



Three Essays on Economics of Innovation, Technology, and Policy

Citation

Zhuo, Ran. 2023. Three Essays on Economics of Innovation, Technology, and Policy. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37375540>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available. Please share how this access benefits you. [Submit a story](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



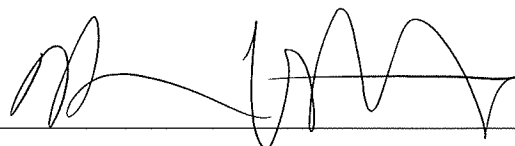
DISSERTATION ACCEPTANCE CERTIFICATE

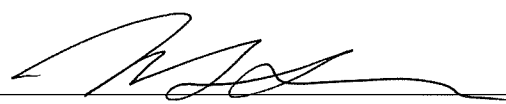
The undersigned, appointed by the Committee for the
PhD in Business Economics have examined a dissertation
entitled

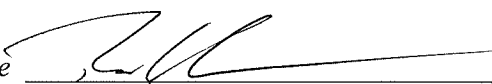
Three Essays on Economics of Innovation, Technology, and Policy

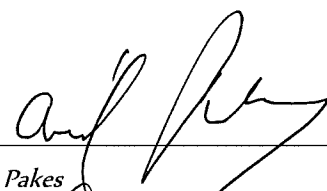
Presented by **Ran Zhuo**

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 
Shane Greenstein, Chair

Signature 
Myrto Kalouptsidi

Signature 
Robin Lee

Signature 
Ariel Pakes

Signature 
Elie Tamer

Date: April 4, 2023

Three Essays on Economics of Innovation, Technology, and Policy

A dissertation presented

by

Ran Zhuo

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Business Economics

Harvard University

Cambridge, Massachusetts

April 2023

© 2023 Ran Zhuo
All rights reserved.

Three Essays on Economics of Innovation, Technology, and Policy

Abstract

My dissertation examines decision- and policy-making in the innovation and technology ecosystem and during the innovation process. I conduct three studies using novel, highly granular data and rigorous empirical methods.

Innovation is inherently risky, and a key question in managing the innovation process is how innovators should allocate resources when information is scarce and environments constantly change. In chapter one, I analyze how large, NIH-supported scientific labs traded off short-term productivity gains (via exploiting safe projects) against information acquisition that could improve long-term productivity (via exploring high-variance projects). I use detailed data on daily resource allocation to over 300,000 research projects from 2000–2015 and estimate a dynamic structural decision model. Results show the labs strongly valued exploration, which had a large and significant positive impact on their productivity. These findings suggest real-world decision-making sophistication and implications for how funding agencies should evaluate innovation productivity.

Innovation generates new business models and regulatory challenges. Chapter two explores whether the EU's General Data Protection Regulation (GDPR), a landmark internet privacy law, affected investments in internet infrastructure. I use state-of-the-art internet measurement data and a difference-in-differences approach to examine if the regulation's data-related restrictions slowed investments in the interconnectedness of networks in the European Economic Area (EEA) compared to non-EEA OECD networks. Results indicate zero effects on various measures, suggesting the GDPR's short-run consequences were limited to application firms, its intended targets, rather than networks, its unintended targets.

Innovation introduces new products with unclear economic properties and user behavior. Chapter three investigates user decisions around one such product: open-source software. Using data with five million

months of open-source web server usage from over 150,000 US organizations between 2000-2018, I develop hazard models to study determinants of organizations' open-source software upgrade decisions. Results show a significant propensity to upgrade through users' routine administrative processes, with counter-intuitive implications for how innovators (developers) should manage releases of new products (upgrades).

My future works will continue to study a broad range of managerial and policy issues at the intersection of economics, technology, and innovation management.

Contents

Title Page	i
Copyright	ii
Abstract	iii
Table of Contents	v
List of Tables	viii
List of Figures	x
Acknowledgments	xii
Dedication	xiv
Introduction	1
1 Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs	6
1.1 Introduction	6
1.2 Empirical Setting	10
1.3 Data	14
1.3.1 Input Allocation	16
1.3.2 Observed output	19
1.3.3 Posterior Beliefs about Output	20
1.4 Model	25
1.4.1 Objective Function	25
1.4.2 Intractability of Backward Induction	27
1.4.3 Index Approximation	28
1.4.4 Models of Index Approximation	31
1.5 Model Fitting	34
1.5.1 Estimation	34
1.5.2 Simulation	36
1.6 Model Fitting Results	38
1.6.1 Estimation Results	38
1.6.2 Simulation Results	42
1.7 Counterfactual—The Impact of Exploration on Productivity	47
1.8 Conclusion	49

2	The Impact of the General Data Protection Regulation on Internet Interconnection	51
2.1	Introduction	51
2.2	Internet Interconnection	56
2.3	The GDPR	60
2.4	Data	65
2.5	Empirical Strategy	71
2.6	Results	76
2.6.1	The Number of Agreements between Countries	76
2.6.2	The Number of Agreements between Countries by Agreement Type	79
2.6.3	The Number of IP-Address-Level Interconnection Points per Agreement	82
2.6.4	The Number of Agreements by Networks	86
2.6.5	The Number of Networks	88
2.6.6	Customer Cone of Networks	88
2.6.7	Robustness Checks	89
2.7.	Conclusion	91
3	Upgraded Software and Embedded Improvements: Tracking Vulnerabilities and Bugs on the Web	94
3.1	Introduction	94
3.2	Contributions	97
3.3	Setting	100
3.4	Data	104
3.5	Empirical Strategy	112
3.6	Results	115
3.6.1	Usage of Server Software with Security Vulnerabilities	115
3.6.2	Do Organizations Update Their Software in Response to Vulnerabilities?	120
3.6.3	Do Costs or Benefits Influence Updating Decisions?	128
3.6.4	Do Discoveries of Severe Security Vulnerabilities Instigate Developments of New Versions of Software?	130
3.7	Conclusion	132
	References	134
	Appendix A Appendix to Chapter 1	146
A.1	Data and Variable Construction	146
A.1.1	Project Rationale	146
A.1.2	Lab Funding	147
A.1.3	Matching Projects to UniProt Molecule Information	149
A.1.4	Data Glossary	150
A.2	Constructing Posteriors	160
A.2.1	Random Forest for Trial Success Probabilities	160

A.2.2	Ridge Regression for Citations	165
A.2.3	Ridge Regression for Downloads	168
A.3	Additional Details on Modeling and Estimation Procedure	172
A.3.1	Functional Form of Payoff Function $\pi_{ijt}(\mathbf{a}_{it}, p_{ijt}; \boldsymbol{\theta}_{Xl})$	172
A.3.2	Specifying Choice Set C_{it}	174
A.4	Additional Results	176
Appendix B	Appendix to Chapter 2	186
B.1	Bilateral Bargaining Model	186
B.1.1	Peer-to-Peer Agreements (p2p)	188
B.1.2	Provider-to-Customer Agreements (p2c)	189
B.1.3	Substitution between Agreement Types	190
B.2	Data Appendix	191
B.3	Robustness Checks	195
B.3.1	Results on Levels	195
B.3.2	First Differences	199
Appendix C	Appendix to Chapter 3	203

List of Tables

1.1	Key variables and data sources	15
1.2	Comparison of likelihoods of different models	39
1.3	Estimates of main parameters of interest in UCB+D model	40
1.4	Comparison of simulated outcomes of different models	41
1.5	Counterfactual outcomes, no exploration	48
2.1	Description of variables	68
2.2	Summary statistics	69
2.3	The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty	80
2.4	The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty and agreement type	83
2.5	The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries, by counterparty	87
3.1	Summary of sources of data	106
3.2	Summary statistics of the broad sample of organizations	110
3.3	Summary statistics of the sample of public firms	111
3.4	Cox proportional hazards analysis of updating decisions (broad sample)	122
3.5	Cox proportional hazards analysis of updating decisions (public firms)	127
3.6	Cox proportional hazards analysis of releasing of new minor versions	131
A.1.1	Funding opportunity announcements (FOA) tied to PSI	148
A.1.2	Data glossary	151
A.3.1	Trials included in the reduced choice set for estimation	175
A.4.1	Comparison of likelihoods of different models, JCSG	177
A.4.2	Comparison of likelihoods of different models, MCSG	178
A.4.3	Comparison of likelihoods of different models, NYSGRC	179
A.4.4	Out-of-sample fit of UCB+D model	180
A.4.5	Estimates of parameters in different models, NESG	181
A.4.6	Estimates of main parameters of interest in UCB+D model, other labs	182
A.4.7	Simulated outcomes of UCB+D model, other labs	183
A.4.8	Out-of-sample simulation results for UCB+D model	184

