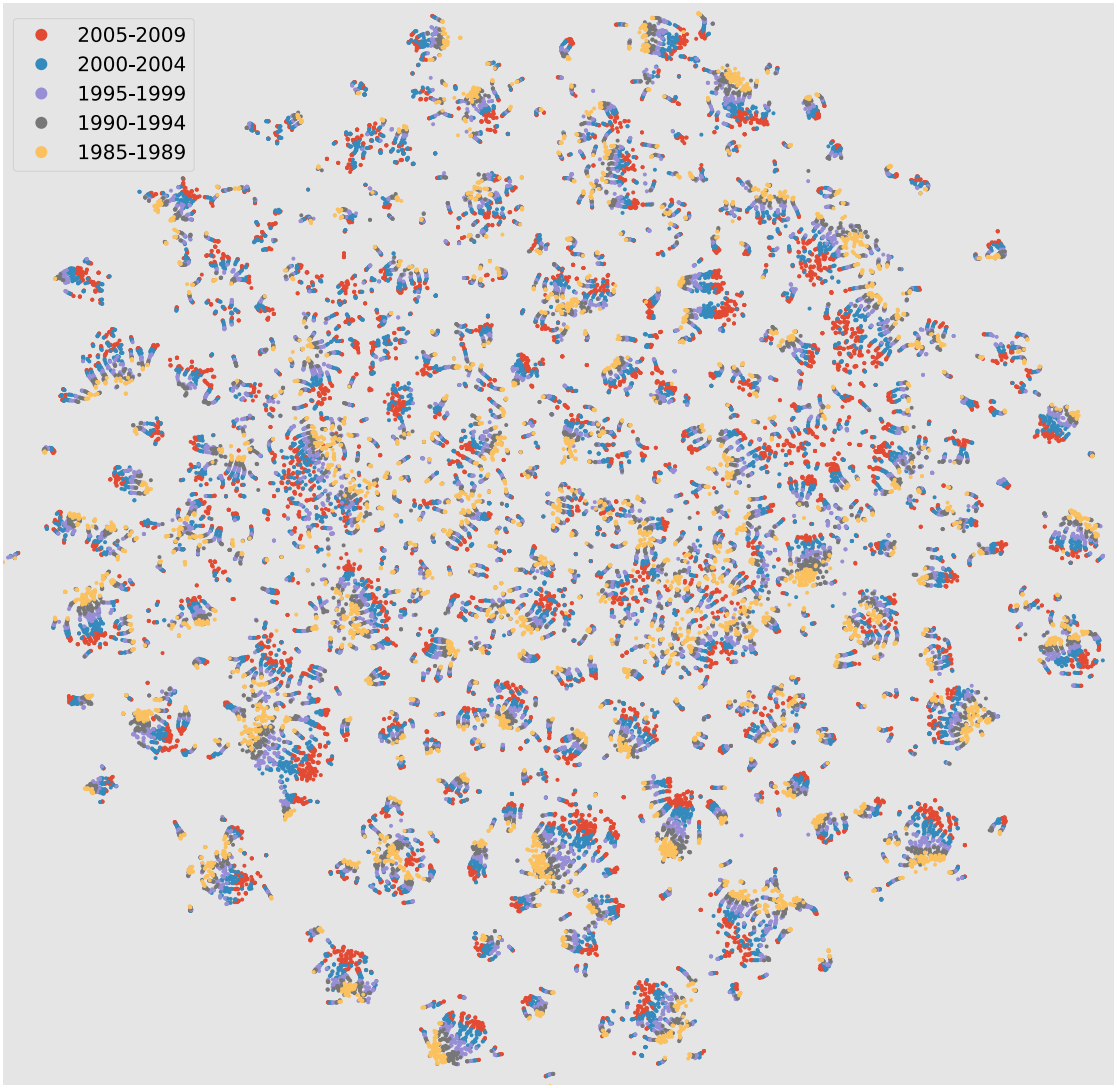


Figure B.7 – **2D Projection of Journal-Year embeddings** Doc2Vec traning on Journal-Year documents. 150 dimensions projected in 2D. Axes (and hence distances) have no intrinsic meaning in this render. Color-coded by groups of 5 years.

B.5 An Extension: Researcher choices as two-period discrete choice model

Motivating the inclusion of dynamics in the study of scientific sub-fields is an easy task. Science is a highly volatile endeavour in which incremental contributions and technological change vastly shape the topics with which it deals the most. In order to instil the inclusion of active choice by the researchers, let us look back at the work of sociologists of science,

following the discourse of the entire chapter.

Once again, sociologists have extensively discussed research choice and problem selection (Busch et al., 1983; Gieryn, 1978; Zuckerman, 1978). Nevertheless, it is mostly the work of Kuhn (1977) and later (Bourdieu, 1975) that defined the *"interplay"* between tradition and innovation (Kuhn) [succession and subversion for Bourdieu] and provided a framework for the consolidation of new fields.[11] Albeit with some particularities, they both argued that scientists were trained in the existing corpus of knowledge in a field. They then faced the strategic choice between converging or deviating from the tradition. Along these lines, recent empirical work by Foster et al. (2015) has tested the dichotomy between innovation and consolidation with conclusive results.

Economic models of research have also explored researcher choices and given academic freedom a high consideration. Aghion et al. (2008) model the process of innovation, stating as a key assumption the creative independence of researchers. For them, the decision lies between a practical and an alternative strategy entirely up to the academics to follow. Similarly, the research cycles model by Bramoulle and Saint-Paul (2010) allows scientist to allocate their choice of effort between *"the exploitation of existing fields and the invention of new ones"* following Kuhn's model of scientific evolution.

As an extension of 2.3, I present a single-agent two-period choice model for the individual researcher inspired by the two-period self-employment model by Humphries (2019). The researcher has now an active choice in the decision to either innovate or consolidate (explore or exploit). Science is still organised around communities of topic speciality, to which a researcher belongs. The strategic action is presented by a class of dynamic discrete choice model, in which researchers devote their time to either task in each time period, depending on their expectations and own abilities. With this formulation, it is possible to interpret the rationale between the researchers' career and their discrete choices between exploration and exploitation within their knowledge area.

Let us assume only non-pecuniary benefits to the research choice. The per-period payoff of choosing consolidation (the average returns as described in sub-section 2.3) is given by:

$$b(\theta, h_{exploit,t}, h_{innovate,t}) + \epsilon_{ex,t} \tag{B.3}$$

where b is the average returns to the contributions in the sub-field, which depends on ability θ and, given the time component, we incorporate the accumulated experience h in exploitation or exploration research $h_{ex,t}, h_{inn,t}$ respectively. $\epsilon_{ex,t}$ is the idiosyncratic shock to average returns. On the other hand, the per-period returns to exploration are given by:

$$\zeta(\theta, h_{ex,t}, h_{inn,t}) E_t^{\alpha} - E_t + \gamma E_1 + \epsilon_{inn,t} \tag{B.4}$$

[11] While Kuhn's book was published later, most of the research had been published before Bourdieu's contribution; hence the "later".

where ζ is the weight of the individual characteristics of the researchers (ability θ and accumulated experience h) to the exploration Effort E. For simplicity, let us assume the *investment* in effort has a Cobb-Douglas productivity with elasticity α, and there is an upfront reward (kind of a risk premium) γE_1 for undertaking innovative activities. The key assumption here is that successful innovations will receive a lasting reward for their effort to innovate, parametrised by $\gamma < 1$ for each speciality. It works as a recognition prize for the successful innovator, independent of the contribution size —i.e. the Cobb-Douglass return on Effort times the productivity of the researcher.[12] The researcher starting off the exploration of a novel research question or an alternative formulation, will face larger efforts to attract both resources and support to these novel ideas. Hence the γE_1 term as a sunk cost.

Taking these payoffs, we construct the indirect utility function for a researcher as a 2-period Bellman Equation (recursive expression of the dynamic programming problem) with a discounting factor β for the (indirect) utility of the future periods. Hence, for period one, the researcher's expected utility of consolidation (exploitation) is:

$$V_{1,exploit}(E_1,\theta,0,0) = b(\theta,0,0) + \epsilon_{ex,t} + \beta\mathbb{E}\left[V(\theta,E_1,1,0)\right] \tag{B.5}$$

and for innovation (exploration):

$$V_{1,explore}(E_1,\theta,0,0) = \zeta(\theta,0,0)E_1^{\alpha} - (1-\gamma)E_1 + \epsilon_{inn,t} + \beta\mathbb{E}\left[V(\theta,E_1,0,1)\right] \tag{B.6}$$

Taking the first order condition from equation B.6 we can derive the optimal level of Effort E_1^*:

$$\alpha\zeta(\theta,0,0)E_1^{\alpha-1} = (1-\gamma) + \beta\frac{\partial}{\partial E_1^*}\mathbb{E}\left[V(\theta,E_1,0,1)\right] \tag{B.7}$$

And, assuming that $\epsilon_{ex,t}$ and $\epsilon_{inn,t}$ both have Type-I Extreme Value Distributions, the binomial choice problem can be expressed using a *logistic choice model* and equation B.7 can be written as [13]:

$$\begin{aligned}\alpha\zeta(\theta,0,0)E_1^{\alpha-1} &= 1+\gamma(-1+\beta Pr(\text{inn}\mid E_1,0,1)\\ &\approx 1-\gamma\cdot Pr(\text{ex}\mid E_1,0,1)\end{aligned} \tag{B.8}$$

so:

$$E_1^* = \left[\frac{\alpha\zeta(\theta,0,0)}{1+\gamma(-1+\beta Pr(\text{inn}\mid E_1,0,1)}\right]^{\frac{1}{1-\alpha}} \tag{B.9}$$

and:

$$E_2^* = \left[\frac{\alpha\zeta(\theta,h_{ex,1},h_{inn,1})}{1-\gamma}\right]^{\frac{1}{1-\alpha}} \tag{B.10}$$

[12] γ may also be understood as an enabler. Some communities have more traditional views (more risk-averse or sceptical). Hence, a lower γ signals less recognition, or lower propensity to, for example, obtain future grants. Low relative γ is thus a "penalty".

[13] See Appendix B.6 for details

Discussion

In the two-period model, the optimal allocation of effort on the first period E_1^* depends on the probability that the researcher will consolidate in subsequent periods. If the indirect utility function for exploitation in $t = 2$ has a larger value than for innovation (conditional on innovating at $t = 1$), the researcher will allocate a greater effort to innovation, knowing she will later consolidate the acquiered knowledge. Similarly, one's self productivity influences the optimal level of effort. On the contrary, fields in which cumulative innovative experience has a large weight on the outcome, call for lower optimal initial levels of effort. Intuitively, the model suggests a researcher will place her odds based on the speciality characteristics. In those knowledge areas where mistakes are costly —experience is paramount— researchers will choose either low innovation efforts early in their careers or consolidation research.

On the second period, the optimal allocation of effort E_2^*, should the researcher choose to innovate, depends on the intrinsic value given by the field to innovation γ. A researcher may forego exploratory research early in their careers if their payoffs are lower in exploitation. Alternatively, if the community penalises risky activities (low γ). She might find an optimal strategy to take more risky research at the end of the career when experience compensates the skewness in preferences towards traditional avenues of investigation.

An external positive shock to the sub-field γ may cause exploration-intensive periods. For instance, in the outbreak of a new disease, rapid scientific advancements have a longstanding effect on the reputation of the scientist, increasing the value of the innovation career-path. Such a shock should be short but involve relatively large efforts. Nonetheless, a negative shock, such as risk-averse funding schemes, will hinder the utility value of exploration, ultimately leading researchers into exploitation. Monetary incentives lacking the adequate reward mechanisms for scientists will enlarge the pool of contributors to consolidation research. This outcome is particularly relevant in funding policy design: lump-sum increases in research expenditure might lead to decreasing productivity with an ill-designed incentive scheme. They, alone, do not suffice to guarantee great leaps forward.

The model's implications can be summarised as follows:

- Researchers who expect to capitalise on early-career innovations (exploit later on), will devote greater efforts if they choose to be disruptive.
- A higher productivity premium (due to either ability or experience) leads to a higher allocation of efforts to innovation. Hence, traditional specialities will see late-stage researchers with a propensity to innovate.
- External shocks to the reward scheme may be the cause of innovation/exploitation. A sudden change in γ will affect the optimal effort levels and the propensity of researchers to choose either.
- Consolidation will be preferred by the majority in mature fields where the risk premium

is low.

- Entering late-career innovation is only possible in those specialities where experience in exploitation and exploration are close substitutes. This transferable skill is particularly symbolic in, perhaps, the social sciences.

