Essays on Utilizing Data Analytics and Dynamic Modeling to Inform Complex Science and Innovation Policies

Arash Baghaei Lakeh

Dissertation submitted to the faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Industrial and Systems Engineering

Navid Ghaffarzadegan, Chair

Jason P Kelly

James Kong

Keyvan Vakili

April 20, 2018

Blacksburg, Virginia

Keywords: Science policy, data analytics, topic modeling, system dynamics, health studies, behavioral and social sciences, funding institutions, science philanthropy, science workforce

Essays on Utilizing Data Analytics and Dynamic Modeling to Inform Complex Science and Innovation Policies

Arash Baghaei Lakeh

ABSTRACT

In many ways, science represents a complex system which involves technical, social, and economic aspects. An analysis of such a system requires employing and combining different methodological perspectives and incorporation of different sources of data. In this dissertation, we use a variety of methods to analyze large sets of data in order to examine the effects of various domestic and institutional factors on scientific activities. First, we evaluate how the contributions of behavioral and social sciences to studies of health have evolved over time. We use data analytics to conduct a textual analysis of more than 200,000 publications on the topic of HIV/AIDS. We find that the focus of the scientific community within the context of the same problem varies as the societal context of the problem changes. Specifically, we uncover that the focus on the behavioral and social aspects of HIV/AIDS has increased over time and varies in different countries. Further, we show that this variation is related to the mortality level that the disease causes in each country. Second, we investigate how different sources of funding affect the science enterprise differently. We use data analytics to analyze more than 60,000 papers published on the subject of specific diseases globally and highlight the role of philanthropic money in these domains. We find that philanthropies tend to have a more practical approach in health studies as compared with public funders. We further show that they are also concerned with the economic, policy related, social, and behavioral aspects of the diseases. We uncover that philanthropies tend to mix and combine approaches and contents supported both by public and private sources of funding for science. We further show that in doing so, philanthropies tend to be closer to the position held by the public sector in the context of health studies. Finally, we find that studies funded by philanthropies tend to receive higher citations, and hence have higher impact, in comparison to those funded by the public sector. Third, we study the effect of different schemes of funding distribution on the career of scientists. In this study, we develop a system dynamics model for analyzing a scientist's career under different funding and competition contexts. We investigate the characteristics of optimal strategies and also the equilibrium points for the cases of scientists competing for financial resources. We show that a policy to fund the best can lead scientists to spend more time on writing proposals, in order to secure funding, rather than writing papers. We find that when everyone receives funding (or have the same chance of receiving funding) the overall optimal payoff of the scientists reaches its highest level and at this optimum, scientists spend all their time on writing papers rather than writing proposals. Our analysis suggests that more egalitarian distributions of funding results in higher overall research output by scientists. We also find that luck plays an important role in the success of scientists. We show that following the optimal strategies do not guarantee success. Due to the stochastic nature of funding decisions, some will eventually fail. The failure is not due to scientists' faulty decisions, but rather simply due to their lack of luck.

Essays on Utilizing Data Analytics and Dynamic Modeling to Inform Complex Science and Innovation Policies

Arash Baghaei Lakeh

GENERAL AUDIENCE ABSTRACT

Science helps us understand the world and enables us to improve how we interact with our environment. But science itself has also been the subject of inquiry by philosophers, sociologists, economists, historians, and scientists. The goal in the investigations of science has been to better understand how scientific advances occur, how to foster innovation, and how to improve the institutions that push science forward. This dissertation contributes to this area of research by asking and responding to several questions about the science enterprise. First, we study how communities of scientists in different parts of the world look at the seemingly same problem differently. We use a computational method to read through a large set of publications on the topic of HIV/AIDS (which includes more than 200,000 papers) and uncover the topics of these papers. We find that in the context of HIV/AIDS, contributions of behavioral and social scientists have increased over time. Moreover, we show that the share of these contributions in any counties' total research output differs significantly. We further find that there is a significant relationship between one country's rate of death, due to HIV/AIDS, and the share of behavioral and social studies in the overall research profile of that country on the topic of HIV/AIDS. Second, we investigate how different sources of research funding affect scientific activities differently. Specifically, we focus on the role of philanthropic money in science and its effect on the content and impact of research studies. In our analysis, we rely on computational techniques that distinguishes between different themes of research in the studies of a few diseases and also different statistical methods. We find that philanthropies tend to have a more practical approach to health studies as compared with public sources of funding. Meanwhile, we find that they are also concerned with the economic, policy related, social, and behavioral aspects of the diseases. Moreover, we show that philanthropies tend to mix and combine approaches and contents supported both by public and private sources of funding for science. We find that, in doing so, philanthropies tend to be closer to the position held by the public sector in the context of health studies. Finally, we show that studies funded by philanthropies tend to receive higher citations. This finding suggests that these studies have a higher impact in comparison to those funded by the public sector. Third, we study how different mechanisms for distributing research funding among scientists can affect their career and success. Many scientists should spend time on both writing papers and research grant proposals. In this work, we aim at understanding how a scientists should allocate her time between these two activities to maximize her career long number of papers. We develop a small mathematical model to capture the mechanisms related to the research career of a scientist in an academic setting. Then, for different schemes of funding distribution, we find the scientist's time allocation that maximizes the number of papers she publishes over her career. We find that when funding is being allocated to the best scientists and best grant proposals, scientists' best strategy is to spend more time on writing research grant proposals rather than papers. This decreases the total number of papers published by the scientists over their career. We also find that luck is important in determining the career success of scientists. Due to errors in evaluation of proposal qualities, a scientist may fail in her career regardless of whether she has followed the best strategy that she could.

Dedication

To *Gohartaj* and *Bahman*, my selfless and compassionate *maman* and *baba*;

To my sweetest and smartest *Shadi*, the happiness of my life;

And to *Reza*, my most amazing and supportive brother**.**

Acknowledgement

Like many things that we do in life, writing a doctoral dissertation cannot be done in isolation and without the tremendous influence and support of others. Part of this support comes from the relationships that we shape with the goal of receiving such influence, like the student-advisor relationship. Another source of support is nested in our relationships with our loved ones: our parents, significant others, and friends. Here, I would like to pause for a moment and try my best to show my appreciation of the wonderful people in my life who have had a positive influence on me and supported me while writing this dissertation. I am aware that this is not an exhaustive list. Not only might I have forgotten some people that supported me during these past years, but also I might not be able to name those who influenced me without my conscious awareness. Hence, the limitations of my memory and my understanding of what makes me, *me,* have made the acknowledgment section to be incomplete. Yet, this is my best shot at it.

First, I would like to sincerely thank my advisor and mentor, Dr. Navid Ghaffarzadegan. I met Navid years ago when I was an undergraduate student. At the time, Navid was a PhD student and I had the chance of attending a talk he gave on modeling people's decision making. I can remember how much I enjoyed listening to him, talking about his research. I was not sure if I could ever work on such an interesting project myself. Little did I know that I would have this chance and I would actually work with Navid during my PhD studies. I learned a lot from Navid and enjoyed so much collaborating with him. He truly is a fine person, a passionate researcher, and a knowledgeable teacher. I am thankful for all the support that I received and all the things that I learned from him. I am also grateful for all the coffees he bought me during the past couple of years!

I also want to thank my dissertation committee members: Dr. Jason Kelly, Dr. Keyvan Vakili, and Dr. James Kong. In all of the meetings that I have had with them, I felt nothing but their intention of helping me to become a better researcher. I took Dr. Kelly's course on quantitative methods and the basis of my statistical knowledge was shaped in that class. I am very thankful for all the things that I learned from him. I would also like to specially thank Dr. Vakili for his helpful comments on the earlier versions of the essays reported in the next chapters.

During my doctoral dissertation, I had the chance to interact with and learn from many other faculty members within and outside Virginia Tech. I specially thank Dr. Hazhir Rahmandad for trusting in me and admitting me into the Industrial and Systems Engineering program. My admission to this program at Virginia Tech has been a turning point in my career and I appreciate Hazhir for offering me this great opportunity. I am also thankful to Dr. Griffin Weber for his comments on my first essay and allowing me to have access to the data that I used in my second essay. I worked on two client oriented projects in collaboration with Dr. Chris Wernz. Both of these experiences were very valuable for me. I am thankful to him for these opportunities. I am also grateful to Dr. Alejandro Salado for letting me be involved in an innovative project as a teaching assistant and a research collaborator. I would also like to thank Dr. Niyousha Hosseinichimeh, Dr. Dick Larson, and Dr. Joshua Hawley for their helpful comments on my works during the past few years.

I am very grateful to my lovely wife, Shadi. We walked this road together. I am the luckiest person to have had such a smart, passionate, caring, sweet, and bighearted companion in this journey. Not every day of a PhD student is filled with aspiration and an optimistic view of the future. It is as simple as this: one of those gloomy days of my doctoral studies could have taken me down if it was not for Shadi. Even though we were not physically close to each other for the most part of

our doctoral studies, I always felt like she was right beside me and helping me to move forward and to think of the bright future. Thank you *azizam*!

I do not know the words that can describe how much I am grateful to my parents and my brother for their constant support of me over the past three decades of my life. Without them, there would be no *me* and hence no doctoral dissertation. Thank you mom, dad, and Reza from the bottom of my heart! I also would like to thank my sister-in-law, Shabnam, for her kind support. My parents-in-law, and Ehsan and Pegah (my brother- and sister-in-law) have also been very supportive of me, which I appreciate from the bottom of my heart.

I have been very lucky to have wonderful, smart, and fun lab mates during my doctoral studies. I would like to specially thank Armin Ashouri Rad, Mohammad Jalali, Sarah Mostafavi, Ran Xu, Alba Rojas-Cordova, Hui Zhang, Maryam Andalib, Nasibeh Azadeh Fard, Jose Guevara Maldonado, Kyle Gentle, Alireza Ebrahimvandi, and Negar Darabi.

I have always been blessed with having the greatest friends in my life. Without them, I would not be the person that I am today and certainly could not go through my doctoral studies as I did. I am grateful to my friends in Blacksburg. Specially, I would like to thank Ebrahim Ahmadisharaf, Nasrin Alamdari, Ehasn Asghari Ghara, Ahmadreza Azizi, Kaveh Bastani, Vahid Bateni, Sorour Ekhtiari Amiri, Giulio Menciotti, Hadi Parsian, Kaveh Rahimi, Elaheh Raisi, Ehsan Rashedi, Alejandra Rosado, Arash Sarshar, and Alireza Sedighi. Also, many thanks goes to my friends in Pittsburgh. I am especially grateful to Majid Darvishan, Laleh Gharanjik, Fattaneh Jabbari, Salim Malakuti, Mostafa Mirshekari, Ali Pakzad, Zahra Rahimi, Maryam Shabani, Amin Tajgardoon, Sareh Yousefzadeh, and Azarin Zarasi. Last, but certainly not least, I would like to thank my dear friend Alireza Faghaninia. Not only did he (and his wife, Agnes Treneyi) drive more than ten hours to help me move from Illinois to Virginia, he has also been an emotional support for me during my studies at Virginia Tech.

Finally, I want to thank many users on the Stack Overflow website, who answered my technical questions during the past years and made my life much easier.

Table of Contents

List of Figures ..... **ix**
List of Tables ..... **xi**
Chapter 1: Introduction **1**
Global trends and regional variations in studies of HIV/AIDS 2
What roles does philanthropic money play in health research? 3
Recipe for success in academia: a dynamic model of scientists' career 3
Research approaches 4
References 4
Chapter 2: Global Trends and Regional Variations in Studies of HIV/AIDS **7**
Abstract 7
Introduction 7
Data: HIV/AIDS Publications between 1985 and 2012 8
Increasing Focus on BSS Aspects of HIV/AIDS 9
Regional Variation in BSS Focus of HIV/AIDS Studies 12
Discussion 13
Methods 14
Topic Modeling 14
Panel Regression Analysis 16
References 17
Acknowledgements 17
Appendix A 19
Chapter 3: What role does the philanthropic money play in health research? **32**
Abstract 32
Introduction 32
Different Sources of Funding for Science 33
Methods and Materials 38
Data 38
Topic Modeling 40
Results 42
Same Problems Different Approaches 42
Philanthropic vs Public and Private Funding Sources 45
Scientific Impact 46
Discussion and Conclusions 51

Acknowledgements 54
References 54
Chapter 4: Recipe for success in academia: a dynamic model of scientists' career 57
Abstract 57
Introduction 57
Success in Science and the Career of Scientists 58
Model 62
Distribution of Funding and the Optimal Strategy 65
Absolute Evaluation Based Distribution of Funding 65
Relative Evaluation Based Distribution of Funding 70
Luck and the Best Strategy 72
Discussion and Conclusions 74
Acknowledgements 77
References 77
Appendix B 82
Appendix C 83
Chapter 5: Conclusions 87
References 89

List of Figures

Figure 1.1: Dissertation overview 2

Figure 2.1: Trends in HIV/AIDS Publications: (a) global trend of publication; (b) average number of authors per paper; (c) number of countries contributing to HIV/AIDS research 8

Figure 2.2: Share of aggregate BSS topics (left axis) and five individual topics (right axis) in HIV/AIDS research over time 9

Figure 2.3: Share of BSS research in all HIV/AIDS papers of different countries. Note: For example, share of BSS research for Ethiopia is shown at 51% meaning that 51% of HIV/AIDS research content published by Ethiopian authors is on BSS topics. Only countries are shown that have published more than 100 papers during the period 1985 to 2012. Color coded by total number of HIV/AIDS papers 10

Figure 2.4: Regional variation in trends of BSS topics in the context of HIV/AIDS research. NA: North America, WE: Western Europe, AS: South and Southeast Asia, AF: Sub-Saharan Africa 11

Figure 2.5: BSS focus of publications from different countries (average during the period 2006 to 2010) compared to their HIV/AIDS mortality rate in 2005 and GDP per capita (average during 2006 to 2010). Each dot corresponds to a country (Red: North America & Western Europe; Yellow: Southern & Southeast Asia; Gray: Sub-Saharan Africa; Blue: Rest of the world). The solid lines show the correlation lines. Countries that are two standard deviation further from the correlation line are labeled 12

Figure 3.1: Sources of global funding for studies of certain diseases 34

Figure 3.2: Distribution of the global research funding for certain diseases during 2007-2016 .. 35

Figure 3.3*:* Three possible relationship between philanthropic (N), Public (G), and Private (P) funders 36

Figure 3.4: Trends of Publication. (a) Number of papers published on different diseases annually. (b) Number of countries that have published at least one paper on each of the diseases. Light Blue: HIV/AIDS; Yellow: Tuberculosis; Orange: Malaria; Gray: Pneumonia; Dark Blue: Neglected Tropical Diseases 38

Figure 3.5: Funding organizations indexed by Web of Science. (a) Percentage of papers with at least one organization indexed. (b) Count of papers with different number of funding organization indexed. 39

Figure 4.1: Conceptual model 63

Figure 4.2: Career trajectory of scientists: (a) Few potential career trajectories. (b) Expected career long payoff for different strategies, (c) Relative standard deviation of expected payoffs for different strategies 67

Figure 4.3: Effect of funding threshold and error in evaluation of proposal quality on (a) optimal strategy, r*, and (b) optimal payoff, $\pi*$. Different colors represent varying levels of error in evaluation of proposal quality: green (δ=0), orange (δ=0.25), gray (δ=0.5), yellow (δ=0.75), and dark blue (δ=1). (c) The relative standard deviation of career-long payoff is shown as a function of threshold and evaluation error 69

Figure 4.4: Effect of different inequality factors on the optimal strategy (a) and career-long payoff (b) of scientists working under competition.
Blue: α=0.3; Orange: α=0.5; and Gray: α=0.7 71

Figure 4.5: The mixed strategy equilibrium under the Winner Takes All type of competition for different values of α. The blue line shows the probability of the strategy to almost invest all resources on proposal writing (very small r). The orange line shows the probability of the strategy to invest all resources on research 72

Figure 4.6: Effect of 25% proposal evaluation error on variation of expected career-long payoff.
Blue: α = 0.3; Orange: α = 0.5; and Gray: α = 0.7 73

Figure 4.7: Distribution of career-long payoff under the optimal strategy and in existence of 25% evaluation error. (a) Quality-threshold based with threshold set at 0.5 (b) Competition based with inequality factor set at 1 74

Figure 5.1: Growth of the science enterprise 87

List of Tables

Table 2.1: Panel data regression results for percentage of BSS topics 16

Table A1: The detailed information of LDA topics 19

Table A2. Comparison of clusters for three slices of data and the original analysis. The table shows the number of shared unique words in the set of unique top words ($\boldsymbol{S_{ij}}$). The $\boldsymbol{r_{ij}}$ (normalized number of shared unique words) values are reported in parentheses. The gray cells show the corresponding cluster for each slice to the one of the original analysis based on the values of $\boldsymbol{r_{ij}}$ 27

Table A3**:** Panel data regression results for BSS topics generated by top 11 topics per paper 28

Table A4**:** Panel data regression results for percentage of BSS topics 28

Table A5**:** Panel data regression results for BSS topics including the time variables 29

Table A6**:** Panel data regression results for subset of publications without any cross-country collaboration 31

Table 3.1: The top 10 philanthropic and private funding agencies 40

Table 3.2: Top five topics in the studies of different diseases 42

Table 3.3*:* List of topics that studies funded only by philanthropies focused more in comparison to those funded only by public agencies 43

Table 3.3*:* Relationship of contents funded by philanthropies as oppose to those funded by government and for-profits 46

Table 3.4*:* Mean and variance of citations for papers written in the context of different diseases and funded by different organization types 47

Table 3.5*:* Results of citation analysis through negative binomial regression model 47

Table 3.6*:* Factors influencing chances of a paper to have a high impact. Results are rendered from a logistic regression model 49

Table 3.7*:* Results of citation analysis through negative binomial regression model (considering the effect of content and high impact papers) 50

Table 4.8*:* Parameter values used in paper 66

Chapter 1: Introduction

How can complex science and innovation policies be informed by employing data analytics and simulation techniques? This question summarizes the general theme of this doctoral dissertation. Combining methods from data analytics, econometrics, and mathematical modeling enables us to conduct an in-depth analysis of complex socio-technical systems. In this dissertation, we use such methods to answer several important and complex questions in the domain of science and innovation.

Science production occurs through complicated and time consuming processes. In these processes, scientists work together and produce outputs that are many times innovative and unexpected. These outcomes are difficult to predict and scientists' success depends on a wide range of technical, social, and institutional factors. In many ways, science production represents a complex, large, socio-technical system which includes human, technology, resources, and interactions of these sub-systems.

A wide range of studies have focused on understanding and improving scientific activities. Productivity of scientists, collaboration patterns, scientists' "birth rate", migration, effects of funding on research activities, and interdisciplinary or transdisciplinary research are some of the major topics of research in this domain [1]–[8]. A common theme in this line of research is the complexity of predicting (and influencing) science production as affected by funding sources [9], [10]. Our research will move in the same direction, trying to offer an in-depth understanding of science production as a complex system.

We analyze the science enterprise from three perspectives: national, organizational, and individual. Our main goal is to examine effects of various domestic and institutional factors on scientific activities. We pursue this goal by conducting three related research projects reported in three essays. The overview of the dissertation is presented in Figure 1.1.

First, we ask how different settings can alter the way we look at the same problem. We focus on a certain problem, i.e. HIV/AIDS, and examine how different approaches to resolving this issue have evolved over time. Specifically, we investigate whether there is a relationship between proximity to a health problem (different context) and the way that the problem is being perceived and tackled by the science enterprise (different approaches). Second, we ask how different sources of funding shape science differently. We take few fields of scientific research in the context of health and investigate how differences in funding sources are related to differences in topics and approaches. Specifically, we focus on the effect of science philanthropy. We examine how this source of funding is different from public sources in terms of the content that their money produce in science. We further analyze the scientific impact of philanthropies' spending in health research and development. Finally, we focus on the effect of different schemes of funding distribution on the research career of scientists. We model the long-term success of a scientist as a function of their time allocation between writing grant proposals and papers. We investigate how different schemes of funding distributions affect the long-term success of scientists and the impact on science enterprise. A range of methods including topic modeling, econometrics, and mathematical modeling and simulation will inform these studies.

In the following I offer more details about each essay, and the general research approach in this dissertation.

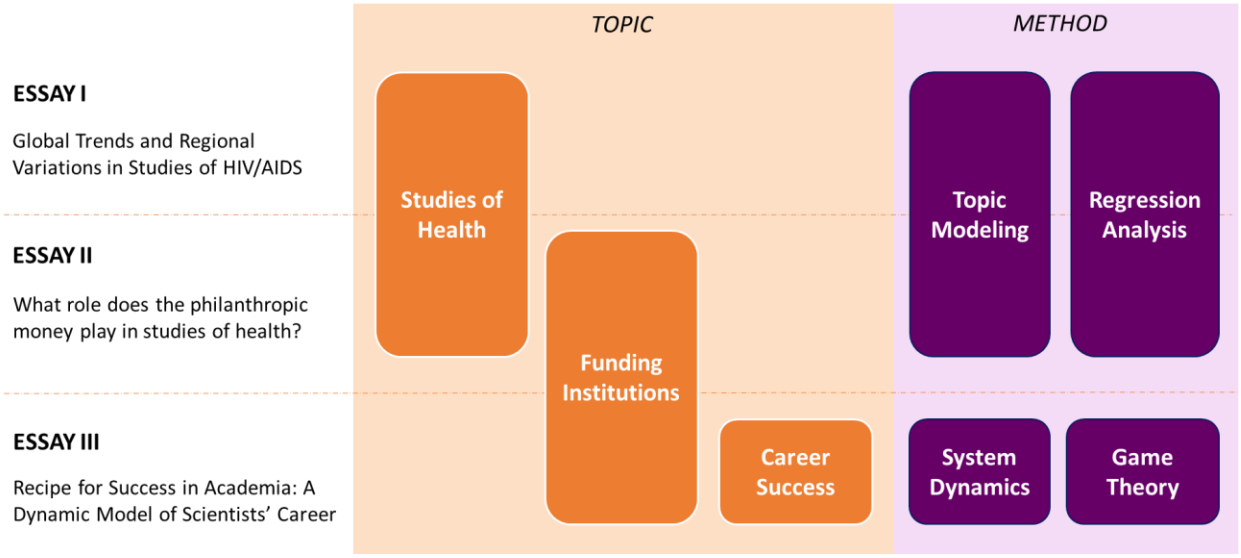


Figure 1.1: Dissertation overview.

Global trends and regional variations in studies of HIV/AIDS

In the first essay (chapter 2), we study how the context of a problem shape the response of the science enterprise to it. There are two competing arguments about the relationship between societal context and science. In one view, scientific paradigms are perceived to be shaping autonomously and without any connection with social contexts in which the scientists are working [11]. This view suggests that a shared understanding about a specific problem will eventually inform all the relevant scientific activities. Another view maintains that the supply of science will reconcile with the societies' demands [12], [13]. In other words, what a society perceives about a problem and how the science enterprise responds to it converge. In this view, we expect that as the context of a problem changes, so does the scientific response to it. In this study we take the case example of HIV/AIDS as a global challenge and investigate how the science enterprise in different countries has responded to it.

In this study, we look at the share and variation of the contribution of behavioral and social sciences in studies of HIV/AIDS. The complexity of the HIV/AIDS problem goes beyond the biomedical aspects of the disease. Many challenges are behavioral and social, such as public awareness of the disease [14], risk perception [15], high-risk behaviors [16], willingness to be tested [17], social stigma [18], and treatment adherence [19]. The infectious nature of the disease also adds to the social network complexity of the problem and to infectious patterns [20]. Numerous multi-disciplinary studies have been conducted to uncover these socio-behavioral aspects of the disease.

The major question in this study is how the contributions of behavioral and social sciences to the health studies evolve over time and vary in different parts of the world. We investigate the shifts

in the scientific community's focus on behavioral and social sciences aspects of HIV/AIDS through a textual analysis of more than 200,000 HIV/AIDS publications in the Scopus data set.

What roles does philanthropic money play in health research?

In the second study (chapter 3), we focus on the topic of funding in science. Specifically, we study the effects of philanthropic money in health studies. Our focus is on research activities related to few diseases that are affecting people in developing countries disproportionately, i.e. HIV/AIDS, Tuberculosis, Malaria, Pneumonia, and Neglected Tropical Diseases.

The effect of philanthropic money is overshadowed by funding from public and private sources [21]. However, their contribution to the research and development is not small. In the case of some diseases (such as Pneumonia), philanthropic fund contributes up to a third of global research expenditure, according to the G-FINDER data base [22]. Moreover, in the context of major research universities in the United States, philanthropies provide up to 30% of the overall research funding [21]. However, few studies have investigated the role of this source of funding on the science enterprise.

In this study, we look at two dimensions of philanthropies' role in science: content and impact. Our benchmarks in this study are the public and private sources of funding and we highlight the differences between philanthropic and these sources of funding. First, we find the focus of philanthropies when investing on health studies. Second, we investigate the relationship between philanthropic, public, and private sources of funding in terms of the contents they tend to support. Finally, we analyze the impact of studies funded by philanthropies through using citations as a proxy, and in comparison with public funding sources. Our study is informed by the publication data (more than 60,000 papers) as reported on the Web of Science data set.

Recipe for success in academia: a dynamic model of scientists' career

In the third study (chapter 4), we focus on the career of scientists and their success. Today, scientists need to spend a significant amount of time on writing grant proposals to secure funding for their research labs [23]–[25]. With the persistent decrease in chances of receiving funding over the recent years, the burden has become heavier on scientists to invest more time on writing higher quality proposals in the hope of receiving funding for their research [26]. But there is an important question that remains unanswered: how much of a scientist's research time should be allocated to writing grant proposals in order to maximize her career-long publications?

In this study, we build a simple mathematical model of scientist's research activities. In this model, any scientist has to allocate their limited time between two activities: writing grant proposals and writing papers. Everything else kept constant, more time on paper results in more research output, and more time on proposal results in more new funding. If we consider the case of a scientist who aims at maximizing their research payoff over their career, spending time on writing papers provides them with immediate benefit aligning with this long-term goal. However, focus on writing papers means less time on proposals. Research output depends on time spent on writing papers as well as funding. Moreover, proposal competitiveness depends on one's past accomplishments.

We analyze how different schemes of funding distribution affect the optimal strategy and career-long success of scientists. We evaluate different factors such as the effect of randomness in evaluation of grant proposals and competition on receiving funding. Our model offers insights into how such institutional and contextual factors affect the career of a scientist and her optimal or equilibrium strategy.

Research approaches

In this dissertation, we use a wide range of methods to tackle our research questions. In a very aggregate view, as shown in Figure 1.1Figure , we use data analytics tools and regression analysis in essays 1 and 2. In essay 3, we conduct our analysis by using mathematical dynamic modeling and game theory approaches.

In essays 1 and 2, we use a topic modeling technique to read through two large sets of publication data, each with hundreds of thousands of papers. Specifically, we use a Bayesian statistical technique, known as Latent Dirichlet Allocation (LDA), that helps find the latent topics that generate a set of documents [27]. The underlying assumption of this method is that documents consist of probability distributions over a set of topics, while topics are probability distributions over the set of unique words that generate the entire set of documents. The output of an LDA implementation over a set of documents includes two pieces: a probability distribution over topics for each of the documents and a probability distribution over all unique words within the data set for each of the topics. The first output helps us know which topics are generating a given document and the second output helps us define each of the topics. LDA has been applied in a wide range of studies, including science and innovation studies [28]–[31].

We also use a range of different econometric analysis in this dissertation. In Essay 1, panel data regression analysis is used to analyze the relationship between countries' focus on the socio-behavioral aspects of HIV/AIDS and the level of HIV/AIDS mortality rate over time. In Essay 2, we use negative binomial and logistic regression models to analyze the citations of papers funded by different sources of funding.

In essay 3, we create a dynamic model for analyzing a scientist's career. We focus on the strategies that a scientist can choose to allocate her limited research time between different two distinct activities that are crucial in nowadays' science enterprise: writing papers and writing grant proposals. Our goal is to find the optimal strategy that the scientist can follow to maximize her long-term research output under different organizational contexts. We further numerically find the Nash equilibrium strategy for a population of scientists competing with each other for funding.

In the next tree chapters, I present each of the essays. The final chapter of this dissertation summarizes and integrates the main findings of this dissertation, discusses its scientific contributions, and its potential broader impacts on the science community and the society.

References

[1] B. Alberts, M. W. Kirschner, S. Tilghman, and H. Varmus, "Rescuing US biomedical research from its systemic flaws.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 16, pp. 5773–7, 2014.

[2] N. Ghaffarzadegan, J. Hawley, R. Larson, and Y. Xue, “A Note on PhD Population Growth in Biomedical Sciences,” *Syst. Res. Behav. Sci.*, vol. 32, pp. 402–405, 2015.

[3] R. C. Larson, N. Ghaffarzadegan, and M. G. Diaz, “Magnified Effects of Changes in NIH Research Funding Levels.,” *Serv. Sci.*, vol. 4, no. 4, pp. 382–395, 2012.

[4] R. C. Larson, N. Ghaffarzadegan, and Y. Xue, “Too many PhD graduates or too few academic job openings: The basic reproductive number R0 in academia,” *Syst. Res. Behav. Sci.*, vol. 31, no. 6, pp. 745–750, 2014.

[5] D. J. H. Mathews, G. D. Graff, K. Saha, and D. E. Winickoff, “Access to stem cells and data: persons, property rights, and scientific progress.,” *Science*, vol. 331, no. 6018, pp. 725–7, 2011.

[6] G. Paraje, R. Sadana, and G. Karam, “Increasing international gaps in health-related publications,” *Science (80-.).*, vol. 308, no. 5724, pp. 959–960, 2005.

[7] M. S. Teitelbaum, “Structural Disequilibria in Biomedical Research,” *Science (80-.).*, vol. 321, no. 5889, pp. 644–645, 2008.

[8] S. Wuchty, B. F. Jones, and B. Uzzi, “The Increasing Dominance of Teams in Production of Knowledge,” *Science (80-.).*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[9] H. Hur, N. Ghaffarzadegan, and J. Hawley, “Effects of government spending on research workforce development: Evidence from biomedical postdoctoral researchers,” *PLoS One*, vol. 10, no. 5, pp. 1–16, 2015.

[10] K. Vakili, A. M. McGahan, R. Rezaie, W. Mitchell, and A. S. Daar, “Progress in human embryonic stem cell research in the United States between 2001 and 2010,” *PLoS One*, vol. 10, no. 3, pp. 1–8, 2015.

[11] P. Bourdieu, *Science of Science and Reflexivity*. The University of Chicago Press, 2004.

[12] D. Sarewitz and R. A. Pielke, “The neglected heart of science policy: reconciling supply of and demand for science,” *Environ. Sci. Policy*, vol. 10, no. 1, pp. 5–16, 2007.

[13] E. C. McNie, “Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature,” *Environ. Sci. Policy*, vol. 10, no. 1, pp. 17–38, 2007.

[14] J. Vandemoortele and E. Delamonica, “The ‘Education Vaccine’ Against HIV,” *Curr. Issues Comp. Educ.*, vol. 3, no. 1, pp. 6–13, 2000.

[15] H. Kohler, J. R. Behrman, and S. C. Watkins, “Social Networks and HIV / AIDS Risk Perceptions,” *Demography*, vol. 44, no. 1, pp. 1–33, 2007.

[16] J. Fisher and W. Fisher, “Changing AIDS risk behavior,” *Psychol. Bull.*, vol. 111, no. 3, pp. 455–74, 1992.

[17] S. C. Kalichman and L. C. Simbayi, “HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a black township in Cape Town, South Africa.,” *Sex. Transm. Infect.*, vol. 79, no. 6, pp. 442–7, 2003.

[18] R. Parker and P. Aggleton, “HIV and AIDS-related stigma and discrimination: A conceptual framework and implications for action,” *Soc. Sci. Med.*, vol. 57, no. 1, pp. 13–24, 2003.