A.4.9	Counterfactual outcomes, no exploration, other labs	185
B.2.1	List of Countries and Territories	194
B.3.1	The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty	195
B.3.2	The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty and agreement type	196
B.3.3	The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries, by counterparty	197
B.3.4	The GDPR's impact on additional outcomes	198
B.3.5	The GDPR's impact on the number of agreements by EEA countries, by counterparty . . .	199
B.3.6	The GDPR's impact on the number of agreements by EEA countries, by counterparty and agreement type	200
B.3.7	The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA countries, by counterparty	201
B.3.8	The GDPR's impact on additional outcomes	202
C.1	Variable definitions	203

List of Figures

1.1	Three-dimensional structure of a protein molecule	11
1.2	Lab makes decisions about allocating resources across projects in each period	12
1.3	Collection and analysis of information guide allocation decisions	13
1.4	Distribution of input allocation across projects	17
1.5	Funding of the four largest labs	19
1.6	Observed output of the four largest labs	21
1.7	Estimated posterior beliefs about output	24
1.8	Simulated distributions of input allocation across projects	43
1.9	Simulated number of unique publications under different models	44
1.10	Characteristics of trials allocated in simulations, static model versus UCB+D model	46
1.11	Simulated number of unique publications, no exploration	48
2.1	Four layers of the Internet	57
2.2	Data collection at the internet layer	66
2.3	Three subsamples for the analysis on the country pair level	77
2.4	Number of interconnection agreements by EEA and non-EEA OECD countries by counter-party	78
2.5	Number of agreements by EEA and non-EEA OECD countries by agreement type	81
2.6	Average log number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries by counterparty	85
2.7	Average log number of interconnection agreements by networks in EEA and non-EEA OECD countries	86
2.8	Average log number of networks in EEA and non-EEA OECD countries	89
2.9	Average log customer cone of networks in EEA and non-EEA OECD countries	90
3.1	Fraction of organizations using Apache versions with severe security vulnerabilities	116
3.2	Fraction of organizations using Apache versions with severe security vulnerabilities that were publicly known	117
3.3	Fraction of organizations using Apache versions with severe security vulnerabilities that were already fixed in new releases	118
3.4	Heterogeneity in the use of Apache versions with severe security vulnerabilities that were already fixed in new releases	119

3.5	Fraction of organizations using Apache versions with severe security vulnerabilities that were already fixed in new releases (breakdown by the number of vulnerabilities)	121
3.6	Continued usage of software versions with and without severe security vulnerabilities . . .	126
A.2.1	Distributions of actual citations and predicted citations	167
A.2.2	Scatterplot of predicted citations against actual citations	167
A.2.3	Average downloads per structure in months since publication	170
A.2.4	Distributions of predicted $\hat{E}(\Delta download_{iy})$ against actual $\Delta download_{iy}$	171
A.2.5	Scatterplot of predicted $\hat{E}(\Delta download_{iy})$ against actual $\Delta download_{iy}$	171
A.2.6	Binned scatterplot of predicted $\hat{E}(\Delta download_{iy})$ against actual $\Delta download_{iy}$	172
A.4.1	Observed output: number of publications and citations	176

Acknowledgments

I am grateful to the following individuals for their contributions to the completion of this dissertation:

First and foremost, I would like to express my deep appreciation to my dissertation committee chair, Professor Shane Greenstein, for his infinite curiosity and enthusiasm for great research, for his unwavering support, encouragement, and patience with my projects, and for being an exemplar of scholarship, mentorship, and collaboration.

I am also thankful to the members of my dissertation committee: Professor Myrto Kalouptsi, Professor Robin Lee, Professor Ariel Pakes, and Professor Elie Tamer, for their sharp intuition and penetrating insights, and for holding my research to the highest standards and molding me into a better researcher.

I owe a debt of gratitude to my co-authors for various projects: kc claffy, Brad Huffaker, Raviv Murciano-Goroff, Chris Snyder, and Audrey Tiew, for being so much fun to work together with, and for being a source of inspiration.

I would also like to extend my thanks to Dan Andersen, Isaiah Andrews, Erik Bohlin, Tim Bresnahan, Shengmao Cao, Dennis Carlton, Varanya Chaubey, Wes Cohen, Daniel Davis, Pierre Dubois, John Everett, Roderick Fanou, Chaim Fershtman, Samuel Goldberg, Avi Goldfarb, Yannai Gonczarowski, Tianxiao Han, John Hill, Ryan Hill, Charles Hodgson, Ginger Zhe Jin, Garrett Johnson, Louis Kaplow, Max Kasy, Jacqueline Ng Lane, Lucas De Lima, Zhi Lin, Alexander MacKay, Kyle Myers, Jeffrey Nickerson, John Norvell, Frank Pinter, Peter Preusch, Devesh Raval, Tim Simcoe, Ken Simons, Carolyn Stein, Paula Stephan, Jeff Strabone, Stephen Strowes, Senmiao Sun, Wei Yang Tham, Janelle Winston, Nataliya Langburd Wright, Hanbin Yang, Ron Yang, David Zhang, participants of EARIE 2022, IIOC 2023, Innovation & Entrepreneurship Seminar at Max Planck Institute, Harvard Industrial Organization workshop, MIT Junior Researcher Series Talk, NBER Productivity Seminar, TPRC48, WICK#8 Doctoral Workshop, Workshop on Internet Economics at CAIDA, for helpful suggestions on chapters of my dissertation.

I am grateful to Finley Foster, Bob Freeman, Will Lombard, Jen Mucciarone, and Marais Young at Harvard Business School, Jamina Coleman, Jeff Miron, Brenda Piquet, and Ann Richards at the Economics Department of Harvard University, for their assistance on various matters related to research and graduate student life.

To my amazing friends Kun Chen, Shushu Liang, Tianwang Liu, Ayushi Narayan, Alice Wu, and Zhecheng Yao, who haven't already appeared above, thank you for being there for me, always.

I would also like to acknowledge the unwavering support of my parents, for always believing in me and for bearing with me for not being able to visit them during the pandemic.

Lastly, I would like to thank the many people who I have met at Harvard, conferences, virtual venues, and campus visits, for discussing research and sharing words of encouragement and wisdom with me.

To my family and friends

Introduction

During my graduate school years, my diverse interests in economics, technology, digitization, and innovation management have driven me to explore various topics in the innovation ecosystem. To reflect this breadth of interest, I have selected the following three chapters for my dissertation. These chapters delve into different aspects of decision-making and policy formation in the innovation and technology ecosystem.

The first chapter examines resource allocation in publicly-funded scientific projects, while the second chapter analyzes the regulation of digital markets and investments in internet infrastructure. In the third chapter, I focus on user behavior concerning an innovative product and the corresponding managerial implications. While each chapter covers a distinct area, they all contribute to the overall theme of examining vital issues in the innovation process.

To conduct these studies, I have used novel, highly granular data and various rigorous empirical methods. These methods include reduced-form analysis, machine learning, and structural estimation of empirical decision models.

Chapter 1

In the first chapter of my dissertation, titled “Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs,” I explore the critical issue of resource allocation in the innovation process. Allocating resources to multiple concurrent projects is challenging for innovative organizations, particularly when they lack perfect information about which projects will be productive. This problem becomes even more critical as resources are often limited. In this chapter, I focus on understanding how innovative organizations make decisions about allocating resources to different projects under incomplete information.

To investigate this issue, I examine a group of large, NIH-funded scientific labs and their allocation of resources among over 300,000 research projects between 2000-2015. I use daily data on the labs’ resource

allocation decisions and outcomes to estimate a dynamic structural model of their decision-making. My findings show that the labs strongly valued the dynamic incentive to explore. Their decision-making closely resembled that predicted by a modified version of the Upper Confidence Bound (UCB) algorithm, a widely used method in the reinforcement learning and multi-armed bandit literature for making dynamic allocation decisions under incomplete information. Furthermore, my research shows that exploration had a significant impact on the labs' productivity. If they had only focused on exploiting safe projects, their output quantity would have decreased by 51%, and their citations would have decreased by 57%.