This simple model provides a new perspective to the matter in hand: how are fields' contributions distributed, between innovation and consolidation, subversion and tradition. With only preferences from individual researchers but now including a dynamic effect, we derive new policy implications and continue to explain the observations from the empirical literature.

B.6 Discrete Choice Model: Type-I GEV results in logit distribution

The following derivation is based on the textbooks of Amemiya (1990) and Train (2009), which are, in turn, using the first derivation of Logit models for Discrete Choice Theory by McFadden (1974):

In a behavioral model, the agent makes the choice between alternatives selecting the one that provides the greates utility. For agent n there exist $j = 1...J$ alternatives which provide utility U_{nj}. Thus, the agent will chose alternative i if and only if $U_{ni} > U_{nj} \forall i \neq j$. The utility can be decomposed in the observed (indirect) utility V_{nj} and the unobserved (by the researcher) attributes ϵ_{nj} such that $U_{nj} = V_{nj} + \epsilon_{nj}$ where ϵ_{nj} is unknown, and therefore treated as random with (joint) density $f(\epsilon_n)$. Hence, the probability that the decision-maker n selects alternative i is given by:

$$\begin{aligned} P_{ni} = &\quad \text{Prob}(U_{ni} > U_{nj} \quad \forall i \neq j) \\ = &\quad \text{Prob}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj} \quad \forall i \neq j) \\ = &\quad \text{Prob}(\epsilon_{nj} - \epsilon_{ni} < \quad V_{ni} - V_{nj} \quad \forall i \neq j) \\ = &\quad \int_{\epsilon} (\epsilon_{nj} - \epsilon_{ni} < \quad V_{ni} - V_{nj} \quad \forall i \neq j) f(\epsilon_n) d\epsilon_n \end{aligned} \tag{B.11}$$

Where $I(\cdot)$ is an indicator function that takes value 1 when the term is *true* and 0 otherwise. Under certain assumptions of the distribution of the error term ϵ_n, the integral expression in equation B.11 has a closed form or has a numerical solution.

In order to derive equation B.8, we had assumed ϵ_{nj} is distributed following a *Type-I* Extreme Value Distribution (also known as Gumbel distribution). The main reason behind the assumption, is the simplicity in the derivation of the closed form of the integral in equation B.11 under *iid Type-I Extreme Value* errors. Each error term follows then:

$$f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}}$$

and the cumulative distribution is:

$$F(\epsilon_{nj}) = e^{-e^{-\epsilon_{nj}}} \tag{B.12}$$

From B.11 we know that P_{ni} is the cumulative distribution for each ϵ_{nj} evaluated at $\epsilon_{ni} + V_{ni} - V_{nj}$. Since ϵ are iid, the cumulative distribution over all $j \neq i$ is the product of all individual cumulative distributions. Hence, from equation B.12:

$$P_{ni}|\epsilon_{ni} = \prod_{j \neq i} e^{-e^{-\epsilon_{ni}+V_{ni}-V_{nj}}} \tag{B.13}$$

and from equation B.11:

$$P_{ni} = \int_{-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-\epsilon_{ni}+V_{ni}-V_{nj}}} \right) e^{-\epsilon_{ni}} e^{-e^{-\epsilon_{ni}}} d\epsilon_{ni} \tag{B.14}$$

some algebraic transformations, using the exponent of e, yields:

$$\begin{aligned} P_{ni} = & \int_{-\infty}^{\infty} exp\left(-\sum_j -e^{-(\epsilon_{ni}+V_{ni}-V_{nj})} \right) e^{-\epsilon_{ni}} d\epsilon_{ni} \\ = & \int_{-\infty}^{\infty} exp\left(-e^{-\epsilon_{ni}} \sum_j -e^{-(V_{ni}-V_{nj})} \right) e^{-\epsilon_{ni}} d\epsilon_{ni} \end{aligned}$$

And, if we define $t = exp(-\epsilon)$, such that $-exp(-\epsilon)d\epsilon = dt$, we have that:

$$\begin{aligned} P_{ni} = & \int_{\infty}^{0} exp\left(-t \sum_j -e^{-(V_{ni}-V_{nj})} \right) (-dt) \\ = & \int_{0}^{\infty} exp\left(-t \sum_j -e^{-(V_{ni}-V_{nj})} \right) (dt) \\ = & \left. \frac{exp\left(-t\sum_j -e^{-(V_{ni}-V_{nj})}\right)}{-\sum_j -e^{-(V_{ni}-V_{nj})}} \right|_0^{\infty} \\ = & \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \end{aligned} \tag{B.15}$$

which corresponds to a *logit* choice probability as a function of the indirect utilities of the agent.

C Chapter 3: Appendix

C.1 Review Articles

Human embryonic stem cells. Pera, MF et al.; 2000. **Journal Of Cell Science**
The embryonic origins of human haematopoiesis. Marshall, CJ and Thrasher, AJ; 2001. **British Journal Of Haematology**
Multilineage differentiation from human embryonic stem cell lines. Odorico, JS et al.; 2001. **Stem Cells**
The derivation and potential use of human embryonic stem cells. Trounson, A.O.; 2001. **Reproduction Fertility And Development**
Derivation and potential applications of human embryonic stem cells. Gepstein, L.; 2002. **Circulation Research**
Derivation and spontaneous differentiation of human embryonic stem cells. Amit, M and Itskovitz-Eldor, J; 2002. **Journal Of Anatomy**
Characterization and differentiation of human embryonic stem cells. Carpenter, MK et al.; 2003. **Cloning And Stem Cells**
Human embryonic stem cells for cardiovascular repair. Nir, SG et al.; 2003. **Cardiovascular Research**
Derivation, propagation and differentiation of human embryonic stem cells. Conley, BJ et al.; 2004. **International Journal Of Biochemistry and Cell Biology**
The immunogenicity of human embryonic stem-derived cells. Drukker, M and Benvenisty, N; 2004. **Trends In Biotechnology**
Derivation, characterization, and differentiation of human embryonic stem cells. Heins, N et al.; 2004. **Stem Cells**
Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. Rao, M.; 2004. **Developmental Biology**
Immunogenicity of human embryonic stem cells: can we achieve tolerance?. Drukker, M.; 2004. **Springer Seminars In Immunopathology**
Derivation, growth and applications of human embryonic stem cells. Stojkovic, M et al.; 2004. **Reproduction**

Human embryonic stem cells: prospects for development. Pera, MF and Trounson, AO; 2004. **Development**

The promise of human embryonic stem cells. Gerecht-Nir, S and Itskovitz-Eldor, J; 2004. **Best Practice and Research In Clinical Obstetrics and Gynaecology**

Human embryonic stem cells as a model for early human development. Dvash, T and Benvenisty, N; 2004. **Best Practice and Research In Clinical Obstetrics and Gynaecology**

Characterization and culture of human embryonic stem cells. Hoffman, LM and Carpenter, MK; 2005. **Nature Biotechnology**

Differentiation pathways in human embryonic stem cell-derived cardiomyocytes. Lev, S et al.; 2005. **Communicative Cardiac Cell**

Culture development for human embryonic stem cell propagation: molecular aspects and challenges. Rao, BM and Zandstra, PW; 2005. **Current Opinion In Biotechnology**

Genetic manipulation of human embryonic stem cells: A system to study early human development and potential therapeutic applications. Menendez, P et al.; 2005. **Current Gene Therapy**

Hematopoietic development from human embryonic stem cell lines. Wang, L et al.; 2005. **Experimental Hematology**

Human embryonic stem cell-derived oligodendrocyte progenitors for the treatment of spinal cord injury. Faulkner, J and Keirstead, HS; 2005. **Transplant Immunology**

Differentiation of human embryonic stem cells after transplantation in immune-deficient mice. Przyborski, S. A. ; 2005. **Stem Cells**

Human blastocyst culture and derivation of embryonic stem cell lines. Bongso, Ariff and Tan, Shawna; 2005. **Stem Cell Reviews**

Cloned human embryonic stem cells for tissue repair and transplantation. Hwang, Woo Suk et al.; 2005. **Stem Cell Reviews**

A molecular basis for human embryonic stem cell pluripotency. Noggle, Scott A. et al.; 2005. **Stem Cell Reviews**

Human embryonic stem cells - An in vitro model to study mechanisms controlling pluripotency in early mammalian development. Vallier, Ludovic and Pedersen, Roger A.; 2005. **Stem Cell Reviews**

Human embryonic stem cell stability. Hoffman, Lisa M. and Carpenter, Melissa K.; 2005. **Stem Cell Reviews**

Manipulation of the human genome in human embryonic stem cells. Kopper, Oded and Benvenisty, Nissim; 2005. **Stem Cell Reviews**

Human embryonic stem cells - Potential tool for achieving immunotolerance?. Menendez, Pablo et al.; 2005. **Stem Cell Reviews**

Development and differentiation of neural rosettes derived from human embryonic stem cells. Wilson, Patricia G. and Stice, Steve S.; 2006. **Stem Cell Reviews**

Human embryonic stem cells: a journey beyond cell replacement therapies. Menendez, P. et al.; 2006. **Cytotherapy**

The production and directed differentiation of human embryonic stem cells. Trounson, A.; 2006. **Endocrine Reviews**

Evaluating human embryonic germ cells: Concord and conflict as pluripotent stem cells. Turnpenny, Lee et al.; 2006. **Stem Cells**

Hematopoiesis from human embryonic stem cells: Overcoming the immune barrier in stem cell therapies. Priddle, Helen et al.; 2006. **Stem Cells**

Concise review: Scientific and ethical roadblocks to human embryonic stem cell therapy. Gruen, Lori and Grabel, Laura; 2006. **Stem Cells**

Human embryonic stem cells as a powerful tool for studying human embryogenesis. Dvash, Tamar et al.; 2006. **Pediatric Research**

Differences between human embryonic stem cell lines. Allegrucci, C. and Young, L. E.; 2007. **Human Reproduction Update**

Strategies for preventing immunologic rejection of transplanted human embryonic stem cells. Cabrera, C. M. et al.; 2006. **Cytotherapy**

Human embryonic stem cells: Long term stability, absence of senescence and a potential cell source for neural replacement. Zeng, X. and Rao, M. S.; 2007. **Neuroscience**

The human embryonic stem cell-derived cardiomyocyte as a pharmacological model. Harding, Sian E. et al.; 2007. **Pharmacology and Therapeutics**

The regulation of self-renewal in human embryonic stem cells. Avery, Stuart et al.; 2006. **Stem Cells And Development**

Concise review: No breakthroughs for human mesenchymal and embryonic stem cell culture: Conditioned medium, feeder layer, or feeder-free; Medium with fetal calf serum, human serum, or enriched plasma; Serum-free, serum replacement nonconditioned medium, or ad hoc formula? All that glitters is not gold!. Mannello, Ferdinando and Tonti, Gaetana A.; 2007. **Stem Cells**

Xeno-free derivation and culture of human embryonic stem cells: current status, problems and challenges. Lei, Ting et al.; 2007. **Cell Research**

Human embryonic stem cells: Current technologies and emerging industrial applications. Ameen, Caroline et al.; 2008. **Critical Reviews In Oncology Hematology**

Immunogenicity of human embryonic stem cells. Grinnemo, Karl-Henrik et al.; 2008. **Cell And Tissue Research**

Critical issues of clinical human embryonic stem cell therapy for brain repair. Li, Jia-Yi et al.; 2008. **Trends In Neurosciences**

Differentiation of embryonic stem cells to clinically relevant populations: Lessons from embryonic development. Murry, Charles E. and Keller, Gordon; 2008. **Cell**

Genetic modification of human embryonic stem cells for derivation of target cells. Giudice, Antonietta and Trounson, Alan; 2008. **Cell Stem Cell**

Deconstructing human embryonic stem cell cultures: niche regulation of self-renewal and pluripotency. Stewart, Morag H. et al.; 2008. **Journal Of Molecular Medicine-Jmm**