[19] J. Gonzalez, F. Penedo, M. Antoni, R. Duran, M. Fernandez, S. McPherson-Baker, G. Ironson, N. Klimas, M. Fletcher, and N. Schneiderman, “Social Support, Positive States of Mind, and HIV Treatment Adherence in Men and Women Living With HIV/AIDS.,” *Heal. Psychol*, vol. 23, no. 4, pp. 413–418, 2004.

[20] A. Neaigus, S. R. Friedman, R. Curtis, D. C. Des Jarlais, R. Terry Furst, B. Jose, P. Mota, B. Stepherson, M. Sufian, T. Ward, and J. W. Wright, “The relevance of drug injectors’ social and risk networks for understanding and preventing HIV infection,” *Soc. Sci. Med.*, vol. 38, no. 1, pp. 67–78, 1994.

[21] F. Murray, “Evaluating the Role of Science Philanthropy in American Research Universities,” *Innov. Policy Econ.*, vol. 13, no. 1, pp. 23–60, 2013.

[22] “G-FINDER.” [Online]. Available: http://policycures.org/gfinder.html.

[23] R. J. Daniels, “A generation at risk: Young investigators and the future of the biomedical workforce,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 2, pp. 313–318, 2015.

[24] D. L. Herbert, A. G. Barnett, P. Clarke, and N. Graves, “On the time spent preparing grant proposals: an observational study of Australian researchers,” *BMJ Open*, vol. 3, no. 5, p. e002800, 2013.

[25] R. S. Decker, P. Investigator Leslie Wimsatt, A. G. Trice, and J. A. Konstan, “A PROFILE OF FEDERAL-GRANT ADMINISTRATIVE BURDEN AMONG FEDERAL DEMONSTRATION PARTNERSHIP FACULTY A Report of the Faculty Standing Committee of the Federal Demonstration Partnership,” no. January, 2007.

[26] F. C. Fang and A. Casadevall, “Reforming science: Structural reforms,” *Infect. Immun.*, vol. 80, no. 3, pp. 897–901, 2012.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.

[28] T. L. Griffiths and M. Steyvers, “Finding scientific topics.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101 Suppl, pp. 5228–35, 2004.

[29] S. Kaplan and K. Vakili, “The double-edged sword of recombination in breakthrough innovation,” *Strateg. Manag. J.*, vol. 36, pp. 1435–1457, 2015.

[30] J. Adams and R. Light, “Mapping interdisciplinary fields: Efficiencies, gaps and redundancies in HIV/AIDS research,” *PLoS One*, vol. 9, no. 12, pp. 1–13, 2014.

[31] R. Light and jimi adams, “Knowledge in motion: the evolution of HIV/AIDS research,” *Scientometrics*, vol. 107, no. 3, pp. 1227–1248, 2016.

Chapter 2: Global Trends and Regional Variations in Studies of HIV/AIDS[1]

Abstract

We conduct textual analysis of a sample of more than 200,000 papers written on HIV/AIDS during the past three decades. Using the Latent Dirichlet Allocation method, we disentangle studies that address behavioral and social aspects from other studies and measure the trends of different topics as related to HIV/AIDS. We show that there is a regional variation in scientists' approach to the problem of HIV/AIDS. Our results show that controlling for the economy, proximity to the HIV/AIDS problem correlates with the extent to which scientists look at the behavioral and social aspects of the disease rather than biomedical.

Introduction

Since it was first detected in the early 1980s, AIDS has been studied by scientists around the world. Biomedical and epidemiological scientists have made substantial progress in understanding and controlling the disease over time. In the early years, HIV was discovered to be the cause of AIDS[1]–[3]. Later, scientists uncovered the modes of transmission and developed blood tests for diagnosing HIV[1]–[3]. By the early 1990s, the first antiviral drugs were developed to suppress HIV[1]–[3]. Together, these developments and efforts to educate vulnerable populations about the transmission mechanisms slowed down the progress of disease during the mid-1990s, at least in the United States. Yet, it was also in the decade of the 1990s that the HIV/AIDS epidemic reached its peak in Sub-Saharan Africa[1]. Eventually, as a result of global and regional efforts to control the epidemic, both the HIV incidence and AIDS-related death rates have dropped after the year 2000[4].

The complexity of the HIV/AIDS problem goes beyond biomedical aspects of the disease. Many challenges are behavioral and social, such as public awareness of the disease[5], risk perception[6], high-risk behaviors[7], willingness to be tested[8], social stigma[9], and treatment adherence[10]. The infectious nature of the disease also adds to the social network complexity of the problem and to infectious patterns[11]. Numerous multi-disciplinary studies have been conducted to uncover these socio-behavioral aspects of the disease.

From the early years, behavioral studies have shed light on the psychological factors corresponding to risky behaviors that can lead to HIV infection[12]. Along the same line of research, preventive behavioral studies have focused on reducing the chances of being infected by HIV and alleviating the adverse outcomes for patients with HIV[13]. These studies have examined how people can avoid risky behaviors by becoming more knowledgeable, motivated, and capable in the context of avoiding HIV infection[7]. HIV/AIDS is a disease with strong social stigma attached to it in many societies. Sociologists and anthropologists have contributed to HIV/AIDS research by proposing theories and models to understand and overcome this stigma[9]. Overall, the behavioral and social sciences (BSS) studies have helped in shaping a better understanding of the disease, recognizing people with the highest risk, designing preventive programs, and maintaining medical treatment for current patients.

[1] This essay is published in the journal of *Scientific Reports*.

Cumulatively, studies conducted by scientists within the fields of biomedical sciences and BSS have shaped our understanding of HIV/AIDS as a global problem. Despite this ostensible shared understanding, the knowledge production process orbiting around the problem of HIV/AIDS has happened incrementally over decades of research and within the context of scientific enterprise in many countries around the world. Yet, little is known about the trends of research and shifts in focus of researchers as related to the problem of HIV/AIDS. In this study, we investigate shifts in the scientific community's focus on BSS aspects of HIV/AIDS through a textual analysis of more than 200,000 HIV/AIDS publications in the Scopus data set. Our goal is to evaluate global and regional trends and potential variations in research. One finding of our study is that, in the case of HIV/AIDS, the share of BSS research has been increasing over time and, controlling for the economy, geographical proximity to the problem is associated with more BSS studies.

Data: HIV/AIDS Publications between 1985 and 2012

We constructed a dataset of all academic papers on the topic of HIV/AIDS as reported in the Scopus data set, totaling 264,102 papers. These papers were published on a variety of subjects related to HIV/AIDS during the period 1985 to 2012. Our criteria for inclusion of the papers in our dataset were existence of the words "HIV" or "AIDS" in the title or abstract of papers. For the analysis, we used the abstract of the papers, year of publication, and location of authors based on their institutional affiliations. Locations of authors were used to assign a paper to one or multiple countries. For example, a paper with authors from three different countries was assigned to those three countries. In the first stage, we excluded papers lacking an abstract, which reduced our dataset to a total of 209,608 papers.

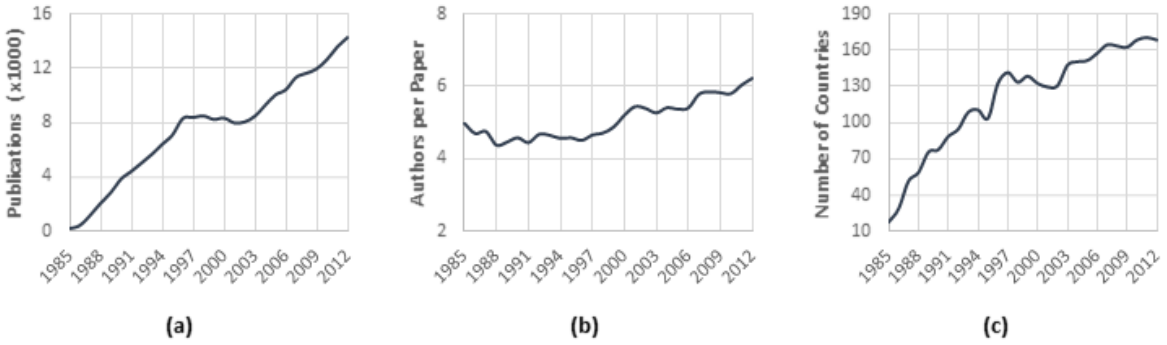


Figure 2.1: Trends in HIV/AIDS Publications: (a) global trend of publication; (b) average number of authors per paper; (c) number of countries contributing to HIV/AIDS research.

We first conducted a set of descriptive analysis at the aggregate level, as reported in Figure 2.1. Figure 2.1a shows that the annual publication of scientific research on HIV/AIDS has been increasing since 1985 except for the period between 1996 and 2001. In 2012, the annual publication rate was about 14,250 papers per year, which shows almost 80% growth in a decade. Figure 2.1b reports collaboration patterns. The average number of authors per paper increased over time, indicating that research in the field has become more collaborative. In particular, the average number of authors per paper increased from about 4.2 in the 1980s and 1990s to about 6.2 authors per paper in 2012. Moreover, as Figure 2.1c shows, more countries have been involved in these studies over time, and it is fair to say that the problem of HIV/AIDS has become a global field of research. As the figure shows, scientists in 168 countries had at least one publication on this topic in 2012.

Increasing Focus on BSS Aspects of HIV/AIDS

In the next step, we conducted textual analysis, looking at the abstract of HIV/AIDS publications over time. Methodologically, we implemented the Latent Dirichlet Allocation (LDA) on our data set. More information about LDA is provided in the method section; in short, in the LDA method, topics are identified based on the frequency of appearance and co-appearance of different words within all the documents. Our textual analysis of all paper abstracts shows an increasing trend in the share of BSS among HIV/AIDS publications (Figure 2.2). In 2012, the share of BSS topics in HIV/AIDS research publications was more than 40%.

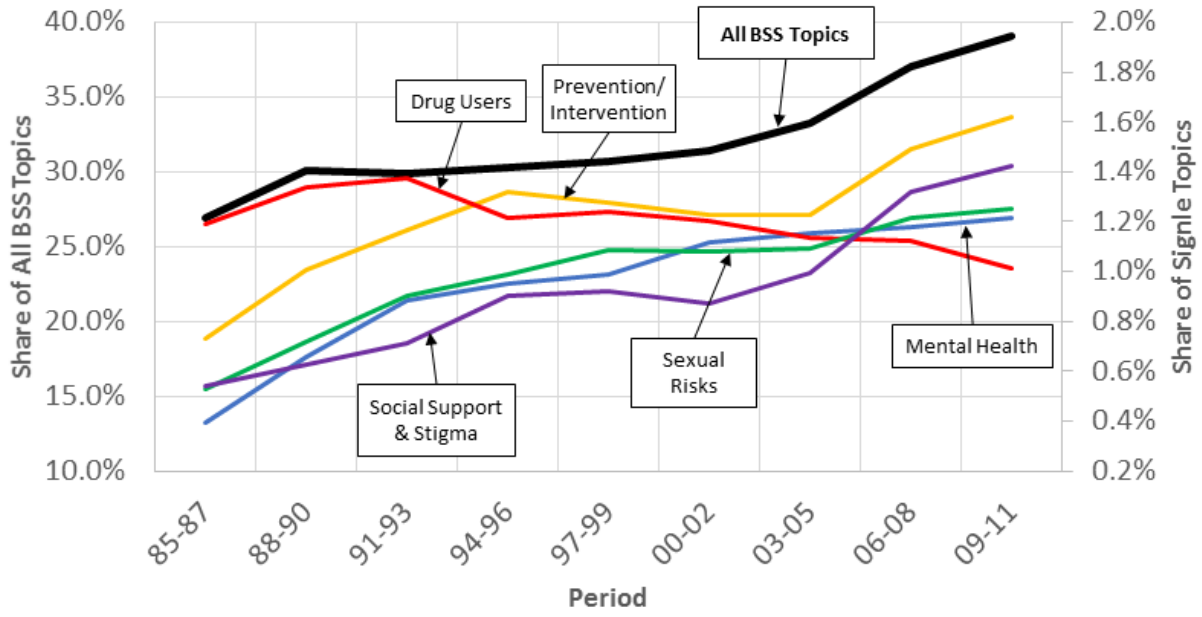


Figure 2.2: Share of aggregate BSS topics (left axis) and five individual topics (right axis) in HIV/AIDS research over time.

We also tracked the trends of individual BSS topics over time. As examples, Figure 2.2 shows the share of five individual topics in HIV/AIDS research (i.e., Prevention/Intervention, Mental Health, Sexual Risks, Social Support & Stigma, and Drug Users); the remainder are reported in the SI. Some of these topics are gaining increasing attention in the scientific community. For example, the Prevention/Intervention, Mental Health, Sexual Risks, and Social Support & Stigma topics are all increasingly studied. Some other BSS topics, such as Drug Users show a declining trend in the share of publications in HIV/AIDS field since the beginning of the 21st century.

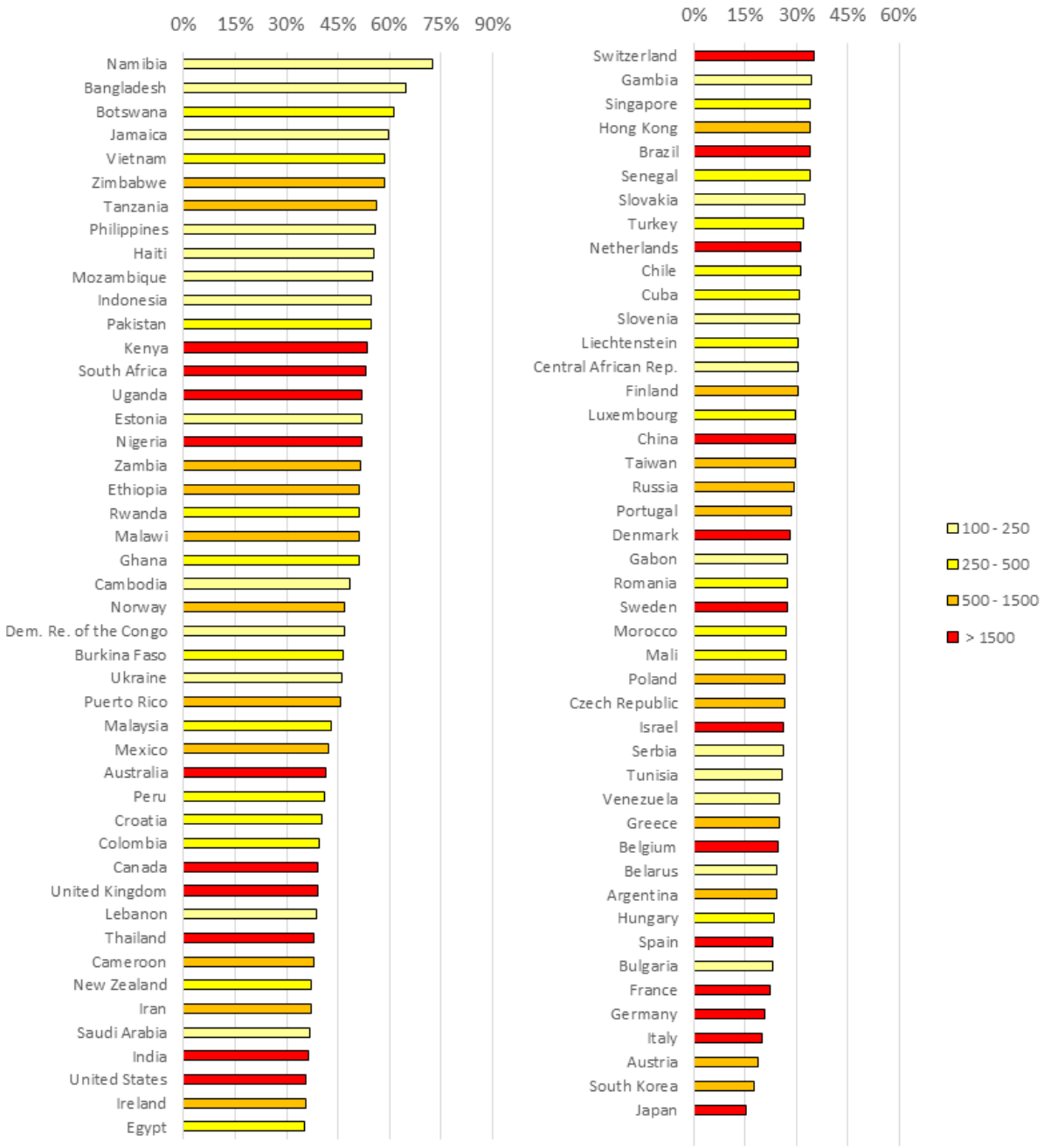


Figure 2.3: Share of BSS research in all HIV/AIDS papers of different countries. Note: For example, share of BSS research for Ethiopia is shown at 51% meaning that 51% of HIV/AIDS research content published by Ethiopian authors is on BSS topics. Only countries are shown that have published more than 100 papers during the period 1985 to 2012. Color coded by total number of HIV/AIDS papers.

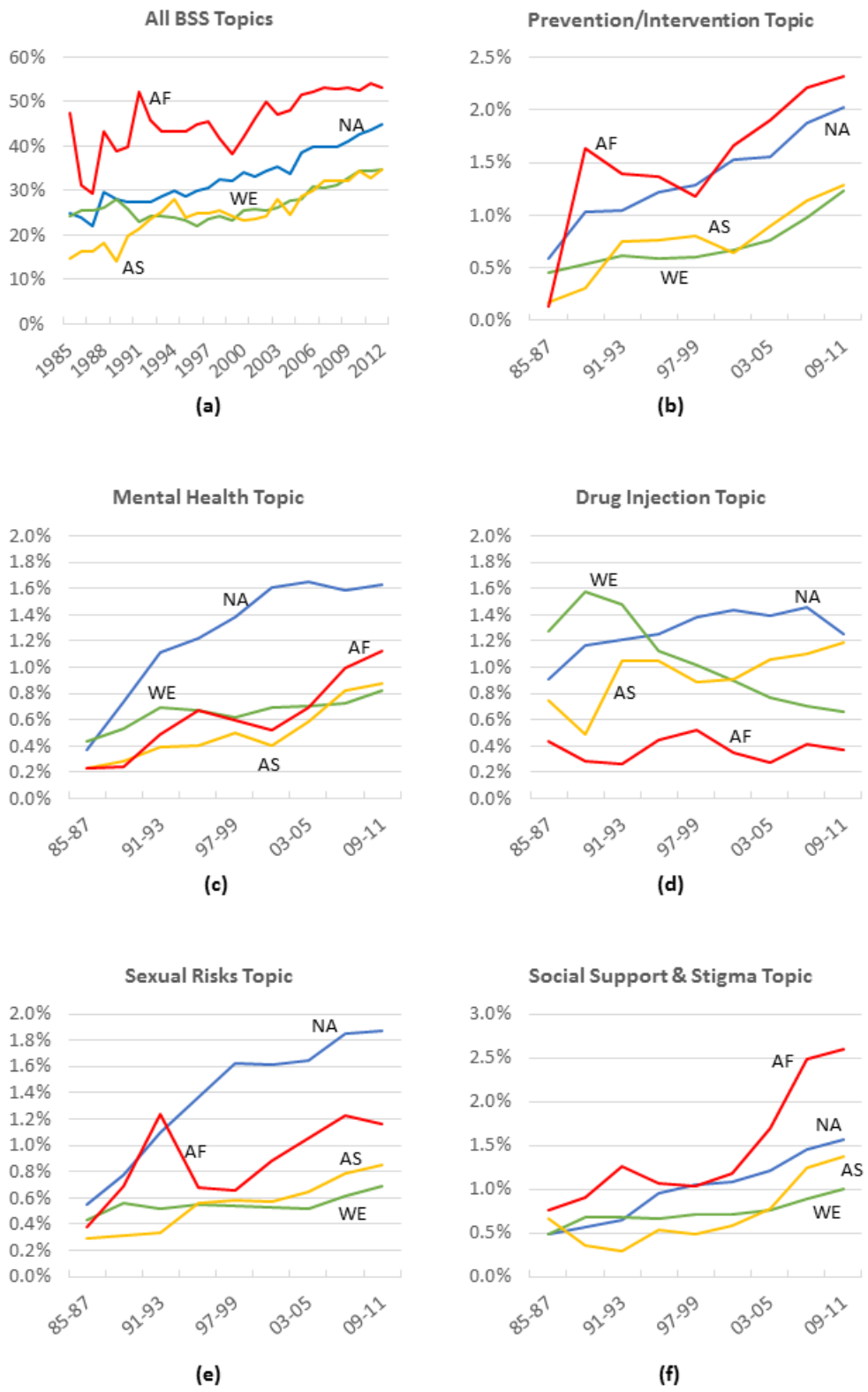


Figure 2.4: Regional variation in trends of BSS topics in the context of HIV/AIDS research. NA: North America, WE: Western Europe, AS: South and Southeast Asia, AF: Sub-Saharan Africa.

Regional Variation in BSS Focus of HIV/AIDS Studies

Different countries have shown different level of focus on BSS in their HIV/AIDS research portfolio, as Figure 2.3 illustrates. The share of BSS in HIV/AIDS publications from Namibia, Bangladesh, and Botswana is above 60%. Meanwhile, this proportion is less than 20% in Austria, South Korea, and Japan.

We observe a regional variation in BSS study trends, as Figure 2.4 shows. Figure 2.4a shows varying regional trajectories in BSS studies of HIV/AIDS. In Sub-Saharan Africa, for instance, far more attention has been paid to BSS studies than elsewhere. The share of BSS studies among all HIV/AIDS research publications of Sub-Saharan African countries has been above 50% since 2005. In other regions, the share of BSS research has been increasing. The growth rate has been faster in South and Southeast Asia (with an average growth rate of 0.66% per year) and North America (with an average growth rate of 0.73% per year) compared with Western Europe (with an average growth rate of 0.34% per year).

Regional variation also exists among individual topics. Figures 2.4b-f demonstrate these variations for five sub-topics of BSS (i.e., Prevention/Intervention, Mental Health, Sexual Risks, Social Support & Stigma, and Drug Users). We observe that Sub-Saharan Africa and Northern America have shown more interest in Intervention/Prevention than Western Europe and South and Southeast Asia. On the topic of mental health and sexual risks as related to HIV/AIDS, the attention of North American scientists has been the highest in comparison to other regions. We observe that Western Europe showed a high interest in topic of Drug Users in the earlier years, but this interest has recently faded. All regions have shown an increasing interest in the Social Stigma topic since 1985. The trend has been increasing at a higher rate for Sub-Saharan Africa.

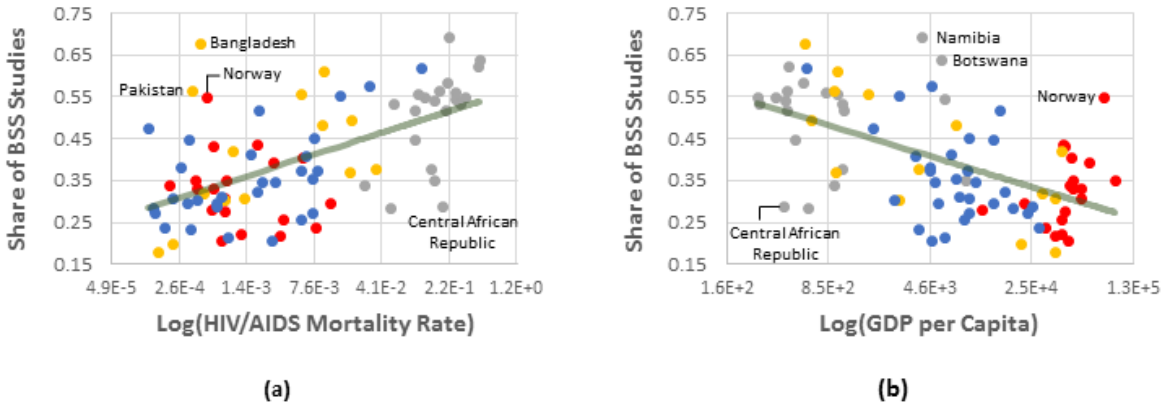


Figure 2.5: BSS focus of publications from different countries (average during the period 2006 to 2010) compared to their HIV/AIDS mortality rate in 2005 and GDP per capita (average during 2006 to 2010). Each dot corresponds to a country (Red: North America & Western Europe; Yellow: Southern & Southeast Asia; Gray: Sub-Saharan Africa; Blue: Rest of the world). The solid lines show the correlation lines. Countries that are two standard deviation further from the correlation line are labeled.

Why does this variation between different regions in terms of their focus on the BSS aspects of HIV/AIDS exist? To study what drives BSS focus in HIV/AIDS research, we perform a panel data regression analysis. The factors we include in our analysis are the HIV/AIDS mortality rate and

GDP per capita. We expect HIV/AIDS mortality to be associated with how researchers and science policy makers perceive the problem of HIV/AIDS. We represent the magnitude of the problem by HIV/AIDS mortality rate using the HIV mortality data from the "Global burden of disease study 2013" [Available from http://ghdx.healthdata.org/global-burden-disease-study-2013-gbd-2013-data-downloads]. Our BSS-related question is whether the higher rate of the problem is associated with higher BSS studies.

One other potential explanation for research focus is the research capacity or financial resources available for research. Biomedical research is more costly. By having more resources available for research, we expect that affluent countries be more capable of conducting biomedical research. This can affect the relative share of BSS in research portfolio of these countries. In our analysis, we control for GDP per capita as a proxy for countries' economic prosperity.

Our regression analysis shows there is a positive association between countries' HIV/AIDS mortality rate and the ratio of BSS topics in their research publications on HIV/AIDS ($p<0.001$). The detailed results are reported in the methods section. For the sake of simplicity, Figure 2.5 depicts the correlation ($\rho = 0.58$ for relationship between percentage of BSS and HIV/AIDS mortality rate and $\rho = -0.56$ for relationship between percentage of BSS and GDP per capita). Our analysis also provides evidence that supports a negative correlation between GDP per capita in countries and the ratio of BSS topics in their publications ($p<0.001$).

Discussion

By constructing and analyzing a large data set of more than 200,000 papers, and by implementing a statistical natural language processing method (Latent Dirichlet Allocation), we were able to investigate the trends of studies in HIV/AIDS research in different regions. We found interesting trends in behavioral and social studies of HIV/AIDS. We observe that there is variation among different countries in terms of their focus on HIV/AIDS. The BSS focus has, on average, been the highest in Sub-Saharan Africa. Furthermore, we find an association between HIV/AIDS mortality and the percentage of BSS topics in HIV/AIDS research.

If we assume scientists' responses to a problem are informed by the nature of the problem they observe in their proximity, our findings will have several implications. The scientific community's focus on the "same" medical disease may vary across different regions and times, and this could speak to the varying nature of a seemingly similar disease in different regions and different time periods. If one assumes that a society's understanding of a problem materializes in the form of scientific publications, our findings imply that the problem of HIV/AIDS is conceived to be more behavioral and social in places with higher mortality rates. The takeaway from this finding for science policy makers is that in defining priorities of research for a specific problem, the problem should not be looked at from an isolated perspective. To the contrary: it is of the utmost importance to nail down the problem in the context of each country and set priorities based on that perspective.

Our findings can be seen in the light of several prior science policy studies. The concept of "reconciling supply of science with societies' demand"[14], [15] implies that scientists' approach to study a disease can partially reflect societal demand. Under this concept, a higher focus on the BSS aspects of HIV/AIDS in a country implies that those aspects of HIV/AIDS are perceived to be more critical in that region. Such premise seems reasonable: the broadly adverse social, cultural, and economic effects of HIV/AIDS on regions with high prevalence has made it a challenging epidemic to be tackled in those regions[16]. On the other hand, in low prevalence

countries, more focus on biomedical studies may reflect the different nature of the same disease in those regions. In countries with less significant health challenges we often observe higher literacy, better quality of life, and more awareness about health. Consequently, less stress is placed on behavioral and social interventions than biomedical solutions. This premise is in line with medicalization and biomedicalization of health as studied by sociologists of health (e.g.,[17]). We have also shown a negative association between GDP per capita and the ratio of BSS research in the contest of HIV/AIDS. This result is in line with previous studies that have shown affluent countries publish more biomedical papers[18]. Furthermore, our findings resonate with recent arguments on significant contributions of behavioral and social sciences to health[19].

We focused on journal publications, but we admit that they are not the ideal representation of all research activities around a topic. Some research studies may not turn to journal publications and some may fail to get published. Different journals and different fields also have different standards and norms for publications. We also acknowledge that GDP per capita is not an accurate proxy for research spending of different countries. However, as used in other past science policy studies (e.g.,[18]), GDP is a well-established measure for economy with relatively accurate data. Many other measures (such as research spending) are inconsistently calculated for different countries and, moreover, are not available for all countries. We think these limitations can be addressed in other future studies. Our study was a first major step and benefited from incorporating data analytics to uncover behavioral and social science contributions to health. We report details of our constructed data set in SI.

Methods

Topic Modeling

Different techniques can be used to extract the BSS theme within publications on HIV/AIDS. Simple approaches include inference solely based on keywords and authors' departmental affiliations. These are easy to implement but have limitations. Assuming keywords are a sufficient description of content, and that different keywords have the same weight in describing an entire paper, is too simplistic. Departmental affiliation is not necessarily an accurate indicator of an author's research; for example, faculty members in a public health department may do different types of research and address different topics. Furthermore, the relative weight of each keyword or contribution of each author in a given publication is unclear; is a paper with three authors, each one affiliated with three different departments of psychology, medicine, and public health, a BSS paper? If so, what topic within BSS? One needs to go deeper into the content of the paper to answer these questions.

Advancements in the data sciences help us conduct a more precise analysis. We address this methodological issue by using a Bayesian statistical technique known as Latent Dirichlet Allocation (LDA) that helps find the latent topics that generate a set of documents [20]. The underlying assumption of this method is that documents consist of probability distributions over a set of topics, while topics are probability distributions over the set of unique words (or corpus) that generate the entire set of documents. The output of an LDA implementation over a set of documents includes two pieces: a probability distribution over topics for each of the documents and a probability distribution over all unique words within the data set for each of the topics. The first output helps us know which topics are generating a given document and the second output helps us define each of the topics. LDA has been applied in a wide range of studies, including science and innovation studies [21]–[24]. For more information on the method, see [20].

In this paper, we use the abstracts of papers as the set of documents to implement LDA. After removing the common words in English by using a standard stop list, we generated the corpus of unique words that generates all abstracts in our data set. The size of this corpus was 17,783,690 unique words. In our analysis we limited the set of unique words only to the ones that have been repeated in at least 100 abstracts. By applying this rule, the size of our corpus was reduced to 11,192 unique words. To implement LDA, three parameters need to be specified: the number of topics, a topic smoothing coefficient, and a term smoothing coefficient. To have a model with a better fit, using large number of topics is recommended. This, however, poses the problem of confronting topics that are indistinguishable in a meaningful way for human understanding [25]. In some prior topic modeling applications in the context of science and innovation [21], [22], researchers have chosen to use 100 topics for their analysis. Following their lead, we also chose 100 topics to implement LDA in our data set. For topic and term smoothing parameters, 0.1 has typically been chosen as the default value [21], [22]. However, since we are looking for more finely separated topics and since the topic space of our documents is narrower (all papers are about HIV/AIDS), we reduced the term smoothing parameter to 0.01. This helps us have more discrete topics [21]. We used the python LDA 1.0.4 package to perform our analysis [Available at https://pypi.python.org/pypi/lda.].

Many of the topics generated by the LDA do not necessarily show a discipline or field of research. In fact, each abstract is a mixture of all the topics with different probabilities. However, we expect that, on average, topics belonging to BSS studies will have higher probabilities assigned to them in BSS publications. Based on this assumption, we anticipate that pairs of topics related to a specific body of knowledge are more likely to have co-occurrence probability (defined as the multiplication of the individual probabilities) in a publication compared with a pair of topics each related to a different fields of study. By making this assumption, we can look at the co-occurrence network of the topics, as uncovered by LDA, to find fields of research within our publication data set.