These findings are interesting and important for several reasons. First, extensive theoretical literature on multi-armed bandits has recognized the importance of exploration in resource allocation for innovation and has theorized about optimal exploitation-exploration tradeoff. However, little research has studied empirically how innovators make this tradeoff in messy real-world settings. This paper overcomes the data and computational challenges and shows sophistication in real-world decision-making in innovative organizations.

Furthermore, the finding that exploration significantly improves innovation productivity has important policy implications for the innovation industry, which has a large spending of over \$500 billion per year in the US. If these results extend to the whole industry in any degree, exploration of ideas makes a substantial contribution to innovation productivity through the channel of improving the resource allocation process, and has generated hundreds of billions of dollars of economic value through this channel. Therefore, funding agencies evaluating innovation productivity should encourage exploration and information sharing about project productivity among innovators.

Finally, this paper also contributes methodologically to the empirical literature on single-agent dynamic choices. Innovators typically allocate resources among lots of projects over horizons that last years. Workhorse estimation methods for dynamic choice models do not work well with large choice sets and long horizons due to the curse of dimensionality. This paper develops the index approximation approach to model dynamic allocations, which overcomes the computational challenges in estimation and allows estimation for problems that consist of thousands of periods and millions of choices per period.

Chapter 2

The second chapter, titled “The Impact of the General Data Protection Regulation on Internet Interconnection,” examines the effects of regulation of digital markets on internet infrastructure investment. Innovation, particularly in the digital ecosystems, presents new business models and regulatory challenges. Regulations aimed at protecting consumers and disciplining one set of players in the ecosystem may have unintended consequences for others. Therefore, it is crucial to study the impact of new regulations on different parts of the value chain to enable a well-rounded welfare analysis.

The study investigates whether the EU’s landmark internet privacy law, the GDPR, had unintended consequences on investments in internet infrastructure. The Internet consists of thousands of independently operated networks, interconnected using bilaterally negotiated data exchange agreements. The GDPR imposes strict restrictions on how application firms should handle personal data of EEA residents. Empirical evidence shows that the GDPR brings significant costs to application firms and reduces data usage in the EEA relative to other markets.

With coauthors Bradley Huffaker, kc claffy, and Shane Greenstein, I use a difference-in-differences approach to examine whether this decline in derived demand for data exchange negatively impacted investment in interconnection by networks in the EEA relative to networks in non-EEA OECD countries following the announcement and implementation of the GDPR. We employ a large sample of interconnection agreements between networks globally from 2015 to 2019, collected through state-of-the-art internet measurement tools. Our findings show precisely zero effects of the GDPR on numerous measures of interconnectedness, including the number of observed interconnection agreements, the inferred agreement types, and the number of observed IP-address-level interconnection points per agreement. We also find economically small effects of the GDPR on the entry and the observed number of customers of networks.

These findings are noteworthy for two reasons. First, the GDPR is the most significant privacy policy since the commercialization of the Internet in the 1990s, and it was widely speculated that it would have a profound impact on the operation of the Internet and the digital economy. Our results allow us to conclude that the GDPR’s short-term effects were concentrated on its intended targets, the application firms, and did not spill over to unintended targets, the networks. This is an important policy implication for regulators in many parts of the world considering the implementation of similar internet privacy laws.

In addition, our paper contributes to the literature by presenting data of unprecedented scale and scope.

While a theoretical literature has explored questions regarding network operators and interconnection agreements, empirical research on this topic has been limited. Across the academic and policy arenas, there has been a significant lack of well-measured data describing the interconnectivity and traffic flow in the Internet, which has attracted considerable attention in issues such as net neutrality, international trade in digitally delivered goods, market power of big technology firms, and privacy regulations. Our data represent a valuable step towards addressing this data gap.

Chapter 3

In the third chapter, titled “Upgraded Software and Embedded Improvements: Tracking Vulnerabilities and Bugs on the Web,” my coauthors Raviv Murciano-Goroff, Shane Greenstein, and I investigate user behavior regarding the adoption, use, and upgrading of open-source software in the context of web servers. Open-source software, a product of innovation, poses unique economic properties and user behavior that are not yet well understood. Our earlier work has revealed significant heterogeneity in the propensity to adopt, use, and upgrade open-source web servers.

Using data containing five million months of open-source web server usage at over 150,000 US organizations from 2000 to 2018, we explore the determinants of organizations’ decisions to upgrade their open-source software. Specifically, we examine quasi-natural experiments that occurred after the appearance of severe security vulnerabilities and find significant variance in the firm-specific attentiveness to upgrade software in response to a vulnerability. To address the reverse causality issue, we develop hazard model approaches that account for firm-specific proclivities to upgrade and fit a subset of data with a range of explanatory variables.

We find a significant propensity to upgrade because a new version has appeared, suggesting that many users gain protection against vulnerabilities as a byproduct of routine administrative processes that support maintaining frontier software. Users accelerate upgrading when their web server supports electronic commerce instead of merely an informational or coordinating role. However, we also find deceleration when users support an incredibly complex IT operation, which is consistent with models of high costs of disrupting operations.

This paper makes a significant contribution to the software security literature by presenting empirical evidence of the distribution of vulnerabilities and the factors affecting this distribution in widely used open-source software. Our results have important implications for prioritizing the disclosure of security

vulnerabilities and software updates. Contrary to common assumptions, our analysis suggests that software developers may want to publicize available updates while providing less information about the vulnerabilities. This finding has important practical implications for how software developers manage the release of new products and communicate about flaws in existing products.

Moving Forward

In my future research, I will continue to study a broad range of managerial and policy issues at the intersection of economics, technology, and innovation management. Specifically, I am interested in the allocation of innovation resources and investment in new technologies in organizations. As follow-on research to the first chapter of my dissertation, I am currently working on improving resource allocation in innovative organizations. I am exploring various policy instruments, from changing the scale and duration of the resource allocation process to the use of information systems and data analytics.

Furthermore, my future work related to open-source software will examine the incentives to invest in open-source software and the policy implications from the management and security perspectives. Additionally, I am conducting early-stage research on investments and strategic decision-making in high-tech manufacturing industries that experience frequent and expensive technology turnovers.

Chapter 1

Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs

1.1 Introduction

Innovators often need to make a tradeoff between exploitation and exploration in allocating innovation resources. How they do it has a large impact on their productivity. Exploitation prioritizes safe projects that innovators deem the most productive in the present. Exploration prioritizes high-variance projects that would yield more information to improve allocation and productivity in the future. Innovators need to make this tradeoff because they have incomplete information about which projects will be productive. The purpose of exploration is to acquire information about productivity—by allocating input to the projects that innovators have poor information about and observing those projects' output. Extensive theoretical literature, starting with seminal works of Gittins (1979) and Weitzman (1979), has recognized the importance of exploration in resource allocation for innovation and has theorized about optimal exploitation-exploration tradeoff.

Little research has studied empirically how innovators make this tradeoff. One reason is a lack of data with sufficient granularity, which needs to contain resources allocated to, and output produced by, every research project in an innovator's choice set. Another reason is the computational challenge. Innovators typically allocate resources among lots of projects over horizons that last years. Workhorse estimation methods for dynamic choice models, developed by an empirical literature starting with Pakes (1986) and Rust (1987), do not work well with large choice sets and long horizons due to the curse of dimensionality.

Employing novel data and proposing a new estimation approach, this paper studies empirically how a group of large scientific labs made the exploitation-exploration tradeoff. I find that a simple model captures the labs’ decision-making and that the estimates of the model’s free parameters suggest the labs explored extensively. I also find that, under counterfactual simulations, had the labs not explored, their output quantity and citations would have decreased substantially.

These answers are nonobvious and important. They confirm intuition from theoretical literature about the value of exploration for resource allocation under incomplete information. They also suggest sophistication in real-world decision-making—the labs took into account the dynamic incentives to explore and acquire information in their decision-making, as they should. Moreover, they have policy implications for the large innovation industry in which US spending per year exceeds \$500 billion (Boroush, 2021). If these results extend to the whole industry in any degree, exploration of ideas makes a substantial contribution to innovation productivity through the channel of improving the resource allocation process, and has generated hundreds of billions of dollars of economic value through this channel.