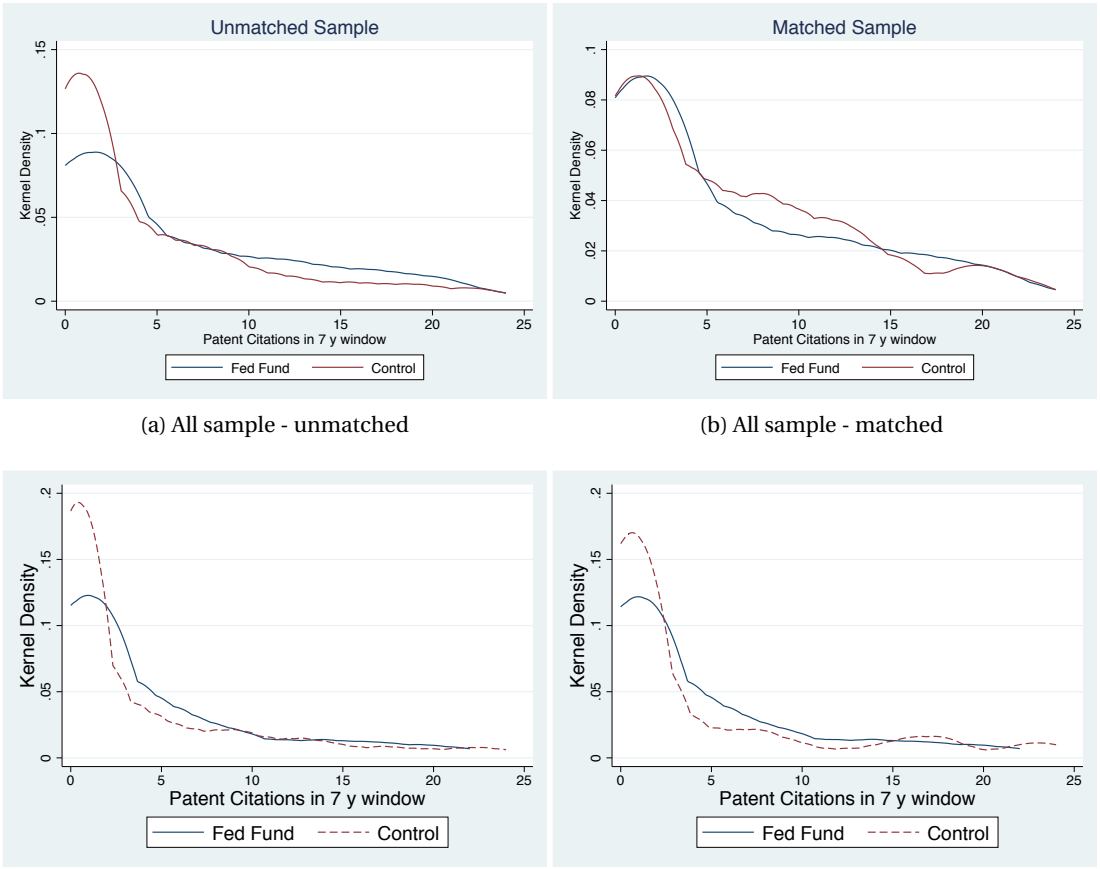
Human embryonic stem cells: origins, characteristics and potential for regenerative therapy. Mountford, J. C.; 2008. **Transfusion Medicine**

The tumorigenicity of human embryonic stem cells. Blum, Barak et al.; 2008. **Advances In Cancer Research, Vol 100**

Neural Differentiation of Human Embryonic Stem Cells. Dhara, Sujoy K. and Stice, Steven L.;

2008. **Journal Of Cellular Biochemistry**
Human embryonic stem cells for cardiomyogenesis. Habib, Manhal et al.; 2008. **Journal Of Molecular And Cellular Cardiology**
Human Embryonic Stem Cell Differentiation Toward Regional Specific Neural Precursors. Erceg, Slaven et al.; 2009. **Stem Cells**
Surface marker antigens in the characterization of human embryonic stem cells. Wright, Andrew J. and Andrews, Peter W.; 2009. **Stem Cell Research**
Challenges of Stem Cell Therapy for Spinal Cord Injury: Human Embryonic Stem Cells, Endogenous Neural Stem Cells, or Induced Pluripotent Stem Cells?. Ronaghi, Mohammad et al.; 2010. **Stem Cells**
The human sperm epigenome and its potential role in embryonic development. Carrell, Douglas T. and Hammoud, Saher Sue; 2010. **Molecular Human Reproduction**
Differentiation of Human Embryonic Stem Cells to Cardiomyocytes for In Vitro and In Vivo Applications. Vidarsson, Hilmar et al.; 2010. **Stem Cell Reviews And Reports**
Translational potential of human embryonic and induced pluripotent stem cells for myocardial repair: Insights from experimental models. Kong, Chi-Wing et al.; 2010. **Thrombosis And Haemostasis**
Human motor neuron generation from embryonic stem cells and induced pluripotent stem cells. Nizzardo, M. et al.; 2010. **Cellular And Molecular Life Sciences**
Human embryonic stem cells: Derivation, culture, and differentiation: A review. Vazin, Tandis and Freed, William J.; 2010. **Restorative Neurology And Neuroscience**
Cardiac regeneration using human embryonic stem cells: producing cells for future therapy. Wong, Sharon S. Y. and Bernstein, Harold S.; 2010. **Regenerative Medicine**
Potential of Human Embryonic Stem Cells in Cartilage Tissue Engineering and Regenerative Medicine. Toh, Wei Seong et al.; 2011. **Stem Cell Reviews And Reports**
New Approaches in the Differentiation of Human Embryonic Stem Cells and Induced Pluripotent Stem Cells toward Hepatocytes. Behbahan, Iman Saramipoor et al.; 2011. **Stem Cell Reviews And Reports**
Biomaterials for the Feeder-Free Culture of Human Embryonic Stem Cells and Induced Pluripotent Stem Cells. Higuchi, Akon et al.; 2011. **Chemical Reviews**
The tumorigenicity of human embryonic and induced pluripotent stem cells. Ben-David, Uri and Benvenisty, Nissim; 2011. **Nature Reviews Cancer**
Differentiation of Human Embryonic Stem Cells and Induced Pluripotent Stem Cells to Cardiomyocytes A Methods Overview. Mummery, Christine L. et al.; 2012. **Circulation Research**

C.2 Matching



(a) All sample - unmatched

(b) All sample - matched

(c) Pre-ban - unmatched

(d) Pre-ban - matched

Figure C.1 – **CEM Matching Densities:** Comparison of Matched and Unmatched densities for Patent Citations to root articles in a 7-year window. For the full sample, CEM-matching improves the L_1 imbalance metric from 0.6828 down to 0.3354. There are just 22 unmatched samples out of 244 treated elements. For the pre-ban, the L_1 imbalance metric from 0.7056 down to 0.3871.

C.3 Model Selection

The following tables and figures support the choice of a Negative Binomial Regression instead of a Poisson Model Regression in Chapter 3.

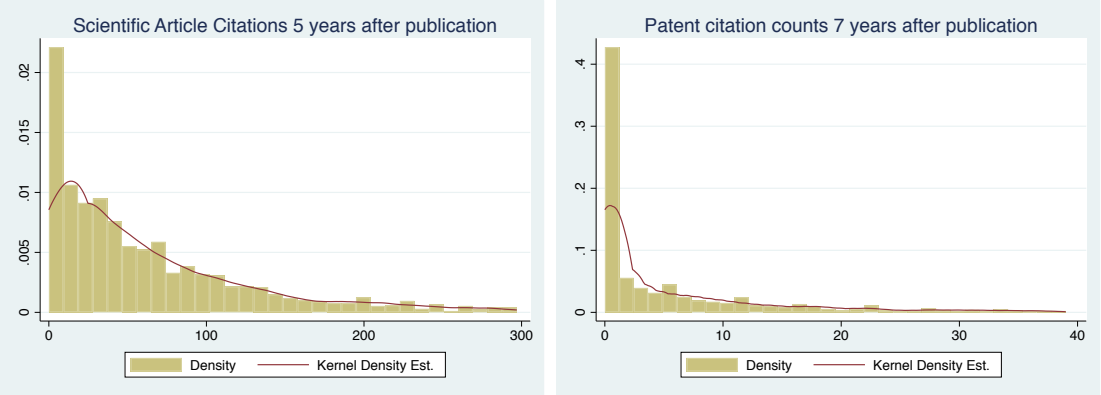


Figure C.2 – **Density Plots for Citation Counts** showing overdispersion of the count variables.

Table C.1 – Count Model Statistics

	SciCit (1)	PatCit7y (2)	PatPatCit7_5y (3)
	Poisson model statistics DV = Citations from Articles (1) and Patents (2,3)		
AIC	50571.0	18733.4	61755.4
BIC	50660.7	18823.1	61845.0
Log-likelihood	−25265.5	−9346.7	−30857.7
Chi2	251.7	108.2	246.5
Pearson	82617.6	27510.0	142723.5
Deviance	46900.7	16507.9	59575.1
	Zero Inflated Poisson model statistics DV = Citations from Articles (1) and Patents (2,3)		
AIC	37475.31	15486.07	43208.29
BIC	37578.45	15589.22	43311.43
Log-likelihood	−18714.65	−7720.034	−21581.14
chi2	235.67	108.33	238.66
	Negative Binomial model statistics DV = Citations from Articles (1) and Patents (2,3)		
AIC	7219.6	4711.9	5065.3
BIC	7309.3	4801.6	5155.0
Log-likelihood	−3589.8	−2335.9	−2512.7
chi2	260.7	115.3	130.7
Pearson	615.7	713.9	502.4
Deviance	796.0	752.6	699.2
Alpha	1.171***	1.916***	4.075***
	(0.217)	(0.179)	(0.431)
Observations	656	656	656

This table presents the model statistics for Poisson, Negative Binomial and Zero Inflated Poisson models, in order to determine the best fit to the data. The dependent variables are Citation Counts from Scientific Articles in a 5 year window after publication; from Patents in a 7 year window; and from 2nd degree Patents in a 7+5 year window. The models fitted correspond to that presented in 3.4

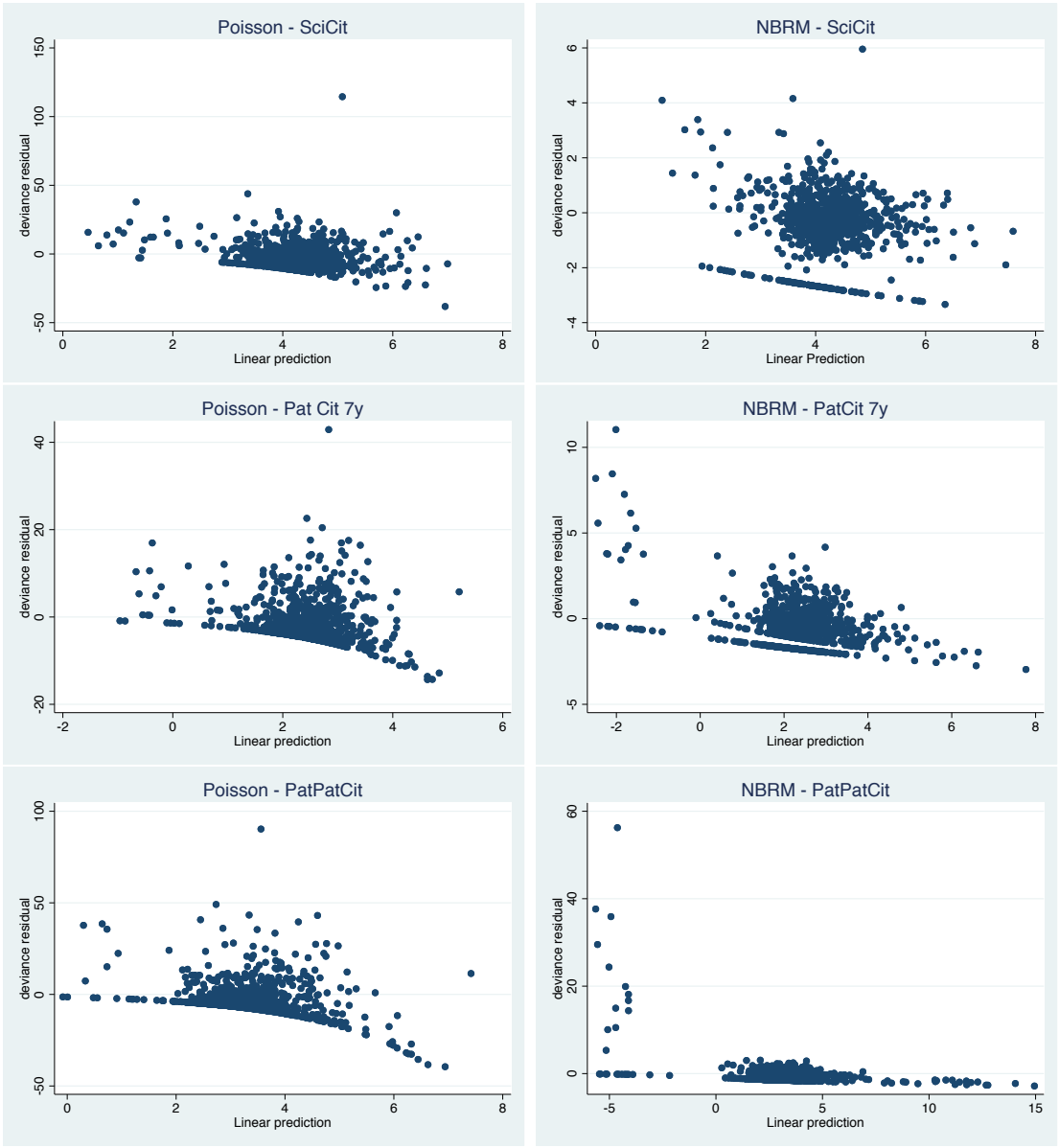


Figure C.3 – **Residual Deviance** plots for Poisson Regression Model (PRM) and Negative Binomial Regression Model (NBRM)