We generate the topic network by considering the five most probable topics of each document. Each of these five have, on average, a probability higher than 5%. Moreover, their average cumulative probability is greater than 70%. In our topic network, each node represents a topic. Topics are connected to each other based on their co-occurrence probability in all papers. The Louvain clustering [26] was implemented on this network to find the communities of topics corresponding to fields of research. Three clusters were found. Cluster 1 includes topics related to medical case studies; Cluster 2 includes BSS topics; and cluster 3 includes biomedical topics. The individual topics and these clusters are reported in the SI. The robustness of these clusters against variability of input data was confirmed through sensitivity analysis for random sub-samples of papers. Results are presented in the SI. As another robustness check, we also generated a network with the eleven most probable topics of each paper. Each of these eleven topics, on average, have probabilities higher than 1%. Performing Louvain clustering on this network provided us with very similar clusters.

Finally, we would like to mention that a benefit of using LDA is to count for multidisciplinary works. Since in this method a topic of a paper is a probability distribution over different potential topics, a paper can be considered as x% BSS and y% biomedical. These percentages are estimated based on the words used in the abstracts.

Panel Regression Analysis

To investigate the variation among different countries in terms of their focus on BSS aspects of HIV/AIDS, we performed a panel data regression analysis. As mentioned earlier, we had access to publication data for the period 1985 to 2012. The HIV/AIDS mortality rate estimates are available for the years 1990, 1995, 2000, 2005, 2010, and 2013. In our analysis, we regressed the average share of socio-behavioral topics during years *i+1* to *i+5* on the HIV/AIDS mortality rate of countries in year *i* controlling for their average GDP per capita during those years. Overall, our analysis includes 84 panels (countries with more than 100 papers on the topic HIV/AIDS during the period 1985-2012) for which we have four observations (*i* took the values 1990, 1995, 2000, and 2005).

Table 2.1: Panel data regression results for percentage of BSS topics.

		Dependent Variable: Percentage of BSS topics in HIV/AIDS Research	
		Model 1	Model 2
Independent Variables	Log(HIV/AIDS Mortality)	0.016**	0.014**
		(0.003)	(0.002)
	Log(GDP per Capita)	-0.027**	-0.023**
		(0.005)	(0.003)
	Intercept	0.679**	0.606**
		(0.009)	(0.026)
	$F(2, 83)$	12042.55***	
	R^2	0.2807	
	$Wald\ \chi^2$		172.17***

* p < 0.05 ** p < 0.001 *** p < 0.0001

Standard errors are presented in parentheses.

Table 2.1 reports the summary of our regression analysis. We used a pooled least squares regression with Driscoll-Kraay standard errors [27], [28] in our main model (Model 1). Due to the nature of our problem, we are concerned about heteroscedasticity and autocorrelation in our regression analysis. However, the selected model is robust against heteroskedasticity and autocorrelation. Other potential models such as Huber-White and Newey-West have also been used to confirm our regression results (see the results in SI). Due to the small number of observations per panel, we did not use a fixed effect model in our analysis, as we did not have enough variation. We tested for multicollinearity in our data and did not find an evidence for a potential problem (un-centered VIF: 1).

As a robustness check for our model selection, we performed the same analysis in Model 2 with another approach. In Model 2, we used a generalized least square model with heteroskedastic panels and first order autoregression. As Table 2.1 shows, the results from models 1 and 2 are very similar. However, since it is recommended to use the generalized least square model when the number of observations per panel is larger than the number of panels [21], we selected Model

1 as our primary model in this paper. Overall, our models offer consistent results showing a positive association between HIV/AIDS mortality rate of countries and their focus on BSS aspects of HIV/AIDS. Our models also provide evidence to support a negative association between GDP per capita of countries and share of BSS in their HIV/AIDS publications.

Acknowledgments

The National Institute of General Medical Sciences and the Office of Behavioral and Social Sciences Research of the National Institutes of Health (NIH) supported this work (Grant 2U01GM094141-05). We thank Richard Larson (MIT), Keyvan Vakili (London Business School), Michael Spittel (NIH), Stephen Marcus (NIH), and Griffin Weber (Harvard) for helpful comments.

References

[1] K. A. Sepkowitz, “AIDS - The First 20 Years,” *he New Engl. J. Med.*, vol. 344, no. 23, pp. 1764–1772, 2001.

[2] A. S. Fauci, “HIV SPECIAL HIV and AIDS : 20 years of science,” *Nat. Med.*, vol. 9, no. 7, pp. 839–843, 2003.

[3] R. C. Gallo, “A reflection on HIV/AIDS research after 25 years.,” *Retrovirology*, vol. 3, no. May 1982, p. 72, 2006.

[4] “World Health Organization fact sheet on HIV/AIDS.” [Online]. Available: http://www.who.int/mediacentre/factsheets/fs360/en/.

[5] J. Vandemoortele and E. Delamonica, “The ‘Education Vaccine’ Against HIV,” *Curr. Issues Comp. Educ.*, vol. 3, no. 1, pp. 6–13, 2000.

[6] H. Kohler, J. R. Behrman, and S. C. Watkins, “Social Networks and HIV / AIDS Risk Perceptions,” *Demography*, vol. 44, no. 1, pp. 1–33, 2007.

[7] J. Fisher and W. Fisher, “Changing AIDS risk behavior,” *Psychol. Bull.*, vol. 111, no. 3, pp. 455–74, 1992.

[8] S. C. Kalichman and L. C. Simbayi, “HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a black township in Cape Town, South Africa.,” *Sex. Transm. Infect.*, vol. 79, no. 6, pp. 442–7, 2003.

[9] R. Parker and P. Aggleton, “HIV and AIDS-related stigma and discrimination: A conceptual framework and implications for action,” *Soc. Sci. Med.*, vol. 57, no. 1, pp. 13–24, 2003.

[10] J. Gonzalez, F. Penedo, M. Antoni, R. Duran, M. Fernandez, S. McPherson-Baker, G. Ironson, N. Klimas, M. Fletcher, and N. Schneiderman, “Social Support, Positive States of Mind, and HIV Treatment Adherence in Men and Women Living With HIV/AIDS.,” *Heal. Psychol*, vol. 23, no. 4, pp. 413–418, 2004.

[11] A. Neaigus, S. R. Friedman, R. Curtis, D. C. Des Jarlais, R. Terry Furst, B. Jose, P. Mota, B. Stepherson, M. Sufian, T. Ward, and J. W. Wright, “The relevance of drug injectors’ social and risk networks for understanding and preventing HIV infection,” *Soc. Sci. Med.*, vol. 38, no. 1, pp. 67–78, 1994.

[12] J. A. Kelly, D. A. Murphy, K. J. Sikkema, and S. C. Kalichman, “Psychological interventions to prevent HIV infection are urgently needed: new priorities for behavioral research in the second decade of AIDS.,” *Am. Psychol.*, vol. 48, no. 10, pp. 1023–1034, 1993.

[13] J. a Kelly and S. C. Kalichman, “Behavioral research in HIV/AIDS primary and secondary prevention: recent advances and future directions.,” *J. Consult. Clin. Psychol.*, vol. 70, no. 3, pp. 626–639, 2002.

[14] D. Sarewitz and R. A. Pielke, “The neglected heart of science policy: reconciling supply of and demand for science,” *Environ. Sci. Policy*, vol. 10, no. 1, pp. 5–16, 2007.

[15] E. C. McNie, “Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature,” *Environ. Sci. Policy*, vol. 10, no. 1, pp. 17–38, 2007.

[16] P. Piot, M. Bartos, P. D. Ghys, N. Walker, and B. Shwartlander, “The global impact of HIV,” *Nature*, vol. 410, no. April, pp. 968–973, 2001.

[17] A. Clarke, J. SHIM, M. L, F. J, and F. J, “Biomedicalization: Technoscientific transformations of health, illness, and U. S. biomedicine: American Sociological Review,” vol. 68, no. 2, pp. 161–194, 2003.

[18] P. Vinkler, “Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non-EU countries,” *Scientometrics*, vol. 74, no. 2, pp. 237–254, 2008.

[19] H. Hur, M. A. Andalib, J. A. Maurer, J. D. Hawley, and N. Ghaffarzadegan, “Recent trends in the U.S. Behavioral and Social Sciences Research (BSSR) workforce,” *PLoS One*, vol. 12, no. 2, pp. 1–18, 2017.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.

[21] T. L. Griffiths and M. Steyvers, “Finding scientific topics.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101 Suppl, pp. 5228–35, 2004.

[22] S. Kaplan and K. Vakili, “The double-edged sword of recombination in breakthrough innovation,” *Strateg. Manag. J.*, vol. 36, pp. 1435–1457, 2015.

[23] J. Adams and R. Light, “Mapping interdisciplinary fields: Efficiencies, gaps and redundancies in HIV/AIDS research,” *PLoS One*, vol. 9, no. 12, pp. 1–13, 2014.

[24] R. Light and jimi adams, “Knowledge in motion: the evolution of HIV/AIDS research,” *Scientometrics*, vol. 107, no. 3, pp. 1227–1248, 2016.

[25] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *Adv. Neural Inf. Process. Syst. 22*, pp. 288--296, 2009.

[26] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” 2008.

[27] D. Hoechle, “Robust standard errors for panel regressions with cross-sectional dependence,” *Stata J.*, vol. 7, no. 3, pp. 281–312, 2007.

[28] J. C. Driscoll and A. C. Kraay, “Consistent covariance matrix estimation with spatially dependent data,” *Rev. Econ. Stat.*, vol. 80, no. 4, pp. 549–560, 1998.

Appendix A

LDA Topics

In Table A1, we report the topics from our topic modeling analysis. For each topic we report the top 20 most probable words. The clusters in which each topic belongs to is also reported.

Table A1: The detailed information of LDA topics.

Topic	Top 5 Cluster	Top 11 Cluster	Top 20 Words
1	2	2	support, living, people, social, family, hiv, stigma, hiv/aids, study, disclosure, care, status, members, families, aids, hiv-positive, caregivers, health, discrimination, work
2	1	1	oral, malaria, diarrhea, infection, intestinal, hiv, gastrointestinal, lesions, candidiasis, infections, mucosal, diarrhoea, associated, chronic, saliva, parasites, common, sp, stool, cryptosporidium
3	2	2	cost, costs, hiv, estimated, treatment, life, number, impact, cost-effectiveness, model, data, year, analysis, total, methods, coverage, effectiveness, compared, health, estimates
4	2	2	health, hiv/aids, aids, countries, public, policy, development, global, epidemic, international, national, government, people, access, economic, developing, social, policies, prevention, rights
5	1	1	patients, hiv-infected, study, hiv-positive, patient, hiv, methods, compared, clinical, infected, conclusions, higher, conclusion, treated, objective, background, included, hiv-negative, performed, aim
6	3	3	cells, cell, nk, dendritic, dc, expression, surface, molecules, dcs, cd4, receptor, human, adhesion, role, receptors, hiv, expressed, target, dc-sign, complement
7	2	2	care, health, services, medical, providers, service, primary, access, treatment, hiv, management, quality, system, clinics, patient, facilities, clinic, healthcare, provide, public
8	2	2	mortality, disease, diseases, deaths, countries, death, health, morbidity, infectious, developing, infections, major, population, people, burden, hiv/aids, worldwide, life, problem, rates
9	3	3	cells, cd4+, cell, lymphocytes, t-cell, cd8+, cd4, cd8, blood, expression, activation, peripheral, lymphocyte, immune, memory, subsets, flow, individuals, infection, increased
10	3	3	ccr5, cxcr4, receptor, entry, hiv-1, chemokine, receptors, coreceptor, infection, cells, cd4, r5, cell, chemokines, x4, strains, rantes, viral, human, virus

11	2	2	exposure, risk, hiv, transmission, workers, infection, health, occupational, care, blood, dental, exposures, hepatitis, control, medical, procedures, injuries, prophylaxis, exposed, pep
12	3	3	mice, cells, human, bone, marrow, vivo, cell, model, mouse, vitro, stem, hematopoietic, growth, murine, transgenic, normal, rats, studies, days, transplantation
13	2	2	sexual, adolescents, young, youth, sexually, health, adolescent, reproductive, sex, family, contraceptive, transmitted, years, pregnancy, risk, girls, age, contraception, intercourse, education
14	3	3	data, analysis, set, method, methods, based, model, study, models, prediction, developed, correlation, algorithm, validity, approach, predictive, predicted, system, test, reliability
15	1	1	plasma, concentrations, dose, drug, concentration, ritonavir, pharmacokinetic, pharmacokinetics, study, oral, administration, daily, days, indinavir, auc, clearance, doses, day, saquinavir, exposure
16	3	3	development, review, therapeutic, potential, novel, viral, target, understanding, agents, molecular, targets, approaches, drugs, drug, discovery, design, strategies, approach, including, biological
17	1	1	ci, risk, cohort, incidence, mortality, survival, follow-up, months, death, interval, confidence, associated, rate, study, years, ratio, hazard, time, compared, rates
18	2	2	ci, associated, factors, risk, odds, analysis, regression, study, ratio, logistic, multivariate, confidence, age, interval, association, aor, methods, adjusted, conclusions, prevalence
19	1	1	aids, disease, progression, hiv, infection, time, early, diagnosis, clinical, years, individuals, survival, course, late, associated, rapid, death, stage, long-term, infected
20	1	1	levels, serum, higher, level, plasma, concentrations, lower, increased, compared, correlated, high, elevated, low, measured, correlation, controls, concentration, values, normal, markers
21	2	2	increased, increase, time, changes, number, period, rate, decreased, decrease, change, observed, rates, incidence, decline, increasing, increases, year, trends, proportion, remained
22	1	1	patients, hospital, study, cases, years, age, medical, period, clinical, diagnosis, retrospective, january, objective, total, clinic, methods, records, university, diagnosed, included
23	1	1	renal, disease, patients, risk, kidney, hiv, failure, hypertension, cardiovascular, heart, function, cardiac, associated, chronic, hiv-infected, dysfunction, transplantation, diabetes, factors, transplant

24	1	2	children, hiv-infected, age, months, infected, years, child, infants, adults, pediatric, hiv, mothers, born, uninfected, infection, paediatric, adult, childhood, life, study
25	3	3	cells, virus, hiv-1, cell, infected, replication, viral, infection, human, blood, peripheral, mononuclear, culture, pbmc, primary, lines, macrophages, immunodeficiency, cultures, lymphocytes
26	1	1	patients, cmv, infections, pneumonia, pcp, infection, prophylaxis, opportunistic, carinii, pneumocystis, cytomegalovirus, aids, herpes, toxoplasmosis, retinitis, disease, hiv-infected, bacterial, respiratory, zoster
27	2	2	hiv, rights, reserved, article, review, published, discussed, current, clinical, treatment, issues, presented, university, literature, society, press, management, paper, science, elsevier
28	3	3	activity, inhibition, antiviral, inhibited, inhibitory, replication, vitro, inhibit, hiv-1, anti-hiv, potent, activities, cells, concentrations, compounds, virus, human, effects, concentration, g/ml
29	3	3	production, ifn-, expression, cytokines, cytokine, cells, tnf-, factor, il-2, levels, increased, ifn, response, tnf, monocytes, il-6, necrosis, macrophages, il-10, mrna
30	1	1	cd4, count, cell, counts, cd4+, lt, patients, lymphocyte, cells/, median, gt, cells/mm3, hiv-infected, clinical, hiv, t-cell, baseline, low, cells/mm, lower
31	2	2	syphilis, transmitted, sexually, genital, std, infection, women, infections, sti, hiv, stds, hsv-2, herpes, transmission, diseases, stis, prevalence, chlamydia, semen, vaginal
32	1	1	lymphoma, ks, kaposi, sarcoma, ebv, cases, patients, hhv-8, lymphomas, nhl, disease, associated, chemotherapy, non-hodgkin, tumor, b-cell, primary, virus, tumors, herpesvirus
33	2	2	medical, physicians, guidelines, patient, ethical, practice, issues, general, problems, health, informed, recommendations, decision, care, legal, consent, medicine, physician, public, questions
34	3	3	resistance, mutations, drug, mutation, hiv-1, rt, resistant, virus, susceptibility, variants, genotypic, viruses, mutant, associated, wild-type, strains, reverse, isolates, drug-resistant, phenotypic
35	2	2	women, alcohol, states, african, black, hiv, american, united, percent, reported, white, risk, rates, abuse, substance, persons, data, differences, population, hispanic
36	1	1	fat, patients, metabolic, lipodystrophy, insulin, cholesterol, associated, hiv-infected, lipid, glucose, increased, changes, body, mitochondrial, antiretroviral, lipoatrophy, therapy, syndrome, levels, resistance

37	1	1	zidovudine, azt, combination, lamivudine, zdv, tenofovir, therapy, nucleoside, didanosine, treatment, stavudine, ddi, drug, 3tc, abacavir, drugs, d4t, toxicity, antiretroviral, tdf
38	2	2	intervention, participants, study, baseline, group, follow-up, months, trial, design, control, hiv, program, interventions, randomized, measures, outcome, compared, conclusions, objective, methods
39	2	2	africa, south, rural, african, sub-saharan, hiv, prevalence, areas, countries, uganda, urban, high, kenya, epidemic, region, district, tanzania, southern, study, population
40	3	3	membrane, fusion, cell, cells, protein, proteins, peptide, surface, uptake, transport, membranes, lipid, peptides, cellular, plasma, microscopy, intracellular, delivery, particles, viral
41	3	3	gag, protein, rna, virus, rev, hiv-1, viral, proteins, domain, type, immunodeficiency, particles, human, assembly, region, processing, sequence, mutant, cleavage, amino
42	2	2	prevalence, years, hiv, age, population, data, older, males, higher, aged, rates, females, infection, adults, women, general, seroprevalence, incidence, rate, study
43	2	2	risk, sexual, behavior, hiv, behaviors, factors, interventions, prevention, condom, behavioral, perceived, associated, social, risky, sex, change, relationship, self-efficacy, study, reduction
44	1	1	group, subjects, groups, compared, control, controls, study, higher, hiv-positive, hiv+, differences, difference, individuals, versus, lower, hiv-negative, healthy, hiv-, test, statistically
45	3	3	binding, structure, residues, protein, site, peptide, hiv-1, structures, interactions, structural, complex, molecular, interaction, protease, affinity, complexes, conformation, crystal, amino, conformational
46	2	2	hiv, trials, clinical, vaccine, phase, trial, efficacy, effective, vaginal, development, studies, transmission, potential, prevention, microbicide, prevent, microbicides, safety, protection, gel
47	1	1	patients, hiv, aids, infection, seropositive, asymptomatic, subjects, stage, disease, clinical, hiv-seropositive, seronegative, stages, cdc, symptomatic, individuals, control, hiv-seronegative, early, group
48	1	1	haart, therapy, antiretroviral, active, highly, patients, hiv-infected, hiv, treatment, receiving, era, initiation, individuals, introduction, immune, treated, reconstitution, associated, anti-retroviral, decreased
49	1	1	treatment, therapy, drug, drugs, antiretroviral, clinical, effects, agents, therapeutic, combination, effective, hiv, therapies, management, regimens, efficacy, resistance, side, interactions, treatments

50	1	1	brain, csf, cns, system, central, nervous, dementia, neurological, fluid, cerebrospinal, pml, astrocytes, aids, encephalitis, disease, neuronal, microglia, hiv, disorders, cerebral
51	2	2	studies, data, review, trials, evidence, literature, published, criteria, identified, included, search, articles, outcomes, systematic, methods, clinical, conclusions, interventions, controlled, reports
52	1	1	weeks, patients, study, adverse, week, trial, treatment, events, randomized, placebo, group, daily, baseline, efficacy, safety, regimen, arm, subjects, received, nvp
53	3	3	compounds, activity, derivatives, synthesized, acid, anti-hiv, inhibitors, synthesis, series, compound, potent, hiv-1, analogues, novel, antiviral, active, activities, prepared, group, rights
54	3	3	assay, sensitivity, detection, assays, samples, specificity, test, method, methods, tests, laboratory, diagnostic, sensitive, clinical, rapid, standard, performance, testing, diagnosis, laboratories
55	1	1	hcv, hepatitis, hbv, virus, liver, infection, patients, chronic, coinfection, hiv, coinfected, hbsag, co-infected, anti-hcv, genotype, co-infection, fibrosis, cirrhosis, ribavirin, prevalence
56	1	1	tuberculosis, tb, cases, treatment, patients, pulmonary, mycobacterium, control, infection, high, active, disease, isoniazid, hiv, diagnosis, positive, sputum, tuberculin, case, culture
57	1	1	protease, inhibitors, inhibitor, reverse, transcriptase, pi, antiretroviral, nucleoside, drugs, pis, nnrti, therapy, regimens, drug, non-nucleoside, patients, combination, indinavir, treatment, ritonavir
58	1	1	viral, load, rna, plasma, copies/ml, patients, therapy, levels, virological, baseline, antiretroviral, suppression, response, vl, failure, virologic, undetectable, loads, median, log
59	2	2	model, data, models, time, method, estimates, population, dynamics, methods, based, approach, estimate, parameters, paper, distribution, analysis, proposed, mathematical, number, rate
60	3	3	gp120, envelope, antibodies, hiv-1, binding, gp41, peptide, neutralizing, v3, glycoprotein, antibody, peptides, cd4, env, neutralization, region, virus, human, epitope, epitopes
61	3	3	subtype, hiv-1, sequences, subtypes, sequence, strains, genetic, analysis, isolates, viruses, region, env, diversity, regions, variants, phylogenetic, gene, recombinant, identified, samples
62	3	3	immune, infection, viral, host, system, response, hiv, role, mechanisms, infections, immunity, virus, disease, replication, responses, pathogenesis, factors, cellular, chronic, understanding

63	1	1	cancer, women, hpv, cervical, anal, human, infection, lesions, carcinoma, risk, hiv-positive, types, cancers, squamous, papillomavirus, screening, intraepithelial, hiv-infected, hiv, neoplasia
64	2	2	social, hiv/aids, study, qualitative, interviews, paper, women, context, analysis, health, focus, article, cultural, people, gender, understanding, experiences, issues, relationships, findings
65	1	1	patient, case, report, diagnosis, presented, man, reported, cases, infection, lesions, rare, revealed, treatment, developed, clinical, hiv, acute, fever, history, symptoms
66	3	3	method, ph, conditions, chromatography, mass, concentration, gel, acid, human, analysis, liquid, mm, water, phase, temperature, developed, solution, high, range, detection
67	3	3	dna, pcr, chain, reaction, polymerase, samples, detected, gene, analysis, detection, blood, genotype, amplification, proviral, study, individuals, polymorphisms, allele, sequences, presence
68	2	2	drug, users, injection, idus, injecting, hiv, risk, drugs, intravenous, treatment, idu, abuse, cocaine, methadone, heroin, sharing, needle, prison, syringe, syringes
69	1	1	patients, pulmonary, diagnosis, lung, tuberculosis, clinical, infection, disease, cases, mycobacterium, chest, diagnostic, positive, respiratory, mac, culture, sputum, avium, findings, symptoms
70	1	1	weight, body, loss, vitamin, mass, nutritional, growth, food, hiv-infected, status, intake, deficiency, wasting, supplementation, bone, low, increased, associated, bmi, muscle
71	2	2	prevention, program, health, community, programs, interventions, project, hiv, education, evaluation, training, activities, implementation, national, intervention, hiv/aids, effective, development, support, programme
72	2	2	men, sex, sexual, msm, hiv, partners, risk, gay, anal, homosexual, reported, intercourse, unprotected, male, heterosexual, bisexual, circumcision, hiv-positive, partner, behavior
73	3	3	dna, rt, reverse, rna, hiv-1, transcriptase, integration, enzyme, integrase, activity, rnase, polymerase, strand, site, viral, synthesis, primer, transcription, sequence, transfer
74	3	3	nef, hiv-1, viral, protein, proteins, vpr, replication, virus, vif, human, cells, cellular, cell, host, cycle, vpu, nuclear, interaction, infectivity, factor
75	3	3	vaccine, responses, vaccines, immune, vaccination, response, immunization, dna, antibody, immunity, mice, recombinant, induced,

			immunized, antigens, influenza, immunogenicity, mucosal, cellular, humoral
76	3	3	responses, ctl, class, epitopes, response, hla, peptides, t-cell, cytotoxic, cell, mhc, hiv-specific, peptide, epitope, cd8+, specific, viral, immune, control, gag
77	2	2	sex, women, sexual, condom, partners, condoms, partner, female, workers, risk, hiv, men, reported, male, sexually, couples, clients, transmitted, transmission, prevention
78	3	3	hiv-1, infection, hiv-2, infected, type, hiv-1-infected, virus, human, immunodeficiency, individuals, primary, study, hiv-1-positive, findings, hiv-1-seropositive, uninfected, early, anti-hiv-1, dual, seronegative
79	1	1	patients, factor, anemia, severe, complications, platelet, infection, hiv, surgery, hemophilia, surgical, thrombocytopenia, treatment, iron, anaemia, haemophilia, associated, bleeding, patient, risk
80	2	2	women, transmission, pregnant, pregnancy, infants, maternal, mothers, hiv, delivery, infant, birth, risk, mother, vertical, mother-to-child, perinatal, breastfeeding, child, milk, breast
81	1	1	tissue, lymph, cells, cases, nodes, lesions, tissues, lymphoid, node, leishmaniasis, biopsy, situ, changes, biopsies, detected, staining, findings, observed, skin, revealed
82	1	1	positive, negative, test, samples, hiv, tested, tests, western, antibody, elisa, blot, antibodies, assay, blood, specimens, confirmed, sera, serum, eia, screening
83	2	2	depression, symptoms, life, health, mental, quality, physical, scores, psychiatric, hiv, psychological, associated, cognitive, study, disorders, scale, functioning, depressive, measures, illness
84	2	1	blood, donors, transfusion, donor, hiv, risk, screening, virus, hepatitis, units, products, donations, transmission, infections, safety, donation, recipients, viruses, plasma, transfusions
85	3	3	antibodies, antibody, antigen, p24, igg, serum, sera, human, antigens, detected, virus, htlv-i, specific, presence, iga, igm, elisa, monoclonal, titers, assay
86	1	1	disease, diseases, clinical, patients, infection, infections, syndrome, manifestations, common, skin, disorders, associated, hiv, diagnosis, conditions, inflammatory, infectious, chronic, reactions, cases
87	2	2	cases, aids, hiv, epidemic, transmission, countries, reported, population, infection, europe, spread, number, surveillance, infected, people, china, infections, country, heterosexual, thailand

88	3	3	apoptosis, cells, cell, protein, activation, expression, effects, induced, death, kinase, signaling, pathway, role, gp120, stress, increased, human, receptor, activity, pathways
89	2	2	hiv, infection, infected, individuals, hiv-infected, persons, transmission, primary, early, increased, infections, acute, uninfected, high, evidence, increase, studies, number, suggests, associated
90	1	1	patients, infections, isolates, candida, fluconazole, species, bacterial, infection, albicans, isolated, fungal, meningitis, strains, cryptococcal, antifungal, bacteria, clinical, neoformans, pneumococcal, candidiasis
91	1	1	pain, patients, imaging, ct, neuropathy, ocular, peripheral, findings, hiv, retinal, lesions, magnetic, examination, visual, nerve, mri, resonance, eye, loss, clinical
92	3	3	role, studies, evidence, factors, data, play, findings, association, effects, well, mechanisms, influence, differences, potential, direct, remains, support, provide, hypothesis, suggested
93	2	2	art, adherence, antiretroviral, treatment, therapy, medication, patients, arv, initiation, monitoring, outcomes, medications, patient, settings, regimen, receiving, regimens, failure, hiv, clinical
94	1	1	patients, treatment, months, therapy, treated, response, days, median, follow-up, received, three, weeks, patient, time, duration, survival, range, complete, outcome, period
95	3	3	tat, transcription, hiv-1, expression, protein, gene, ltr, binding, nf-, rna, promoter, activation, human, tar, activity, cells, terminal, transcriptional, long, repeat
96	2	2	hiv, testing, test, screening, tested, counseling, counselling, voluntary, routine, positive, vct, clinic, tests, clinics, rapid, offered, uptake, care, clients, services
97	2	2	knowledge, hiv/aids, aids, students, education, attitudes, study, respondents, survey, questionnaire, level, school, awareness, educational, health, high, people, sample, attitude, conducted
98	3	3	virus, siv, immunodeficiency, infection, macaques, human, animals, simian, infected, rhesus, monkeys, viruses, model, humans, viral, fiv, animal, macaque, species, shiv
99	1	1	virus, immunodeficiency, human, hiv, acquired, syndrome, aids, infection, type, infected, deficiency, immune, -infected, virus-infected, agent, -positive, persons, report, -related, virus/acquired
100	3	3	gene, cells, expression, vector, vectors, genes, cell, human, protein, expressed, lentiviral, virus, recombinant, system, transduction, expressing, transfer, transduced, lines, hiv-1

Robustness check of clusters under variability of data

In order to check for the robustness of our topic clusters under the variability of input data (i.e. abstracts), we randomly sliced our data set of abstracts into three slices containing 70064, 69685, and 69859 abstracts. Then, for each category, we followed the procedure of topic modeling, creating the networks of topics, and clustering the topics. In this procedure, we used the same parameters as the original analysis, the only difference being the input data. The clustering algorithm, provided us with three clusters for each slice of data. In order to match the clusters for each slice of data with clusters in the original analysis, we took the top 10 words of each topic and found the set of unique top words for each cluster. Afterwards, we calculated the following similarity index for each pair of clusters from each slice and the original clusters.

$$r_{ij} = \frac{S_{ij}^2}{N_i \widehat{N}_j}$$

In the equation above, r_{ij} is the similarity of cluster *i* from the original analysis and cluster *j* of the slice k (where *i*, *j*, and *k* = {1, 2, 3}). Also, N_i is the number of items in the set of unique top words for cluster *i* of the original analysis and $\widehat{N}_j$ is the number of items in the set of unique top words for cluster *j* of the slice *k*. Number of shared words in these two sets are captured by S_{ij}. Comparison of the clusters for three slices of data is shown in Table A2. We find that for each slice, any of the clusters is highly similar to only one of the clusters from the original analysis. Hence, the clusters of topics used in this paper as the main item of analysis is robust to the variation in input data.

Table A2: Comparison of clusters for three slices of data and the original analysis. The table shows the number of shared unique words in the set of unique top words ($\boldsymbol{S_{ij}}$). The $\boldsymbol{r_{ij}}$ (normalized number of shared unique words) values are reported in parentheses. The gray cells show the corresponding cluster for each slice to the one of the original analysis based on the values of $\boldsymbol{r_{ij}}$.

		Slice 1			**Slice 2**			**Slice 3**		
	Cluster:	1	2	3	1	2	3	1	2	3
	$\widehat{N}$:	*230*	*228*	*227*	*258*	*263*	*221*	*268*	*236*	*244*
Original Analysis	Cluster 1 (*N = 242*)	179 (0.58)	44 (0.03)	51 (0.05)	195 (0.61)	45 (0.03)	34 (0.02)	174 (0.47)	48 (0.04)	57 (0.05)
	Cluster 2 (*N = 231*)	47 (0.04)	33 (0.02)	174 (0.58)	59 (0.06)	192 (0.61)	24 (0.01)	53 (0.04)	185 (0.63)	28 (0.01)
	Cluster 3 (*N = 206*)	34 (0.02)	179 (0.68)	19 (0.01)	59 (0.06)	22 (0.01)	146 (0.47)	51 (0.05)	22 (0.01)	151 (0.45)

Robustness check of regression results to clustering

In Table A3, we report results of a regression analysis based on the clusters generated by considering the top 11 topics in each paper. Our results hold under new clusters.

Table A3: Panel data regression results for BSS topics generated by top 11 topics per paper.

		Dependent Variable: Percentage of BSS topics in HIV/AIDS Research
Independent Variables	Log(HIV/AIDS Mortality)	0.016**
		(0.003)
	Log(GDP per Capita)	-0.027**
		(0.005)
	Intercept	0.679**
		(0.009)
	$F(2, 83)$	12042.55***
	R^2	0.281

* p < 0.05 ** p < 0.001 *** p < 0.0001

Standard errors are presented in parentheses.