The empirical setting of this paper is a group of large structural biology labs funded by an NIH program of \$1.3 billion over a horizon of sixteen years. This is an ideal setting to study because it is realistically complex and affords highly granular data. Moreover, the labs thoroughly documented how they used an AI-assisted approach to analyze information observed during past allocations and to inform future decision-making, providing guidance for modeling. A few additional NIH policy features, such as restricting the pool of projects the labs could choose from, also make this setting particularly clean for model identification.

The main data is the labs’ resource allocation and output across projects. I observe, at the daily frequency, the input allocated to, and output produced by, each research project that each lab ever attempted over the sixteen-year horizon. A project in this setting is the determination of the three-dimensional structure of a protein molecule, and the molecule uniquely identifies the project. A unit of input is a distinct experimental trial on the project. The output is whether the trial produced a publication of the structure and the number of citations and downloads of that publication. Those are *observed* output from input actually allocated. The labs instead made allocation decisions based on their *beliefs* about output from potential allocations. To construct the labs’ posterior beliefs about output, a key model ingredient, I made a best-effort replication of the labs’ AI-assisted approach of belief formation, which was based on analyzing information observed during past allocations.

To recover how the labs made the exploitation-exploration tradeoff, I estimate a dynamic model of the

labs' decision-making, assuming the labs used a simple index to approximate the complex underlying value function. This kind of index is well-studied in theory (Lai & Robbins, 1985; Agrawal, 1995; Katehakis & Robbins, 1995; Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012) and well-used in practice (Yang et al., 2020; Scale, 2021; Zhang, 2021) but has not been applied to estimation of empirical models of decision-making. Under my model, in each period a lab first analyzes past information to form posterior beliefs about output from potential allocations. The lab then uses this posterior to compute a simple index to approximate the value function associated with allocating input to each project. The lab then allocates input to projects with the highest index values. The index is simple to compute because computation only requires information available to the labs in the current period. Workhorse dynamic models based on backward induction, on the other hand, tie the value function to the evolution of information in *future* periods and, consequently, suffer from the curse of dimensionality as the number of periods grows. Due to its computational simplicity, index approximation is also a more empirically plausible model of the labs' decision-making than backward induction.

My model modifies a well-used index called Upper Confidence Bound (UCB). Intuitively, UCB approximates the upper confidence bound of an allocation's posterior expected payoff. Allocating input to a project that the lab has poor information about has a high UCB. As the lab allocates more input to the project and has better information about its productivity, the UCB of further allocations decreases. To proxy for the unobserved factors during the allocation process, such as project-specific learning staying with individual researchers over time, I make a modification to the baseline UCB index. The modification is an additional term that captures time-discounting of the value of older projects that the lab has attempted in the distant past. As robustness checks, I specify many alternative models with other types of indices discussed in the theoretical literature. These include the seminal Gittins (1979) index, which prescribes optimal decisions for some stylistic dynamic allocation problems.

I validate the model in two steps. In the first step, I estimate the free parameters in the model by maximizing the log likelihood of the observed allocation decisions. Identification of model parameters is based on revealed preferences and is very intuitive. I gauge model fit by computing the predicted likelihood of making the observed allocation decisions at convergence of the model. In the second step, I use the estimated parameters from the model to forward simulate the labs' entire history of input allocation and output. I compare the patterns of input allocated and output from the simulated data to those from the actual data to determine whether the model could generate patterns similar to those in the actual data. I repeat this two-step

procedure for the alternative models to benchmark the fit of my main model against these alternatives.

I find my main model fits the data extremely well and captures the labs' decision-making. During the maximum likelihood step, the main model by far has the smallest magnitude of log likelihood at convergence among the many alternative models I tested. With the same number of parameters, its log likelihood at convergence is only 59% to 72% of the second best fit model across labs. On average, the main model predicts a 70% to 84% likelihood of allocating trials that were actually allocated, and a smaller than 0.5% likelihood of allocating trials that were actually *not* allocated, far better than any alternative model. During the simulation step, the main model generates input allocation patterns and output that are very similar to those in the data. For all labs, the deviations of the simulated output from the actual output are within 10%. The alternative models fail to generate patterns matching the data as closely.

Based on this well-fitting model, I find that the labs explored extensively and that exploration had a large impact on the labs' productivity. Based on the estimates of the free parameters of the model, I am able to reject the hypothesis that the labs did not explore at a 95% confidence level for all labs. Moreover, I am able to reject the hypothesis that there were no unobserved factors that impacted the labs' incentives to explore a project over time at a 95% confidence level. Counterfactual simulations show that exploration boosted the labs' productivity substantially. Had the labs not explored, they would have missed many low-hanging fruit and misallocated resources towards less productive projects. As a result, their output quantity would have decreased by 51%, and their citations would have decreased by 57%. The decrease is equivalent to forgoing at least \$650 to \$720 million of economic value.

This paper contributes to the empirical literature on science of science. This literature has studied extensively the nature of idea development and the value of experimentation (Cohen & Levinthal, 1989; Henderson & Cockburn, 1996; Azoulay et al., 2011; Ederer & Manso, 2013; Ganglmair et al., 2019; Krieger, 2021; Hoelzemann et al., 2022; Khmel'nitskaya, 2022; Lane et al., 2022), but few works have examined exploration of ideas as a micro-determinant of productivity at this level of detail. This paper is a rigorous empirical counterpart to the theoretical literature on resource allocation for innovation (Arrow, 1962; Roberts & Weitzman, 1981; Bergemann & Välimäki, 1996), and its findings add to the discussion of the impact of policies on innovation productivity (Jaffe, 2002; Furman & Stern, 2011; Azoulay, 2012; Cantoni & Yuchtman, 2014; Lane et al., 2015; Azoulay et al., 2019; Myers, 2020).

This paper also contributes to the empirical literature on single-agent dynamic choices. This literature has relied on recursive and simulation methods for estimation (Pakes, 1986; Rust, 1987; Hotz & Miller,

1993; Rust, 1994; Timmins, 2002; Aguirregabiria & Mira, 2010) which face significant computational challenges with large state-spaces and long horizons. This paper develops the index approximation approach to model dynamic allocations, which overcomes the computational challenges in estimation. An extensive theoretical literature on dynamic allocation inspires this approach. Theoretical literature has often called dynamic allocation problems like that in our setting “multi-armed bandits” and has studied models of optimal exploitation-exploration tradeoff in the form of indices. Gittins & Jones (1979), Weitzman (1979), Lai & Robbins (1985), Whittle (1988), Bergemann & Välimäki (1996), Bolton & Harris (1999), Auer et al. (2002), Keller et al. (2005), Bubeck & Cesa-Bianchi (2012), and Russo et al. (2017) are some works in this literature. A couple of empirical works have analyzed bandit-like problems in other economic settings (Miller, 1984; Erdem & Keane, 1996; Crawford & Shum, 2005; Dickstein, 2018; Li et al., 2020; Caria et al., 2020). The index approximation approach formalizes many methodological insights from those earlier papers.

The rest of this paper is organized as follows. Section 1.2 discusses the empirical setting. Section 2.4 describes the data. Section 1.4 builds the model. Section 1.5 describes model fitting. Section 2.6 shows model fitting results. Section 1.7 shows counterfactual results. Section 2.7 concludes.

1.2 Empirical Setting

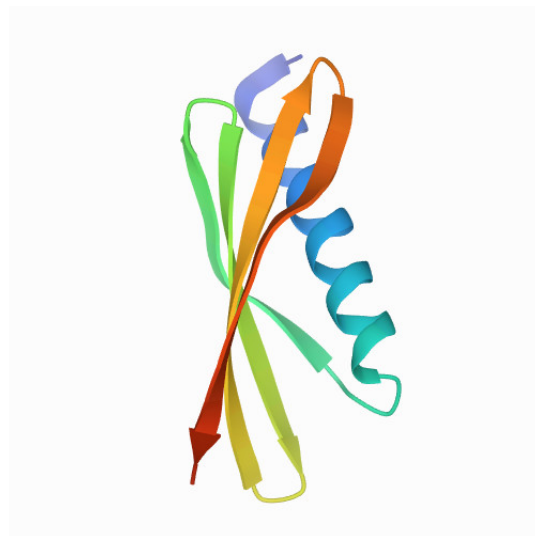
The empirical setting of this paper is a group of large structural biology labs funded by a \$1.3 billion NIH program from 2000–2015. This is an ideal setting to study for many reasons. First, this setting captures the complexity of decision-making typical in resource allocation for innovation and affords highly granular data. Second, the labs thoroughly documented how they used information observed during past allocations to inform future decision-making, providing guidance for modeling. A few additional NIH policy features also make this setting particularly clean for model identification.