C.4 Full Tables from Chapter 3

Table C.2 – Full Table 3.4. NB, hESC=1

	Patent Citations		2nd Degree Pat Cit	
Federal Funding	1.579***	2.805***	3.859***	7.253***
	(0.483)	(0.674)	(0.766)	(1.295)
Ban	2.462***	3.746***	4.075***	6.984***
	(0.379)	(0.584)	(0.604)	(1.122)
Federal Funding × Ban	−2.070***	−3.229***	−4.368***	−7.588***
	(0.524)	(0.688)	(0.801)	(1.283)
1997	−0.403	−0.355	0.207	−0.0496
	(0.612)	(0.555)	(0.722)	(0.704)
1998	4.208***	4.534***	7.097***	8.622***
	(0.669)	(0.786)	(0.831)	(1.146)
1999	1.567***	2.466***	3.747***	6.021***
	(0.568)	(0.635)	(1.049)	(1.358)
2000	1.790***	3.887***	3.927***	8.089***
	(0.449)	(0.694)	(0.731)	(1.267)
2001	0.113	0.110	0.396	0.622
	(0.599)	(0.554)	(0.618)	(0.558)
2002	0.145	−0.257	0.685*	0.129
	(0.344)	(0.342)	(0.405)	(0.334)
2003	0.118	0.353	0.257	0.724**
	(0.337)	(0.287)	(0.362)	(0.304)
2004	−0.00666	−0.0198	0.0814	0.183
	(0.344)	(0.270)	(0.381)	(0.296)
2005	−0.181	0.0504	−0.318	0.0431
	(0.313)	(0.260)	(0.354)	(0.265)
2006	(.)	(.)	(.)	(.)
Number of Authors		0.0549*		0.0892**
		(0.0315)		(0.0369)
JIF		0.0622***		0.115***
		(0.0181)		(0.0271)
Reprint Author in USA		0.499		0.568
		(0.307)		(0.370)
At least one Author in USA		0.384		0.387
		(0.381)		(0.456)
CoAuthor from fav. country		0.448**		0.459*
		(0.213)		(0.267)
International collab.		0.0973		0.318
		(0.246)		(0.333)
Corporate		0.469**		0.804***
		(0.223)		(0.254)
Last Auth Age		0.000739		0.00517
		(0.00864)		(0.0106)
Reprint Auth Age		−0.0256**		−0.0274**
		(0.01000)		(0.0132)
Constant	0.568**	−2.376***	−0.235	−5.806***
	(0.266)	(0.613)	(0.546)	(1.235)
$log(\alpha)$	0.751***	0.636***	1.504***	1.397***
	(0.0950)	(0.0939)	(0.103)	(0.107)
Observations	655	655	655	655
Log-Likelihood	−2369.8	−2331.1	−2545.2	−2510.2

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.3 – Full Table 3.5. NB-Reg Sim-hESC>0.35

	NB Reg		Unit-offset-log (OLS)	
	Pat Cit	2nd Pat. Cit.	log(PatCit)	log(PatPatCit)
Federal Funding	0.537	0.860	0.210	0.456
	(0.483)	(0.578)	(0.444)	(0.692)
Ban	1.131**	0.0751	0.776	0.0438
	(0.568)	(0.628)	(0.661)	(1.181)
Federal Funding × Ban	−1.058**	−1.312**	−0.818*	−1.293*
	(0.497)	(0.580)	(0.473)	(0.726)
1997	−0.191	−0.157	0.0490	−0.0621
	(0.683)	(0.747)	(0.801)	(1.250)
1998	1.475**	1.820**	0.555	0.645
	(0.698)	(0.789)	(0.845)	(1.359)
1999	−0.0154	−0.781	0.233	−0.654
	(0.561)	(0.818)	(0.676)	(1.160)
2000	1.197**	0.541	1.280**	1.397
	(0.533)	(0.562)	(0.638)	(1.031)
2001	0.796**	1.486***	0.692	1.087*
	(0.341)	(0.407)	(0.494)	(0.642)
2002	−0.0476	0.260	0.373	0.944**
	(0.275)	(0.304)	(0.289)	(0.386)
2003	0.615**	1.082***	0.606**	0.904**
	(0.284)	(0.287)	(0.281)	(0.372)
2004	0.349	0.678**	0.298	0.553
	(0.268)	(0.306)	(0.316)	(0.429)
2005	0.321	0.484*	0.247	0.205
	(0.238)	(0.264)	(0.256)	(0.340)
2006	(.)	(.)	(.)	(.)
Number of Authors	0.0440	0.0625**	0.0647**	0.0884**
	(0.0270)	(0.0317)	(0.0256)	(0.0381)
JIF	0.0733***	0.126***	0.0378**	0.0331
	(0.0227)	(0.0305)	(0.0172)	(0.0279)
Reprint Author in USA	0.466	0.556	0.644**	0.779**
	(0.303)	(0.367)	(0.258)	(0.356)
At least one Author in USA	0.512	0.604	0.490	0.819*
	(0.373)	(0.447)	(0.337)	(0.447)
CoAuthor from fav. country	0.373*	0.360	0.330	0.654**
	(0.202)	(0.243)	(0.208)	(0.292)
International collab.	0.124	0.525	0.0697	−0.179
	(0.215)	(0.320)	(0.241)	(0.356)
Corporate	0.544***	0.958***	0.452**	0.660**
	(0.210)	(0.245)	(0.223)	(0.293)
Last Auth Age	0.00631	0.0123	0.00685	0.0132
	(0.00829)	(0.00896)	(0.00943)	(0.0131)
Reprint Auth Age	−0.0268***	−0.0319***	−0.0227***	−0.0267**
	(0.00750)	(0.00977)	(0.00841)	(0.0122)
Constant	−0.0814	0.642	−0.454	−0.268
	(0.577)	(0.701)	(0.652)	(1.188)
$log(\alpha)$	0.711***	1.376***		
	(0.106)	(0.105)		
Observations	677	677	677	677
Adjusted R^2			0.177	0.196
Log-likelihod	−2486.2	−2753.5	−1153.4	−1367.2

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.4 – Unit-offset Log OLS Treatment Federal Funding, hESC=1

	Patent Citations		2nd Degree Pat Cit	
Federal Funding	0.346	0.487	0.800	0.900
	(0.431)	(0.454)	(0.686)	(0.717)
Ban	1.226***	2.094***	1.789***	2.595***
	(0.282)	(0.375)	(0.465)	(0.763)
Federal Funding × Ban	−0.836*	−1.081**	−1.379*	−1.609**
	(0.464)	(0.475)	(0.731)	(0.759)
1997	−0.154	0.460	0.492	1.246
	(0.511)	(0.510)	(0.808)	(0.814)
1998	1.926	2.106*	3.418*	3.545**
	(1.345)	(1.180)	(1.860)	(1.694)
1999	0.845*	1.840***	0.952	1.992**
	(0.470)	(0.411)	(0.982)	(1.007)
2000	0.799**	2.089***	1.774***	3.198***
	(0.385)	(0.599)	(0.607)	(0.925)
2001	−0.454	−0.491	−0.327	−0.352
	(0.725)	(0.687)	(0.906)	(0.848)
2002	0.343	−0.00122	0.878	0.462
	(0.372)	(0.329)	(0.578)	(0.487)
2003	0.109	0.249	0.130	0.337
	(0.276)	(0.271)	(0.436)	(0.406)
2004	0.0375	−0.0322	−0.0183	−0.0951
	(0.308)	(0.253)	(0.487)	(0.417)
2005	−0.246	−0.0842	−0.411	−0.230
	(0.247)	(0.235)	(0.403)	(0.355)
2006	(.)	(.)	(.)	(.)
Number of Authors		0.0823***		0.112***
		(0.0258)		(0.0426)
JIF		0.0360***		0.0363
		(0.0122)		(0.0222)
Reprint Author in USA		0.693***		0.832**
		(0.252)		(0.344)
At least one Author in USA		0.634**		0.946**
		(0.297)		(0.424)
CoAuthor from fav. country		0.647***		0.992***
		(0.200)		(0.287)
International collab.		−0.150		−0.364
		(0.224)		(0.353)
Corporate		0.0677		0.278
		(0.212)		(0.304)
Last Auth Age		−0.00401		0.00514
		(0.00791)		(0.0139)
Reprint Auth Age		−0.0172**		−0.0131
		(0.00815)		(0.0121)
Constant	1.000***	−1.557***	0.636**	−2.702***
	(0.189)	(0.416)	(0.271)	(0.820)
Observations	655	655	655	655
Adjusted R^2	0.061	0.195	0.072	0.185
Log-Likelihood	−1133.3	−1078.0	−1356.5	−1309.3

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.5 – Full Table 3.6 (part 1); NBReg hESC=1

	Research Institute Patents		Private Sector Patents	
	Non-US	US	Non-US	US
Federal Funding	0.141	1.834***	3.947***	2.373***
	(1.606)	(0.646)	(1.223)	(0.802)
Ban	7.564***	3.593***	3.213***	3.510***
	(2.492)	(0.675)	(1.047)	(0.836)
Federal Funding × Ban	−0.734	−2.142***	−4.507***	−2.837***
	(1.609)	(0.654)	(1.212)	(0.822)
1997	4.262	0.995	−21.47***	0.246
	(3.327)	(0.766)	(0.785)	(0.681)
1998	6.571***	4.664***	4.838***	4.377***
	(2.407)	(0.867)	(1.075)	(1.011)
1999	6.292**	2.584***	−0.201	2.715***
	(2.700)	(0.572)	(0.704)	(0.877)
2000	7.615***	3.055***	3.934***	3.694***
	(2.690)	(0.891)	(1.288)	(0.859)
2001	−0.518	0.205	1.026	−0.0904
	(0.560)	(0.604)	(0.676)	(0.592)
2002	−1.368***	0.343	0.428	−0.616
	(0.502)	(0.537)	(0.417)	(0.394)
2003	0.0875	0.679*	0.930**	0.175
	(0.370)	(0.354)	(0.407)	(0.395)
2004	−0.510	0.0304	0.216	−0.148
	(0.360)	(0.328)	(0.464)	(0.372)
2005	−0.175	0.0980	0.263	−0.0270
	(0.301)	(0.322)	(0.388)	(0.350)
2006	(.)	(.)	(.)	(.)
Number of Authors	0.0678	0.0374	0.0830*	0.0587
	(0.0437)	(0.0351)	(0.0500)	(0.0392)
JIF	0.137***	0.0750***	0.0814***	0.0297
	(0.0361)	(0.0166)	(0.0282)	(0.0258)
Reprint Author in USA	0.930**	0.430	0.102	0.417
	(0.385)	(0.377)	(0.376)	(0.354)
At least one Author in USA	0.169	0.189	0.681	0.443
	(0.433)	(0.462)	(0.485)	(0.473)
CoAuthor from fav. country	0.505*	0.0300	0.494*	0.501**
	(0.292)	(0.301)	(0.293)	(0.231)
International collab.	−0.312	0.0484	−0.0187	0.287
	(0.291)	(0.252)	(0.408)	(0.290)
Corporate	0.153	0.298	0.635**	0.617**
	(0.295)	(0.255)	(0.310)	(0.266)
LastAuthAge	0.00732	0.00630	0.0135	−0.00500
	(0.0109)	(0.0118)	(0.0164)	(0.0121)
CorrAuthAge	−0.0294*	−0.0358***	−0.0280	−0.0247*
	(0.0167)	(0.0124)	(0.0200)	(0.0127)
Constant	−9.021***	−3.939***	−4.794***	−2.573***
	(2.718)	(0.767)	(1.196)	(0.787)
$log(\alpha)$	0.765***	0.901***	1.226***	1.160***
	(0.220)	(0.166)	(0.149)	(0.126)
Observations	655	655	655	655
Log-Likelihood	−894.6	−1218.3	−1040.4	−1777.2

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.6 – 3.6 (part 2); NBReg hESC=1