Regression estimates by using alternative models

We have explained in the manuscript why we have selected the Driscoll-Kraay model as our main analysis tool. Here, in Table A4, we present the results of regression analysis with two other widely used methods: the White-Huber (Model 3) and Newey-West (Model 4) standard errors. As it can be seen, our results holds in both models.

Table A4: Panel data regression results for percentage of BSS topics.

	Dependent Variable: Percentage of BSS topics in HIV/AIDS Research	
	Model 3	Model 4

Independent Variables	Log(HIV/AIDS Mortality)	0.016**	0.016**
		(0.003)	(0.004)
	Log(GDP per Capita)	-0.012*	-0.027**
		(0.006)	(0.005)
	Intercept	0.554**	0.679**
		(0.054)	(0.041)
	$F(2, 237)$		50.6***
	$Wald\ \chi^2$	32.73***	

* p < 0.05 ** p < 0.001 *** p < 0.0001

Standard errors are presented in parentheses.

Regression estimates when controlling for time and its interactions

In Table A5, we report the results of our regression analysis (using Huber-White standard errors) including time periods, and the interactions of time period and the other two variables. As it can be observed, while our findings still hold under this condition, neither of terms with time periods were found to be significant.

Table A5: Panel data regression results for BSS topics including the time variables.

		Dependent Variable: Percentage of BSS topics in HIV/AIDS Research
Independent Variables	Log(HIV/AIDS Mortality)	0.013*
		(0.005)
	Log(GDP per Capita)	-0.033**
		(0.007)
	Time Period	
	(1996-2000)	0.019
		(0.052)
	(2001-2005)	-0.012
		(0.061)
	(2006-2010)	0.064
		(0.067)

Time Period X Log(HIV/AIDS Mortality)	
(1996-2000)	-0.007 (0.006)
(2001-2005)	-0.003 (0.006)
(2006-2010)	-0.004 (0.006)
Time Period X Log(GDP per Capita)	
(1996-2000)	-0.009 (0.007)
(2001-2005)	0.000 (0.008)
(2006-2010)	-0.003 (0.009)
Intercept	0.685^{**} (0.066)
$Wald\ \chi^2$	187.25^{***}
R^2	0.31

* p < 0.05 ** p < 0.001 *** p < 0.0001

Standard errors are presented in parentheses.

Regression estimates for papers without cross-country collaboration

In Table A6, we report the results of regression analysis by only considering papers without cross-country collaboration. As shown in Table A6, our findings reported in the main manuscript holds in this subset of data as well.

Table A6: Panel data regression results for subset of publications without any cross-country collaboration.

		Dependent Variable: Percentage of BSS topics in HIV/AIDS Research
Independent Variables	Log(HIV/AIDS Mortality)	0.016**
		(0.003)
	Log(GDP per Capita)	-0.026**
		(0.005)
	Intercept	0.661**
		(0.025)
	$F(2, 83)$	1755.00***
	R^2	0.2218

* p < 0.05 ** p < 0.001 *** p < 0.0001

Standard errors are presented in parentheses.

Chapter 3: What role does the philanthropic money play in health research?

Abstract

Philanthropies have been funding research in many disciplines including global health research. In this paper, we analyze how the flow of philanthropic money influences scientific publications of five major global health challenges: HIV/AIDS, Tuberculosis, Malaria, Pneumonia, and Neglected Tropical Diseases. We look at the journal publications across the globe, related to these disease, and published during 2008-2015 with known funding sources (n=62,168). By conducting textual analysis, we find the themes of research that philanthropies tend to fund more in health studies. Our analysis illustrates that some research themes such as cost-effectiveness, health policy, prevalence estimation, health services, intervention and prevention programs, and social aspects of diseases have received higher attention in studies funded by philanthropies. We next use the output from our textual analysis to investigate the relationship between philanthropic, public, and private sources of funding. We find that philanthropies' approaches in funding health research cover topics funded by both government and private agencies. Finally, we evaluate the impact of the papers in terms of citations. We find that in most diseases studies funded by philanthropies tend to receive more citations in comparison to those funded only by governments. Our study sheds light on the effect of philanthropic money in science and have several implications for scientific institutions.

Introduction

Financial support for science comes from different sources. In an aggregate level, these sources can be categorized in three main groups of public agencies, private firms, and philanthropic institutions. Public funding agencies include governments (at local, state, or national levels) and intergovernmental panels. This source of funding constitutes the largest share of expenditure on science in the modern era [1]. Private sources of funding include for-profit firms that invest in research and development, for example pharmaceutical companies that invest in drug development. Philanthropic sources of funding consists of individuals and organizations that contribute resources to scientists throughout their career [2].

Different science funders have diverse goals, strategies, and priorities [3]. They shape science by providing funding to scientists who work on specific projects and therefore help maintaining or creating new scientific paradigms through new discoveries and manipulation of scientific fields [4], [5]. Since the source of funding and expectations of funding agencies can be influential in shaping future research, it is important to study the impacts of funding sources. One approach for doing so is to analyze and distinguish between the scientific outputs. i.e., journal publications, supported by different agencies.

Differences of public and private sources of science funding can be conceptualized from an economic point of view. Intuitively, private firms are expected to invest on projects that can help their own firms. They exert their objective by supporting applied research projects related to their products [6]. It is argued that private supporters of science are prone to overlooking basic research

problems and long-term scientific initiatives that work on risky projects [1]. But, basic research can ensure a sustainable growth for societies and in some cases, such as health studies, it can have dramatic effects on people's everyday lives. To avoid market failure, and by considering science as a public good, public institutions such as governments have also become involved in funding science, especially after the World War II [1]. Public sources tend to fund basic research and areas of inquiry that do not provide enough incentives for private actors to invest in.

The role of philanthropies in science tend to be more complex to model and understand. Given their altruistic endeavors, one might expect that they seek to invest in problems benefiting the society and not their own profits. Moreover, many philanthropies have practical goals that are not long term and some are interested in achieving concrete results in a narrowly defined area [7], [8]. Philanthropies' complex motivations make it difficult to understand their focus areas. Moreover, given their limited resources as opposed to public sources of funding, it is not clear whether they are capable of supporting high impact scientific works or not.

In this study, we investigate the role of philanthropic funds in scientific inquiry. We focus on studies of health and evaluate how philanthropic actors supported different projects in studies of certain diseases. We specifically ask what areas of inquiry attract the focus of philanthropies more. Furthermore, we draw the relationship between philanthropic, public, and private sources of science funding. Finally, we evaluate the scientific impact of the philanthropies by analyzing the citations of papers funded by them.

Different Sources of Funding for Science

Funding for science has increased over the past century [1]. Many governments around the world contribute financial and other resources to support scientists and their studies. From a development perspective, public spending on science and innovation has been considered as a policy tool for growth [1]. However, governments and public agencies are not the only sources of funding for science. Private firms are eager to spend on research and development to secure profit. Also, nonprofit or philanthropic money has been one of the main sources of funding for science since centuries ago and even before the industrial revolution [1], [2].

Different funding sources can shape science differently. For example, when State funding was dominate in the United Sates, universities tend to focus more on local research problems. Later, the focus of research institutions changed due to the increasing influence of federal grants [1]. Private and philanthropic institutions potentially have different agenda which can influence scientific enterprise likewise. Although the amount of funding provided by these sources is usually smaller than the government funding, their impact is not negligible.

Contributions from philanthropies constitute up to 30% of research funding in major research universities in the United States [9]. Yet, few studies have attempted to quantitatively analyze the effects of this funding source on the science enterprise. To better understand what goals the philanthropists are pursuing when funding science, we need to better understand how they perceive and employ scientific research and discovery in the context of their domain of expertise. One specific area that has been popular among philanthropists is public health. Medical research, specifically, has long been an attractive area for philanthropists to invest and since a century ago, philanthropies have been amongst the most important players in the global public health issues [10], [11].

The focus of many philanthropies, in the context of public health, have been global health challenges as related to the poorest communities around the world. The health conditions and diseases which many people around the world are suffering from, are not being addressed at a proportional weight globally. This problem is often called the “10/90 gap”: only about 10 percent of global health resources are being spent on diseases that challenges the livelihood of 90 percent of the world population [12]. Many philanthropies, such as the Bill & Melinda Gates Foundation, are specifically focusing on addressing such diseases [11].

One important feature of the largest and most successful philanthropies in the context of global health is their focus on the knowledge creation process [11], [13]. These foundations rely on scientific research to improve their tools in fighting against the diseases and health challenges in poor communities around the world. Figure 3.1 shows the percentage of philanthropic funds in the pool of global expenditure on research activities related to a few diseases during 2007-2016. The data for this figure is provided by the G-FINDER data set [14]. As it can be seen, in the context of Tuberculosis, Malaria, and Pneumonia, about a quarter of all research funding or more is being provided by philanthropies.

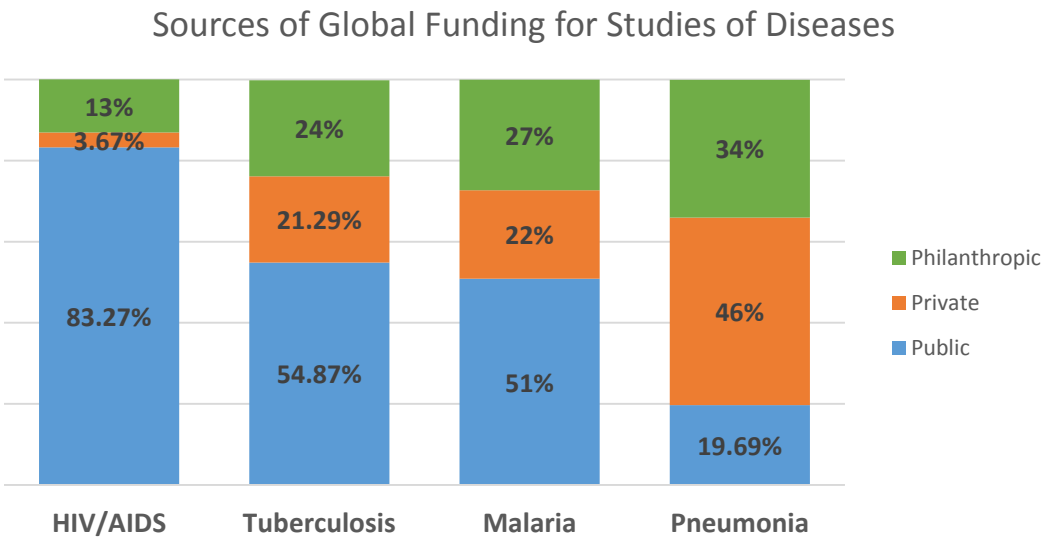


Figure 3.1: Sources of global funding for studies of certain diseases.

In this study, we focus on the effects of a few most important philanthropies on the scientific studies of a handful of diseases. There are many individuals that donate their money to support scientific research in the context of public health. However, about 80% of all donations for research is coming from institutions rather than individuals [6]. Moreover, institutional philanthropism can push an agenda in the scientific community more easily compared with individual altruistic philanthropism [15]. Based on this rational, we limit our analysis to the philanthropic institutions that has supported a considerable number of research publications as a proxy for their research spending. We acknowledge that our results will be biased due to this selection and our findings from this analysis may not be applicable to the effect of individual philanthropists on science.

Our study is limited to research activities on a handful of diseases. Specifically, we focus on HIV/AIDS, Tuberculosis, Malaria, Pneumonia, and Neglected Tropical Diseases. There is a considerable amount of burden caused by these diseases globally. Annually 1, 2, and 3 million people are dying because of Malaria, Tuberculosis, and HIV/AIDS respectively [11]. Moreover, most of the burden from these diseases are in developing countries with less resources at disposal for conducting relevant research. Hence, we expect the philanthropies to be interested in funding research projects related to these diseases. By limiting the focus of our study to a manageable number of philanthropies and fields of research, we can conduct an in-depth investigation of the effect of philanthropic money on science.

Our goal in this paper is to shed light on the role of philanthropies in shaping science in the context of health studies. We pursue this goal by asking three questions regarding the characteristics of publications funded by philanthropies as compared to scientific papers funded by other sources of funding.

Question 1: What areas of scientific inquiry do philanthropies focus more in comparison to public funders?

Funders of health studies provide financial support for projects aiming at different aspects of diseases. Figure 3.2 shows the percentage of funding spent globally on different areas of inquiry within the context of four diseases during 2007-2016. The data for this figure is provided by the G-FINDER data set [14]. There are variations in the amount of money spent on different aspects of the diseases. For example, majority of funding spent on HIV/AIDS is aimed at vaccine research while Pneumonia studies are mostly funded for drug development. Funding in studies of Tuberculosis and Malaria is more evenly distributed between basic research, drugs, and vaccines.

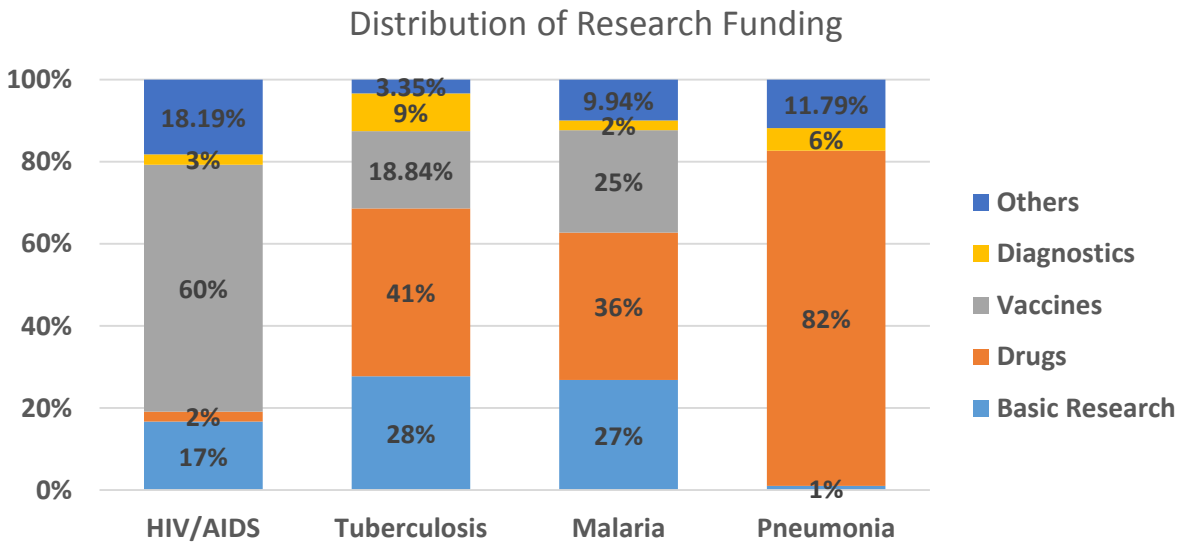


Figure 3.2: Distribution of the global research funding for certain diseases during 2007-2016.

The variation in the distribution of funding between different research areas might be partly correlated with the sources of funding supporting studies of each disease. Given differences of motivation and strategies, different funding sources may tend to focus on specific research areas. In this study, we investigate how the studies funded by philanthropies are different from those funded by public funders in terms of their contents. We ask what themes of research are more likely to be present in papers that are supported by philanthropies as compared with papers funded by government agencies.

Philanthropies tend to focus on diseases that affect people in developing countries disproportionately and receive insufficient funding from other sources. But within the context of such diseases, philanthropies are expected to prioritize research programs differently when compared with public funders. For example, it is believed that philanthropies tend to focus on implementable and measurable strategies [16]. It is also argued that they tend to have a technical focus on public health issues [13].

Question 2: What is the relationship between philanthropic, public, and private funders in the context of health studies?

Understanding the role of philanthropies in science is not trivial. Their priorities do not fit the conventional framework of basic vs practical research that is usually used to understand differences between public and private funders. In this paper, we look at the role of philanthropies as a player in the context of global health studies with respect to public and private sources of funding. We hypothesize three possible scenarios depicted in Figure 3.3.

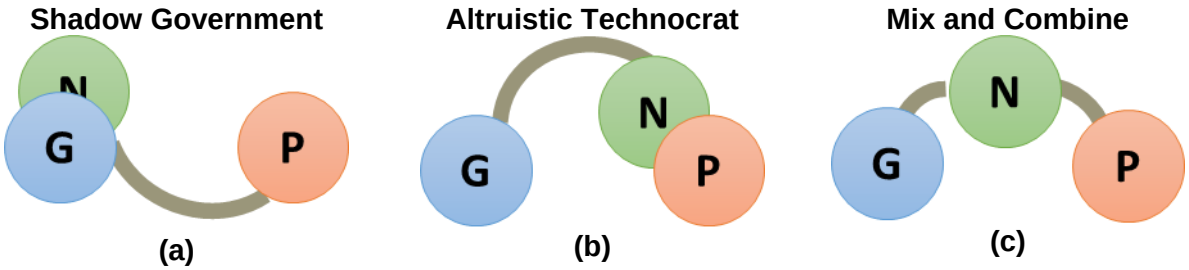


Figure 3.3: Three possible relationship between philanthropic (N), Public (G), and Private (P) funders.

The first hypothesis is that the focus of philanthropies and public funders are close to each other in the context of health studies. We labeled this hypothesis as *shadow government*, and the assumption is that the goals and strategies of philanthropies and governments converge. This hypothesis is based on a model of relationship between nonprofit sector and governments that highlights a complementary cooperation between these two [17]. Under this view, philanthropies and government share the same goals but may take different approaches of problem solving and delivering public goods and services. Based on this view, we expect to observe a higher level of intimacy between the philanthropic and public research compared with the private research.

Second, we hypothesize that philanthropies tend to have a focus closer to that of the private sector. We label this model as *altruistic technocrat*. A common feature of the philanthropic actions is that they intend to employ a business-oriented approach in addressing health issues. This has led them not only to try to apply the existing scientific knowledge in fighting the diseases but also to support researchers around the world to come up with implementable solutions for health problem [11], [15]. Such business-oriented approaches may translate into a research focus more similar to the private funders of health studies rather than the public ones.

Finally, we hypothesize that philanthropies cover and mix the areas of inquiry supported by the public and private sectors. This view maintains that philanthropies tend to focus on large-scale and long-term research initiatives aimed at disease eradication while pushing for practical and technical solutions to disease control.

The relationship between the philanthropies, public and private sources of funding depends on the nature of research problems and the failure of public or private sectors in providing sufficient amount of funding. We evaluate this relationship in the context of different diseases that highlight such differences. For example, the focus on basic and vaccine research is very high in HIV/AIDS studies. While drug development is the dominant research portfolio in studies of Pneumonia. On the other hand, public funding spending is very high in the context of HIV/AIDS and relatively low in the context of Pneumonia. Our goal is to illustrate how the relationship between different bodies of scientific funding varies in the context of diverse diseases.

Question 3: What is the impact of scientific studies funded by philanthropies?

After analyzing the contents of research studies funded by philanthropies and highlighting their relationship with public and private funders, we focus on the scientific impact of philanthropies. The role and impact of large philanthropies in shaping the global public health initiatives is well documented [11], [16], [18]. Moreover, in the context of scientific inquiry, it is shown that specific schemes of funding provided by prestigious nonprofit agencies (such as the Howard Hughes Medical Institute) tend to produce scientific publications with higher impact [19]. However, the potential large-scale effect of philanthropic funding on the scientific impact is not studied. We aim at conducting such analysis in this paper.

The selection process through which funding agencies distribute their funding affect the overall shape of science [20]. Since World War II, the competitive proposal-based funding distribution scheme has dominated these decision processes [21]. Philanthropies have some advantages and disadvantages, compared with public funders, in their selection process. On one hand, philanthropies often lack or have less access to the institutional infrastructure of public funding agencies (such as the peer-review boards) [22]. This can reduce the philanthropies' ability to distribute their resources efficiently. Poor judgment of grant proposals may inhibit philanthropies to grant their fund to the best research projects and end up with lower impact.

On the other hand, philanthropies face less constraints in their funding decisions for several reasons. First, they have no obligation to distribute their resources fairly. In other words, philanthropies can be more selective in choosing which projects to support and can distribute their funding highly non-egalitarian [18]. Second, it is argued that federal funding agencies are more risk averse and interested in supporting incremental research studies. These institutions tend to focus more on low-risk and adaptive discoveries rather than revolutionary high-risk ones.

Philanthropies, on the other hand, are only responsible to their boards when deciding on spending money. As a result, they are free to pursue riskier paths in the domain of scientific discovery [6], [8], [9], [16]. Therefore, we expect the philanthropies to be more likely to fund high impact studies. By investing in such studies, we expect that on average studies funded by philanthropies receive more citations.

In the following sections, we first introduce the data set that is going to inform our study. Some descriptive statistics and trends are presented to provide a better understanding of the problem. Next, we explain how we are going to prepare this data for our analysis. A topic modeling method will be used to read through our data set. After the initial set up of our data, we use statistical methods to find the answers to our research questions. Finally, the results will be discussed and the potential implications will be outlined.

Methods and Material

Data

We use journal publication data as reported on the Web of Science (WoS) data set. Our data set consists of all papers from WoS that include any of the following terms within their title, abstract, or keywords: HIV, AIDS, Tuberculosis, Malaria, Pneumonia, or Neglected Tropical Diseases (NTDs). Through this data set, we have access to a variety of information for each of the publications including title, abstract, author information, funding acknowledgement, and citation records.

The data set includes 638,505 publications. Among these, 367,531 papers were related to HIV/AIDS, 106,676 papers focused on Tuberculosis, 68,288 on Malaria, 85,726 on Pneumonia, and 2,843 papers were related to NTDs. The overall publication trends on these diseases are illustrated in Figure 3.4. As the figure shows, the publication rate has increased over the past three decades for all of the diseases. Moreover, the quest to overcome the burden of these diseases has become a global challenge, reflected in the number of countries who contribute to the research output on these diseases.

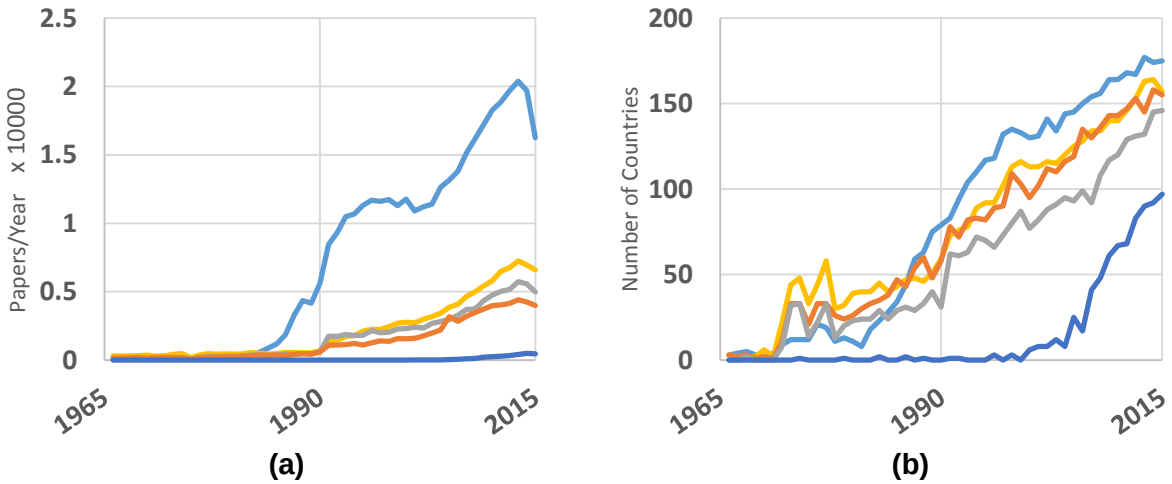


Figure 3.4: Trends of Publication. (a) Number of papers published on different diseases annually. (b) Number of countries that have published at least one paper on each of the

diseases. Light Blue: HIV/AIDS; Yellow: Tuberculosis; Orange: Malaria; Gray: Pneumonia; Dark Blue: Neglected Tropical Diseases.

Since August 2008, the WoS has begun indexing funding information from the funding acknowledgement section of papers[1]. In our data set, the funding information for about 45% of papers published between 2008 and 2015 is indexed. The trend for ratio of papers with indexed funding information is shown in Figure 3.5a. The ratio of papers with indexed funding information has increased from about 40% in 2009 to close to 60% in 2015. Since the focus of our analysis is on the effect of different sources of funding for science, we limit our analysis to the papers with at least one funding agency indexed by the WoS. This bounds our focus to 119,088 papers; all were published between 2008 and 2015.

Among papers with indexed funding information, about the third are supported by only one organization. However, there are some extreme cases where a paper is funded by many organizations. For example, the paper titled "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013" by Vos *et al*. is authored by 680 researchers and 89 funding organizations have supported this work. But in general, papers funded by more than 10 organizations are rare (i.e., less than 3% of the dataset). Figure 3.5b shows the number of papers with *k* distinguished funding agencies for varying values of *k*.

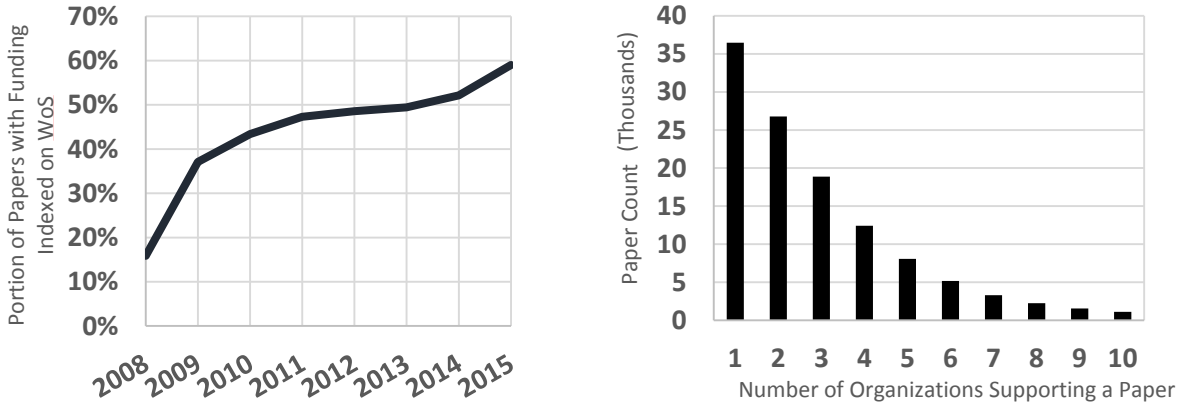


Figure 3.5: Funding organizations indexed by Web of Science. (a) Percentage of papers with at least one organization indexed. (b) Count of papers with different number of funding organization indexed.

There are 117,439 unique organization names indexed in our data set as supporters of the research papers. Further information about these organizations are not available in the WoS. We limit our focus to 299 organizations that have at least supported 100 papers in our data set. This enables us to manually look up necessary information for these organizations from their websites and/or Wikipedia pages. We then categorize funding organizations in three groups: public, philanthropic/nonprofit, and private. Overall, we find 229 public agencies, 35

[1] http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/

philanthropies, and 35 private organizations. In our data set, 9,837 and 4,668 papers are respectively funded by at least one philanthropic or private organization. Moreover, there are 54,605 papers that were supported by at least one public agency. Number of unique papers in this data set is 62,168 papers. Table 3.1 shows the top 10 philanthropic and private organizations that supported research papers in our data set.

Based on this categorization, papers can be labeled into seven groups. Papers funded only by public, philanthropic, or private agencies are labeled as **G**, **N**, or **P** respectively. Those funded by public agencies and philanthropies are labeled as **GN**, a combination of funding by public and private agencies labeled as **GP**, and **NP** denotes papers funded by private and philanthropic organizations. Finally, papers that are supported by all types of organizations are labeled as **GNP**.

Table 3.1: The top 10 philanthropic and private funding agencies.

Non-profit/Philanthropic Organizations		**For-profit Firms**	
Name	**Number of Papers**	**Name**	**Number of Papers**
Wellcome Trust	3365	**Pfizer**	1625
Bill & Melinda Gates Foundation	2994	**Gilead Sciences**	1541
Howard Hughes Medical Institute	437	**GlaxoSmithKline**	1417
Doris Duke Charitable Foundation	349	**Bristol-Myers Squibb**	1265
American Heart Association	269	**Merck & Co.**	1177
Robert Wood Johnson Foundation	251	**Abbott Laboratories**	998
Michael Smith Foundation for Health Research	195	**Boehringer Ingelheim**	783
The Royal Society	174	**Roche Holding AG**	752
Pasteur Institute	173	**ViiV Healthcare**	567
American Cancer Society	164	**Tibotec**	480

Topic Modeling

We use topic modeling to analyze the content of papers in our data set. Specifically, we implement Latent Dirichlet Allocation (LDA) which is a Bayesian statistical technique for natural language processing. The output from an LDA model helps us uncover the latent topics within a set of documents, i.e. abstracts of the publications in our case. Topic modeling has been used widely in studies of science and innovation [23]–[26]. More information about the method can be found in [27].

Other methods and sources of data may be used to analyze the contents of the papers such as analysis of the keywords of papers. While such approach is simple to implement, it does not provide the precision we need for our analysis. Most importantly, when looking at the set of keywords for any given paper, it is unclear how important each of the keywords are in describing the overall content of a paper. Specifically, in the case of interdisciplinary studies, one cannot make sure how much of a paper is related to any of the fields/disciplines that are crossed in one study just by looking at the keywords. The LDA model resolves this problem. In this method, the content of each paper is modeled as a probability distribution over a set of (latent) topics. These

topics are themselves probability distributions over the set of unique words that make up all the documents.

The two output from the LDA model helps us uncover the themes of research in our data set and relate each of the papers to those themes. For each paper, the LDA output provides us with a probability distribution over topics. This information can be used to find the focus of each paper, and the share of each topic in the knowledge generation process. Moreover, the LDA output defines each topic as a probability distribution over the corpus of unique words within the documents. This helps us to understand each topic and distinguish between them. These outputs are generated by the LDA model in an unsupervised process after an initial setup.

Three parameters need to be determined in an LDA implementation: number of topics, topic smoothing parameter, and term smoothing parameter. Generally, larger number of topics provide a model that fits the data better. However, a large number of topics makes them undistinguishable to human readers [28]. Therefore, and by following the prior topic-modeling studies of science and innovation ([29]–[31]), we use a moderate number of topics for our data set, i.e. 200 topics. The topic smoothing parameter changes the shape of the topic probability distribution for papers. When the topic smoothing parameter is set very low, only few topics will have a relatively large probability value and the rest will be assigned probabilities close to zero. Higher values of topic smoothing parameter allows multiple topics to have a relatively high probability values assigned to them in a given paper. The term smoothing term alters the shape of the word probability distributions for topics. The lower is the term smoothing term, the more distant would the topics be. Following the prior topic modeling studies of science we use 0.1 and 0.01 as our topic and term smoothing coefficients.

The process of implementing LDA is straight forward. There are 212,745 abstracts in our data set for papers published during or after 2008. We start by generating the corpus of unique words in this set of documents. We remove the very common English words from these abstracts by using a standard stop list and further limit our analysis to the words that have appeared in at least 100 abstracts. This will reduce the size of our corpus to 15,599 unique words. We then use the python LDA 1.0.4 package to perform the topic modeling analysis.

The five topics with highest cumulative probabilities, within the context of each disease, are shown in Table 3.2. For each topic, the top five words are presented. Different themes of research in studies of these diseases can be observed here. In the context of HIV/AIDS such themes include understanding the social, behavioral, and epidemiological aspects of the disease, studying the virus's replication process, and investigating the prevention mechanisms. As far as the Tuberculosis, themes of research such as bacterial studies, drug resistance, test and diagnosis mechanisms, investigation of enzymes and case studies are evident. Studies on Malaria include topics such as parasites, mosquitos, treatments, and spread control techniques. In the context of Pneumonia, research themes include different types of pneumonia, community based interventions, and surgeries. Finally, research themes such as different types of diseases, development of drug and screening techniques, and health policy can be observed in studies of the NTDs.