A particularly clean feature of this setting is that research projects are clearly defined and distinct from each other, making it easy to compile highly granular data that associates input and output with specific projects. Moreover, the labs recorded input allocation in discrete units, making it easy to track the amount of input allocated to each project.

A project in this setting is the determination of the three-dimensional structure of a protein molecule, and the molecule uniquely identifies the project. These projects are important research. Knowing a protein’s structure is often critical for developing drugs that target the protein. Proteins consist of building blocks

called amino acids. These amino acids are arranged into a chain, which folds up onto itself, creating a three-dimensional structure (Hill & Stein, 2020, 2021). Figure 1.1 shows one structure, published by one of the labs in my data. Publications of protein structures have aided the development of therapeutics such as oncology drugs (Van Montfort & Workman, 2017) and COVID-19 vaccines (Wrapp et al., 2020).

Figure 1.1: Three-dimensional structure of a protein molecule



Note: The Northeast Structural Genomics Consortium (NESG), one of the large labs in my data, published this structure in 2014 (Guan et al., 2014).

A unit of input in this setting is a distinct experimental trial of the project. Each trial is a bundle of labor, capital, and materials associated with working on the molecule. A trial proceeds in multiple sequential stages. The outcome of each stage is either a success or a failure. If a trial succeeds in all stages, it produces a unit of output—a publication of the molecule’s structure. Trials are very risky. 98% of trials in my data failed at various stages, and there were large variations in trial outcomes even within projects.¹ Output is thus a random variable. I will assume that a trial stops if and only if the trial fails, which means the labs do not actively kill ongoing trials. This assumption significantly simplifies the allocation problem and is mostly

¹Much of the work in these trials involves trial-and-error of different combinations of experimental conditions for physicochemical reactions. As a result, researchers often need to perform many trials for a project before successfully producing a structure for publication. Luck is an important element in the process (Chruszcz et al., 2008).

supported by the patterns in the data.²

This setting captures a realistically complex resource allocation problem over a long horizon. Figure 1.2 illustrates the decisions a structural biology lab faces. In each period, the lab decides to allocate units of input among a portfolio of research projects. The portfolio contains older projects the lab has worked on in the past, and the lab can draw new projects to include in the portfolio. The objective is to maximize over the horizon welfare from the output. The lab or the NIH determines the welfare from each unit of output according to some welfare weights unknown to us the economists. A key challenge in the lab's decision-making is that trial outcomes are very noisy and the lab has incomplete information about which trials are likely to produce output.

Figure 1.2: Lab makes decisions about allocating resources across projects in each period

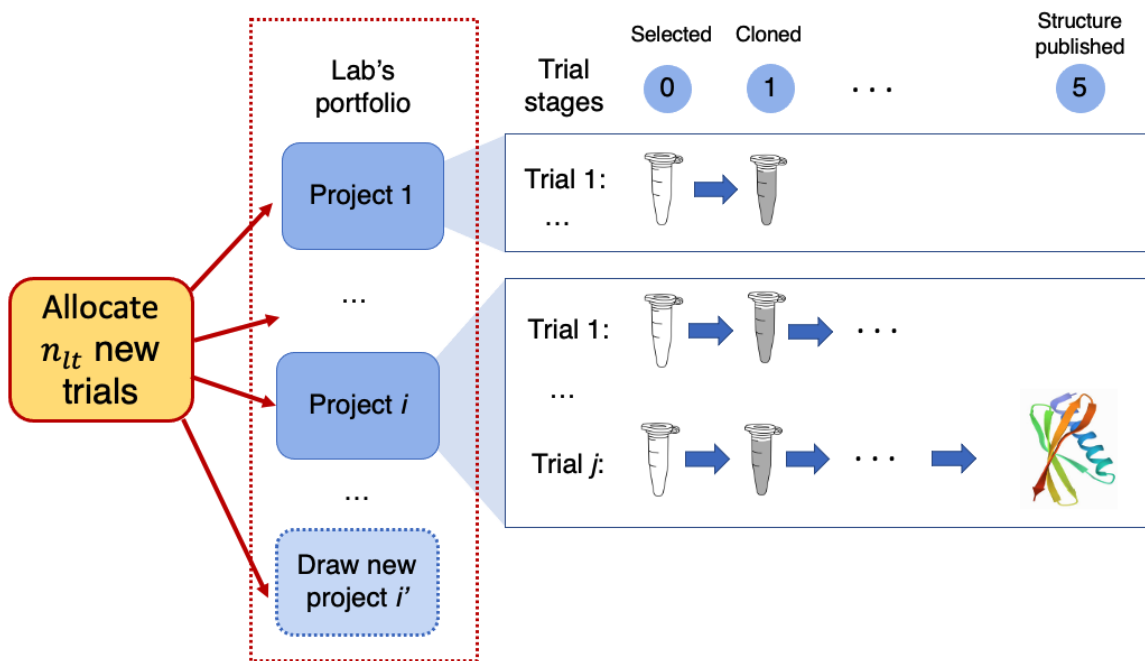


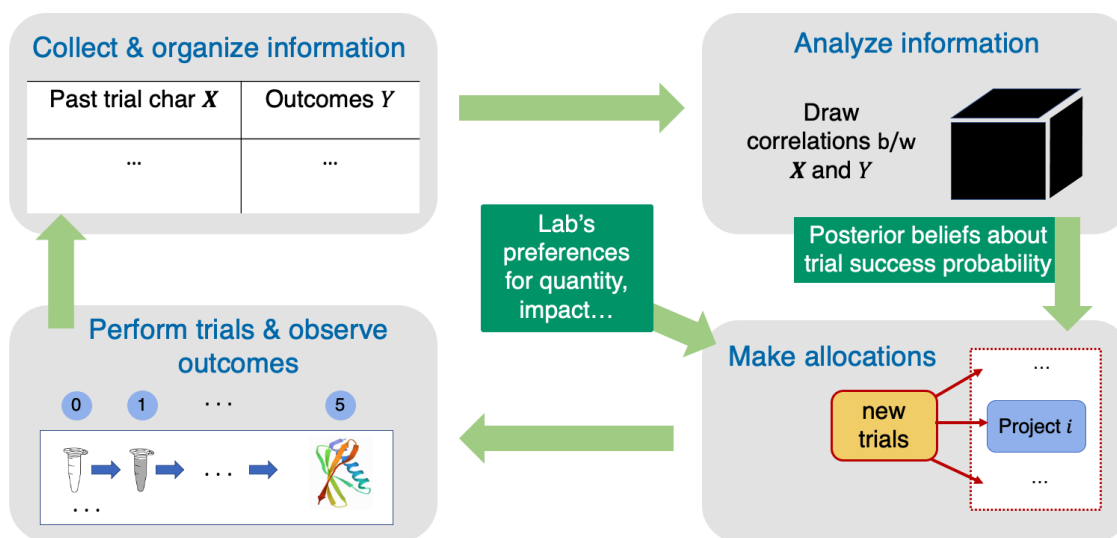
Image credits: The images of PCR tubes are from labicons.net.

The labs used information they observed during past allocations to inform decision-making and thoroughly

²For 15% of the trials, I observe the reasons for trial terminations. The reasons are mostly exogenous, e.g., "expression failed", "purification failed", "poor diffraction" and so on. In some cases, the trial termination reason is "duplicate target found" when the labs found a highly similar project was already published. Upon further investigation, I find the ongoing trials in most of those cases were allowed to keep going. Some of those trials were even successful at all stages and produced publications.

documented the process, providing guidance for modeling for this paper (Slabinski et al., 2007a,b; Jaroszewski et al., 2008; Price et al., 2009a,b; Babnigg & Joachimiak, 2010; Jahandideh et al., 2014). Figure 1.3 illustrates the process. As a lab performs trials and observes their outcomes, it collects and organizes the information about the trials' characteristics and outcomes. The lab then analyzes that information, fitting machine learning models to draw correlations between trial characteristics and outcomes. The fitted models allow the lab to predict a new trial's probability of producing output based the trial's characteristics. That prediction forms the lab's posterior belief about the trial's output. The lab then allocated trials based its posterior beliefs about different trials' output and its welfare weights. As the lab performs more trials and observes more information over time, it refits its machine learning models periodically.