	Research Institute Patents		Private Sector Patents	
	Non-US	US	Non-US	US
Federal Funding	−0.395	0.0160	−0.275	−0.131
	(0.270)	(0.244)	(0.349)	(0.271)
Ban Short(2001-2003)	8.042***	3.639***	0.978	1.970***
	(2.698)	(0.877)	(0.960)	(0.751)
Federal Funding × (2001-2003)	−0.854*	−0.884*	−0.394	−0.527
	(0.513)	(0.465)	(0.498)	(0.448)
1997	4.847*	1.867*	−17.72***	0.892
	(2.657)	(1.009)	(0.934)	(0.916)
1998	6.846***	4.017***	1.958*	2.853***
	(2.461)	(0.983)	(1.182)	(1.055)
1999	6.631***	2.860***	0.249	1.747**
	(2.523)	(0.920)	(1.146)	(0.860)
2000	7.890***	2.536**	0.651	2.084***
	(2.741)	(1.021)	(1.007)	(0.682)
2001	−0.684	−0.598	0.0557	−0.315
	(0.604)	(0.585)	(0.659)	(0.577)
2002	−1.509***	−0.409	−0.464	−0.789**
	(0.531)	(0.483)	(0.440)	(0.377)
2003	(.)	(.)	(.)	(.)
2004	7.164***	2.642***	0.144	1.449**
	(2.676)	(0.867)	(0.917)	(0.659)
2005	7.541***	2.728***	0.198	1.529**
	(2.676)	(0.868)	(0.887)	(0.683)
2006	7.660***	2.661***	0.0123	1.641**
	(2.634)	(0.869)	(0.905)	(0.749)
Number of Authors	0.0704	0.0407	0.0948*	0.0558
	(0.0432)	(0.0349)	(0.0569)	(0.0419)
JIF	0.142***	0.0637***	0.0397	0.0118
	(0.0370)	(0.0187)	(0.0395)	(0.0357)
Reprint Author in USA	0.848**	0.358	0.205	0.405
	(0.368)	(0.376)	(0.383)	(0.351)
At least one Author in USA	0.222	0.177	0.526	0.379
	(0.421)	(0.471)	(0.537)	(0.510)
CoAuthor from fav. country	0.470	−0.00334	0.523	0.543**
	(0.297)	(0.302)	(0.335)	(0.247)
International collab.	−0.368	0.0635	0.0807	0.218
	(0.284)	(0.265)	(0.431)	(0.313)
Corporate	0.138	0.300	0.560	0.617**
	(0.296)	(0.263)	(0.343)	(0.285)
LastAuthAge	0.00981	0.00931	0.0116	−0.00154
	(0.0110)	(0.0124)	(0.0171)	(0.0129)
CorrAuthAge	−0.0306*	−0.0380***	−0.0297	−0.0251*
	(0.0163)	(0.0119)	(0.0199)	(0.0131)
Constant	−9.251***	−3.024***	−1.354	−0.640
	(2.859)	(0.986)	(0.927)	(0.662)
$log(\alpha)$	0.744***	0.904***	1.296***	1.192***
	(0.210)	(0.167)	(0.139)	(0.125)
Observations	655	655	655	655
Log-Likelihood	−893.1	−1222.3	−1055.4	−1787.0

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.7 – OLS - Topical Variety (spread); Treat=Fed Fund., hESC=1

	STD	Variation		
	(1)	(2)	(3)	(4)
Fed Fund	0.0643***	0.0814***	0.0826***	0.0876***
	(0.00662)	(0.0193)	(0.0196)	(0.0221)
Ban	−0.0706***	−0.190***	−0.191***	−0.197***
	(0.00848)	(0.0249)	(0.0249)	(0.0233)
Fed Fund × Ban	−0.0425***	−0.0420**	−0.0427**	−0.0470**
	(0.00670)	(0.0198)	(0.0198)	(0.0224)
CoAuthor fav. country	−0.00145		−0.00818*	0.0136
	(0.000973)		(0.00474)	(0.0186)
Fed Fund × CoAuthor fav. country				−0.157***
				(0.0313)
Ban × CoAuthor fav. country				−0.0258
				(0.0194)
Fed Fund × Ban × CoAuthor fav. country				0.170***
				(0.0326)
1997	−0.00542	−0.0355	−0.0375	−0.0490**
	(0.00816)	(0.0257)	(0.0258)	(0.0248)
1998	0.00275	−0.00900	−0.0102	−0.0253
	(0.0108)	(0.0279)	(0.0279)	(0.0281)
1999	−0.0138*	0.00837	0.00818	−0.00103
	(0.00821)	(0.0273)	(0.0275)	(0.0257)
2000	−0.00952	−0.0332	−0.0347	−0.0477
	(0.00909)	(0.0315)	(0.0317)	(0.0312)
2001	0.0561***	−0.0705***	−0.0714***	0.125***
	(0.00665)	(0.0271)	(0.0272)	(0.0147)
2002	0.0356***	0.104***	0.104***	0.104***
	(0.00178)	(0.0113)	(0.0111)	(0.0111)
2003	0.0150***	0.0418***	0.0417***	0.0421***
	(0.00116)	(0.00616)	(0.00608)	(0.00609)
2004	0.00104	0.00519	0.00527	0.00529
	(0.000994)	(0.00487)	(0.00483)	(0.00484)
2005	−0.00291***	−0.00566	−0.00621	−0.00619
	(0.000898)	(0.00458)	(0.00458)	(0.00458)
2006	(.)	(.)	(.)	(.)
Number of Authors	−0.0000129		0.000702	0.000620
	(0.000128)		(0.000549)	(0.000545)
At least one Author in USA	−0.000512		−0.000761	−0.00427
	(0.00154)		(0.00738)	(0.00564)
International collab.	0.000544		0.000984	0.00248
	(0.00138)		(0.00584)	(0.00548)
Corporate	−0.000191		0.00492	0.00458
	(0.000916)		(0.00421)	(0.00428)
Constant	0.168***	0.459***	0.459***	0.466***
	(0.00842)	(0.0246)	(0.0249)	(0.0232)
Observations	806	806	806	806
Adjusted R^2	0.833	0.614	0.613	0.617
Log-Likelihood	2275.5	1141.3	1144.0	1149.3

Standard errors in parentheses adjusted for heteroskedasticity
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.5 Marginal Effects on Academic Publications

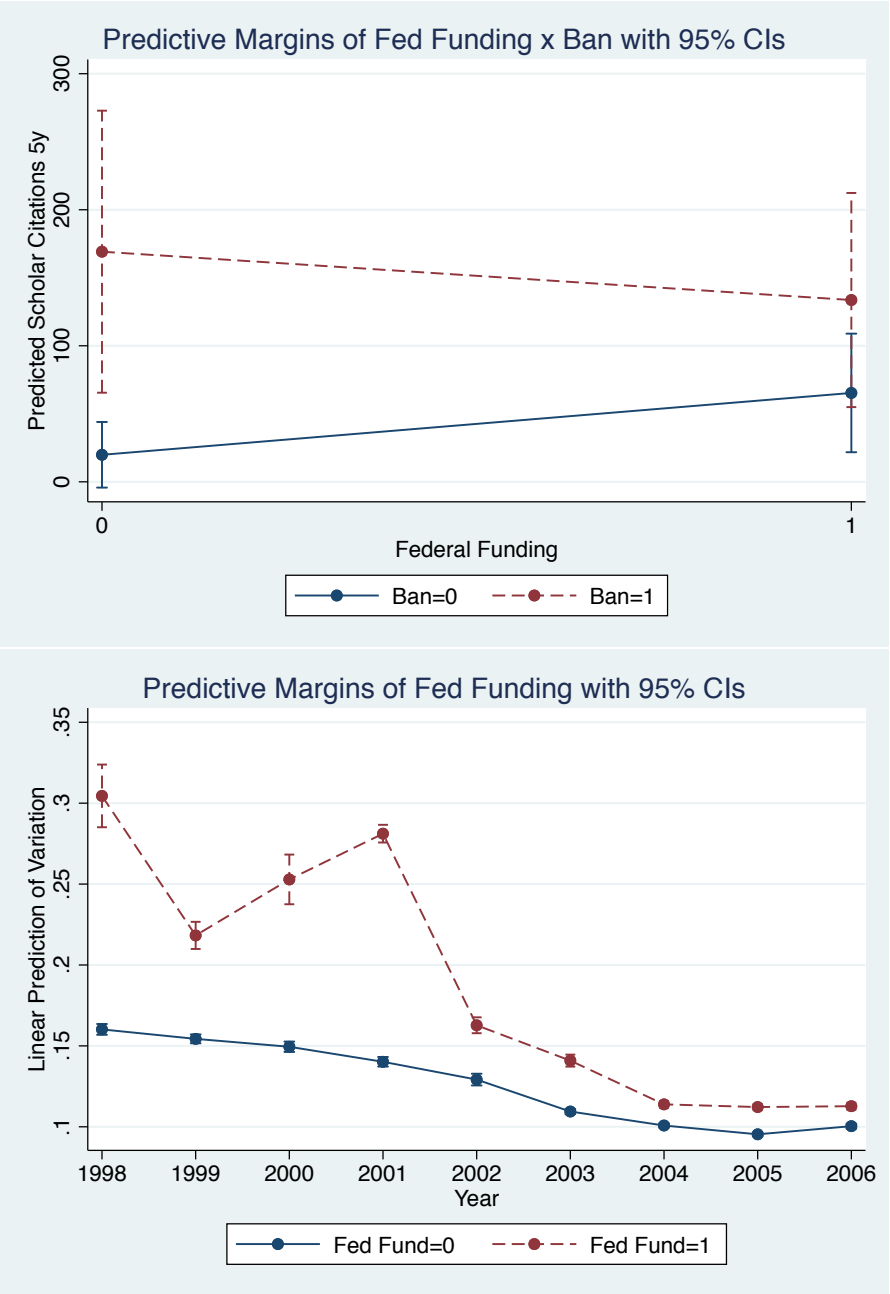


Figure C.4 – **Marginal effects** of Scholar Citations and Variation within hESC publications

C.6 Extending the analysis: close substitutes

The evidence so far has pointed towards increased academic freedom as a source of higher impact. We have, however, only examined publications that are hESC-related, with the treatment

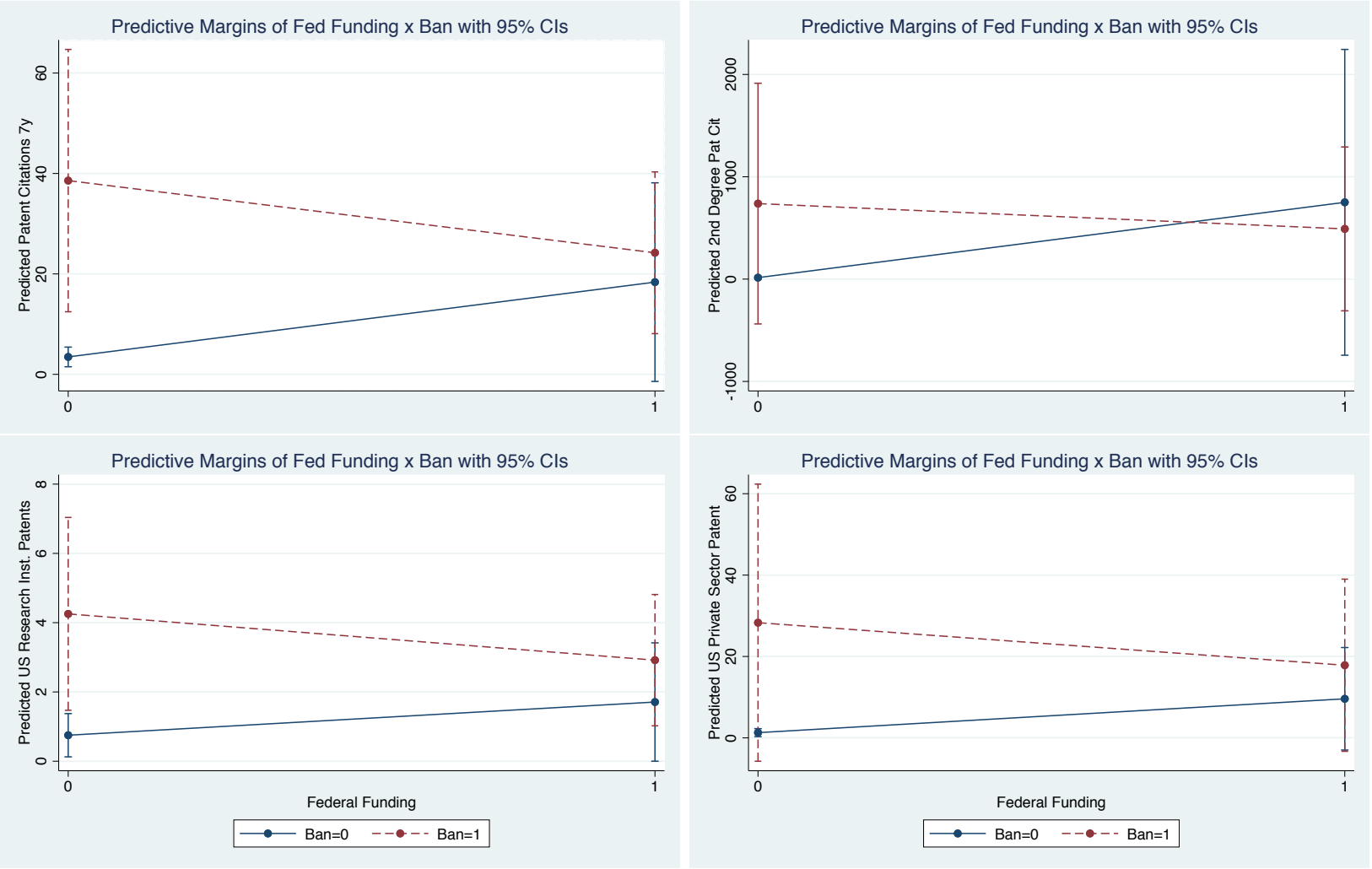


Figure C.5 – **Marginal effects of Federal Funding × Ban**: Marginal effects of the interaction terms on Patent Citations, 2nd degree Patent Citations, Research Institute Patent Citations and Private-sector Patent Citations

focus on the source of the grants. In this section, we attempt to uncover trends in areas that are close substitutes. For this, we include the whole sample of 1885 articles and re-split the sample for three different analysis. Because there is no presumption that researchers *actually* moved from one field to another, nor we cover the individual affiliations, it should be clear that this section only documents correlations and not causal effects. However, we believe these are interesting trends that might shed new light on the events, are encourage future work.