Table 3.2: Top five topics in the studies of different diseases.

Disease	Top Topics
HIV/AIDS	social, article, paper, aids, health
	men, sex, msm, sexual, hiv
	sexual, condom, behavior, sex, risk
	hiv-1, cells, viral, replication, infection
	hiv, aids, people, prevention, living
Tuberculosis	tb, tuberculosis, patients, pulmonary, cases
	tuberculosis, mycobacterium, mtb, mycobacterial, bovis
	tuberculosis, drug, resistance, mdr-tb, isoniazid
	test, tst, positive, tb, tuberculosis
	enzyme, acid, enzymes, pathway, activity
Malaria	parasite, plasmodium, malaria, falciparum, parasites
	malaria, falciparum, plasmodium, vivax, parasite
	mosquito, mosquitoes, anopheles, vector, malaria
	malaria, falciparum, treatment, antimalarial, sp
	malaria, control, nets, net, coverage
Pneumonia	antibiotic, antibiotics, aureus, mrsa, infections
	patients, pneumonia, cap, mortality, score
	patients, care, icu, vap, intensive
	patients, surgery, complications, surgical, postoperative
	lung, pulmonary, interstitial, pneumonia, fibrosis
NTDs	disease, diseases, tropical, neglected, leishmaniasis
	health, policy, national, global, development
	drugs, drug, development, therapeutic, agents
	health, countries, public, global, developing
	Inhibitors, compounds, target, novel, screening

Results

Same Problems Different Approaches

What is the distinguished focus of philanthropies in tackling each of the diseases? We answer this question by analyzing the relationship between probabilities assigned to topics (from our topic modelling analysis) and funding sources of each paper.

For each disease, we focus on the subset of *major* topics from the whole 200 topics. We define major topics, within the context of studies of each disease, to be those that have at least a probability of 25% assigned to them in one paper related to that disease. This helps us focus on the major topics in studies of each disease. To find which funding sources tend to focus more on any of these topics, we run a series of simple regression models. The dependent variables are the probability of topic *i* in every paper. We use a categorical variable as the independent variable which includes different modes of funding (i.e., **G**, **N**, **P**, **NP**, **GN**, **GP**, and **GNP**). The regression models that we run are as following:

$$P(Topic_i) = \beta_0 + \boldsymbol{\beta} Funding$$

Number of observations for each regression model is the total number of papers written on each of the diseases. We use a base significant level of 0.01 and adjust it based on the number of the major topics for each of diseases. Table reports the topics that have a significantly higher probability in papers supported only by philanthropies as compared to those supported only by public agencies.

Table 3.3: List of topics that studies funded only by philanthropies focused more in comparison to those funded only by public agencies.

Area	Topic	HIV/AIDS	Tuberculo	Malaria	Pneumoni	NTDS
Policy, Management, and Economics	costs, cost, cost-effectiveness, economic, estimated	*	*	*	*	*
	health, countries, public, global, developing	*	*	*	*	
	health, policy, national, global, development	*	*	*	*	
	will, control, strategies, effective, strategy	*	*	*		*
	management, practice, guidelines, physicians, clinical	*	*	*		
	products, product, production, medicines, market		*	*	*	
	decision, making, aids, decisions, process	*				
Epidemiology and Population Studies	mortality, death, deaths, morbidity, rates	*	*	*	*	
	women, infants, pregnancy, maternal, pregnant	*	*	*	*	
	africa, south, african, sub-saharan, countries	*	*			
	children, years, age, pediatric, child	*			*	
	health, household, status, households, income	*	*			
	incidence, rates, rate, increased, period	*		*		
	prevalence, population, data, survey, high	*		*		
	transmission, contact, infection, spread, individuals	*		*		
	ci, risk, incidence, cohort, hr			*		
	sex, china, workers, female, work	*				
	studies, trials, review, data, evidence			*		
	study, areas, rural, india, urban	*				
	years, age, older, adults, aged	*				
Prevention, Diagnosis, and Treatment	health, services, care, service, facilities	*	*	*	*	
	interventions, prevention, community, programs, program	*	*	*	*	
	trial, events, adverse, randomized, placebo	*	*	*		
	vaccine, vaccination, vaccines, bcg, protection	*	*	*		
	art, antiretroviral, therapy, initiation, treatment	*	*			
	culture, positive, sputum, negative, samples	*	*			
	diagnosis, clinical, fever, diagnostic, symptoms	*		*		
	guidelines, criteria, recommendations, clinical, based		*	*		
	hiv, testing, screening, test, tested	*	*			
	intervention, control, participants, trial, baseline		*	*		
	adherence, medication, arv, treatment, study			*		
	assay, test, tests, detection, rapid		*			
	care, health, medical, providers, patient		*			

	concentrations, dose, plasma, concentration, pharmacokinetic			*		
	hiv, aids, people, prevention, living	*				
	hospital, hospitals, patients, hospitalization, emergency			*		
	malaria, control, nets, net, coverage			*		
	malaria, falciparum, treatment, antimalarial, sp			*		
	patients, therapy, failure, response, treatment		*			
	sensitivity, specificity, predictive, diagnostic, accuracy			*		
	therapy, haart, antiretroviral, active, highly		*			
Social, Behavioral and Cultural	interviews, qualitative, study, participants, conducted	*	*	*		
	social, article, paper, aids, health	*	*	*		
	knowledge, survey, study, participants, respondents	*		*		
	quality, scores, life, scale, score			*		
	women, men, hiv, partner, violence	*				
Other Complications	severe, clinical, disease, cm, associated	*				
	cancer, hpv, cervical, women, anal	*				
	hiv, testing, screening, test, tested			*		
	hiv, vaginal, prep, circumcision, genital	*				
	malaria, falciparum, plasmodium, vivax, parasite	*				
	malaria, falciparum, treatment, antimalarial, sp				*	
	pneumoniae, pneumococcal, pneumonia, disease, streptococcus	*				
	tb, tuberculosis, patients, pulmonary, cases	*				
	vitamin, deficiency, iron, anemia, supplementation			*		
Biomedical	malaria, falciparum, plasmodium, vivax, parasite			*		
	pneumoniae, pneumococcal, pneumonia, disease, streptococcus				*	

* Probability of the topic was significantly higher in studies funded only by philanthropies as compared with those funded only by governments.

Studies funded by philanthropies tend to focus more on the economics of the diseases and the costs associated with medications. We see this focus consistently in studies of all diseases in our data set. These studies also pay more attention to the public health policies both at the national and global level. Furthermore, philanthropies invest in studying the effective strategies for controlling diseases, and management issues in clinical practice. In the context of Tuberculosis, Malaria, and Pneumonia, there has also been a push by philanthropies to study the market of medication production.

When investigating epidemiology of the diseases, philanthropies tend to have a higher focus on estimating the mortality rates and burden of the diseases. In these studies, we also observe a higher share for the topic that is related to analyzing the health of pregnant women and infants. The regional trends of the diseases and local factors particularly associated with sub-Saharan Africa, China, and India have been investigated at a higher degree by philanthropies.

Philanthropies have focused on health services and healthcare facilities when it comes to prevention, diagnosis, and treatment of the diseases. Furthermore, we find that the topic related to community-based intervention and prevention programs have received a higher attention by philanthropies. Also, analyzing drug trials has been focused more by philanthropies in the context

of HIV/AIDS, Tuberculosis, and Malaria. In studies of these diseases, more attention has also been given to vaccines and vaccination by philanthropists.

Social, behavioral, and cultural aspects of the diseases have also been in the center of philanthropies' attention. Specifically, in the context of HIV/AIDS, Tuberculosis, and Malaria, philanthropists tend to focus more on the social aspects of these diseases and the qualitative analysis of their effects on society as compared with governments.

Philanthropic vs Public and Private Funding Sources

Next, we investigate the relationship between philanthropic and public agencies on one hand and private agencies on the other hand. We use the major topics, for each disease, to analyze how similar the contents of papers funded by philanthropies are to those of the public and private funders. For each type of funding organization, and within the context of each disease, we calculate the sum for share of all the major topics and normalize them. By doing this, we are able to vectorize the knowledge space for each of the funding organization types in the context of each disease.

We use the log transformation of topic vectors for each funding agency and run a linear regression model to carry out our analysis. The model is as following:

$$\mathbf{X}_N = \beta_N + \beta_G \mathbf{X}_G + \beta_P \mathbf{X}_P$$

In this model, $\mathbf{X}_N$, $\mathbf{X}_G$, and $\mathbf{X}_P$ are log-transformed vector variables that respectively represent the knowledge space of philanthropies, governments, and for-profits. The coefficients β_G and β_P capture the relationship between the philanthropies' knowledge space and that of public and private funders, respectively. The coefficient β_N is a constant term, capturing characteristics of the philanthropies' knowledge space that can be best explained independent of public and private funders.

Our hypotheses regarding the relationship between philanthropies and other major sources of funding in science can be tested with the regression model presented above. If only $\beta_G > 0$, we show that philanthropic and public funders tend to have the same focus on the issues and we find support for our shadow government hypothesis. When only $\beta_P > 0$, our altruistic technocrat hypothesis is supported, i.e. we show that philanthropies tend to focus on issues in the same manner as the private funders. Finally if both of these coefficients are greater than zero, we show that philanthropies mix and combine the knowledge funded by both the public and private agencies.

We estimate the values of β_N, β_G, and β_P by running a robust linear regression model for each disease. The number of observations for each model equals to the number of major topics for that model. The results are presented in Table 3.3.

Table 3.3: Relationship of contents funded by philanthropies as oppose to those funded by government and for-profits.

	HIV/AIDS	Tuberculosis	Malaria	Pneumonia	NTDs
β_G	0.846*** (0.051)	0.803*** (0.037)	0.552*** (0.046)	0.743*** (0.056)	0.650*** (0.062)
β_P	0.111* (0.044)	0.171*** (0.042)	0.395*** (0.043)	0.179*** (0.049)	0.387*** (0.043)
β_N	-0.248 (0.224)	-0.113 (0.199)	-0.255 (0.170)	-0.400 (0.216)	0.197 (0.198)
N	185	171	142	163	77
$F(2, N-3)$	299.36***	474.13***	484.11***	275.57***	381.24***
R^2	0.739	0.835	0.839	0.748	0.848

* p-value < 0.05 ** p-value < 0.01 *** p-value < 0.001
Robust standard errors are reported in the parentheses.

We observe a strong support for the hypothesis that philanthropies mix and combine knowledge across studies of all diseases in our data set. Not only philanthropists tend to fund studies closely similar to those funded by public funders, they are also interested in the type of scientific inquiry pursued by private sector.

We further check whether the focus of philanthropies is significantly closer to the public rather than private funders. In other words, we ask where the philanthropies' position is as they mix and combine knowledge supported by public and private agencies. We answer this question by performing Wald tests on the parameters of our regression models to see if the β_G and β_P coefficients are significantly different from each other. We find that the philanthropic sector's focus is significantly closer to public funders in the context of all diseases ($p < 0.001$ for all the models). It is unclear from our analysis and data set, whether philanthropies have followed the scientific push of the governments in these areas or whether the reverse has been the case. However, we can conclude from our analysis that the scientific approaches of philanthropies lies between the public and private actors while being closer to the public funders.

Scientific Impact

We use citations received by papers as a proxy to evaluate scientific impact. Our goal is to assess the scientific impact of philanthropies by comparing the received citations of the papers they have funded with those funded by governments. We use a negative binomial model to conduct our analysis. The same method has been used in the prior studies to analyze citations [32].

Negative binomial regression is a good candidate for analysis of over-dispersed count data. This is the case where the variance of the count data is larger than its mean. The mean and variance of citation counts for papers written in the context of different diseases and supported by different source of funding are presented in Table 3.4. It is evident that the variance of citations are larger than their mean in all categories. Therefore, the negative binomial regression is a proper model to be used in this paper. By implementing the negative binomial regression, we model both the

mean and dispersion of the selected dependent variable. Hence, not only we investigate what factors are associated with the mean of the dependent variable, but we also assess the significant factors related to the variance of the dependent variable.

Table 3.4: Mean and variance of citations for papers written in the context of different diseases and funded by different organization types.

		HIV/AIDS		Tuberculosis		Malaria		Pneumonia		NTDs	
		Mean	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**
Funding Source	**G**	13.67	999.02	13.73	1333.86	12.93	410.91	14.13	769.28	12.89	551.40
	N	15.42	1426.08	16.50	860.46	17.31	1428.61	17.46	2693.17	12.06	319.03
	P	14.15	1088.32	18.59	1172.76	10.36	130.37	19.28	2739.63	9.29	278.37
	NP	32.13	3624.68	34.87	3198.84	26.17	1436.15	97.63	29512.27	5	21
	GN	23.38	2849.72	27.05	5328.80	23.89	410.91	30.27	16945.65	28.42	3283.15
	GP	22.44	6055.13	42.07	40566.89	12.40	269.19	23.98	1361.74	14	173.75
	GNP	47.22	21523.64	64.82	27512.45	52.81	22525.99	30.27	3054.56	198.4	150578.3

The dependent variable in our models is the count of citations received by the paper. The main independent variable is the funding source. We have created this variable in a categorical form: the base level is **G** and varying forms of funding organizations constitute other levels of the variable. We also control for some factors that is shown to have effect on the number of citations [32], [33]. First, we control for the age of the papers. We include the number of years since a paper's publication time and also its squared age. We further include variables to control for number of authors on a paper and whether the paper involves international collaboration. Same variables are included when modeling the dispersion of citations. The results of the regression models are presented in Table 3.5.

Table 3.5: Results of citation analysis through negative binomial regression model.

		Model 1				
		HIV/AIDS	Tuberculosis	Malaria	Pneumonia	NTDs
Mean						
Funding	N	0.121** (0.042)	0.242*** (0.064)	0.251*** (0.052)	0.133 (0.125)	0.256* (0.123)
	P	0.055 (0.039)	0.246** (0.094)	-0.200† (0.105)	0.327*** (0.058)	-0.067 (0.249)
	NP	0.773*** (0.217)	0.586* (0.289)	0.559*** (0.166)	1.114* (0.458)	0.124 (0.464)
	GN	0.417*** (0.031)	0.550*** (0.049)	0.449*** (0.045)	0.248* (0.106)	0.285** (0.109)
	GP	0.269*** (0.051)	0.790*** (0.245)	-0.114 (0.172)	0.499*** (0.114)	0.301* (0.145)
	GNP	0.436*** (0.098)	0.513** (0.162)	0.505* (0.254)	0.924*** (0.272)	0.831* (0.345)
Age		1.281*** (0.022)	1.195*** (0.054)	1.226*** (0.039)	1.253*** (0.059)	1.147*** (0.075)
Age^2		-0.093*** (0.002)	-0.083*** (0.006)	-0.089*** (0.004)	-0.090*** (0.006)	-0.093*** (0.012)

Number of Authors		0.050*** (0.002)	0.026*** (0.004)	0.039*** (0.003)	0.046*** (0.005)	0.029*** (0.008)
International		0.025 (0.019)	0.136*** (0.035)	0.049† (0.029)	0.189*** (0.056)	0.157† (0.082)
Constant		-1.456*** (0.049)	-1.127*** (0.108)	-1.280*** (0.089)	-1.322*** (0.124)	-0.195 (0.121)
Dispersion						
Funding	N	0.021 (0.060)	0.122 (0.082)	0.228** (0.078)	0.188 (0.150)	0.113 (0.172)
	P	0.153*** (0.046)	0.115 (0.104)	-0.175 (0.201)	0.036 (0.073)	-0.071 (0.366)
	NP	0.123 (0.177)	0.819 (0.510)	-0.265 (0.233)	0.584 (0.399)	-0.315 (1.013)
	GN	-0.028 (0.040)	-0.089 (0.062)	0.097 (0.068)	-0.084 (0.167)	-0.337† (0.196)
	GP	0.095 (0.062)	0.437* (0.202)	-0.198 (0.299)	0.153 (0.131)	-1.094 (0.810)
	GNP	-0.105 (0.107)	-0.096 (0.246)	0.359 (0.245)	0.043 (0.246)	-0.237 (0.349)
Age		-0.449*** (0.034)	-0.531*** (0.087)	-0.464*** (0.066)	-0.571*** (0.083)	-0.117 (0.172)
Age^2		0.043*** (0.004)	0.053*** (0.010)	0.042*** (0.007)	0.055*** (0.009)	0.027 (0.023)
Number of Authors		0.007*** (0.001)	0.002 (0.002)	0.006*** (0.001)	0.004 (0.003)	0.008*** (0.001)
International		-0.049* (0.025)	-0.060 (0.048)	-0.151*** (0.043)	0.078 (0.073)	-0.325* (0.133)
Constant		1.028*** (0.073)	1.175*** (0.175)	0.945*** (0.147)	1.261*** (0.172)	-0.188 (0.283)
N		35229	11798	9525	6005	746
χ^2_{Wald}		11568.44***	3526.52***	3580.80***	2025.29***	732.31***
Pseudo R^2		0.0527	0.0535	0.0575	0.0531	0.0864

† p < 0.1 * p < 0.05 ** p ≤ 0.01 *** p ≤ 0.001
Robust standard errors are reported in the parentheses.

Papers funded by philanthropies have received more citations in comparison to those funded by public agencies. This can be consistently seen in studies of HIV/AIDS, Tuberculosis, Malaria, and NTDs. However, such citation advantage is not evident in the context of Pneumonia studies. Moreover, we observe that papers funded simultaneously by public agencies and other types of organizations, on average, receive higher citations in comparison to papers that received only funding from public agencies.

One potential reason for the citation advantage of papers funded by philanthropies is that they are funding more innovative and high impact projects [6], [8], [9], [16], [19]. In order to capture and control for this effect, we first check whether papers funded by philanthropies have a higher chance of being among the high impact papers and then control for whether a paper is high impact. We define a high impact paper as one that has received citation counts in the top 5% percentile of its cohort.

We use logistic regression models to evaluate the effect of different factors on whether a paper has become high impact or not. Table 3.6 reports the results. Papers that are only supported by philanthropic funds have a higher chance of being high impact when compared to those who have only received funding from public agencies. This is true for papers written in the context of HIV/AIDS, Tuberculosis, Malaria, and NTDs. Now, the question is whether the overall citation advantage of papers supported by philanthropies (shown in Table 3.5) can be explained with the higher chances of philanthropies' papers to be high impact. In order to check this possibility, we re-run the negative binomial regression model explained above but this time add a dummy variable to denote the papers that have high impact. We further, control for the content of papers to capture the differences of citation distributions between different fields of study within the context of one disease. The results are reported in Table 3.7.

Table 3.6: Factors influencing chances of a paper to have a high impact. Results are rendered from a logistic regression model.

		Model 2				
		HIV/AIDS	Tuberculosis	Malaria	Pneumonia	NTDs
Funding	N	0.191† (0.101)	0.584*** (0.143)	0.523*** (0.127)	0.014 (0.261)	1.091** (0.409)
	P	0.333*** (0.099)	0.692** (0.261)	-1.208 (1.015)	0.817*** (0.151)	0.255 (1.229)
	NP	1.704*** (0.356)	1.481* (0.670)	1.306* (0.627)	2.401** (0.874)	-
	GN	0.775*** (0.070)	0.941*** (0.120)	1.061*** (0.117)	0.297 (0.282)	0.673 (0.518)
	GP	0.462*** (0.114)	1.038*** (0.292)	0.406 (0.602)	0.730** (0.237)	-
	GNP	0.490* (0.233)	1.105* (0.544)	1.378** (0.494)	1.656** (0.598)	2.669* (1.198)
Number of Authors		0.087*** (0.004)	0.062*** (0.007)	0.066*** (0.006)	0.102*** (0.009)	0.079*** (0.023)
International		0.091† (0.051)	0.279** (0.092)	0.095 (0.112)	0.256* (0.127)	0.473 (0.421)
Constant		-3.779*** (0.046)	-3.762*** (0.081)	-3.904*** (0.100)	-4.149*** (0.121)	-4.365*** (0.398)
N		35229	11798	9525	6005	734
χ^2_{Wald}		886.00***	254.81***	320.08***	210.29***	34.19***
Pseudo R^2		0.0585	0.0533	0.0821	0.0857	0.1166

† p < 0.1 * p < 0.05 ** p ≤ 0.01 *** p ≤ 0.001

Even after considering the effect of high impact papers and controlling for the content of papers, studies funded by philanthropies tend to have a higher citation counts, compared with those funded only by public agencies. This effect can be observed in the context of HIV/AIDS, Tuberculosis, and Malaria studies. Our analysis shows that in the context of these diseases, the citation advantage of philanthropies does not only come from philanthropists' support of high impact projects. The case of NTDs, however, shows an alternative scenario. When controlling for the content of papers and high impact publications, the citation advantage of philanthropies

diminishes in this scenario. This result suggests that the citation advantage of philanthropic funds in the context of NTDs is related to their project selection.

Consistent with prior findings [33], we show that larger teams tend to have a higher scientific impact. But papers authored by larger team sizes also tend to have a higher variation in number of citations they receive. The effect of international scientific collaboration is more complex. In Table 3.5, we observe that cross-country collaboration has a negative effect on the count of citations, at least in the context of HIV/AIDS, Malaria, and NTDs. However, when we control for the content of research studies and the effect of high impact papers, cross-country collaboration tends to boost scientific impact across studies of HIV/AIDS, Tuberculosis, Malaria, and Pneumonia (Table 3.7). Papers co-authored by International collaborators also tend to have a smaller variation in the number of citations that they receive on average. We further find that, in the context of HIV/AIDS, Tuberculosis, and Malaria, papers co-funded by public and philanthropic funders tend to have a smaller variation in expected citations.

Table 3.7: Results of citation analysis through negative binomial regression model (considering the effect of content and high impact papers).

		Model 3				
		HIV/AIDS	Tuberculosis	Malaria	Pneumonia	NTDs
Mean						
Funding	N	0.101*** (0.022)	0.109** (0.035)	0.100*** (0.025)	0.050 (0.051)	-0.095 (0.089)
	P	0.058* (0.026)	-0.025 (0.076)	-0.106 (0.089)	0.096* (0.039)	-0.163 (0.231)
	NP	0.220† (0.119)	-0.024 (0.247)	0.475*** (0.142)	-0.019 (0.338)	0.367 (0.297)
	GN	0.186*** (0.017)	0.246*** (0.027)	0.192*** (0.025)	0.106* (0.055)	0.187† (0.097)
	GP	0.160*** (0.028)	0.164† (0.087)	-0.317** (0.100)	0.214*** (0.056)	0.290 (0.195)
	GNP	0.244*** (0.055)	0.286 (0.195)	0.048 (0.131)	0.365* (0.174)	-0.011 (0.377)
Age		1.367*** (0.012)	1.294*** (0.021)	1.278*** (0.022)	1.332*** (0.030)	1.202*** (0.057)
Age^2		-0.101*** (0.001)	-0.094*** (0.002)	-0.093*** (0.002)	-0.098*** (0.003)	-0.106*** (0.009)
Number of Authors		0.029*** (0.001)	0.021*** (0.003)	0.025*** (0.002)	0.026*** (0.004)	0.007*** (0.001)
International		0.053*** (0.011)	0.089*** (0.019)	0.108*** (0.020)	0.067* (0.027)	0.184** (0.069)
High Impact		1.885*** (0.019)	1.850*** (0.032)	1.752*** (0.032)	1.882*** (0.046)	1.608*** (0.117)
Content		*Controlled*	*Controlled*	*Controlled*	*Controlled*	*Controlled*
Constant		-1.802*** (0.106)	-1.518*** (0.169)	-1.934*** (0.124)	-1.566*** (0.203)	-0.370† (0.205)
Dispersion						

Funding	N	-0.013 (0.043)	0.104 (0.070)	0.001 (0.054)	0.091 (0.101)	-0.391 (0.246)
	P	0.200*** (0.042)	0.078 (0.134)	-0.065 (0.235)	0.001 (0.066)	-0.746 (0.785)
	NP	-0.309 (0.224)	1.065† (0.623)	-0.107 (0.286)	0.419 (0.311)	-11.951*** (0.290)
	GN	-0.229*** (0.035)	-0.330*** (0.062)	-0.095† (0.057)	-0.101 (0.129)	-0.126 (0.246)
	GP	-0.043 (0.053)	0.018 (0.173)	-0.617* (0.277)	-0.006 (0.115)	-0.526 (0.864)
	GNP	-0.267* (0.107)	0.345 (0.272)	-0.102 (0.270)	0.000 (0.330)	0.025 (0.628)
Age		-0.134*** (0.033)	-0.225*** (0.066)	-0.247*** (0.064)	-0.321*** (0.083)	0.262 (0.278)
Age^2		0.015*** (0.003)	0.022*** (0.007)	0.024*** (0.007)	0.032*** (0.009)	-0.012 (0.035)
Number of Authors		0.006*** (0.001)	0.004** (0.001)	0.008*** (0.001)	-0.001 (0.004)	-0.018** (0.006)
International		-0.128*** (0.020)	-0.216*** (0.036)	-0.213*** (0.041)	-0.151* (0.054)	-0.412* (0.199)
Constant		-0.249*** (0.077)	-0.018 (0.151)	0.073 (0.148)	0.217 (0.187)	-1.199* (0.505)
N		35229	11798	9525	6005	746
χ^2_{Wald}		49966.44***	18106.42***	14541.46***	8831.90***	38525.56***
Pseudo R^2		0.1181	0.1253	0.1191	0.1203	0.1506

† p < 0.1 * p < 0.05 ** p ≤ 0.01 *** p ≤ 0.001

Discussion and Conclusions

Science philanthropism is one of the major sources of funding in science. Yet the role of this source of funding is less studied compared with public and private funders. Philanthropists play a critical role specifically in the context of global health challenges. In these areas, philanthropic organizations are spending significantly large donations to support health research in order to achieve bold goals such as to "cure, prevent, or manage all diseases by the end of this century[2]". Funding from public sources, on the other hand, has stagnated over the past decade and competition has become tougher [6], [20].

Different sources of funding have different priorities and their choices can affect science [16]. By providing funding to specific areas of inquiry, endorsing certain methods and approaches, and rewarding some individual scientists and institutions, funding agencies shape science and scientific paradigms [4]. In this sense, understanding the effects of different sources of funding is a crucial science policy problem. In this study, we aim at investigating the effect of philanthropic money in science. We ask and answer questions that help us illustrate how the flow of funding

[2] https://www.theguardian.com/technology/2016/sep/21/mark-zuckerberg-priscilla-chan-end-disease

from philanthropies sway the scientific research directions in the context of health challenges and their impact on science enterprise.

We look at two dimensions of philanthropies' role in science: content and impact. Our benchmarks in this study are the public and private sources of funding and we highlight the differences between these sources of funding. First, we find the topics that philanthropies tend to have a higher focus on when investing in health studies, as compared with public funders. Second, we investigate the relationship between philanthropic, public, and private sources of funding in terms of the contents they tend to fund. Finally, we analyze the impact of studies funded by philanthropies by using citations as a proxy to impact.

Our results imply that philanthropies tend to have a more practical approach to health studies compared with public funders. We find this practical approach in different areas of inquiry and across different diseases. They tend to focus more on costs, cost-effectiveness, health policy and management. They also tend to have a higher concentration on health services, community-based intervention and prevention programs, and vaccination. We also find that they tend to be concerned with social and behavioral context of the diseases. Some scholars consider philanthropists as business-oriented organizations that tend to push technical solutions for health problems [18]. Our results suggest that this view does not provide a complete image of science philanthropism in the context of health studies. While we show that philanthropies do focus on practical medical solutions, we also observe that they consider the non-medical aspects and contexts in their fight against diseases.

We show that philanthropies mix and combine scientific contents supported by public and private sectors in the context of global health challenges. Our results are consistent with prior studies that argued philanthropies fill the gap created by other sectors [7]. We further show that in combining knowledge from these sources, philanthropies tend to be closer to the position of public sector on studies of public health challenges. There is some variation, in our results, on how close philanthropies are to public sector in supporting health studies. Such variation might be attributed to different factors such as epidemiological aspects or research stages of the diseases. Future studies are needed to examine why there is variation in philanthropies' positioning in covering the topics funded by public and private sectors.

We also show that studies funded by philanthropies tend to receive higher citations, and hence have higher impact, in comparison to those funded by public sector. This phenomenon is observed within the context of HIV/AIDS, Tuberculosis, Malaria, and Neglected Tropical Diseases. The philanthropic impact boost is shown not to be because of the contents of studies that are funded by philanthropic sector. Also, we have tested whether few high impact papers funded by philanthropies are increasing the average citation effect. Our results show that the effect is persistence in the context of HIV/AIDS, Tuberculosis, and Malaria even after controlling for the content and exceptionally highly cited papers.

The only exception to the philanthropic impact advantage is Pneumonia studies. Studies of Pneumonia are different from the other diseases in at least two aspects. First, the focus of Pneumonia studies on basic research is significantly lower in comparison to other diseases (Figure 3.2). Second, Pneumonia is the only disease in our data set that received less public funding than private funding (Figure 3.1). Based on these differences, one potential explanation for lack of philanthropic impact boost is that the research stage of this disease is not in the best orientation with philanthropies' capabilities. In fact, in the context of Pneumonia research, studies

funded by private sector have a higher citation count with respect to those funded by public sources. This strengthen the possibility that Pneumonia is best controlled through biomedical research funded by the private sector.

Our results have several implications for scientific institutions, funding agencies, and science policy makers. First, scientific institutions should consider incentivizing their researchers to seek for philanthropic funding. Given the decreasing chances of receiving federal funding for research through grant proposal submission process, this alternative source can be credible and effective for scientists' research [34]. Moreover, as we have shown in our analysis, philanthropic funding is associated with an impact boost compared with public funding. Seeking for philanthropic funding might be encouraged through different channels such as universities' collaboration with philanthropies, scientists' intra and across institution collaboration, and providing information to scientists regarding funding opportunities from philanthropies.

Second, philanthropies should focus on the areas of inquiry with significant public funding or where they have the power to attract public funds. Philanthropies tend not to have a high scientific impact when focusing on problems abandoned by public sources.

Third, policy makers should consider encouraging and supporting philanthropism in science. Our results show that philanthropic money tends to mic and combine the areas of inquiry supported by public and private sources of funding. To enforce this role, the impact of scientific philanthropic should be acknowledged and supported.

Our study have several limitations. First, in our analysis we rely on the self-reported funding acknowledgement of publications' authors. This poses our analysis to the self-selection bias of these authors. Although we do not have any reason to believe that this shortcoming may have affected our results, the generalizability of our findings should be done with extra caution as we did not have access to all papers funded by all different funding sources.

Second, we do not have access to the financial data of grants supporting publications in our analysis. Our study could be fed by the dollar value information of funding grants. However, this information is not available on the funding acknowledgement section of the papers. Subsequently, we have considered all funding agencies as equal when analyzing the effect of different sources of funding. Such approach has obvious shortcomings as it considers a hundred thousand dollar funding equal to a multimillion dollar grant. Moreover, it is unclear how much money has actually been spent on a research project related to the papers in our data set. The use of the money is also latent in our data set. For example, we cannot be sure whether funding has been used to purchase equipment, conduct experiment, or pay for the tuition and stipend of graduate students.

Third, the time-span of our data set is short and limited. Therefore, we could not generate causal inferences about the changes in fields of study due to different sources of funding. Ideally, it would be great to analyze how philanthropic money causes change in studies of different diseases and how the career of authors supported by these funding changes over time. Unfortunately, our data limitations do not let us conduct these analyses.