Figure 1.3: Collection and analysis of information guide allocation decisions



Several NIH policy features simplify this setting and mitigate worries about modeling and identification, one of them being limits on the labs' choices of projects. The NIH periodically drew families of novel molecules and solicited nominations of molecules from the biomedical research community for the labs to work on. The labs could also pursue projects of their own interest but had to communicate with the NIH about those projects well in advance (NIGMS, 2007a,b, 2011b). Those processes limited the projects in the labs' choice sets, allowing me to construct choice sets that align well with the actual ones.

Another feature is the collaborative nature of the grant program, which mitigates worries about competition between the labs confounding with the exploitation-exploration tradeoff. The NIH funded the program, called

the Protein Structure Initiative (PSI), through a collaborative U01 mechanism, rather than the competitive R01 mechanism. Throughout the program, the labs faced little competition for funding (and research questions due to the NIH assigning projects).³

The NIH also evaluated the labs' productivity periodically based on a set of metrics,⁴ which limited what the labs placed nonzero welfare weights on. Metrics included the quantity of structures published, novelty, biomedical importance, human proteins, eukaryotic proteins,⁵ and membrane proteins.⁶

Moreover, an exogenous change of NIH's preferences over the metrics in 2009 provides a nice robustness check. A well-fitting model of the labs' decision-making should capture this shift in its estimates of the labs' welfare weights. Before 2009, the NIH had a strong preference for the quantity of publications and novelty (NIGMS, 2008a, 2011a). By mid-2008, the lack of emphasis on biomedically important projects had sparked heated debates in the community (Petsko, 2007; Moore, 2007); this prompted the NIH to boost its preference for biomedically important projects since 2009 (NIGMS, 2007c, 2008b, 2009a,b). To facilitate the change, the NIH partnered with outside researchers to identify biomedical important projects and gave more attention to biomedical importance in its evaluation process.

In addition, the NIH's close involvement in those labs makes it reasonable to assume away the principal-agent problem, which further simplifies the setting.

1.3 Data

The main data in this paper is those structural biology labs' resource allocation and output across projects. I observe, at the daily frequency, the input allocated to and output produced by each research project each lab

³A more typical lab faces additional tradeoffs due to competition. Hill & Stein (2021) found structural biology labs traded off quality of research and speed of publication because of competition to establish priority in publication. However, they also found the structural biology labs in my data (labelled as "structural genomics (SG)" labs in their paper) were not as motivated by competition to sacrifice quality for speed.

⁴The NIH published statistics on those metrics online. The archived versions of those publications can be found by searching the urls <http://targetdb.pdb.org/metrics/milestonestables.html> and <http://targetdb.pdb.org/Metrics/SummaryTable.html> on the Internet Archive.

⁵Eukaryotes include all living organisms other than the eubacteria and archaeobacteria. An eukaryote is an organism consisting of a cell or cells in which the genetic material is DNA in the form of chromosomes contained within a distinct nucleus.

⁶Membrane proteins are proteins found in the cell membrane. Membrane proteins are particularly hard for structure determination due to their physicochemical properties.

Table 1.1: Key variables and data sources

Variable	Variable Description	Source
Input allocation		
i	Unique identifier of a project or molecule	Berman et al. (2017)
j_i	Unique identifier of the j th trial of project i	Berman et al. (2017)
n_{lt}	Number of trials to allocate at lab l on day t	Berman et al. (2017)
$novel_i$	Binary = 1 if project i was novel	Berman et al. (2017)
$prevStruct_{iy}$	Number of published structures in the same protein family as i by year y	EMBL-EBI (2021)
$biomed_i$	Binary = 1 if project i was biomedically important	Berman et al. (2017)
$prevPub_{iy}$	Number of publications (not limited to structures) on molecule i by year y	UniProt Consortium (2021)
$human_i$	Max percentage identity of i to any human molecule	UniProt Consortium (2021)
$eukaryote_i$	Max percentage identity of i to any eukaryotic molecule	UniProt Consortium (2021)
$membrane_i$	Binary = 1 if molecule i is a membrane protein	UniProt Consortium (2021)
$funding_{ly}$	Funding in dollars lab l received in year y	NIH (2019), NIH (2021)
Observed output		
Y_{ijt}	Binary = 1 if trial j_i on day t succeeded	Berman et al. (2017)
Y_{ijkt}	Binary = 1 if stage k of trial j_i on day t succeeded	Berman et al. (2017)
$citation_{iy}$	five-year citations and mentions of project i published in year y	Varadi et al. (2020)
$download_{im}$	Number of downloads of project i in month m (between Aug 2007 and Nov 2013)	wwPDB (2013)
Posterior beliefs about output		
$\hat{E}_{\tilde{F}_t}(p_{ijt})$	Best-effort replication of the labs' posterior expectation of the probability of success of trial j_i on day t	Appendix A.2.1
$\widehat{Var}_{\tilde{F}_t}(p_{ijt})$	Posterior variance of the probability of success of trial j_i on day t	Appendix A.2.1
$\hat{E}(citation_{iy})$	Expectation of five-year citations and mentions of project i published in year y	Appendix A.2.2
$\hat{E}(download_{iy})$	Expectation of five-year downloads of project i published in year y	Appendix A.2.3

Note: This table only summarizes key variables and data sources. For a full description of all data sources and variables, please see Appendix A.1.

ever attempted over the horizon. Highly granular data is available thanks to an NIH requirement for data collection and sharing.⁷ Available in this data are all projects in a lab's portfolio, all trials performed on each project, dates and outcomes of stages of each trial, output of structures (if any), and identifiers of projects and structures to link to other bioinformatics databases. Labs could see each others' data through a shared information system at all times.

This section describes the key variables on input allocation, observed output, and the labs' posterior beliefs about output. Table 1.1 shows a list of the key variables and their sources. Appendix A.1 describes variable construction and shows additional variables.

1.3.1 Input Allocation

The input allocation data consists of units of input (trials) each lab allocated to individual projects on each day, and those projects' and trials' characteristics.

A project is identified by the distinct biological molecule it studied. I use the letter i to denote a project. The data contains 335,553 projects.

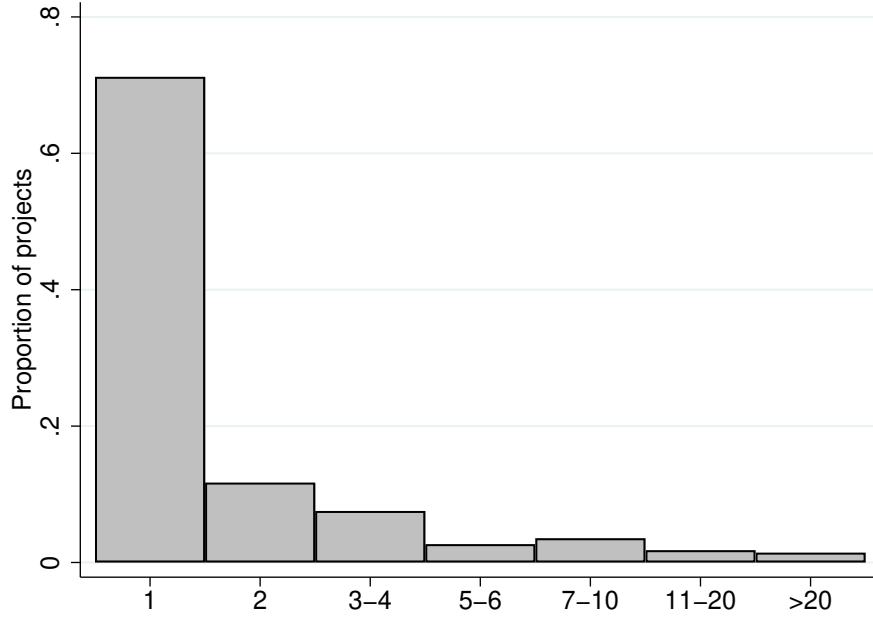
A trial is identified by its association to a specific project and the order in which the trial occurred. I use j_i to denote the j th trial of project i . The data contains 961,260 unique trials and captures 3,783,026 records of trial progress updates at the daily level. 99% of these records were between 2000 and 2015. Figure 1.4 shows a distribution of the number of trials allocated to each project. The distribution has a prominent right skew. 71% of projects had only a single trial. In contrast, less than 4% of projects had more than 10 trials. This pattern is consistent with extensive exploration, where the labs sampled projects they had poor information about in input allocation to learn which of those projects would be promising.⁸

Different labs in this data operated on very different scales, and I will focus on analyzing the four largest

⁷This data includes 27 labs that received NIH funds and 13 international labs that contributed data on a voluntary basis. A lab was typically a consortium that had sub-labs handling different stages of the trials. The number of sub-labs captured in this data is 147.