The prior art (econometric analysis) that has studied the effects of the ban (Furman et al., 2012; Huang and Jong, 2019) has used, in both cases a very similar model. The authors explained the effect of the ban on US science and firms using a regression in the lines of:

$$\begin{aligned} \text{CITES}^r_{it} = \quad & \epsilon_{it} + \gamma_i + \beta_t + \text{age}^r_{it} + \\ & \alpha_0(\text{hESC}_i \cdot 2001_{it}) + \alpha_1(\text{hESC}_i \cdot (t > 2001_{it})) + \\ & \phi_0(\text{US}^r_t \cdot \text{hESC}_i \cdot 2001_{it}) + \phi_1(\text{US}^r_t \cdot \text{hESC}_i \cdot (t > 2001_{it})) \end{aligned} \tag{C.1}$$

where *CITES* is a per-year count of citations (or projects) to focal publications from two different stacks: US and non-US; hESC is a dummy for hESC-related articles; 2001 and > 2001 are time dummies that cover the specified periods; age represents the years since the publication of the focal article; and γ are publication fixed effects. Therefore, ϕ_0 and ϕ_1 are the coefficients of interest, indicating the marginal impact of the policy intervention on US citations or projects.

We imitate their analyses but using Patent Citations. For this, we drop all publications from 2001 onward, and we count arrival of (patent) citations yearly and until 2009. Following the literature, we split the post-ban effect into three time-periods to examine the average treatment effect along time. Table C.8 presents the results. Without going too deep in the analysis, the evidence points in the same direction as previous work. The drop in citations (and overall industry reaction) is slightly delayed, drops a few years into the ban, and recovers to previous levels in the long run. There is, however, a surprising trend. The average treatment effect immediately after the ban, represented by the coefficient USA×hESC×(2002-2003) is led by patent citations from non-profit research institutions, while the private sector leads the recovery (in the long run) represented by USA×hESC×(2006-2009).

One possibility is that, in the aftermath of the ban, researchers turned their efforts into patenting. The reasons behind this could be multiple. On the one hand, limited access to funds may have encouraged patenting activity in order to attract private funds. While the moratorium introduced uncertainty in the hESC landscape, it was acknowledged as an exciting area of research with a huge prospect for private capital. On the other hand, the limitation in the direction and tools might have encouraged more *applied research* — i.e. closer to patents — with the existing cell lines. Discussions with researchers active in the field during the ban have highlighted a *rush* for patenting in order to get priority. Many of the resulting patents, however, were unfruitful due to the low maturity of the field in general. This trend could have multiplied the patenting behaviour post-ban while slowing down research productivity.

Table C.8 – Patent citation flows to pre-2001 scientific articles

	Conditional fixed effects negative binomial, stacked DV = Patent Cites (Granted USPTO)		
	Private Sector	Research Inst.	All Patents
	(1)	(2)	(3)
hESC×2001	0.183	−0.299	−0.359
	(0.281)	(0.460)	(0.248)
hESC×(2002-2003)	1.084***	−0.142	0.483*
	(0.238)	(0.373)	(0.214)
hESC×(2004-2005)	0.177	1.007*	0.356
	(0.309)	(0.411)	(0.234)
hESC×(2006-2009)	−0.102	−0.626	−0.500*
	(0.295)	(0.414)	(0.244)
USA×hESC×2001	0.783***	0.120	0.704***
	(0.212)	(0.468)	(0.191)
USA×hESC×(2002-2003)	0.570***	1.810***	0.835***
	(0.118)	(0.258)	(0.108)
USA×hESC×(2004-2005)	0.0618	0.00625	0.0503
	(0.186)	(0.225)	(0.145)
USA×hESC×(2006-2009)	1.197***	0.799***	1.046***
	(0.146)	(0.176)	(0.112)
CoAuthor Private Affil	−0.342	0.458	−0.0737
	(0.197)	(0.265)	(0.196)
	(0.301)	(0.516)	(0.249)
N	1730	1730	1730

Standard errors in parentheses adjusted for heteroskedasticity

Models include constant, hESC*Year FEs, article age FEs and article FEs.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C.8 points towards patents (and applied research) might be the close-substitute activities that researchers endured. To further examine this possibility, we incorporate contemporaneous publications. We estimate the following model on the non-federally-funded subsample:

$$E\left[y_{it}|X_{it}\right] = f\left[\epsilon_{it}; \quad \beta_0 + \beta_1 \text{BAN}_t + \beta_2 \text{hESC}_i + \beta_3 \text{hESC}_i \times \text{BAN}_t + \delta_t\right] \tag{C.2}$$

The estimation results are displayed on Table C.9, and point in the same direction. While hESC-related research significantly increased the patent citations on average, the effect is not significant for citations from US patents from research institutes. However, it is positive and significant on average during the ban. That is, during the ban, non federally-funded hESC-related publications received more attention from research institute patents compared to non-hESC (and also non-federally-funded) publications. This result provides insinuating evidence that they were more closely linked with the innovations that cite them. Finally, we compare

Table C.9 – NB Regression Patent by Origin; Fed Fund=0

	Research Institute		Private Sector	
	Non-US	US	Non-US	US
hESC	0.868***	0.150	0.635**	0.801***
	(0.322)	(0.334)	(0.262)	(0.272)
Ban	1.831**	−0.380	0.718	0.305
	(0.798)	(0.449)	(0.519)	(0.485)
hESC × Ban	0.262	0.652*	−0.112	0.266
	(0.380)	(0.387)	(0.328)	(0.332)
Year FE	*Yes*	*Yes*	*Yes*	*Yes*
Article Controls	*Yes*	*Yes*	*Yes*	*Yes*
$log(\alpha)$	1.494***	1.581***	1.549***	1.841***
	(0.119)	(0.0845)	(0.0897)	(0.0631)
Observations	1212	1212	1212	1212
Adjusted R^2				
Log-Likelihood	−1045.5	−1513.4	−1549.0	−2187.8

Standard errors in parentheses adjusted for heteroskedasticity

Including unreported constant and controls for JIF, USA author, Private-sector affiliated Author, Author from Favorable Country

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the interaction term *hESC*×*Ban* for research institute patent counts by the source of funding and by counterfactual similarity. In order to do so, we increasingly limit the sample by the similarity to hESC value (the variable we used for the robustness check, represented in Figure 3.3). This way, the average treatment effect compares to a counterfactual that is increasingly similar to hESC-research, and thus, closer to being an outside option for researchers. Figure C.6 shows the coefficient of the interaction term with the 95% confidence interval. The effect is driven by non-federally funded research, and it increases as the counterfactual becomes

more confined.

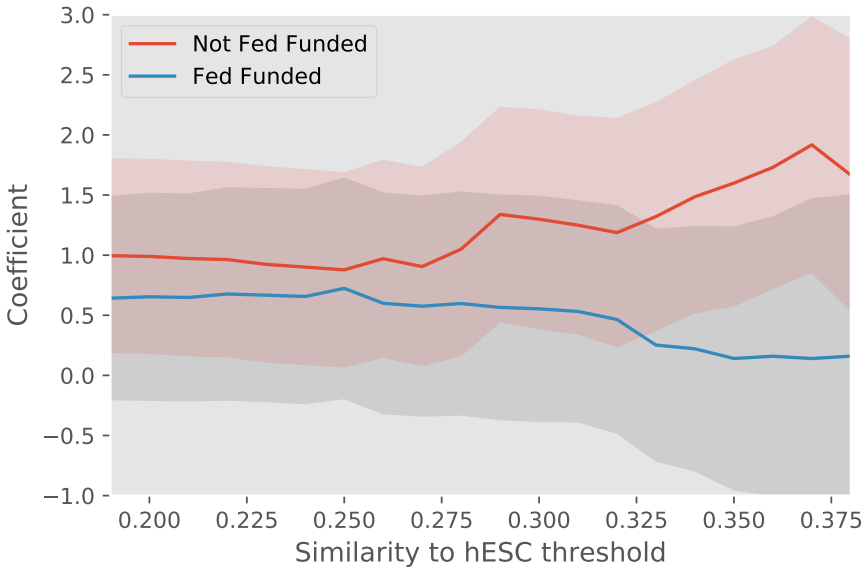


Figure C.6 – **Close substitutes analysis**: Not significantly different

Bibliography

ACI (2020). Cancer facts and figures 2020. *American Cancer Society*.

Aghion, P., Dewatripont, M., and Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics*, 39(3):617–635.

Aghion, P. and Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, 60(2):323–51.

Agrawal, A., Fu, W., and Menzies, T. (2018). What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, 98:74 – 88.

Agrawal, A., McHale, J., and Oettl, A. (2017). How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy*, 46(4):853–867.

Aguirregabiria, V. and Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38 – 67. Structural Models of Optimization Behavior in Labor, Aging, and Health.

Ahmadpoor, M. and Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587.

Ai, Q., Yang, L., Guo, J., and Croft, W. B. (2016). Analysis of the paragraph vector model for information retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 133–142, New York, NY, USA. Association for Computing Machinery.

Amemiya, T. (1990). *Discrete Choice Models*, pages 58–69. Palgrave Macmillan UK, London.

Apesteguia, J. and Maier-Rigaud, F. P. (2006). The role of rivalry: Public goods versus common-pool resources. *Journal of Conflict Resolution*, 50(5):646–663.

Arora, A., Belenzon, S., and Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, 39(1):3–32.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pages 609–626. National Bureau of Economic Research, Inc.

Ayoubi, C., Barbosu, S., Pezzoni, M., and Visentin, F. (2020). What matters in funding: The value of research coherence and alignment in evaluators' decisions. MERIT Working Papers 010, United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT).

Ayoubi, C., Pezzoni, M., and Visentin, F. (2019). The important thing is not to win, it is to take part: What if scientists benefit from participating in research grant competitions? *Research Policy*, 48(1):84 – 97.

Azoulay, P., Doran, K., and MacGarvie, M. (2018). Best practices for funding early careers of scientists: Evidence and unanswered questions. (21579).

Azoulay, P., Fons-Rosen, C., and Zivin, J. S. G. (2015a). Does science advance one funeral at a time? Working Paper 21788, National Bureau of Economic Research.

Azoulay, P., Furman, J. L., Krieger, and Murray, F. (2015b). Retractions. *The Review of Economics and Statistics*, 97(5):1118–1136.

Azoulay, P., Graff Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules. *The Review of Economic Studies*, 86(1):117–152.

Azoulay, P., Graff Zivin, J. S., and Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3):527–554.

Azoulay, P., Graff Zivin, J. S., and Wang, J. (2010). Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589.

Azoulay, P. and Li, D. (2020). Scientific grant funding. Working Paper 26889, National Bureau of Economic Research.