This work also motivates future studies and invites researchers to work on further research questions regarding the role of philanthropic money in science. It is important to understand and distinguish between the mechanisms and processes that philanthropies use to select projects they fund and compare that with public sources of funding. The results of such inquiry increases public awareness about the effects and roles of different organizations in shaping science. It also

pushes for more transparency in the selection of research areas funded by the philanthropic sector. The variation in the relationship between philanthropic, public, and private funders of research should be investigated to provide an understanding about the underlying mechanisms leading to such variation in the context of more diseases. Moreover, the effect of philanthropic funding on the science enterprise in developing countries should be investigated. Here, we analyzed the effect of philanthropies from a global point of view. However, recent studies have shown that the context matters in complex socio-medical problems [31]. It is important to study whether philanthropies' effect on science enterprise is the same in developed and developing countries.

In this paper, we provide a large-scale analysis of science philanthropy with focus on scientific publications in the context of healthcare. Our results shed light on the growing role of philanthropic money in science. We identify the themes of research that have received higher attention by philanthropies compared with public funders. These not only include technical and practical themes of research but also economic, and policy related topics. Furthermore, we show that philanthropies mix and combine topics supported both by public and private agencies. Finally, we find evidence that papers funded by philanthropies are more likely to receive higher citation counts. The higher impact was consistently found in diseases where public funding tend to be dominating the global research and development spending.

Acknowledgements

The National Institute of General Medical Sciences and the Office of Behavioral and Social Sciences Research of the National Institutes of Health (NIH) supported this work (Grant 2U01GM094141-05). We thank Griffin Weber for providing us with the publication data from Web of Science dataset.

References

[1] P. E. Stephan, *How economics shapes science*. Cambridge: Harvard University Press, 2012.

[2] S. Kearney, F. Murray, and M. Nordan, "A new vision for funding science," *Stanford Soc. Innov. Rev.*, pp. 50–55, 2014.

[3] R. F. Viergever and T. C. C. Hendriks, "The 10 largest public and philanthropic funders of health research in the world: What they fund and how they distribute their funds," *Heal. Res. Policy Syst.*, vol. 14, no. 1, 2016.

[4] T. Kuhn, *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[5] P. Bourdieu, "The specificity of the scientific field and the social conditions of the progress of reason," *Soc. Sci. Inf.*, vol. 14, no. 6, pp. 19–47, 1975.

[6] E. M. Ohman, P. S. Douglas, L. B. Dean, and G. S. Ginsburg, "Philanthropy for Science," *Circ. Res.*, vol. 119, no. 10, pp. 1057–1059, 2016.

[7] J. Youde, "Private actors, global health and learning the lessons of history," *Med. Confl. Surviv.*, pp. 1–18, 2016.

[8] O. T. Flatto, “The case of philanthropy: bringing scientists and philanthropic donors together, for good,” *Dis. Model. Mech.*, vol. 8, no. 9, pp. 1011–1012, 2015.

[9] F. Murray, “Evaluating the Role of Science Philanthropy in American Research Universities,” *Innov. Policy Econ.*, vol. 13, no. 1, pp. 23–60, 2013.

[10] W. Scaife, “Venturing into Venture Philanthropy: Is More Sustainable Health and Medical Research Funding Possible Through Venture Philanthropy and Social Entrepreneurship?,” *J. Nonprofit Public Sect. Mark.*, vol. 20, no. 2, pp. 245–260, 2008.

[11] A. Birn, “Philanthrocapitalism , past and present: The Rockefeller Foundation, the Gates Foundation, and the setting (s) of the international/global health agenda,” *Hypothesis*, vol. 12, no. 1, pp. 1–27, 2014.

[12] P. J. Garcia and W. H. Curioso, “Strategies for aspiring biomedical researchers in resource-limited environments,” *PLoS Negl. Trop. Dis.*, vol. 2, no. 8, pp. 274–276, 2008.

[13] A. E. Birn, “Gates’s grandest challenge: Transcending technology as public health ideology,” *Lancet*, vol. 366, no. 9484, pp. 514–519, 2005.

[14] “G-FINDER.” [Online]. Available: http://policycures.org/gfinder.html.

[15] J. Eckl, “The power of private foundations: Rockefeller and Gates in the struggle against malaria,” *Glob. Soc. Policy*, vol. 14, no. 1, pp. 91–116, 2014.

[16] K. R. W. Matthews and V. Ho, “The grand impact of the Gates Foundation,” *EMBO Rep.*, vol. 9, no. 5, pp. 409–412, 2008.

[17] S. R. Smith and K. A. Grønbjerg, “Scope and theory of government-nonprofit relations,” in *The nonprofit sector: A research handbook*, 2nd ed., 2006, pp. 221–242.

[18] J. Youde, “The Rockefeller and Gates Foundations in Global Health Governance,” *Glob. Soc.*, vol. 27, no. 2, pp. 139–158, 2013.

[19] P. Azoulay, J. Graff Zivin, and G. Manso, “Incentives and Creativity : Evidence from the Howard Hughes Medical Investigator Program,” *Rand J. Econ.*, vol. 42, no. January, pp. 527–554, 2011.

[20] F. C. Fang and A. Casadevall, “Reforming science: Structural reforms,” *Infect. Immun.*, vol. 80, no. 3, pp. 897–901, 2012.

[21] R. J. Daniels, “A generation at risk: Young investigators and the future of the biomedical workforce,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 2, pp. 313–318, 2015.

[22] J. L. Wheeler, S. A. Rum, and S. M. Wright, “Philanthropy, Medical Research, and the Role of Development,” *Am. J. Med.*, vol. 127, no. 10, pp. 903–904, 2014.

[23] C. A. Hart and S. Kariuki, “Antimicrobial resistance in developing countries.,” *BMJ*, vol. 317, no. 7159, pp. 647–50, 1998.

[24] K. R. Foster and H. Grundmann, “Do we need to put society first? The potential for tragedy in antimicrobial resistance,” *PLoS Med.*, vol. 3, no. 2, pp. 0177–0180, 2006.

[25] R. Agarwal and A. Ohyama, “Industry or Academia, Basic or Applied? Career Choices and Earnings Trajectories of Scientists,” *Manage. Sci.*, vol. 59, no. 4, pp. 950–970, 2013.

[26] R. Light and jimi adams, “Knowledge in motion: the evolution of HIV/AIDS research,”

Scientometrics, vol. 107, no. 3, pp. 1227–1248, 2016.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.

[28] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *Adv. Neural Inf. Process. Syst. 22*, pp. 288--296, 2009.

[29] S. Kaplan and K. Vakili, “The double-edge sword of recombination in breakthrough innovation,” *Strateg. Manag. J.*, vol. 36, pp. 1435–1457, 2015.

[30] T. L. Griffiths and M. Steyvers, “Finding scientific topics.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101 Suppl, pp. 5228–35, 2004.

[31] A. Baghaei Lakeh and N. Ghaffarzadegan, “Global Trends and Regional Variations in Studies of HIV/AIDS,” *Sci. Rep.*, vol. 7, 2017.

[32] J. Wang, R. Veugelers, and P. Stephan, “Bias against novelty in science: A cautionary tale for users of bibliometric indicators,” *Res. Policy*, vol. 46, no. 8, pp. 1416–1436, 2017.

[33] S. Wuchty, B. F. Jones, and B. Uzzi, “The Increasing Dominance of Teams in Production of Knowledge,” *Science (80-.).*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[34] P. S. Hinds and M. S. Weaver, “Can Cancer Research Careers and Study Findings Be Credible if Funded by Philanthropy?,” *Cancer Nurs.*, vol. 38, no. 3, pp. 165–166, 2015.

Chapter 4: Recipe for success in academia: a dynamic model of scientists' career

Abstract

Scientists continuously face a critical resource allocation question: Should they spend their limited time more on writing proposals or on writing academic papers? Conceptually, this is a workforce development question concerned with resource allocation strategies under constrained resources. We develop a mathematical model of the career-long success of a scientist as a function of her temporal resource allocation decision between writing funding proposals and papers. We simulate and analyze different modes of research production over one's career, and explore optimal solutions. We show that, first, optimal strategy for allocation of one's labor to writing papers is different for various funding distribution schemes. Specifically, we show that funding schemes aiming at supporting the best proposals will reinforce time spent on proposal writing and subsequently imped researchers' performance. Second, we find that the optimal strategy changes as receiving funding becomes more uncertain. With higher stochasticity in funding allocation, the optimal time on writing grant proposals increases which negatively affects the entire science system's performance. Third, we investigate the case of competing rational scientists all pursuing optimal resource allocation strategies, and show that the Nash equilibrium strategies are different for the two cases of 'Winner Takes Some' and 'Winner Takes All' competitions. In the case of 'Winner Takes All', there is a mixed-strategy equilibrium. This potentially explains variation in the practice of professorship. Finally, we show that even under the optimal strategy, a considerable proportion of scientists end up failing. The results have several implications for science policy and scientific workforce development.

Introduction

Consider Mary, a junior scientist who is recruited as an assistant professor after years of graduate studies and professional training. She potentially starts her academic job with a few publications and some ideas on how to pursue her research further. As part of her startup package, her university offers Mary a start-up funding that can help her conducting research for a limited time. Mary realizes that she needs to secure external funding to sustainably pursue her own research agenda. This requires writing grant proposals for the funding agencies that support research projects related to her area of expertise. Both Mary's former advisor and her current mentor tell her that having a brilliant idea is not enough to have her proposals funded; she also needs to maintain a strong record of publications to increase her chances of receiving funding. Funding needs publications; publications need funding. A puzzling question is: how should she allocate her limited research time between writing grant proposals and papers?

Today's high cost of research and the dim chances of receiving funding elevate the urgency of this question. In academia, laboratory equipment and the high-skilled labor (including graduate students and post-docs) add to the increasing cost of research. Running a research laboratory at a public university is estimated to cost, on average, $550,000 annually [1]. Competition for

receiving funding has also increased over time. In the context of National Institutes of Health as a case example, the success rate of research proposals has fallen by more than half during the past half century [2].

Universities and other academic institutions do not afford to support all researchers with their own internal funding sources. Many governments around the world make an effort to provide federal funding for science. In the United States, following the Second World War, the federal government ramped up its spending on research and development [1]. In the following decades, various agencies were founded with the aim of supporting scientists' research in the academic settings. Allocation of funding based on the research grant proposals, written and submitted by the scientists, soon became the norm in these agencies [3].

Today, scientists need to spend a significant amount of time on writing grant proposals to secure funding for their research labs [3]–[5]. With the persistent decrease in chances of receiving funding over the recent years, the burden has become heavier on scientists to invest more time on writing higher quality proposals in the hope of receiving funding for their research [2]. But this question remains unanswered: how much of a scientist's research time should be allocated to writing grant proposals in order to maximize her career-long publications?

In this paper, we develop a system dynamics model to analyze the career-long research output of a scientist as a function of her resource allocation between writing papers and grant proposals. In doing so, we take into account the interconnectedness of scientist's ability to secure funding for her research endeavors and her research output. We, furthermore, incorporate the learning mechanisms and capability development of the scientist over her career and as she engages with different activities.

Success in Science and the Career of Scientists

Understanding the underlying processes of scientific inquiry and the elements of success in academia have been a subject of analysis for scientists and philosophers for many years. Thanks to the availability of bibliometric data and the progress in the field of data analytics, such studies have accelerated in recent years. A major line of research in this area focuses on understanding the characteristics of successful scientific publications as the product of the scientific inquiry.

The innate quality of a paper has shown to be the only significant factor in driving its long-term success and impact. In an analysis of hundreds of thousands of publications in different fields of physics, Wang *et al* show that the only factor that explains the long-term performance of a paper is its *fitness* within the field of study [6]. This finding corroborates with Merton's view on how the science enterprise motivates scientists to do research and publish their results by designing a merit-based incentive system [7]. The fitness of a research publication is a very aggregate concept. What quality makes a paper to be more fitted within a field of science? Relevancy can be one such quality. When considering the impact of a research publication, it is shown that there is a *hotspot* in the time distribution of the prior knowledge that the current research relates to. Publications that cite recent literature but maintain a high variance in the knowledge age of their references tend to have a higher impact [8].

The level of innovation in the work is another driver of publications' impact. It is shown that innovative projects tend to be high risk and high reward in comparison to more conventional and

incremental studies [9], [10]. Recombination of knowledge from relatively distant fields of inquiry has shown to be correlated with higher levels of innovation and impact. In a study of scientific publications, Uzzi *et al* have shown that the highest-impact publications not only include numerous bridges between conventionally connected domains of knowledge, but they also include novel combinations of scientific domains that are less common in a field of study [11]. Similar effects have been observed in the context of patents and studies of innovation [12].

Team collaboration can boost innovation and enhance the long-term impact of scientific publications [13], [14]. Overall, teams are shown to be more inclined to combine distant domains of knowledge [11]. However, the relationship between team size and novelty has an inverted-U shape. As the team size increases, so is the possibility of novel recombination. This effect is due to the growth in diversity of ideas. However, very large teams poses the research project to the obstacles of project management challenges and impede novelty. The effect of team on the success of papers is specifically larger when the team members have diverse knowledge or task expertise [15]. Moreover, different teams have different impacts and not all the teams are the same. For example, it is shown that while smaller teams tend to work on research problems that are novel and disruptive in the context of scientific evolution, larger teams tend to develop and refine the existing ideas [16].

Another approach to understanding what makes a scientific publication successful is to analyze its authors. Scientists' choice of problems and their ability to exploit the nested research opportunities of those problems determine the fate of a scientific publication [10], [17]. Figuring out who will be a successful scientist is important for scientific institutions and a stream of research is dedicated to measuring and prediction of successful scientists based on their prior achievements [18]–[21]. Measures such as number of publications, citation counts, and *h* index are being used in the process of evaluating junior scientists during their tenure-track stage of careers. In one study, Bertismas *et al* use centrality metrics in the networks of citations and co-authorship and propose a model that predicts the success of a scientist given information about their performance in the first five years of their career. They show that their model outperforms the tenure committees in predicting the future success of scientists at least in the field of operations research [22].

Other than individual factors, institutional contexts are shown to affect the success of scientists. For example, scientists who have trained or work in more prestigious institutions tend to be more productive at least in the early stages of their career [3], [23], [24]. Availability of funding also correlates with higher productivity [25]. Moreover, the funding mechanism can affect scientists' success and there is variation in the effect of funding on scientific inquiry. In a comparison study of investigators funded by the Howard Hughes Medical Institute and those funded by the National Institutes of Health, Azoulay *et al* show that funding agencies' differences matter in the success of scientists. They show that scientists who receive funding from agencies that tolerate early failure, hold a long-term view on the success of projects, and provide the investigators with more freedom tend to publish more high-impact scientific papers [26].

Scientists' ability to publish scientific papers with the highest impact, in comparison with their own papers, is shown to be unrelated to their stage of career. In other words, a scientist's most cited paper can be any of her publications regardless of how many papers she has already published [17]. Despite their constant ability of exploiting a research opportunity, scientists' productivity changes over their career. The conventional view about the career trajectory of scientists holds that their productivity will go through a rapid rise and a subsequent slow decline over time. Way

et al challenge this view and show that there is a lot of variation in the career trajectory of scientists. For example, although the productivity peak usually happens during the first five years of scientists' career, there is variation in the exact timing of this peak and the increasing or decreasing productivity rates during scientists' careers [24].

Scientists are the workforce of science enterprise and their career choices and career trajectories shape the future of science and innovation. Some studies have focused on investigating the professional choices of scientists in the context of the careers they pursue, types of problems that they choose to work on, and the external forces that affect these decisions. Bourdiue's field theory of science holds that scientists' choices about the type of questions to work on and their view about the innovative ideas depend on their relative position in the recognition space of a field of science [9]. Moreover, scientists with higher research abilities[1] are more likely to work on basic research rather than applied research in the context of academia. However, such relationship does not exist in the context of industry research jobs [27].

Both internal and external forces affect the career choices of scientists. Two distinctive career paths for scientists are whether joining an academic institution or an industry firm. Such decision has profound effect on the type of the research that scientists can pursue in their career. Academic environment is perceived to value basic research more than applied ones and provide scientists more freedom in their scientific inquiry. On the other hand, industry firms tend to provide more resources for the scientists but also enforce more limitations on their research agenda [28]. Overall, it is shown that scientists who have a stronger "taste for science" (i.e., preference for freedom in choosing research questions and publishing their results) are more likely to join academic institutions [27], [29].

Competition and endogenous characteristics of the job market have an influence on the human capital development of scientists and their careers [27], [30], [31], [32]. Although some scientists do not have the preference of joining academic institutions, many will have no choice but to leave academia because there is not enough academic positions available for everyone. Considering the average for all fields of science in the United States, only 12.8% of PhD graduates can land an academic position assuming a steady state workforce and job market [33]. The shortage of permanent academic positions and the subsequent competition for those positions have lead many PhD graduates to take temporary research positions such as postdocs in their early careers. Effect of different macro policies aiming regarding this temporary workforce of science (such as capping the duration or an increase in government spending) has shown to be heterogeneous, highlighting the inner-sectoral differences of scientific institutions and workforce [28], [34].

During their careers in academia, scientists take different roles and need to perform a variety of tasks at the same time. The role of scientists or faculty members in their research projects and the amount of time they invest in doing research depends on a variety of individual, organizational, and nation-wide factors. Overall, it is shown that faculty members start to take more leadership roles in their research projects as they progress in their careers. Over the length of scientists' careers, the ratio of first-author papers declines while the ratio of last-author papers rises. This transition tend to be faster in more prestigious institutions [24].

[1] Research abilities are measured by 1) time to complete first baccalaureate degree, 2) ranking of the PhD program, 3) whether the scientist received grants during doctoral program, and 4) education levels of parents [27].

Faculty members who are more interested in research, have recently published papers, and have access to funding tend to invest more time in doing research and their research output is on average higher [25], [35]–[37]. Moreover, the time that scientists allocate to doing research changes as they move in their career paths. For example, assistant professors tend to invest more time on research and teaching than service. The research and service time tend to increase during the career of scientists until their mid-career point and then decreases [37].

The time that faculties allocate to research also depends on the institutional and nation-wide factors. The evidence for such factors can be seen in the variation of research time spent by faculties in universities in different countries [35], [36]. Moreover, the amount of time allocated by faculties on research as opposed to teaching and their view on the relationship between research and teaching vary depending on the incentive structure of the institutions in which they work [37]. For example, on average, faculties find their research activities to be complementing their teaching more in research oriented institutions [35].

Although some studies have investigated the multi-task responsibilities of faculty members and their time allocation between these tasks (including research, teaching, and service) [24], [35]–[38], studies that aim at disentangling the different tasks of faculty members in the process of conducting research are scarce. Once exception is the work by Geard and Noble in which they look at the allocation of time spent by the scientists on writing grant proposals versus writing papers. By assuming certain decision rules for scientists, they show that a competitive bidding system for allocation of funding will hurt the efficiency of science as a whole [39].

The research career of a scientist can be modeled by employing a resource-allocation framework. Every scientist possesses a set of resources that can be put in to use during the process of conducting research. Such resources include scientist's time, attention, experience, social network (or collaborators), human resources (including graduate students and postdocs), and laboratory equipment. These resources have different characteristics. For example, some can be produced by others, some depreciate over time, and some may only be possessed by a person to a certain degree because of the specific physical constraints. People not only are different on the basis of the resources they possess, but also with respect to how efficiently they can use them. The strategic decisions of conserving or building these resources over time can determine the variation in scientists' performance over the long run.

Some resources can only be possessed in a finite amount and will deplete as they are being put into use [40], [41]. Time is an example of such resources. There is a limit for the amount of time that everyone has access to and as the time is being used for a purpose the amount of it available for other purposes declines. These resources are essential inputs for cognitive processes [41]. Other type of resources can be possessed with no constraints and do not deplete as being put into use. Although these resources do not play a direct role in cognitive processes that produce desirable outputs, they affect the productivity of people as they engage in cognitive processes [40], [41]. In the context of problem solving, for example, experience is a resource of this type. The more experienced a person is, the higher will be her productivity when investing time in the process of problem solving.

When the resources do not deplete while being used, a resource-allocation problem seems redundant. However, when the resources are exhaustible and when only a limited quantity of them is available, proper allocation can become a challenging problem. The resource-allocation framework has been used in variety of fields to understand the dynamic of success (e.g., in

lifespan development [40]–[42]; strategy [43]; science [39]). In this study, we employ this framework to analyze research career-trajectories of scientists.

The dynamic nature of a career trajectory and the long-term effect of resource allocation decisions are important in modeling the research career of a scientist. For example, it is shown that if a scientist's stock of knowledge and experience passes a tipping point, the scientist can maintain a high level of production for a long time [44]. This highlights the importance of early career success and provide an explanation on why the newest cohort of faculty members are not always the most productive one [45]–[47].

Despite the range of analysis performed to study the performance of scientists and the science enterprise, only few has examined optimal resource allocation with the intention of maximizing long term performance of individuals and organizations. This is specifically important given the diminishing chances of receiving funding from major funding sources [2].

In this paper, we create a system dynamics model for analyzing a scientist's career. We focus on the strategies that a scientist can choose to allocate her limited research time between different two distinct activities that are crucial in nowadays' science enterprise: writing papers and writing grant proposals. Our goal is to find the best strategy that the scientist can follow to maximize her long-term research output under different organizational contexts.

Model

We build a simple system dynamics model of scientist's research activities. Figure 4.1 shows the main logic of the model which we will formulate.

In this model, any scientist has to allocate their limited time between two activities: writing grant proposals and writing papers. Everything else kept constant, more time on paper results in more research output, and more time on proposal results in more new funding. If we consider the case of a scientist who aims at maximizing their research payoff over their career, spending time on writing papers provides them with immediate benefit aligning with this long-term goal. However, the focus on writing papers means less time on proposals. Research output depends on time spend on writing papers as well as funding. Moreover, proposal competitiveness depends on one's past accomplishments.

It is also important to state that both wiring papers and grants proposals have learning curves, and scientists often get better as they perform more. These mechanisms are depicted by two reinforcing loops of learning by doing.

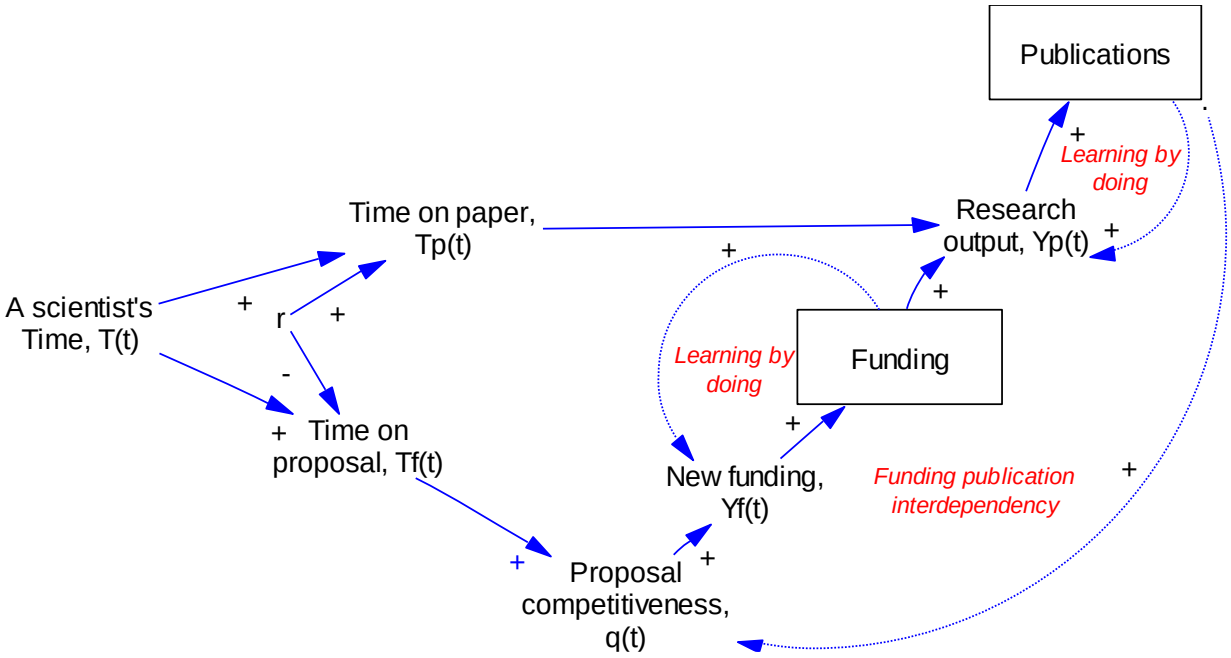


Figure 4.1: Conceptual model.

The difficult task for a scientist is to decide how she wants to build these capabilities over her career while considering their interdependency and inertia. Understanding the different dimensions of this strategic decision making problem is not trivial and an optimal strategy may be affected by the organizational context in which the scientist works in. Here, we use a formal model based on the conceptual model explained here to find this optimal strategy under different institutional settings.

Resource allocation: We assume any scientist have a limited time at each time period to be spent on writing papers or grant proposals. Let $T(t)$ represent the available research time for a scientist at time t, and $r(t)$ the ratio of scientist's total time dedicated to writing papers at time t. We will have:

$$T_P(t) = r(t) \cdot T(t) \qquad (1)$$

$$T_F(t) = (1 - r(t)) \cdot T(t) \qquad (2)$$

where $T_P(t)$ and $T_F(t)$ represent time spent on writing papers and writing funding proposals respectively. It is expected that scientist's time on research change over one's career, as they get assigned to more service activities and potentially pass the tenure stage. Prior studies have shown that the total time dedicated to research by scientists decline over time [36], [47]. Based on the results from these studies ([36]), we use a simple linear function to represent total time for research as following:

$$T(t) = T(t_0)(1 - 0.0077t) \qquad (3)$$

As people spend more resources on writing papers, everything else kept constant, they will write more papers, and get less funding, and vice versa. However, as stated before, funding is expected

to have an indirect effect on writing papers, so they need to keep enough resources for writing proposals as well.

Learning by doing: The scientist learns how to have a higher productivity in writing papers and proposals as she engages with these activities more. In other words, there exist learning curves as related to these two activities. The more a scientist writes papers, the more effective she will be in writing papers in the future and the scientist's labor in the process of writing papers will become more productive as she publishes more papers. This process is in agreement with studies of experiential learning [48]. By following conventional formulations of learning [49], we model this learning process by introducing a learning coefficient, λ, and formulate the effective time dedicated to research ($\hat{T}_P(t)$) and writing proposals ($\hat{T}_F(t)$) as following:

$$\hat{T}_P(t) = T_P(t) \cdot \mathbf{P}(t)^{\lambda} \qquad (4)$$

$$\hat{T}_F(t) = T_F(t) \cdot \mathbf{F}(t)^{\lambda} \qquad (5)$$

In equations 4 and 5, $\mathbf{P}(t)$ and $\mathbf{F}(t)$ are the stocks of scientist's recent publications and funding, respectively. The learning coefficient, λ, takes values between 0 and 1 and captures how much and to what extent learning is important in these two activities. The extreme case of no learning can be modeled by choosing the value of λ to be equal to zero.

Writing papers: In order to do research and write papers one needs to have time and funding resources. Funding is necessary to potentially pay for students' tuition and stipend or laboratory equipment. We use a Cobb-Douglas production function with the constant return to scale and model the publication payoff as following:

$$Y_P(t) = (\hat{T}_P(t))^{\alpha}(K(t))^{1-\alpha} \qquad (6)$$

where $K(t)$ represents funding spending in year t. To keep it simple, and following prior studies, we do not model delays associated with publishing a paper and assume that all papers will be published [47], [50], [51].

We understand that fields differ in terms of relative importance of funding versus professors' direct time toward research. In this equation, the parameter α (or the elasticity factor) represents the relative importance of the scientist's time in comparison to funding in the process of producing publications. Various values of α may be used to model different fields of scientific inquiry with varying levels of dependence on funding. For example, fields such as experimental physics which require expensive equipment and large teams have lower values of α while fields such as theoretical physics which are less dependent on such resources can be modeled with a higher value of α.

Funding: In order to have access to funding, scientists write grant proposals. We assume that the chances of receiving funding for any scientist, at time *t*, depends on the quality of proposal submitted by her during that time period ($q(t)$). We further assume that proposal quality increases by both the amount of effective time spent on writing it, and the researchers' background reflected in their recent publications [52]. Let $\mathbf{P}_{T_p}(t)$ denote publication during the past T_p time. Equation 7 represents such a relationship:

$$q(t) = g_1(\hat{T}_F(t)) \cdot g_2(\mathbf{P}_{T_p}(t)) \qquad (7)$$

The functions g_1 and g_2 represent these effects and need to be strictly increasing with potentially diminishing returns. Without loss of generality, we use the function g as defined below, for both g_1 and g_2 functions, and later conduct sensitivity analysis using various other functions:

$$g(x) = 1 - e^{-x} \qquad (8)$$

Funding will depend on proposal quality (Equation 5) and the funding distribution scheme. For example, factors such as how resources at the government level are allocated among competing proposals, the amount of funding, and accuracy of proposal evaluation affects the funding.

Let, $\mathbf{F}(t)$ denote the stock of funding for a scientist where

$$\mathbf{F}(t) = \sum_{\tau=t-T_f}^{\tau=t-1} Y_F(\tau) \qquad (9)$$

In this equation, T_f is the average funding period, and $Y_F(t)$ is the inflow of funding which depends on the quality of the proposal and the funding scheme and will be defined in the following sections.

As stated, the scientist, as modeled here, engages in two different types of activities: writing papers and writing proposals. The output of these two activities depend on each other's recent outputs. Therefore, a reinforcing loop structure emerges: the more a scientist has recent papers, the higher is her chances of receiving funding, and the more available funding she has access to, the more she can write papers.

Payoff function: As stated we assume professors would like to maximize the total number of papers that they write during their entire career. Following this assumption, considering a career of C years long, the payoff function would be:

$$\pi = \sum_{\tau=1}^{\tau=C} Y_R(\tau) \qquad (10)$$

Distribution of Funding and the Optimal Strategy

Our assumption is that a funding agency decides on how to allocate funding to scientists based on the quality of their proposals. We consider two distribution schemes. First, we investigate the "absolute evaluation" condition were all good proposals get funded. In other words, when the agency does not compare the proposals with each other, but the perceived quality of any proposal is compared with a quality threshold for being funded. Second, we examine a "relative evaluation condition" where proposals compete with each other and based on the quality, funding is allocated to the relatively best proposals.

Absolute Evaluation Based Distribution of Funding

First, we start our analysis by considering the funding distribution scheme based on absolute evaluation of proposals. In this scheme, proposals are being evaluated for their quality and if their perceived quality is larger than a threshold, as set by the funding agency, they will receive funding. For simplicity, we normalize the amount of funding that can possibly be granted to any proposal

to 1. We further assume that funding agency allocates funding to proposals by using the following decision rule:

$$Y_F(t) = \begin{cases} 0 & if\ \hat{q}(t) < \varphi(t) \\ 1 & if\ \hat{q}(t) \geq \varphi(t) \end{cases} \qquad (11)$$

In the equation 11, $\hat{q}(t)$ is the proposal quality as perceived by the funding agency and $\varphi(t)$ is the threshold set by the funding agency for the minimum proposal quality that earns funding. How accurate does the funding agency evaluate the quality of proposals? The error in evaluation of proposals by peer review committees is established in the literature [53]. And if their perception of proposal qualities does not match the actual quality level of proposals, then what is the effect of this error on scientists' optimal strategy? In order to answer to these questions, we introduce an error term to relate the actual proposal quality with the perceived proposal quality as shown in the equation below:

$$\hat{q}(t) = q(t) + \delta\varepsilon \qquad (12)$$

In equation 12, δ captures the amplitude of the error and ε is a random variable drawn from the uniform distribution between -1 and 1.

The career trajectory of the scientist depends on the strategy that she chooses. Next, we simulate the model. The parameters used for the simulation throughout the paper is shown in Table 4.1.