⁸The high proportion of single-trial projects could be consistent with an allocation model that only exploits, but that would require the first trial of each project to provide a lot of information about the project. It is possible that the labs started the first trials of a lot of new projects thinking the projects were high-payoff. Then they learned enough information about each project's potential from a single trial to identify low-payoff projects whose future trials were unnecessary. However, given that trial outcomes were very noisy, it is unlikely that the first trial would convey so much information to generate the observed patterns. Moreover, with updating of posterior beliefs, it is unclear how the labs sustained the belief that new projects were high-payoff under this model over time, given that most new projects turned out to be low-payoff ones after a single trial.

Figure 1.4: Distribution of input allocation across projects



Note: This figure includes all 335,553 projects ever attempted in my data.

labs.⁹ These four labs accounted for 71% of the projects and 85% of the trials in this data. The other labs were considerably smaller and had high variations in their research interests and funding levels.

I use n_{lt} to denote the amount of input available for allocation at lab l on day t and let n_{lt} equal the actual number of trials allocated at the lab on that day. For the four largest labs, n_{lt} has a mean of 35 trials per day per lab.

Several variables capture a lab's rationale for allocating a trial to a project, and these variables correspond to the NIH's evaluation metrics for the labs' productivity, among which a key metric is the novelty of the project. The variable $novel_i$ is a binary and is equal to 1 if the labs cited novelty as a reason to allocate trials to project i in the information system.¹⁰ When the labs cited a project i as being novel, they often emphasized that there were no or few already published structures in the same protein family as i . I therefore construct

⁹They are Joint Center for Structural Genomics (JCSG), Midwest Center for Structural Genomics (MCSG), New York Structural Genomics Research Consortium (NYSGRG), and Northeast Structural Genomics Consortium (NESG).

¹⁰The information system contains textual descriptions of why labs allocated trials to a project. The relevant fields are populated for 84% of projects at the four largest labs. Construction of $novel_i$ and $biomed_i$ is based on keywords in those descriptions. See the full list of keywords and additional details in Appendix A.1.1.

$prevStruct_{iy}$, a continuous variable that captures the changes over year (subscripted with the letter y) of the number of published structures in the same protein family as i .

Another key NIH evaluation metric is the biomedical importance of the project. The variable $biomed_i$ is a binary and is equal to 1 if the labs cited biomedical importance as a reason to allocate trials to project i in the information system.¹¹ As an additional proxy for the biomedical importance of a molecule, I look into the number of publications related to the molecule, including structures and other types of publications. I construct $prevPub_{iy}$, a continuous variable that captures the changes over year of the number of publications on molecule i .

Additional NIH evaluation metrics correspond to whether the project was related to human beings, eukaryotes,¹² and the cell membrane. The variable $human_i$ captures how similar molecule i is to any molecules from human beings.¹³ The variable $eukaryote_i$ captures how similar molecule i is to any molecules from eukaryotes.¹⁴ The variable $membrane_i$ is a binary and is equal to 1 if molecule i is related to the cell membrane.

I also obtain the dollar value of funds each lab received as an alternative measure of input. The variable $funding_{ly}$ captures the total amount of funds lab l received from the NIH in year y . Funding information comes from two sources. First, the NIH released a series of funding opportunity announcements (FOAs) directly tied to the grant program (NIH, 2019),¹⁵ which allows me to search directly all grants associated with those FOAs on NIH RePORT database (NIH, 2021). Second, labs sometimes received supplementary funds from the NIH so I also perform a direct search of the labs' names and abbreviations using RePORT's advanced search functionality. I then aggregate each lab's sum of research grants by year from the search results. Figure 1.5 shows the level of funding for the four largest labs.

¹¹Same as footnote 10.

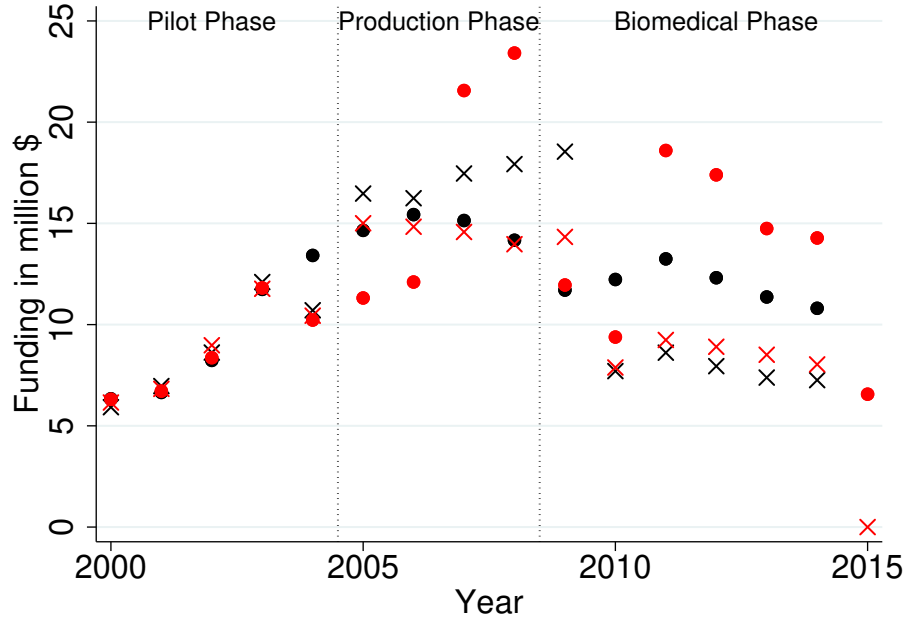
¹²See definition in footnote 5.

¹³When a lab worked on a "human" molecule, often the molecule was actually from bacteria but was very similar to a molecule from human beings and was much easier than the human molecule. Therefore, the right construction for $human_i$ is molecule i 's degree of similarity to human molecules rather than being a human molecule itself. I learned this from a conversation with an NIH program officer in charge of the grant program.

¹⁴Same as footnote 13.

¹⁵See the full list of FOAs in Appendix A.1.2.

Figure 1.5: Funding of the four largest labs



Note: Each series represents one lab.

1.3.2 Observed output

The main measure of output is the number of unique publications. A trial produced a publication if it succeeded in all stages. I use a binary Y_{ijt} to denote the outcome of trial j_i on day t . $Y_{ijt} = 1$ if the trial succeeded and produced a publication. Only 1.6% of trials in the data succeeded. I also observe the outcomes of all stages of each trial and use a binary Y_{ijk_t} to denote the outcome of stage k of trial j_i on day t .¹⁶ I use these intermediate outcomes to construct the labs' posterior beliefs about output.¹⁷

Uniqueness of a publication is important because more publications do not necessarily mean more welfare when lots of publications are duplicates of earlier ones. Duplicate publications happened in our setting because a lab often allocated multiple trials on the same project at roughly the same time and sometimes

¹⁶ $k = 0, 1, 2, 3, 4$. $Y_{ij0t} = 1$ if DNA was successfully cloned. Y_{ij1t} is only defined when $Y_{ij0t} = 1$ and is equal to 1 if protein was successfully expressed. Y_{ij2t} is only defined when $Y_{ij0t} = 1$ and $Y_{ij1t} = 1$ and is equal to 1 if protein was successfully purified. Y_{ij3t} is only defined when $Y_{ij0t}, Y_{ij1t}, Y_{ij2t} = 1$ and is equal to 1 if protein was successfully prepared for studying its structure (through X-ray crystallography or NMR or cryo-EM). Y_{ij4t} is only defined when all previous stages were successful and is equal to 1 if the structure was successfully produced and deposited to the Protein Data Bank (PDB) for publication.