Ballester, O. and Penner, O. (2019). Evolution of Topics and Novelty in Science. In Catalano, G and Daraio, C and Gregori, M and Moed, HF and Ruocco, G, editor, *17TH INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS (ISSI2019), VOL II*, Proceedings of the International Conference on Scientometrics and Informetrics, pages 1606–1611, KATHOLIEKE UNIV LEUVEN, FACULTEIT E T E W, DEKENSTRAAT 2, LEUVEN, B-3000, BELGIUM. Int Soc Scientometr & Informetr, INT SOC SCIENTOMETRICS & INFORMETRICS-ISSI. 17th International Conference of the International-Society-for-Scientometrics-and-Informetrics (ISSI) on Scientometrics and Informetrics, Sapienza Univ Rome, Rome, ITALY, SEP 02-05, 2019.

Banerjee, I., Chen, M. C., Lungren, M. P., and Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest ct cohort. *Journal of Biomedical Informatics*, 77:11 – 20.

Belford, M., Namee, B. M., and Greene, D. (2017). Stability of topic modeling via matrix factorization. *CoRR*, abs/1702.07186.

Bergstrom, C. T., West, J. D., and Wiseman, M. A. (2008). The eigenfactor™ metrics. *Journal of Neuroscience*, 28(45):11433–11434.

Bhattacharya, J. and Packalen, M. (2011). Opportunities and benefits as determinants of the direction of scientific research. *Journal of Health Economics*, 30(4):603 – 615.

Bikard, M., Murray, F., and Gans, J. S. (2015). Exploring trade-offs in the organization of scientific work: Collaboration and scientific reward. *Management Science*, 61(7):1473–1495.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

Blei, D. M., Jordan, M. I., Griffiths, T. L., and Tenenbaum, J. B. (2003a). Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 17–24, Cambridge, MA, USA. MIT Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–44.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Boom, C. D., Canneyt, S. V., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150 – 156.

Borner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Information (International Social Science Council)*, 14(6):19–47.

Boyack, K. W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3):351–374.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., and 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.

Braam, R., Moed, H., and Raan, T. (1991). Mapping of science by combined co-citation and word analysis. i. structural aspects. *JASIS*, 42:233–251.

Bramoulle, Y. and Saint-Paul, G. (2010). Research cycles. *Journal of Economic Theory*, 145(5):1890 – 1920.

Busch, L., Lacy, W. B., and Sachs, C. (1983). Perceived criteria for research problem choice in the agricultural sciences-a research note. *Social Forces*, 62(1):190–200.

Bush, Vannevar, .-. ([1945]). *Science–the endless frontier : a report to the President on a program for postwar scientific research*. Repr. May 1980. [Washington, D.C.] : National Science Foundation, [1980].

Callon, M. and Bowker, G. (1994). Is science a public good? fifth mullins lecture, virginia polytechnic institute, 23 march 1993. *Science, Technology, & Human Values*, 19(4):395–424.

Callon, M., Courtial, J. P., and Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205.

Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. on Knowl. and Data Eng.*, 19(2):261–272.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 288–296, USA. Curran Associates Inc.

Chen, J., Zhou, L., and Pan, S.-y. (2014). A brief review of recent advances in stem cell biology. *Neural Regeneration Research*, 9(7):684–687.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why u.s. manufacturing firms patent (or not). Working Paper 7552, National Bureau of Economic Research.

Council, N. R. and of Medicine, I. (2002). *Stem Cells and the Future of Regenerative Medicine*. The National Academies Press, Washington, DC.

Crane, D. (1969). Social structure in a group of scientists: A test of the "invisible college" hypothesis. *American Sociological Review*, 34(3):335–352.

Cyranoski, D. (2018). How human embryonic stem cells sparked a revolution. *Nature*, 555(7697):428—430.

Dai, A. M., Olah, C., and Le, Q. V. (2015). Document embedding with paragraph vectors. *CoRR*, abs/1507.07998.

Dasgupta, P. and David, P. A. (1994). Toward a new economics of science. *Research Policy*, 23(5):487 – 521. Special Issue in Honor of Nathan Rosenberg.

David, P. (2003). The economic logic of "open science" and the balance between private property rights and the public domain in scientific data and information: A primer. In Council, N. R., editor, *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*. The National Academies Press, Washington, DC.

David, P. A. and Foray, D. (1995). Accessing and expanding the science and technology knowledge base. *STI Review*, (16). OECD periodical.

de Rassenfosse, G. and Higham, K. (2020). Wanted: a standard for virtual patent marking. *Journal of Intellectual Property Law and Practice*, 15(7):544–553.

de Rassenfosse, G. and Verluise, C. (2020). PatCit: A Comprehensive Dataset of Patent Citations. *Zenodo*.

Ding, C., Li, T., and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8):3913–3927.

Dreifus, C. (2006). At harvard's stem cell center, the barriers run deep and wide. *New York Times*.

Eckhouse, S., Lewison, G., and Sullivan, R. (2008). Trends in the global funding and activity of cancer research. *Molecular Oncology*, 2(1):20 – 32.

Fernández-Zubieta, A., Geuna, A., and Lawson, C. (2015). Chapter 1 - what do we know of the mobility of research scientists and impact on scientific production. pages 1 – 33.

Fleming, L., Greene, H., Li, G., Marx, M., and Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141.

Fontana, M., Montobbio, F., and Racca, P. (2019). Topics and geographical diffusion of knowledge in top economic journals. *Economic Inquiry*, 57(4):1771–1797.

Foster, J. G., Rzhetsky, A., and Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908.

Furman, J. L., Murray, F., and Stern, S. (2012). Growing stem cells: The impact of federal funding policy on the u.s. scientific frontier. *Journal of Policy Analysis and Management*, 31(3):661–705.

Gardner, R., Ostrom, E., and Walker, J. M. (1990). The nature of common-pool resource problems. *Rationality and Society*, 2(3):335–358.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111.

Garfield, E., Malin, M. V., and Small, H. R. (1978). Citation data as science indicators. pages 179–208.

Garten, J., Sagae, K., Ustun, V., and Dehghani, M. (2015). Combining distributed vector representations for words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 95–101, Denver, Colorado. Association for Computational Linguistics.

Gersani, M., Brown, J. s., O'Brien, E. E., Maina, G. M., and Abramsky, Z. (2001). Tragedy of the commons as a result of root competition. *Journal of Ecology*, 89(4):660–669.

Gieryn, T. F. (1978). Problem retention and problem change in science. *Sociological Inquiry*, 48(3-4):96–115.

Gintis, H. (2009). *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction - Second Edition*. Princeton University Press, rev - revised, 2 edition.

Glaser, J., Glanzel, W., and Scharnhorst, A. (2017). Same data—different results? towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2):979–979.

Glenisson, P., Glänzel, W., Janssens, F., and De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Inf. Process. Manage.*, 41(6):1548–1572.

Gordon, H. S. (1954). The economic theory of a common-property resource: The fishery. *Journal of Political Economy*, 62(1):124 – 142.

Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? stability analysis for topic models. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Berlin, Heidelberg. Springer Berlin Heidelberg.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Griliches, Z. (1992). The search for r&d spillovers. *The Scandinavian Journal of Economics*, 94:S29–S47.

Hagstrom, W. O. (1974). Competition in science. *American Sociological Review*, 39(1):1–18.

Hall, B. H., Griliches, Z., and Hausman, J. A. (1984). Patents and r&d: Is there a lag? Working Paper 1454, National Bureau of Economic Research.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hecking, T. and Leydesdorff, L. (2018). Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Hegde, D. and Sampat, B. (2015). Can Private Money Buy Public Science? Disease Group Lobbying and Federal Funding for Biomedical Research. *Management Science*, 61(10):2281–2298.

Henderson, R., Jaffe, A., and Trajtenberg, M. (1998). Universities as a source of commercial technology: A detailed analysis of university patenting, 1965-1988. *The Review of Economics and Statistics*, 80(1):119–127.

Higgins, M. J., Stephan, P. E., and Thursby, J. G. (2011). Conveying quality and value in emerging industries: Star scientists and the role of signals in biotechnology. *Research Policy*, 40(4):605–617.

Hjaltason, G. R. and Samet, H. (2003). Index-driven similarity search in metric spaces (survey article). *ACM Trans. Database Syst.*, 28(4):517–580.

Ho, D., Imai, K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., and Wittenburg, P. (2018). Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data.

Holland, S., Lebacqz, K., and Zoloth, L. (2001). The human embryonic stem cell debate: Science, ethics and public good. Basic Bioethics Series. Massachusetts Institute of Technology, USA.

Huang, H. and Jong, S. (2019). Public funding for science and the value of corporate r&d projects; evidence from project initiation and termination decisions in cell therapy. *Journal of Management Studies*, 56(5):1000–1039.

Humphries, J. E. (2019). The causes and consequences of self-employment over the life cycle. (THESIS):103.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

Jacob, B. A. and Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, 95(9):1168 – 1177. Special Issue: The Role of Firms in Tax Systems.

Jaffe, A. B. (2002). Building programme evaluation into the design of public research-support programmes. *Oxford Review of Economic Policy*, 18(1):22–34.

Jaffe, A. B. and de Rassenfosse, G. (2017). Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6):1360–1374.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations*. *The Quarterly Journal of Economics*, 108(3):577–598.

Jefferson, O. A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., Ballagh, A., Moe, J., DiCuccio, M., Ward, K., Bilder, G., Dolby, K., and Jefferson, R. A. (2018). Mapping the global influence of published research on industry and innovation. *Nature Biotechnology*, 36(1):31–39.

Jin, G. Z., Jones, B., Lu, S. F., and Uzzi, B. (2013). The reverse matthew effect: Catastrophe and consequence in scientific teams. Working Paper 19489, National Bureau of Economic Research.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Kaplan, D. (2005). How to improve peer review at nih. *The Scientist*, 19:10.

Kealey, T. and Ricketts, M. (2014). Modelling science as a contribution good. *Research Policy*, 43(6):1014 – 1024.

Kiri, B., Lacetera, N., and Zirulia, L. (2018). Above a swamp: A theory of high-quality scientific production. *Research Policy*, 47(5):827 – 839.

Klavans, R. and Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4):984–998.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kuhn, T. S. (1977). *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press.

Larédo, P. (2015). *Supporting Frontier Research, Which Institutions and Which Processes*, pages 189–205. Springer International Publishing, Cham.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788 EP –.

Lenz, D. and Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE*, 15(1):1–18.

Levine, A. (2004). Trends in the geographic distribution of human embryonic stem-cell research. *Politics and the Life Sciences*, 23(2):40–45.

Levine, A. D. (2011). Policy uncertainty and the conduct of stem cell research. *Cell Stem Cell*, 8(2):132–135.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the nih. *American Economic Journal: Applied Economics*, 9(2):60–92.

Li, D., Azoulay, P., and Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, 356(6333):78–81.

Li, Y., Vanhaverbeke, W., and Schoenmakers, W. (2008). Exploration and exploitation in innovation: Reframing the interpretation. *Creativity and Innovation Management*, 17(2):107–126.

Lockard, A. and Tullock, G. (2001). *Efficient Rent-Seeking: Chronicle of an Intellectual Quagmire*. Springer US.

Lu, K. and Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10):1973–1986.

Löser, P., Schirm, J., Guhr, A., Wobus, A. M., and Kurtz, A. (2010). Human embryonic stem cell lines and their use in international research. *STEM CELLS*, 28(2):240–246.

Manso, G. (2011). Motivating innovation. *The Journal of Finance*, 66(5):1823–1860.

Marburger, III, J. H., Lane, J. I., Shipp, S. S., and Fealing, K. H. (2011). *The Science of Science Policy: A Handbook*. Stanford University Press, Stanford.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87.

Marx, M. and Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594.

Matheson, A. (2008). Corporate science and the husbandry of scientific and medical knowledge by the pharmaceutical industry. *BioSocieties*, 3(4):355–382.

McCormick, J. B., Owen-Smith, J., and Scott, C. T. (2009). Distribution of human embryonic stem cell lines: Who, when, and where. *Cell Stem Cell*, 4(2):107 – 110.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior*. Academic Press.

McKnight, S. L. (2009). Unconventional wisdom. *Cell*, 138(5):817 – 819.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 490–499, New York, NY, USA. Association for Computing Machinery.

Merton, R. and Storer, N. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Phoenix books. University of Chicago Press.

Merton, R. K. (1968). The matthew effect in science. *Science*, 159(3810):56–63.

Michel, J. and Bettels, B. (2001). Patent citation analysis. a closer look at the basic input data from patent search reports. *Scientometrics*, 51(1):185–201.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Moretti, E., Steinwender, C., and Van Reenen, J. (2019). The intellectual spoils of war? defense r&d, productivity and international spillovers. Working Paper 26483, National Bureau of Economic Research.

Murray, F. (2007). The stem-cell market of patents and the pursuit of scientific progress. *New England Journal of Medicine*, 356(23):2341–2343. PMID: 17554114.

Murray, F. (2010). The oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. *American Journal of Sociology*, 116(2):341–388.