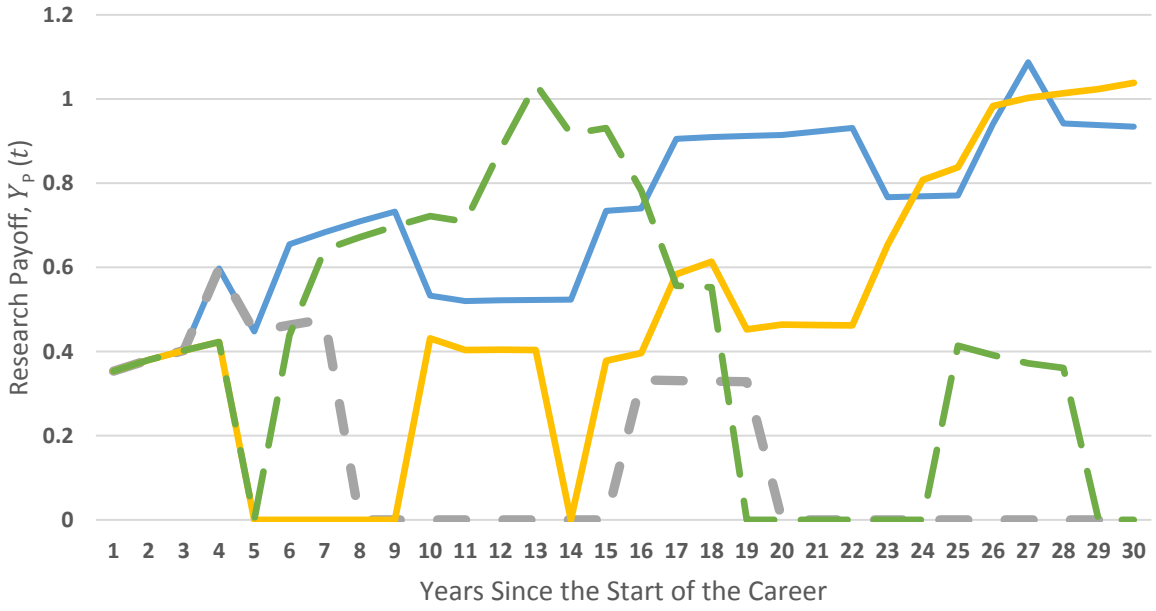
Table 4.1: Parameter values used in paper.

Parameter	Description	Value
$\boldsymbol{\alpha}$	Elasticity factor in the production function of papers. It represents the relative importance of scientists' time in producing papers as compared with funding.	0.5
$\mathbf{P}(t_0)$	Number of papers at the start of career, potentially form PhD or postdoc training	1
$\mathbf{F}(t_0)$	Funding at the start of career, representing start-up package	1
T_p	The period of time in which a paper stays recent and scientifically relevant.	10
T_f	Duration of funding grants.	4
$T(t_0)$	Normalized research time available to the scientist when starting her career as an assistant professor.	1
λ	The learning coefficient.	0.5

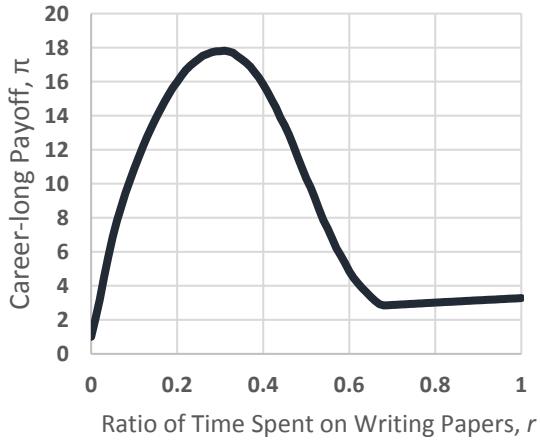
Different modes of career trajectory can be observed in our model. Figure 4.2a depicts few examples for a case example strategy ($r = 0.5$). As it can be seen, some scientists will have a higher productivity in the early years, and some in the later years of their career. Some scientists may experience few years of unproductivity in terms of research publications. The heterogeneity in the career trajectory of the scientists as shown in our model corroborates with recent findings

that challenge the conventional view on the career-trajectory of scientists according to which there is a global pattern of fast increase and slow decrease in productivity of scientists [24].

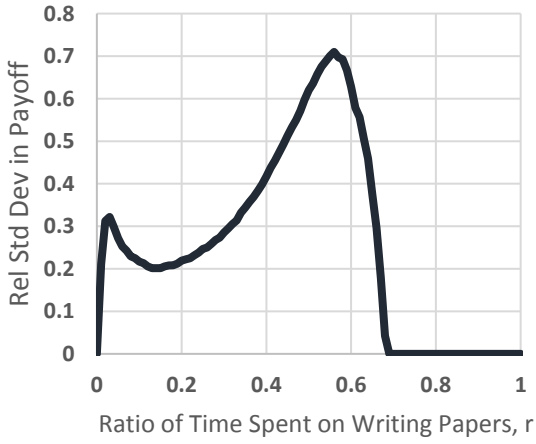
Now, we simulate the model for different values of resource allocation ratio, and for each value, 1,000 times for different random seeds, and investigate career-long payoff and its variation across different runs for various values of time spent on research. First, let's look at the results for two specific values of $\varphi(t)$ and δ before generalizing the results. The simulation results for a case where the threshold ($\varphi(t)$) is set at 0.5 and the evaluation error (δ) is at 0.25 are shown in Figure 4.2. Figure 4.2b illustrates the inverse U-shape relationship between the scientist's expected career-long payoff and her selected strategy. The optimal strategy (r^*) in this scenario is close to 0.31. Any strategy below or above this level, would yield a lower expected payoff for the scientist.



(a)



(b)



(c)

Figure 4.2: Career trajectory of scientists: (a) Few potential career trajectories. (b) Expected career long payoff for different strategies, (c) Relative standard deviation of expected payoffs for different strategies.

The uncertainty about the expected career-long payoff also depends on the selected strategy of the scientist. Figure 4.2c demonstrates the relationship. The maximum relative standard deviation of the expected payoff is found to be at a strategy when the scientist is focusing more on writing papers as compared to the optimal strategy ($r = 0.56$).

We now focus on the optimal career trajectories under different funding distribution conditions. It is important to know how the level of the quality threshold affects the optimal strategy and scientists' career-long payoff. After all, scientists wish to submit proposals that pass the quality check process of funding agencies; as those who have submitted proposals with quality less than $\varphi(t)$ will not receive any funding. Moreover, the effect of uncertainty on optimal strategy should be investigated. We can do so by sweeping different values of δ and looking for changes on the optimal strategy and career-long payoff.

Figure 3 illustrates the optimal strategy, expected career-long payoff, and the relative standard deviation of the career-long payoff for a range of threshold and uncertainty values. In all of the cases, we have considered a constant threshold over the career of the scientist. Overall, as the threshold for receiving funding increases, the optimal strategy of scientists moves toward spending less time on research and more time on proposal writing. This change in optimal strategy means that scientists' long-term payoff decreases as the threshold value increases. Based on these results we propose the following propositions:

> **Proposition 1.1a:** *The optimal amount of time spent by scientists on writing papers is lower when the quality threshold of funding distribution is higher.*
>
> **Proposition 1.1b:** *The maximum career-long payoff of scientists is lower when the quality threshold of funding distribution is higher.*

The interplay of uncertainty and threshold in evaluation of proposals and their effect on scientists' career is more complex. When the threshold is set high, uncertainty in evaluation of proposals have a positive correlation with the optimal strategy. The higher is the error in evaluation of proposals, the more scientists will spend their time on doing research rather than proposal writing. Moreover, the expected payoff of scientists increases as the uncertainty in evaluation of proposals increases. When the threshold is at a low value, however, the scientists' expected payoff increases as the uncertainty decreases while scientists focus more on proposal writing. Assuming a high quality threshold set by most funding agencies, the following propositions can be derived from our model's results:

> **Proposition 1.2a:** *The optimal amount of time spent by scientists on writing papers is lower when the proposal evaluation is more accurate.*
>
> **Proposition 1.2b:** *The maximum career-long payoff of scientists is lower when the proposal evaluation is more accurate.*

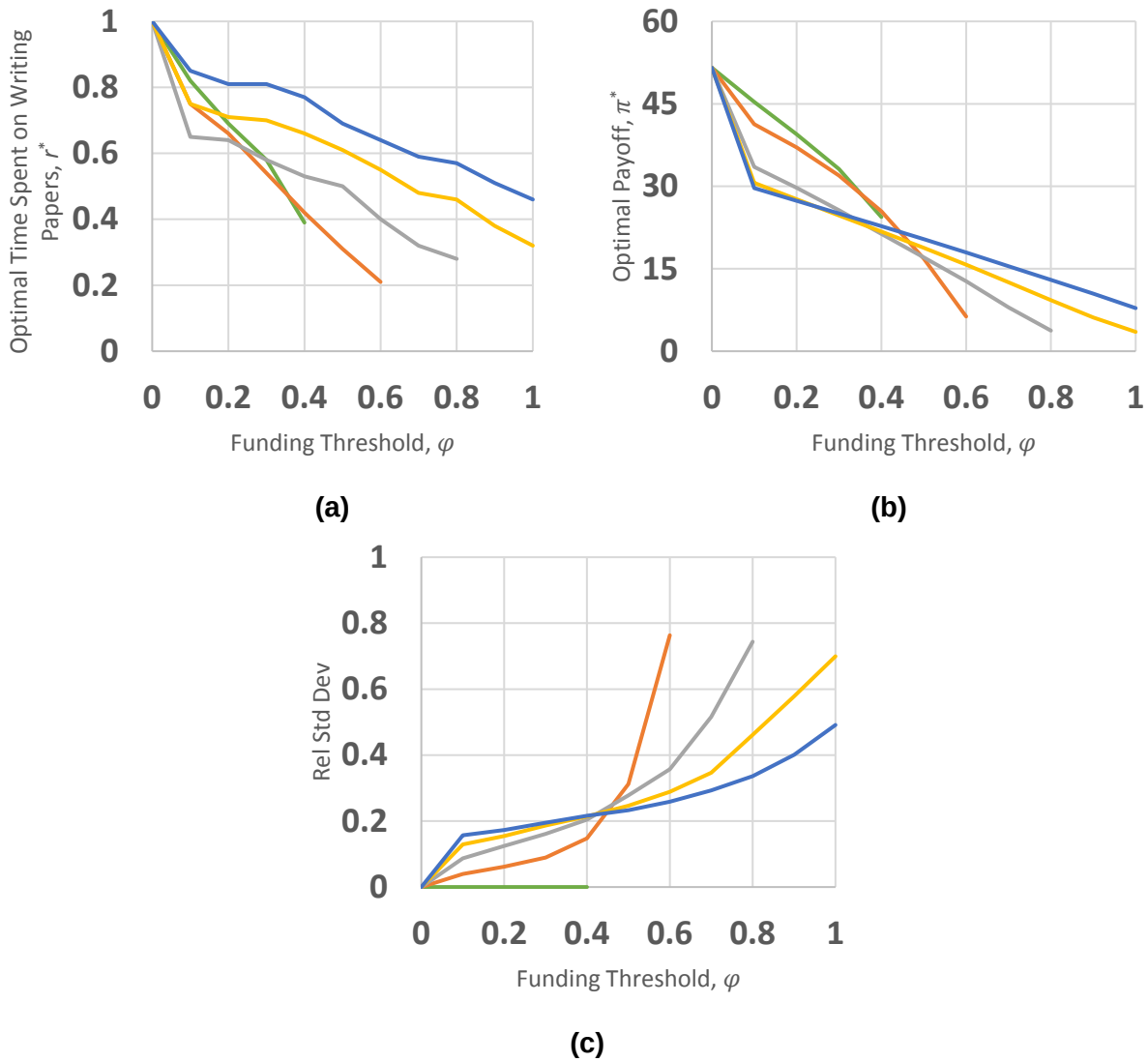


Figure 4.3: Effect of funding threshold and error in evaluation of proposal quality on (a) optimal strategy, r^*, and (b) optimal payoff, π^*. Different colors represent varying levels of error in evaluation of proposal quality: green ($\delta = 0$), orange ($\delta = 0.25$), gray ($\delta = 0.5$), yellow ($\delta = 0.75$), and dark blue ($\delta = 1$). (c) The relative standard deviation of career-long payoff is shown as a function of threshold and evaluation error.

Variation in career-long performance increases as the level of threshold goes higher. Figure 4.3c captures this relationship. By setting a high value for the threshold of proposal quality, the funding agency increases the relative variation of scientists' career-long payoff. The variation that the scientist should expect in her career-long payoff has a positive association with the quality threshold of the funding agency. When the level of threshold is high, the relative standard deviation of payoff has a negative relationship with the evaluation error. Therefore, when the threshold is set at a high value and as the level of evaluation error decreases, the variation in career-long payoff increases. Based on these results, and assuming a high quality threshold, we put forward the following propositions:

Proposition 1.2c: *The variation in the maximum career-long payoff of scientists is higher when the quality threshold of funding distribution is higher.*

Proposition 1.2c: *The variation in the maximum career-long payoff of scientists is higher when the proposal evaluation is more accurate.*

Relative Evaluation Based Distribution of Funding

We now turn our focus to the funding distribution scheme based on relative evaluation of proposals. To model this scheme, we assume that there are *N* scientists who are working on a related topic and are seeking funding from a funding agency. Without loss of generality, we further assume that the funding agency has a total amount of funding dollars equal to *N*. At each time period, the funding agency compares the quality of proposals submitted by the *N* scientists and distribute the total amount of funding between them. We assume that all scientists are acting rational, i.e., they are all trying to maximize their long-term payoff in a selfish manner and by considering the strategies of other scientists.

The funding agency may choose different schemes to distribute the total available funding between the *N* scientists. On the one hand, the funding agency may want to allocate more funding to the best proposals, but on the other hand, it may also want to distribute the funding more evenly between scientists and provide everyone with some opportunity. The major dimension that distinguishes between these schemes of funding distributions is how egalitarian the agency wants to behave. We model different allocation schemes by introducing a variable to capture how unequal the agency wishes to distribute funding, μ (the inequality factor). Using this variable, we model that the funding agency distributes funding at each time period, according to the following equation.

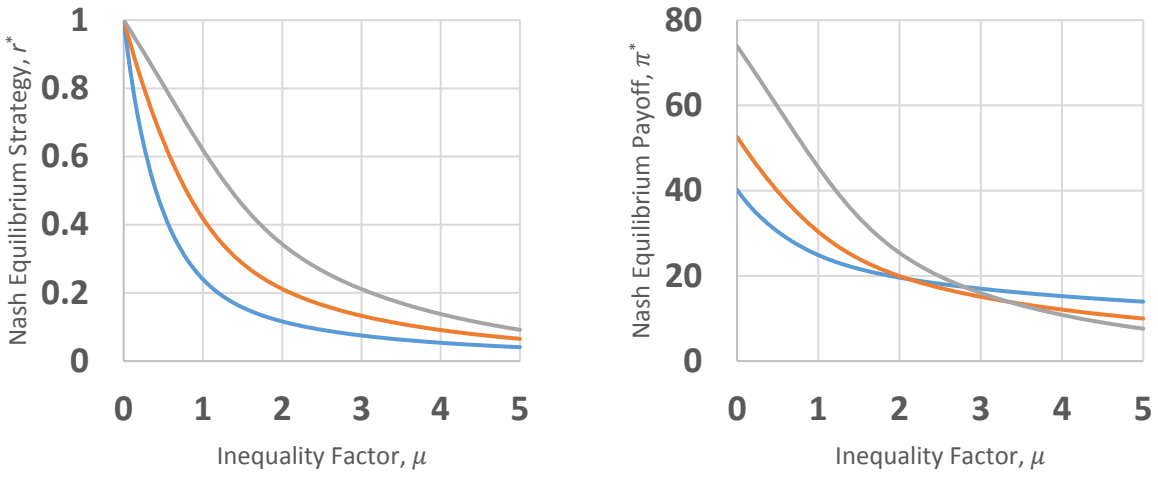
$$Y_{F_i}(t) = \frac{(p_i(t))^{\mu}}{\sum_{j=1}^{j=N}(p_j(t))^{\mu}} N \qquad (13)$$

In this equation, $p_i(t)$ is the proposal quality of the scientist's *i* at time period *t*. It is important to note that when $\mu = 0$, the agency distributes the funding between the scientists regardless of their proposal quality. This is the most egalitarian scheme. As $\mu \to \infty$, the agency allocates all the funding only to the proposals with the highest quality. This is the least egalitarian scheme. In all the cases with a real value of μ, the scientist with the highest proposal quality will receive most of the available funding, but not all of it. For this reason, we label these distribution schemes as Winner Takes Some (WTS) with different inequality factors.

The scientists' optimal strategy negatively correlates with the inequality factor (Figure 4.4). In a completely egalitarian distribution regime ($\mu = 0$), scientists allocate all their time on writing papers. However, as the distribution regime becomes less egalitarian, the optimal strategy suggests that scientists need to invest less and less time on writing papers (asymptotically reaching zero). Note that, in Figure 4.4, each point corresponds to a pure strategy Nash equilibrium of a competitive game for 10 scientists at the specific inequality level. The numerical process for finding the equilibrium points is reported in Appendix B. Moreover, the expected payoff of scientists at equilibrium declines while the inequality factor increases. The following propositions can be generated based on these results:

Proposition 2.1a: *The amount of time spent by scientists on writing papers decreases at equilibrium when the distribution of funding is less egalitarian.*

Proposition 2.1b: T*he career-long payoff of scientists is lower at equilibrium when the distribution of funding is less egalitarian.*



(a) **(b)**

Figure 4.4: Effect of different inequality factors on the optimal strategy (a) and career-long payoff (b) of scientists working under competition. Blue: $\alpha = 0.3$; Orange: $\alpha = 0.5$; and Gray: $\alpha = 0.7$.

We now consider a special case of competition in which the proposals with the maximum quality share all the available funding and the rest receive none. We name this case the Winner Takes All (WTA). The equilibrium strategies for this type of competition are shown in Figure 4.5 for various values of α. Under the WTA competition, specialization becomes vital and hybrid strategies become inefficient. When the relative importance of scientists' temporal resource (α) is low/high in producing publications, the optimal strategy for every scientist is to invest almost all/none of her temporal resources on writing proposals.

We observe a variation in practice for moderate values of α. Under these conditions a pure strategy Nash equilibrium cannot be attained. In other words, one cannot find a strategy that if every scientist follow, no one will have an incentive to deviate from. The proof for the non-existence of such pure strategy equilibrium is provide in Appendix C for a simplified version of the model. Intuitively, we can think of two plausible successful strategies in this condition: to invest almost all resources on proposal writing and receive all the funding at every period, or to invest all resources on research and produce the most papers during the first few years while the initial seed funding is available. The equilibrium for scientists would be to choose either of these two strategies in a fashion that the expected payoff of both become equivalent. For example, when α is close to 0.45, the optimal strategy for scientists is to invest all their resources on writing papers with probability of 0.5 and invest all their resources on writing proposals with probability of 0.5.

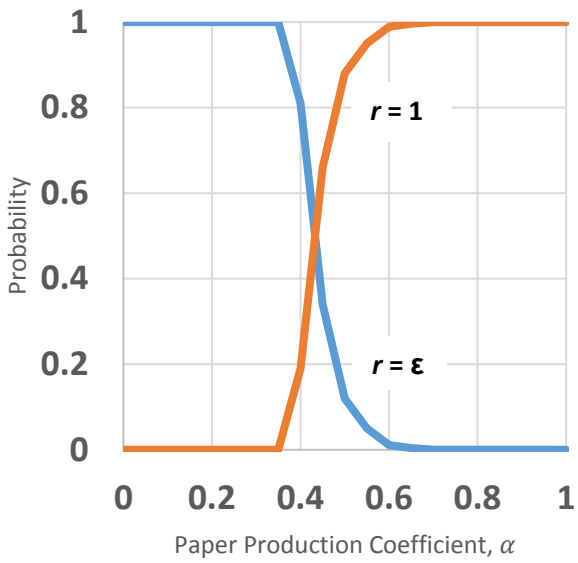


Figure 4.5: The mixed strategy equilibrium under the Winner Takes All type of competition for different values of α. The blue line shows the probability of the strategy to almost invest all resources on proposal writing (very small r). The orange line shows the probability of the strategy to invest all resources on research.

Based on the results from the special case of WTA, we can propose two propositions:

> **Proposition 2.2:** *Under WTA distribution of funding, specialization can be observed and hybrid strategies will become inefficient.*
>
> **Proposition 2.3:** *Under WTA distribution of funding and when scientists' time and funding have similar marginal effect on writing papers, variation in practice can be observed.*

Luck and the Best Strategy

In Figure 4.3a, we have shown the effect of uncertainty and evaluation error on the career-long payoff of scientists in for the context of funding distribution scheme based on absolute evaluation of proposals. We demonstrated that the variation in the career-long payoff increases as the quality standard for funding rises. In Figure 5, we present the result of the same type of analysis in the context of funding distribution based on relative evaluation of proposals. By introducing an error term in the evaluation process of proposals by the funding agencies, we see that the relative standard deviation of career-long payoff increases at the equilibrium as the funding distribution becomes less egalitarian. For conducting this analysis, we replaced the proposal quality terms in Equation 13 with the perceived proposal qualities from Equation 11.

Luck plays a major role in determining the long-term success of scientists in terms of their career-long payoff. The effect is the largest when funding distribution scheme is designed to reward the *best* and is especially important during the early years of career [54], [55]. In both quality-threshold and competition schemes of funding distribution, an unlucky early career in terms of proposal evaluation will hinder funding, slow down research, and further reduces the chances of receiving

funding in the next years. A better luck in the years following the early stage of the career may mitigate the effect of initial bad luck. However, if the distribution scheme is aimed at rewarding the best, i.e. the threshold or the inequality factor is set high, a better luck may not be enough to ensure a funding in the years after the early stage of the career. This is an unfortunate failure for the scientist as she has followed the optimal strategy and yet failed.

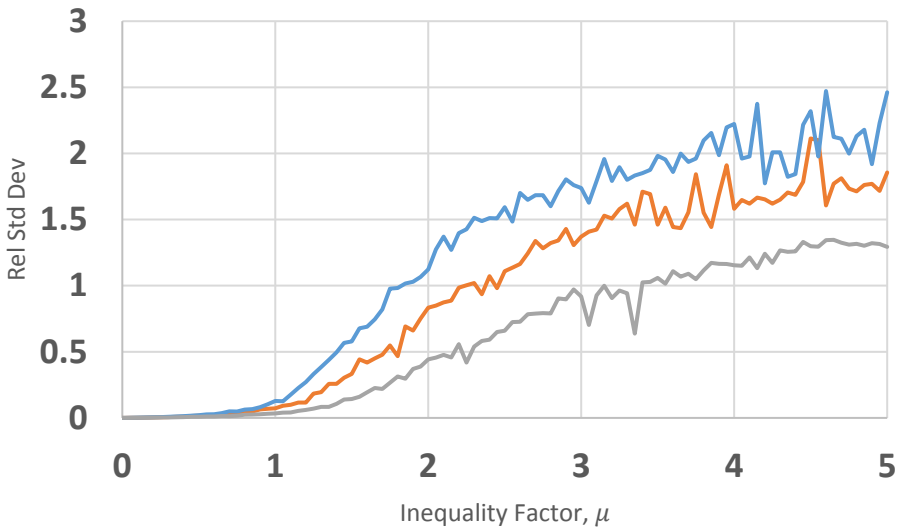
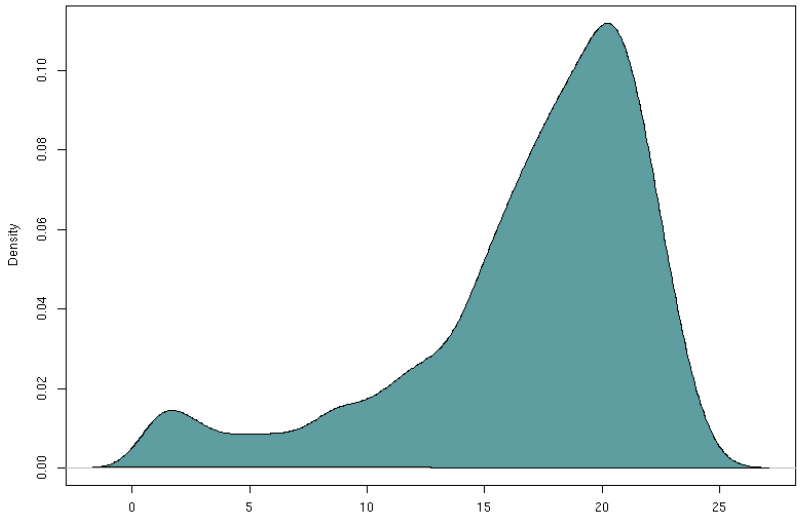
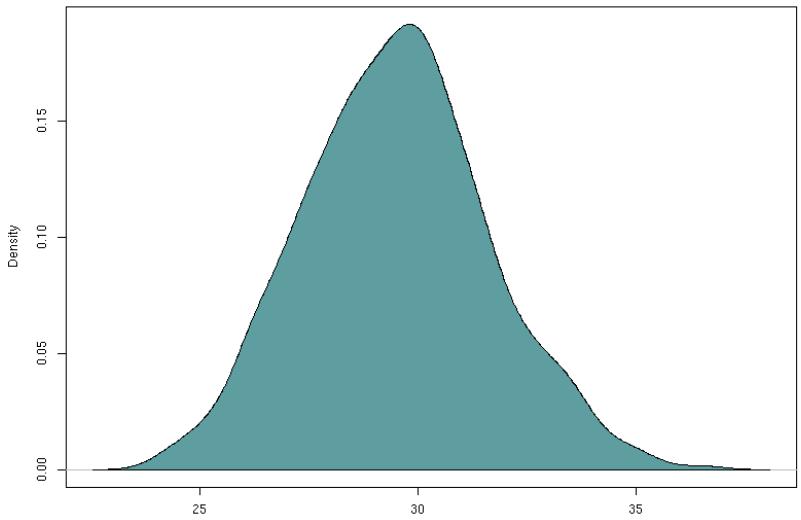


Figure 4.6: Effect of 25% proposal evaluation error on variation of expected career-long payoff. Blue: α = 0.3; Orange: α = 0.5; and Gray: α = 0.7.

Figure 4.6 demonstrates the effect of random error in proposal evaluation processes on the career-long success of scientists. Each plot shows the distribution of career-long payoff for 1,000 simulation runs where the scientist is following the optimal strategy under the existence of up to 25% error in proposal evaluation processes. The only source of discrepancy between the 1,000 runs was the randomness in the calculations of perceived proposal quality. Figure 4.7a shows the funding distribution scenario based on absolute evaluation of proposals with threshold set as 0.5 and Figure 4.7b shows the funding distribution scheme based on relative evaluation of proposals with the inequality factor set as 1. In both cases, there is about 15% chance that the scientist who follows the optimal/equilibrium strategy ends up with the career-long payoff less than one standard deviation below the expected payoff.



(a)



(b)

Figure 4.7: Distribution of career-long payoff under the optimal strategy and in existence of 25% evaluation error. (a) Quality-threshold based with threshold set at 0.5 (b) Competition based with inequality factor set at 1.

Discussion and Conclusions

Different aspects of scientists' career trajectory is studied in the literature. Our study contributes to the analysis of scientists' research career under the influence of different organizational contexts. We propose a simple system dynamics model of scientists' research career by highlighting their resource allocation decisions over the course of their career. What makes our study different from the past studies is proposing a simple theoretical framework for analyzing scientists' career trajectory under different scenarios of funding distribution.

Our analysis sheds theoretical insights into the effect of different organizational aspects of the science enterprise and the career trajectory of academic scientists. First, our model suggests that

a tendency to fund the *best* yields a condition where scientists invest more time on writing proposals rather than writing papers. Although some studies have shown that the process of writing proposals may benefit scientists in their research and in the long-term [56], the excessive amount of time spent by scientists on writing proposals remains alarming. In fact, our results corroborate with studies that have suggested a random distribution of funding among scientists or distribution over the whole population [53], [57], [58]. The novelty of our analysis is that we show while hurting the science advancement, it appears to be optimal at individual level to spend significant time on writing proposals. We show that when everyone receives funding (or have the same chance of receiving funding) the overall payoff of the scientists reaches its highest possible and they spend all their time on writing papers rather writing proposals.

Second, given the constraints of funding agencies in evaluation of grant proposals, evaluation errors are inescapable [53], [59]. However, our model shows that when the threshold for allocating funding to proposals is set high, having evaluation error does not necessarily have a negative impact on the career of scientists. Reversely, we show that as the proposal evaluation becomes more accurate, scientists have to invest more time on writing proposals.

Third, we show that when funding distribution is based on relative evaluation of proposals, a Winner Takes Some (WTS) distribution of funding is more efficient than a Winner Takes All (WTA) scheme. This result is in agreement with recent studies showing that continuous distributions of funding has advantage over dichotomous ones [60], [61]. Under a WTS distribution of funding, scientists spend more time on writing papers and their overall payoff is higher. Prior studies have argued about the positive effects of more diverse and egalitarian distribution of funding [2], [57], [62]–[64]. We show that even if we do not consider different sources of variation in the population of scientists, the more egalitarian distribution of funding leads to a higher overall payoff and higher ratio of scientists' time spent on writing papers. However, we also show that when the production of papers have a high elasticity with respect to scientist's time spent on writing papers, a WTA scheme of funding distribution is also as efficient as a completely egalitarian funding distribution. These results explain why it is efficient for some highly theoretical fields such as mathematics to have a prize-based (WTA) type of funding while it is better to support all scientists (or randomly support some of them) in very explorative and resource exhausting fields such as cosmological physics.

Fourth, we show that following the optimal strategy does not guarantee success. If a scientist's quality of proposal is perceived lower than what actually is, even if she follows the optimal strategy, she may end up with lower than expected career-long payoff. Such failure has nothing to do with what the scientist could do. The failure is rather simply due to the stochastic nature of funding distribution and the proposal evaluation error which seems to be difficult to mitigate with the current methods of evaluation.

Our results have several policy and managerial implications and highlight potential avenues for future research. First, consider a funding distribution scheme based on absolute evaluation of proposals. Assume that due to the limited resources, there is a considerable amount of evaluation error in the process of measuring quality of research proposals. Should the funding agency aim at simplifying the application process in order to make the evaluation easier and more accurate? The answer is no, according to our model. Decreasing the evaluation error, when the threshold of allocating funding is high, puts a burden on the scientists and lead them to spend more time on writing proposals and subsequently decreases their overall payoff. The best course of action under this condition is to improve the accuracy of evaluation process only if the resources allow

for a decrease in funding threshold. Our results show that doing nothing is better, in this case, than increasing the accuracy of quality evaluation while maintaining the same high threshold for allocation of funds.

Second, if efficiency in terms of scientists' overall payoff and the proportion of their time spent on writing papers is of concern, a completely egalitarian distribution of funding is the most efficient one. Not only such scheme enables scientists to invest most of their time on writing research papers, it can also encourage them to work on more risky projects which can be very beneficial to the scientific community over the long-run. Although, our model did not include the heterogeneity of research projects, it can be concluded that at least junior scientists tend to focus on conventional and less risky projects to make sure they can maintain a publication record necessary for securing future funding. This obstacle can be eliminated by employing an egalitarian distribution of funding. More detailed studies are needed to fully investigate the effect of different schemes of funding on the risk-taking behavior of scientists and the type of projects they choose to work on.

Our study has several limitations and opens up new avenues of research to address these limitations. First, in the model presented in this paper, we are assuming a rigid institutional context in which the scientist is working in. For example, and in contrast to the real world, we are assuming that the time that the scientist can spend on research does not increase as she is awarded funding. It is possible for many academic scientists, with access to external funding, to reduce their teaching load and subsequently increase their effective research time. This mechanism can be modeled by introducing new decision variables on how the scientist chooses to spend her available funding. Another way of modeling such phenomena is by analyzing the scientists and their criteria for deciding to use their funding for reducing their teaching load. Such criteria can then be added to the model as an endogenous mechanism.

We also assume no mobility in the career of the academic scientists. Career movements are shown to be common among scientists. On average, scientists move once or twice during their career [65]. Moreover, many scientists are appointed on tenure-track positions where their continued appointment depends on an interim performance assessment. These factors can be added to the analysis to better understand the career-trajectory of scientists under different organizational contexts.

Third, we assume no failure and delay in the research activities of scientists and their publication. In reality not every work gets published and not every work is evaluated equally. The payoff that a scientist receives from a more innovative revolutionary work is higher and at the same time the failure risk associated to it is higher [9], [10]. On the other hand, scientists may choose to work on less risky projects that are adaptive incremental studies with lower payoff. Our study consider that there is only one type of research work that the scientist perform. This limitation can be resolved by giving the modeled scientist the agency to choose the type of research she wants to engage in.

Finally, our model does not consider the complex dynamics of collaboration in science. Previous studies have shown that science has become more collaborative and research projects are conducted in larger teams [13], [66]. Many grant proposals are in fact include teams of scientists rather than individuals. Collaboration for junior scientists can be a resource to overcome their shortage of proposal and paper writing capacities. On the other hand, collaborations themselves are formed based on scientists' potentials and their stock of paper and proposal writing capacity.

Further studies are needed to investigate the effect of such mechanisms on the career-trajectory of scientists.

In this paper, we offered a simple system dynamics model of scientists' career trajectory that highlights the effect of different organizational factors within the science enterprise on the career of scientists. In this study, we focused on varying funding distribution schemes as a case example of an organizational context within the science enterprise. Further studies are needed to empirically test the propositions proposed in this paper and to better inform the future theoretical models. Simple models such as the one developed in this paper can be used to better inform decision makers in drafting new policies or to test the potential effect of their policies on the career of scientists.

Acknowledgements

The National Institute of General Medical Sciences and the Office of Behavioral and Social Sciences Research of the National Institutes of Health (NIH) supported this work (Grant 2U01GM094141-05).

References

[1] P. E. Stephan, *How economics shape science*. Cambridge, MA: Harvard University Press, 2012.

[2] F. C. Fang and A. Casadevall, "Reforming science: Structural reforms," *Infect. Immun.*, vol. 80, no. 3, pp. 897–901, 2012.

[3] R. J. Daniels, "A generation at risk: Young investigators and the future of the biomedical workforce," *Proc. Natl. Acad. Sci.*, vol. 112, no. 2, pp. 313–318, 2015.

[4] D. L. Herbert, A. G. Barnett, P. Clarke, and N. Graves, "On the time spent preparing grant proposals: an observational study of Australian researchers," *BMJ Open*, vol. 3, no. 5, p. e002800, 2013.

[5] R. S. Decker, P. Investigator Leslie Wimsatt, A. G. Trice, and J. A. Konstan, "A PROFILE OF FEDERAL-GRANT ADMINISTRATIVE BURDEN AMONG FEDERAL DEMONSTRATION PARTNERSHIP FACULTY A Report of the Faculty Standing Committee of the Federal Demonstration Partnership," no. January, 2007.

[6] D. Wang, C. Song, and A. L. Barabási, "Quantifying Long-Term Scientific Impact," *Science (80-.).*, vol. 342, no. 6154, pp. 127–132, 2013.

[7] R. K. Merton, "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *Am. Sociol. Rev.*, vol. 22, no. 6, p. 635, 1957.

[8] S. Mukherjee, D. M. Romero, B. Jones, and B. Uzzi, "The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot," *Sci. Adv.*, vol. 3, no. 4, p. e1601315, 2017.

[9] P. Bourdieu, "The specificity of the scientific field and the social conditions of the progress of reason," *Soc. Sci. Inf.*, vol. 14, no. 6, pp. 19–47, 1975.

[10] J. G. Foster, A. Rzhetsky, and J. A. Evans, “Tradition and Innovation in Scientists’ Research Strategies,” *Am. Sociol. Rev.*, vol. 80, no. 5, pp. 875–908, 2015.

[11] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, “Atypical combinations and scientific impact,” *Science (80-.).*, vol. 342, no. 6157, pp. 468–472, 2013.

[12] S. Kaplan and K. Vakili, “The double-edged sword of recombination in breakthrough innovation,” *Strateg. Manag. J.*, vol. 36, pp. 1435–1457, 2015.

[13] S. Wuchty, B. F. Jones, and B. Uzzi, “The Increasing Dominance of Teams in Production of Knowledge,” *Science (80-.).*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[14] P. Azoulay, J. S. G. Zivin, and J. Wang, “Superstar Extinction,” *Q. J. Econ.*, vol. 125, no. 2, pp. 549–589, 2010.

[15] Y. N. Lee, J. P. Walsh, and J. Wang, “Creativity in scientific teams: Unpacking novelty and impact,” *Res. Policy*, vol. 44, no. 3, pp. 684–697, 2015.

[16] L. Wu, D. Wang, and J. A. Evans, “Large Teams Have Developed Science and Technology; Small Teams Have Disrupted It,” 2017.

[17] R. Sinatra, D. Wang, P. Deville, C. Song, and A. L. Barabási, “Quantifying the evolution of individual scientific impact,” *Science (80-.).*, vol. 354, no. 6312, 2016.

[18] A. Mazloumian, “Predicting scholars’ scientific impact,” *PLoS One*, vol. 7, no. 11, p. e49246, 2012.

[19] J. Hönekopp and J. Khan, “Future publication success in science is better predicted by traditional measures than by the h index,” *Scientometrics*, vol. 90, no. 3, pp. 843–853, 2012.

[20] J. Hirsch, “Does the h index have predictive power?,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 49, pp. 19193–19198, 2007.

[21] E. Dorsey, B. Raphael, L. Balcer, and S. Galetta, “Predictors of future publication record and academic rank in a cohort of neurology residents,” *Neurology*, vol. 67, no. 8, pp. 1335–1337, 2006.

[22] D. Bertsimas, E. Brynjolfsson, S. Reichman, and J. Silberholz, “OR Forum—Tenure Analytics: Models for Predicting Research Impact,” *Oper. Res.*, vol. 63, no. 6, pp. 1246–1261, 2015.

[23] D. Crane, “Scientists at major and minor universities: A study of productivity and recognition,” *Am. Sociol. Rev.*, vol. 30, no. 5, pp. 699–714, 1965.

[24] S. F. Way, A. C. Morgan, A. Clauset, and D. B. Larremore, “The misleading narrative of the canonical faculty productivity trajectory,” *Proc. Natl. Acad. Sci.*, pp. E9216–E9223, 2017.

[25] D. Teodorescu, “Correlates of faculty publication productivity: A cross-national analysis,” *High. Educ.*, vol. 39, no. 2, pp. 201–222, 2000.

[26] P. Azoulay, J. Graff Zivin, and G. Manso, “Incentives and Creativity : Evidence from the Howard Hughes Medical Investigator Program,” *Rand J. Econ.*, vol. 42, no. January, pp. 527–554, 2011.

[27] R. Agarwal and A. Ohyama, “Industry or Academia, Basic or Applied? Career Choices

and Earnings Trajectories of Scientists," *Manage. Sci.*, vol. 59, no. 4, pp. 950–970, 2013.

[28] H. Sauermann and P. Stephan, "Conflicting Logics? A Multidimensional View of Industrial and Academic Science," *Organ. Sci.*, vol. 24, no. 3, pp. 889–909, 2013.

[29] M. Roach and H. Sauermann, "A taste for science ? PhD scientists ' academic orientation and self-selection into research careers in industry," *Res. Policy*, vol. 39, no. 3, pp. 422–434, 2010.

[30] N. Ghaffarzadegan, Y. Xue, and R. C. Larson, "Work-education mismatch: An endogenous theory of professionalization," *Eur. J. Oper. Res.*, vol. 261, no. 3, pp. 1085–1097, 2017.

[31] M. A. Andalib, N. Ghaffarzadegan, and R. C. Larson, "The Postdoc Queue: A Labour Force in Waiting," *Syst. Res. Behav. Sci.*, 2018.

[32] H. Hur, M. A. Andalib, J. A. Maurer, J. D. Hawley, and N. Ghaffarzadegan, "Recent trends in the U.S. Behavioral and Social Sciences Research (BSSR) workforce," *PLoS One*, vol. 12, no. 2, pp. 1–18, 2017.

[33] R. C. Larson, N. Ghaffarzadegan, and Y. Xue, "Too many PhD graduates or too few academic job openings: The basic reproductive number R0 in academia," *Syst. Res. Behav. Sci.*, vol. 31, no. 6, pp. 745–750, 2014.

[34] H. Hur, N. Ghaffarzadegan, and J. Hawley, "Effects of government spending on research workforce development: Evidence from biomedical postdoctoral researchers," *PLoS One*, vol. 10, no. 5, pp. 1–16, 2015.

[35] E. E. Gottlieb and B. Keith, "The academic research-teaching nexus in eight advanced-industrialized countries," *High. Educ.*, vol. 34, pp. 397–419, 1997.

[36] P. J. Bentley and S. Kyvik, "Individual Differences in Faculty Research Time Allocations Across 13 Countries," *Res. High. Educ.*, vol. 54, no. 3, pp. 329–348, 2013.

[37] L. D. Singell and J. H. Lillydahl, "Will Changing Times Change the Allocation of Faculty Time?," *J. Hum. Resour.*, vol. 31, no. 2, pp. 429–449, 1996.

[38] A. N. Link, C. A. Swann, and B. Bozeman, "A time allocation study of university faculty," *Econ. Educ. Rev.*, vol. 27, no. 4, pp. 363–374, 2008.

[39] N. Geard and J. Noble, "Modelling academic research funding as a resource allocation problem," 2010.

[40] M. Riediger, S. C. Li, and U. Lindenberger, "Selection, Optimization, and Compensation as Developmental Mechanisms of Adaptive Resource Allocation. Review and Preview," *Handb. Psychol. Aging*, pp. 289–313, 2006.

[41] A. M. Freund and M. Riediger, "What I have and what I do--The role of resource loss and gain throughout life," *Appl. Psychol. An Int. Rev.*, vol. 50, no. 3, pp. 370–380, 2001.

[42] D. Navon, "Resources--A theoretical soup stone?," *Psychol. Rev.*, vol. 91, no. 2, pp. 216–234, 1984.

[43] H. Rahmandad, "Impact of Growth Opportunities and Competition on Firm-Level Capability Development Trade-offs," *Organ. Sci.*, vol. 23, no. 1, pp. 138–154, 2012.

[44] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Persistence and

uncertainty in the academic career," *Proc. Natl. Acad. Sci.*, vol. 109, no. 14, pp. 5213–5218, 2012.

[45] D. Chubin, A. Porter, and M. Boeckmann, "Career patterns for scientists: A case for complementary data," *Am. Sociol. Rev.*, vol. 46, pp. 488–496, 1981.

[46] P. Clarke, D. Herbert, N. Graves, and A. G. Barnett, "A randomized trial of fellowships for early career researchers finds a high reliability in funding decisions," *J. Clin. Epidemiol.*, vol. 69, pp. 147–151, 2016.

[47] B. S. G. Levin and P. E. Stephan, "Research Productivity Over the Life Cycle : Evidence for Academic Scientists," vol. 81, no. 1, pp. 114–132, 1991.

[48] D. A. Levinthal and J. G. March, "The myopia of learning," *Strateg. Manag. J.*, vol. 14, no. 2 S, pp. 95–112, 1993.

[49] A. B. Badiru, "Computational Survey of Univariate and Multivariate Learning Curve Models," *IEEE Trans. Eng. Manag.*, vol. 39, no. 2, pp. 176–188, 1992.

[50] A. C. Miller and S. L. Serzan, "Criteria for identifying a refereed journal," *J. Higher Educ.*, vol. 6, pp. 673–697, 1984.

[51] C. E. Nelson and D. K. Pollock, *Communications among scientists and engineers*. Lexington, MA: D. C. Health - Lexington Books, 1970.

[52] A. Ebadi and A. Schiffauerova, "How to receive more funding for your research? get connected to the right people," *PLoS One*, vol. 10, no. 7, pp. 1–19, 2015.

[53] J. P. A. Ioannidis, "More time for research: Fund people not projects," *Nature*, vol. 477, no. 7366, pp. 529–531, 2011.

[54] D. Kaminski and C. Geisler, "Survival analysis of faculty retention in science and engineering by gender," *Science (80-.).*, vol. 335, pp. 864–866, 2012.

[55] A. Peterson, W.-S. Jung, J.-S. Yang, and H. E. Stanley, "Quantitative and empirical demonstration of the Matthew effect in a study of career longevity," *Proc. Natl. Acad. Sci.*, vol. 108, pp. 18–23, 2011.

[56] C. Ayoubi and F. Visentin, "The Important Thing is not to Win , it is to Take Part : What If Scientists Benefit from Participating in Competitive Grant Races ?," 2017.

[57] K. Vaesen and J. Katzav, "How much would each researcher receive if competitive government research funding were distributed equally among researchers?," *PLoS One*, vol. 12, no. 9, p. e0183967, 2017.

[58] D. Gillies, "Selecting applications for funding: why random choice is better than peer review," *RT. A J. Res. Policy Eval.*, vol. 2, no. May, pp. 64–71, 2014.

[59] D. Kaplan, N. Lacetera, and C. Kaplan, "Sample size and precision in NIH peer review," *PLoS One*, vol. 3, no. 7, pp. 3–5, 2008.

[60] R. Mutz, L. Bornmann, and H. D. Daniel, "Funding decision-making systems: An empirical comparison of continuous and dichotomous approaches based on psychometric theory," *Res. Eval.*, vol. 25, no. 4, pp. 416–426, 2016.

[61] A. De Los Reyes and M. Wang, "Applying psychometric theory and research to developing a continuously distributed approach to making research funding decisions.,"

Rev. Gen. Psychol., vol. 16, no. 3, pp. 298–304, 2012.

[62] A. G. Barnett, P. Clarke, C. Vaquette, and N. Graves, “Using democracy to award research funding: an observational study,” *Res. Integr. Peer Rev.*, vol. 2, no. 1, p. 16, 2017.

[63] J. Rigby and K. Julian, “On the horns of a dilemma: does more funding for research lead to more research or a waste of resources that calls for optimization of researcher portfolios? An analysis using funding acknowledgement data,” *Scientometrics*, vol. 101, no. 2, pp. 1067–1075, 2014.

[64] J. M. Fortin and D. J. Currie, “Big Science vs. Little Science: How Scientific Impact Scales with Funding,” *PLoS One*, vol. 8, no. 6, 2013.

[65] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A. L. Barabási, “Career on the move: Geography, stratification, and scientific impact,” *Sci. Rep.*, vol. 4, pp. 1–7, 2014.

[66] R. R. Hake, “Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses,” *Am. J. Phys.*, vol. 66, no. 1, pp. 64–74, 1998.

Appendix B

The following algorithm was used to find the Nash equilibrium numerically. Assume that there are N scientist competing in a game. Each scientist can choose a strategy, r, between 0 and 1 (including). We discretize the strategy variable, providing each scientist to choose from the following set {0, $1/k$, $2/k$, …, 1}.

Step 1: Generate a random set of strategies for all scientists. Save this set in a vector variable and call it *existingStrategies*.

Step 2: Copy the existingStrategies to a new variable and call it *modifiedStrategies*

Step 3: For each scientist:

Step 3a: Find the best strategy assuming all other scientists are selecting the strategy in *modifiedStrategies.*

Step 3b: Replace the best strategy found in Step 3a with the one in the *modifiedStrategies*

Step 4: If *modifiedStrategies* is different from *existingStrategies* then copy the value of *modifiedStrategies* and paste it into *existingStrategies* and go to Step 2. Otherwise save the *modifiedStrategies* as the Nash equilibrium.

We repeated this algorithm for each setting from different initial conditions to make sure that the equilibrium is unique.

Appendix C

The list of assumptions for the simplified version of the model are as following:

- $Y_t = ERL_t \cdot K_t$
- $T_f = 1$
- $T_p = 1$
- $\lambda = 0$
- $R_t = R$
- $G(x) = x$

Theorem. Under winner takes all condition, no pure strategy Nash equilibrium exists.

Proof. Suppose that there exists a pure strategy Nash equilibrium, $0 < R^* \leq 1$, that all N scientists have selected and every scientist has an expected payoff of $\pi^* = 1 + 30R^*$.

Now consider that *scientist 1* chooses a different strategy, $R = R^* - \epsilon_L$, where $0 < \epsilon_L < R^*$.

Three cases can be considered:

Case L1: $\epsilon_L < 2R^* - 1$

$$\pi_{L1} = 1 + R^* - \epsilon_L + 29N(R^* - \epsilon_L)$$

Note that in this case R^* should be greater than 0.5.

Case L2: $\epsilon_L > 2R^* - 1$

$$\pi_{L2} = 1 + R^* - \epsilon_L + 15N(R^* - \epsilon_L)$$

Case L3: $\epsilon_L = 2R^* - 1$

$$\pi_{L3} = 1 + 15(1 - R^*) + 15N(R^* - \epsilon_L) = 16 - 15R^* + 15NR^*$$

Note that in this case R^* should be greater than 0.5.

It can be shown that for $R^* > 0.5$, Case L1 always dominates Case L3:

$$\lim_{\epsilon_L \to 0} \pi_{L1} > \pi_{L3}$$

$$1 + R^* + 29NR^* > 16 - 15R^* + 15NR^* \Rightarrow R^* > \frac{15N + 15}{44N + 16}$$

Which is always true because $R^* > 0.5$ and $\frac{15N+15}{44N+16} < 0.5$ when $N > 1$.

It can also be shown that for $R^* > 0.5$, Case L1 always dominates Case L2:

$$\lim_{\epsilon_L \to 0} \pi_{L1} > \lim_{\epsilon_L \to 2R^* - 1} \pi_{L2}$$

$$1 + R^* + 29NR^* > 2 - R^* + 15N - 15NR^* \Rightarrow R^* > \frac{15N + 1}{44N + 2}$$

Which is always true because $R^* > 0.5$ and $\frac{15N+1}{44N+2} < 0.5$ since $N > 0$.

Therefor the potential dominant strategies for scientist 1 to deviate from the equilibria is Case L1 when $R^* > 0.5$ and Case L2 when $R^* \leq 0.5$.

Now we check whether any of these two strategies make the scientist 1 better off in comparison with the case of selecting the equilibria. First, consider, $R^* > 0.5$:

$$\lim_{\epsilon_L \to 0} \pi_{L1} > \pi^*$$

$$1 + R^* + 29NR^* > 1 + 30R^* \Rightarrow N > 1$$

Which is always true and hence, when $R^* > 0.5$, scientist 1 is better off to deviate.

Now consider, $R^* \leq 0.5$:

$$\lim_{\epsilon_L \to 0} \pi_{L2} > \pi^*$$

$$1 + R^* + 15NR^* > 1 + 30R^* \Rightarrow N > \frac{29}{15}$$

Therefore as long as $N \geq 2$, scientist 1 is better off to deviate.

Now consider that *scientist 1* chooses strategy, $R = R^* + \epsilon_H$, where $0 < \epsilon_H \leq 1 - R^*$, instead of R^*.

Same as before, three cases can be considered:

Case H1: $\epsilon_H < 1 - 2R^*$

$$\pi_{H1} = 1 + R^* + \epsilon_H + 14N(R^* + \epsilon_H)$$

Note that in this case R^* should be less than 0.5.

Case H2: $\epsilon_H > 1 - 2R^*$

$$\pi_{H2} = 1 + R^* + \epsilon_H$$

Case H3: $\epsilon_H = 1 - 2R^*$

$$\pi_{H3} = 1 + 14(R^* + \epsilon_H) = 15 - 14R^*$$

Note that in this case R^* should be less than 0.5.

It can be shown that for $R^* < 0.5$, Case H1 always dominates Case H3:

$$\lim_{\epsilon_H \to 1-2R^*} \pi_{H1} > \pi_{H3}$$

$$1 + 1 - R^* + 14N - 14NR^* > 15 - 14R^* \Rightarrow R^* < 1$$

Which is true in these cases.

It can also be shown that for $R^* < 0.5$, Case H1 always dominates Case H2:

$$\lim_{\epsilon_H \to 1-2R^*} \pi_{H1} > \lim_{\epsilon_H \to 1-R^*} \pi_{H2}$$

$$1 + 1 - R^* + 14N - 14NR^* > 2 \Rightarrow R^* < \frac{14N}{14N + 1}$$

Which is always true because $R^* < 0.5$ and $\frac{14N}{14N+1} > 0.5$.

Therefor the potential dominant strategies for scientist 1 to deviate from the equilibria is Case H1 when $R^* < 0.5$ and Case H2 when $R^* \geq 0.5$.

Now we check whether any of these two strategies make the scientist 1 better off in comparison with the case of selecting the equilibria. First, consider, $R^* < 0.5$:

$$\lim_{\epsilon_H \to 1-2R^*} \pi_{H1} > \pi^*$$

$$1 + 1 - R^* + 14N - 14NR^* > 1 + 30R^* \Rightarrow R^* < \frac{14N + 1}{14N + 31}$$

Which is always true because $R^* < 0.5$ and $\frac{14N+1}{14N+31} > 0.5$.

Now consider, $R^* \geq 0.5$:

$$\lim_{\epsilon_H \to 1-R^*} \pi_{H2} > \pi^*$$

$$2 > 1 + 30R^* \Rightarrow R^* < \frac{1}{30}$$

Which is not possible as $R^* \geq 0.5$, therefore the scientist 1 is not better off her strategy from the equilibria when $R^* \geq 0.5$.

For the cases where $R^* < 0.5$, scientist 1 has two options to increase her payoff. Either to switch to Case L2 or Case H1. We can show that under certain conditions, Case H1 dominates Case L2:

$$\lim_{\epsilon_H \to 1-2R^*} \pi_{H1} > \lim_{\epsilon_L \to 0} \pi_{L2}$$

$$1 + 1 - R^* + 14N - 14NR^* > 1 + R^* + 15NR^* \Rightarrow R^* < \frac{14N + 1}{29N + 1} < 0.5$$

Finally, we consider the case where $R^* = 0$. Under this condition, we propose that the scientist 1 choose the strategy $R_1 = 1 - \epsilon$ where $0 < \epsilon < 1$. Her payoff would be as following:

$$\pi_1 = 1 + (1 + 28N)(1 - \epsilon)$$

It can easily be shown that it will be profitable for the scientist to defect under this condition as well.

Therefore, for any R^*, scientist 1 would be better off defecting and choosing another strategy, based on the equation below.

$$R_1 = \begin{cases} R^* - \epsilon_L (\epsilon_L \to 0) & if\ 0.5 < R^* \leq 1 & yields\ \pi_1 = 1 + R^* + 29NR^* \\ R^* - \epsilon_L (\epsilon_L \to 0) & if\ \frac{14N+1}{29N+2} \leq R^* \leq 0.5 & yields\ \pi_1 = 1 + R^* + 15NR^* \\ 1 - R^* & if\ R^* \leq \frac{14N+1}{29N+2} & yields\ \pi_1 = 2 - R^* + 14N(1 - R^*) \\ 1 - \epsilon (\epsilon \to 0) & if\ R^* = 0 & yields\ \pi_1 = 2 + 28N \end{cases}$$

Hence, no Nash pure strategy exists. ■

Chapter 5: Conclusions

Science is a growing enterprise. This growth can be observed by following the trend of different measures. Figure 5.1 offers two of them. The annual rate of scientific publications has increased by more than 100% between 1980 and 2010 [1]. In the same period of time, the number of annual doctorate recipients from academic institutions in the United States has grown by more than 50%, according to the NSF Survey of Earned Doctorates.

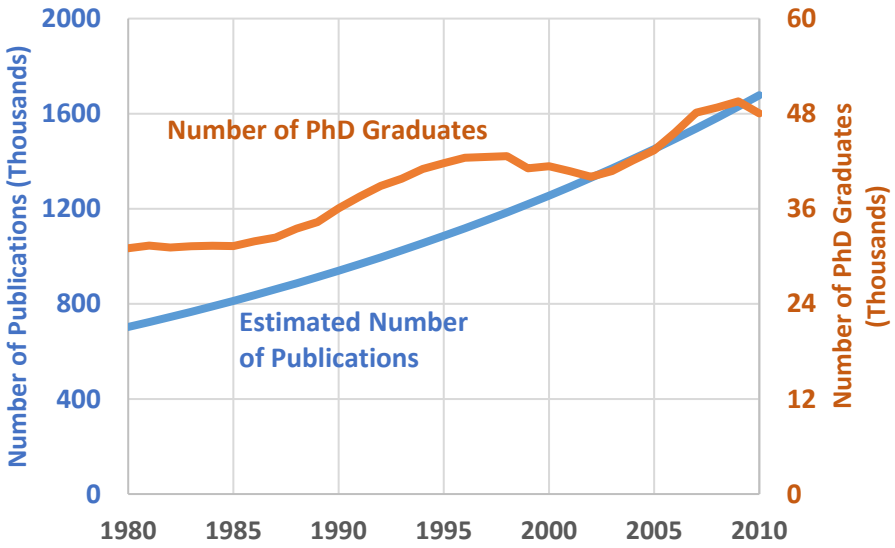


Figure 5.1: Growth of the science enterprise.

Maintaining the rapid growth of science requires a proportional response by societies. Production of scientific output is costly [2]. Yet, resources that are at nations' disposal for dedication to science is limited. This means that societies have to prioritize how they intend to invest in research and development. Here, a science policy question becomes clear: how should resources be distributed to guarantee a sustainable growth in science production that benefits the society in the long-run? To answer this question, a thorough understanding of science itself is necessary.

In many ways, science represents a complex system which involves technical, social, and economic aspects. Analysis of such system requires employing and combining different methodological perspectives and incorporation of different sources of data. To achieve this goal, the use and analysis of scientific publications as byproducts of scientific activity has become widely common in quantitative studies of science. From the classic work of Derek de Solla Price ([3]) to recent large-scale analysis of scientific activities (for a review refer to [4]), publication data has been used to better understand how the institution of science works.

This dissertation contributes to this line of research through providing insights about the effects of different contexts on the institution of science by employing data analytics and simulation models.

In essay 1, we find that the focus of the scientific community within the context of the same problem varies as the societal context of the problem changes. Our findings show that the focus of the science enterprise is not the same in different countries and that the variation is associated with how the problem is affecting that country. Specifically, we find that the focus on the behavioral and social aspects of HIV/AIDS has increased over time and varies in different countries. Further, we showed that this variation is related to the mortality level that the disease causes in each country. Our results suggest that designing a research program for any country should be done by considering the local context of the problem and not from a global perspective.

The financial context of science is investigated in this dissertation. In essay 2, we focused on the role of philanthropic money in studies of health challenges. Specifically, we analyzed two aspects of this role by looking at the content of research funded by philanthropies and the impact of their publications. We highlighted the differences by comparing philanthropic agencies with private and public ones. We find that philanthropies tend to have a more practical approach to health studies as compared with public funders. Meanwhile, we find that they are also concerned with the economic, policy related, social, and behavioral aspects of the diseases. Some scholars consider philanthropists as business-oriented organizations that tend to push technical solutions for health problems [5]. Our results suggest that this view does not provide a complete image of science philanthropism in the context of health studies. To the contrary, we show that philanthropies tend to mix and combine approaches and contents supported both by public and private sources of funding for science. We further show that in doing so, philanthropies tend to be closer to the position held by the public sector in the context of health studies. Finally, we find that studies funded by philanthropies tend to receive higher citations, and hence have higher impact, in comparison to those funded by public sector.

Different schemes of funding distribution affect the career of scientists and the overall output of science. Our model in the third essay suggests that a policy to fund the best can lead scientists to spend more time on writing proposals, in order to secure funding, rather than writing papers. We show that when everyone receives funding (or have the same chance of receiving funding) the overall payoff of the scientists reaches its highest level and they spend all their time on writing papers rather than writing proposals. Our analysis suggests that more egalitarian distributions of funding results in higher overall research output by scientists. We also find that luck play an important role in the success of scientists. We show that following the optimal strategies do not guarantee success. Due to the stochastic nature of funding decisions, some will eventually fail. The failure is not due to scientists' faulty decisions, but rather simply due to their luck, i.e. the proposal evaluation errors.

Our methods in this dissertation enabled us to ask questions within the context of science policy that are not limited to specific conditions or cases. Rather, our studies involved analyses at a high level of aggregation which enabled us to answer our research questions at their generality. In the first study, we analyzed the contributions of behavioral and social sciences to studies of a disease at a global scale. In the second study, we investigated the effect of different funding sources by analyzing publications funded by close to 300 organizations worldwide. Our third study involved analyzing the effect of different theoretical schemes of funding distribution. Data analytics and simulation techniques enabled us to ask and answer such *big* questions.

Data analytics enables us to work with large data sets and extract meaning and insights from these sources of information. In essays 1 and 2, we employed data analytics tools such as topic modeling and social network analysis while using conventional econometric methods such as

panel and negative binomial regression models. These methods enabled us to answer to fundamental questions about the science enterprise by looking at the output of scientific activity. We show that data analytics can help gauging the effect of different factors on scientific publications. Our work fits within the ongoing effort in the field of science of science to use data analytics and analyze large collections of scientific output [4]. In doing so, we relied on using data from different sources such as Scopus, Web of Science, and the World Bank. This dissertation shows the potentials that can be materialized through availability of more data both by scientists and funding agencies.

This dissertation further contributed to the field of science and innovation policy by demonstrating the benefits of building small dynamic models. Small system dynamics model are shown to be effective for public policy [6]. In essay 3, we take this idea and develop a small dynamic model for the career-long success of a scientist. By using this model and finding the optimal strategies and equilibrium points (by using a game theory perspective), we were able to generate a number of important and sometimes counter-intuitive propositions. This work shows how using modeling and simulation methods can help uncover the underlying mechanisms of different behaviors in a complex system, i.e. science in the context of this dissertation. Further studies can benefit from using such approaches in the context of science and innovation.

The studies in this dissertation provide specific implications for science policy makers which are discussed in the previous chapters. The overall dissertation also brings about broader insights important for the society. First, this dissertation emphasizes the social construct characteristics of science and suggests that the society pay attention to the context in which scientists are working. In the first study, we focus on the context of the problem and in the second and third study we concentrate on the funding context of science. The main implication of this work is that the output, efficiency, and contributions of the institution of science should not be understood in isolation from the socially enforced contexts projected on the scientific community.

Second, funding is the fuel for innovation and scientific discovery. There are different ways that a society can support its scientists and to incentivize scientific inquiry. Diversity can be beneficial. In essay 2, we show that different funding bodies focus on a diverse set of priorities when tackling the same problem. This diversity of approaches can benefit the society and generate novel ideas more often and faster than a hypothetical situation where funding is distributed only by one supporter of science. Moreover, in essay 3, we show that distributions of funding that are egalitarian are more efficient for the science and help scientists to have a more successful career in the long-run.

In conclusion, this dissertation contributed to the field of science policy. The studies in the previous chapters are related to science of science, an emerging field of inquiry aiming at using scientific methods to investigate science as a phenomena. This work contributed to this body of literature both in terms of content and application of methods. Moreover, further important research questions are uncovered in this dissertation for the future studies which were proposed in the previous chapters.

References

[1] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” *J. Assoc. Inf. Sci. Technol.*,

vol. 66, no. 11, pp. 2215–2222, 2015.

[2] P. E. Stephan, *How economics shape science*. Cambridge, MA: Harvard University Press, 2012.

[3] D. de Solla Price, *Little Science, Big Science...and Beyond*. Columbia University Press, 1986.

[4] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A.-L. Barabási, "Science of science," *Science (80-.).*, vol. 359, no. 6379, p. eaao0185, Mar. 2018.

[5] J. Youde, "The Rockefeller and Gates Foundations in Global Health Governance," *Glob. Soc.*, vol. 27, no. 2, pp. 139–158, 2013.

[6] N. Ghaffarzadegan, J. Lyneis, and G. P. Richardson, "How small system dynamics models can help the public policy process," *Syst. Dyn. Rev.*, vol. 27, no. 1, pp. 22–44, 2011.