Murugan, V. (2009). Embryonic stem cell research: a decade of debate from bush to obama. *The Yale journal of biology and medicine*, 82(3):101–103.

Narin, F. and Olivastro, D. (1998). Linkage between patents and papers: An interim epo/us comparison. *Scientometrics*, 41(1):51–59.

Nelson, R. R. (1959). The simple economics of basic scientific research. *Journal of Political Economy*, 67(3):297–306.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

NIH (2001). *Stem Cells: Scientific Progress and Future Research Directions.* National Institutes of Health. Department of Health and Human Services, Bethesda, Maryland.

Nitzan, S. (1991). Collective rent dissipation. *Economic Journal*, 101(409):1522–34.

Noyons, E. C. M. and van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an r & d field. *Scientometrics*, 30(1):157–173.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science*, 58(6):1122–1140.

Owen-Smith, J. and McCormick, J. (2006). An international gap in human es cell research. *Nature Biotechnology*, 24(4):391–392.

Packalen, M. and Bhattacharya, J. (2018). Does the nih fund edge science? Working Paper 24860, National Bureau of Economic Research.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Philpott, K., Dooley, L., O'Reilly, C., and Lupton, G. (2011). The entrepreneurial university: Examining the underlying academic tensions. *Technovation*, 31(4):161 – 170. Managing Technology.

Poege, F., Harhoff, D., Gaessler, F., and Baruffaldi, S. (2019). Science quality and the value of inventions. *Science Advances*, 5(12).

Price, J. D. S. D. and Beaver, D. (1966). Collaboration in an invisible college. *The American psychologist*, pages 1011–1018.

Pérez-Castrillo, J. and Verdier, T. (1992). A general analysis of rent-seeking games. *Public Choice*, 73:335–350.

Racherla, P. and Hu, C. (2010). A social network perspective of tourism research collaborations. *Annals of Tourism Research*, 37(4):1012–1034.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, 94(5):1002–1037.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5):S71–S102.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1):4:1–4:38.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

Russo, E. (2005). Follow the money. the politics of embryonic stem cell research. *PLOS Biology*, 3(7).

Ryvkin, D. (2007). Tullock contests of weakly heterogeneous players. *Public Choice*, 132(1/2):49–64.

Salter, A. J. and Martin, B. R. (2001). The economic benefits of publicly funded basic research: a critical review. *Research Policy*, 30(3):509 – 532.

Sampat, B. N. (2010). Lessons from bayh–dole. *Nature*, 468(7325):755–756.

Scott, C. T., McCormick, J. B., and Owen-Smith, J. (2009). And then there were two: use of hesc lines. *Nature Biotechnology*, 27(8):696–697.

Socher, R., Chen, D., Manning, C. D., and Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc.

Stephan, P. (2012). *How Economics Shapes Science*. Harvard University Press.

Steyvers, M. and Griffiths, T. (2013). Probabilistic topic models. In Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors, *Handbook of latent semantic analysis*. Psychology Press.

Suominen, A. and 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.

Swanson, D. R. (1966). Scientific journals and information services of the future. *American psychologist*, 21(11):1005.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861 – 872.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):285 – 305.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Teodoridis, F., Bikard, M., and Vakili, K. (2018). Creativity at the knowledge frontier: The impact of specialization in fast- and slow-paced domains. *Administrative Science Quarterly*, 0(0):1–34.

Thijs, B. (2019). Paragraph-based intra- and inter- document similarity using neural vector paragraph embeddings. pages 1900–1911. ISSI.

Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., and Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147.

Torvik, V. I. and Smalheiser, N. R. (2009). Author name disambiguation in medline. *ACM Trans. Knowl. Discov. Data*, 3(3):11:1–11:29.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition.

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, 5(1):19–50.

Vakili, K., McGahan, A. M., Rezaie, R., Mitchell, W., and Daar, A. S. (2015). Progress in human embryonic stem cell research in the united states between 2001 and 2010. *PLOS ONE*, 10(3):1–8.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *The Journal of Machine Learning Research*.

Vaughan, C., Allen, L., and Chew, M. (2012). Malaria 1990-2009. *WellcomeTrust Portfolio Review*. Portfolio Review: report on human malaria infection research.

Velden, T., Boyack, K. W., Glaser, J., Koopman, R., Scharnhorst, A., and Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics*, 111(2):1169–1221.

Vogel, G. (1999). Capturing the promise of youth. *Science*, 286(5448):2238–2239.

von Eschenbach, A. C. (2003). Nci sets goal of eliminating suffering and death due to cancer by 2015. *Journal of the National Medical Association*, 95(7):637–639.

Wagner, C. S. (2008). *The New Invisible College: Science for Development*. Brookings Institution Press.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., and Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of Informetrics*, 5(1):14 – 26.

Walker, J. M., Gardner, R., and Ostrom, E. (1990). Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management*, 19(3):203 – 211.

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Wertz, D. C. (2002). Embryo and stem cell research in the united states: history and politics. *Gene Therapy*, 9(11):674–678.

Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J., and Campbell, K. H. S. (1997). Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385(6619):810–813.

Yan, E., Ding, Y., Milojević, S., and 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.

Yegros-Yegros, A., Rafols, I., and D'Este, P. (2015). Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PLOS ONE*, 10(8):1–21.

Zuckerman, H. (1978). Theory choice and problem choice in science. *Sociological Inquiry*, 48(3-4):65–95.

OMAR BALLESTER

Curriculum Vitae
July 2020

EPFL-CDM-IIPP
Station 5, ODY 4.16
1015 Lausanne, Switzerland
Phone: +34 627168677
omar.ballester@epfl.ch

EDUCATION:

PhD in Management, **(e. July) 2020**
Thesis: The organisation of Science: Topics, Incentives and Funding
Thesis Advisor: Prof. Gaétan de Rassenfosse

Swiss Program for beginning doctoral students in Economics, **2017**
Concentrations: Microeconomics, Macroeconomics, Econometrics

Postgrad. Introduction to Data Science, Universitat de Barcelona, **2015**
6-month course on the Data Science toolkit, with a final group project on GIS data.

MSc., Management, Univ. Pompeu Fabra (Barcelona School of Management) **2015**
Thesis: *Technology Transfer and Innovation in Spain*
Thesis Advisor: Xavier Castillo, Ph.D.

B.Sc, Physics, Universitat de Valencia, **2014**
Concentrations: Biophysics, Optics
Thesis: *Cell Membrane Rigidity with Optical Tweezers*
Thesis Advisor: Prof. Felix Ritort, Prof. Javier Cervera

Certificate of Higher Education (Cert-HE) in Professional Musicianship, University of Sussex, **2012**
Concentrations: Drums and Modern Music

TEACHING EXPERIENCE:

Teaching Assistant, EPFL, **2016 - present**
Data Science in Practice (Master) (2018-present)
Introduction to Microeconomics (Master) (2017-18)
Principles of Intellectual Property (Master) (2016-17)

Teaching Assistant, UNIGE, **2018 - present**
Mathematics (Master) (2019-present)
Introduction à la Microéconomie (Bachelor) (2018-present)
International Trade (Master) (2018-19)

Private tuition **2004 - 2012**
Private support/reinforcement teaching for students in secondary school.
Physics, Chemistry, Mathematics and English.

WORKING EXPERIENCE:

External Mandate **2019**
External consultant for research project. Text Mining and Topic Modelling of Indian Newspapers for Nathalie Monnet (PhD Candidate at The Graduate Institute, Geneva). Web scraping, parsing, text identification and event classification. Code and resources available on github.

PhD Researcher, École Polytechnique Fédérale de Lausanne **2016 - present**
Junior Researcher and Teaching Assistant (see details in other section)

Data Science & Innovation Consultant, Cuatrecasas (Spain) **2015 - 2016**
Job including two main tasks:

- Operational duties in the Business Intelligence & Reporting domain using SAP tools (Design Studio, Business Objects, BeX query designer). Report-visualization expert.
- Applied research in data-science tools for the legal sector and new measures and analytics for operational excellence including: design of algorithms, queries and development on python environment, sourcing external "off-the-shelf" tools for visual analytics, scientific exploration of assisting tools (or under development) for potential technology transfer.

Research Assistant in Biophysics, Universitat de Barcelona **2013 - 2014**
Interdisciplinary research in collaboration with the Vall d'Hebron Research Institute (oncology). Experimental setup, testing stem cells' elasticity using Optical Tweezers. Processing and analysis of large amounts of data.

Touring and Studio drummer for Papa Topo, Spain, **2012 - 2014**
Also heavily involved in tour management activities.

RESEARCH EXPERIENCE:

PhD Candidate at EPFL in the CDM-IIPP chair **2016-present**
Economics of Science, with a particular interest in STEM Academic careers, topics of interest, mobility and areas of expertise (subfields). Topic modelling of scientific articles with the objective of characterizing researcher's topics. Knowledge-domain visualization and characterization of "maps of science" in Biomedical Research.

Research Assistant in Biophysics, Universitat de Barcelona **2013 - 2014**
Interdisciplinary research in collaboration with the Vall d'Hebron Research Institute (oncology). Experimental setup, testing stem cells' elasticity using Optical Tweezers. Processing and analysis of large amounts of data.

Summer Schools and Workshops (attendee)

CISS Competition and Innovation Summer School Ulcnij, Montenegro (2019)

Junior Researcher Workshop: from science to innovation Max-Planck Institute, Munich (2018)

Barcelona GSE Summer School (2018): Panel Data Linear Analysis, Econometrics of Cross-section Data with Applications, Quantitative Methods of Public Policy Evaluation, Dynamic and Non-linear Panel Data Models.

Economics of Innovation and Technological Change, Lausanne (2017)

European School for Scientometrics, Granada (2016)

Conferences and Workshops (presenter)
"Innovation Stems from Science: Stem-Cell funding shocks and the impact on innovation" – CEMI-IIPP Retreat, St. Luc (March 2020)

"Funding policy uncertainty slows societal impact of research" – WNAC Meeting of SZ. Gerzensee Alumni, Girona (October 2019)

"Evolution of Topics and Novelty in Science. Mapping novel research topics" – GTM2019 – Global Tech Mining Conference, Atlanta (October 2019)

"Research choices in scientific careers: a game of effort and ability" – ATC2019 Atlanta Conference on Science Policy, (October 2019)

"Topic Modelling in Social Sciences" – IV Summer School in Science and Technology Indicators, KUL-EPO, Vienna, (September 2019)

"Evolution of Topics and Novelty in Science. Mapping novel research topics" – ISSI2019 International Conference on Scientometrics & Informetrics, Rome (September 2019)

"Evolution of Topics and Novelty in Science. Mapping novel research topics" – EPFL Workshop on Computational Methods in Social Science (July 2019)

"Research choices in scientific careers: a game of effort and ability" – Barcelona Summer Forum in Economics of Science and Innovation (June 2019)

"Research choices in scientific careers: a game of effort and ability" – CISS Summer School, Ulcnij, Montenegro (May 2019)

"Research choices in scientific careers: a game of effort and ability" - WICK #6 PhD Workshop in Economics of Innovation, Complexity and Knowledge (2019)

"Research choices in scientific careers: a game of effort and ability" – WNAC Meeting of SZ. Gerzensee Alumni, Marrakech (2018)

Publications

"COVID-19: Insights from Innovation Economists" – Science and Public Policy (2020). https://doi.org/10.1093/scipol/scaa028

"Topical Stability: how neural networks defeat traditional Topic Modelling approaches" **In preparation** - Journal of the Association for Information Science and Technology JASIST.

"Evolution of Topics and Novelty in Science. Mapping novel research topics" – ISSI2019 International Conference on Scientometrics & Informetrics Conference Proceedings (2019) Vol. II. p: 1606-1611. ISBN: 978-88-3381-118-5

Service to Profession

Scientific Committee for *17th International Conference on Scientometrics & Informetrics*
Reviewer for *Advances in Complex Systems*
Reviewer for *World Patent Information*
Reviewer for *23rd International Conference on Science and Technology Indicators.*

OTHER (MISC.):

Languages (certificates)
English – (C2, proficiency)
French – (Dalf C1)
German – (B2)
Spanish, Catalan: Native

Other experiences & transversal capabilities:
Amateur musician (drummer & electronic), with several albums out.
AFS-Intercultura (NGO) volunteering (2007-2009)

REFERENCES:

Prof. Gaétan de Rassenfosse – (PhD Supervisor)
gaetan.derassenfosse@epfl.ch
EPFL CDM - IIPP
ODY 2 01.1 (Odyssea) – CH-1015

Dr. Orion Penner - (Analyst - WIPO)
orion.penner@gmail.com

Dr. Christopher Bruffaerts – (External Lecturer – Data Scientist)
christopher.bruffaerts@epfl.ch

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**