¹⁷See Section 1.3.3 and Appendix A.2.

more than one trial succeeded. These successful trials produced structures of the same molecule often of very similar qualities. I will therefore only count the first publication on each project in measuring a lab’s output. The data contains 15,848 publications of structures, among which 10,501 are unique. This represents 15% of world output of structures between 2000 and 2015.

I supplement the main output measure with the number of citations and downloads of each published project. The variable $citation_{iy}$ captures the five-year citations and mentions of a project i published in year y .¹⁸ The variable $download_{im}$ captures the number of downloads of a project i in month m between August 2007 and November 2013.¹⁹

The two measures capture quality in different dimensions. Citations in the academic literature reflect the amount of follow-on research directly built on the published project. Downloads offer a window into the more intangible, such as the amount of interest generated by the publication and the use of the publication in developing new technologies in related fields (for example, DeepMind (2020)).

Figure 1.6 shows the observed output for the four largest labs.

1.3.3 Posterior Beliefs about Output

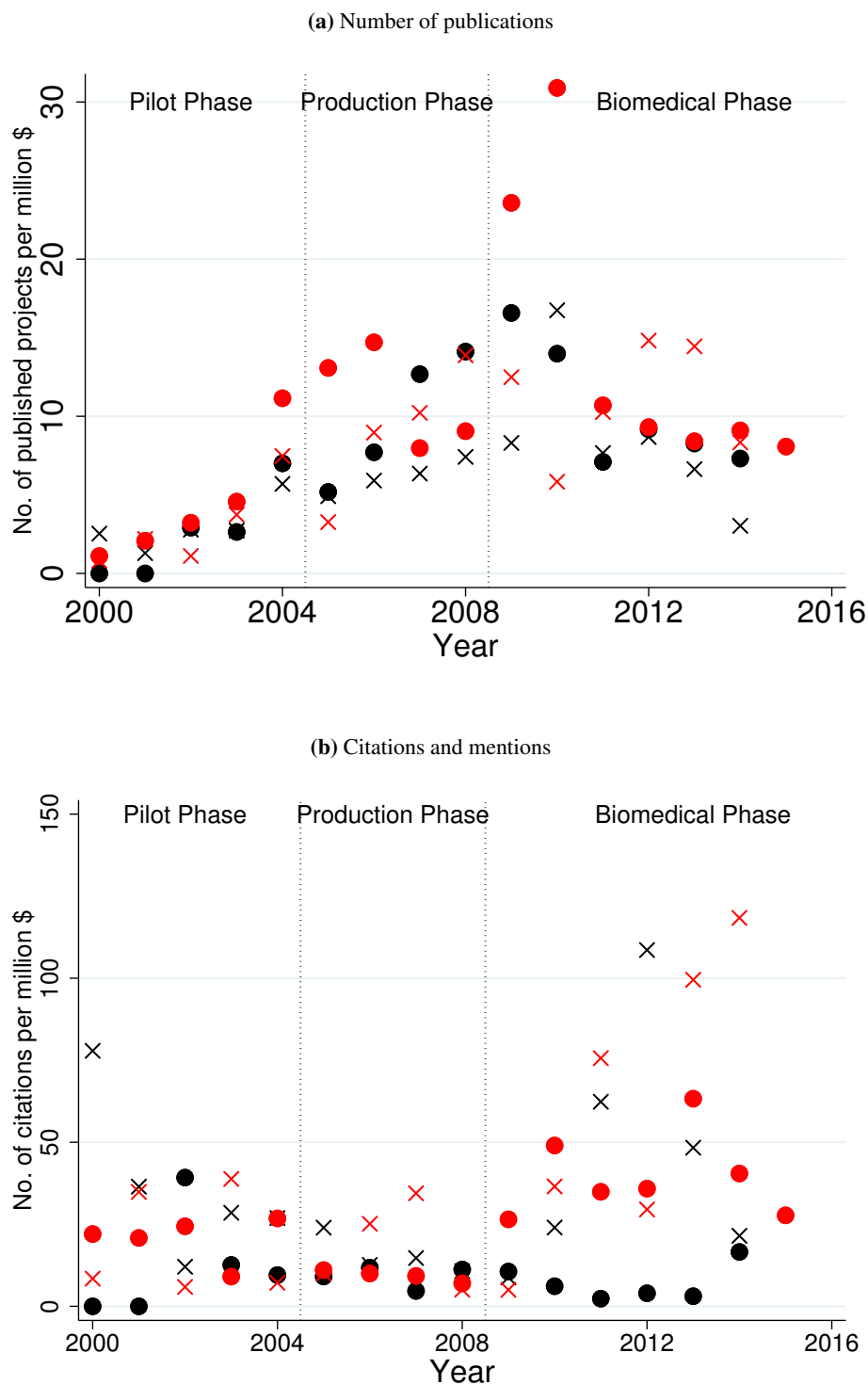
A key model ingredient of this paper is the labs’ posterior beliefs about output from potential allocations at the moment of making allocations.

My measures of the labs’ posterior beliefs about output are based on best-effort replications of the labs’ documented approach of belief formation using information they observed during previous allocations. Variable construction uses several hundreds of variables, coming from a variety of sources. A full list of these variables is available in Appendix A.1.4.

¹⁸When multiple structures of the same project were published, I take the mean of five-year citations and year of publication of those structures. I do not take the sum of five-year citations because those structures were often cited together.

¹⁹This variable is not directly useful because downloads have strong lifecycle trends—downloads were high when a structure just became published then fell over time—and comparing different publications at different stages of their lifecycles is meaningless. In Section 1.3.3, I construct a variable corresponding to the five-year downloads to make comparison across projects meaningful. For that variable, when multiple structures of the same project were published, I take the mean of five-year downloads and year of publication of those structures. I do not take the sum of five-year downloads because those structures were likely downloaded together as they were often cited together.

Figure 1.6: Observed output of the four largest labs



Note: Each series represents one lab. Panel 1.6a shows the number of published structures on unique molecules in a given year divided by the lab's funding in millions in that year. Panel 1.6b shows the number of five-year citations and mentions the published structures in that year generated, divided by the lab's funding in millions in that year.

The four largest labs formed and updated their posterior beliefs about whether a trial would succeed through supervised machine learning (Slabinski et al., 2007a,b; Jaroszewski et al., 2008; Price et al., 2009a,b; Babnigg & Joachimiak, 2010; Jahandideh et al., 2014). This involved using machine learning models to fit observed trial outcomes (successes or failures) on those trials’ characteristics. The fitted model would be able to predict a new trial’s probability of success based on the trial’s characteristics. The labs periodically refitted the models when newer trial data became available to make updates. The labs started using this approach to form posterior beliefs as early as the beginning of 2005, when they had accumulated a considerable amount of trial data in the initial years of operation (Slabinski et al., 2007a,b). Over time and lab-wise, they used a variety of models, ranging from logistic regressions to support vector machine to random forest. Jahandideh et al. (2014) conducted a comparison study and found random forest worked best based on multiple metrics, including prediction accuracy.²⁰

To mimic how the labs updated their posterior for trial success probability, I fit a series of machine learning models over time between 2005 and 2015. To fit a model corresponding to the labs’ posterior beliefs at time t , I use a training sample consisting of trial outcomes realized before t and those trials’ characteristics. The characteristics include the ones the labs used and fall under three categories:

- Physicochemical properties of the molecule based on scientific reasoning.
- Other characteristics of the project, such as novelty, biomedical importance, and the number of prior publications on the molecule.
- Past successes and failures of trials on the same project.

I use random forest to fit each model. Appendix A.2.1 shows the details.

The fitted models allow me to construct two variables, the posterior expectation of the probability of success $\widehat{E}_{\tilde{F}_t}(p_{ijt})$ and the posterior variance $\widehat{Var}_{\tilde{F}_t}(p_{ijt})$. Let the probability of success of a trial j_i at time t be p_{ijt} . The fitted model $\tilde{F}_t$ makes predictions about p_{ijt} based on the trial’s characteristics, and these predictions constitute the posterior belief about p_{ijt} . The fitted model $\tilde{F}_t$ is a random forest consisting of an ensemble of submodels, each called a decision tree. Each decision tree fits the training sample independently and makes an independent prediction $\hat{p}_{ijt}^{tree}$. The collection of these predictions forms the estimated posterior distribution.

²⁰Based on my conversation with a project coordinator at one of the labs, the accumulation of trial data rather than the choice of the machine learning model was the main driver in improving the quality of predictions.

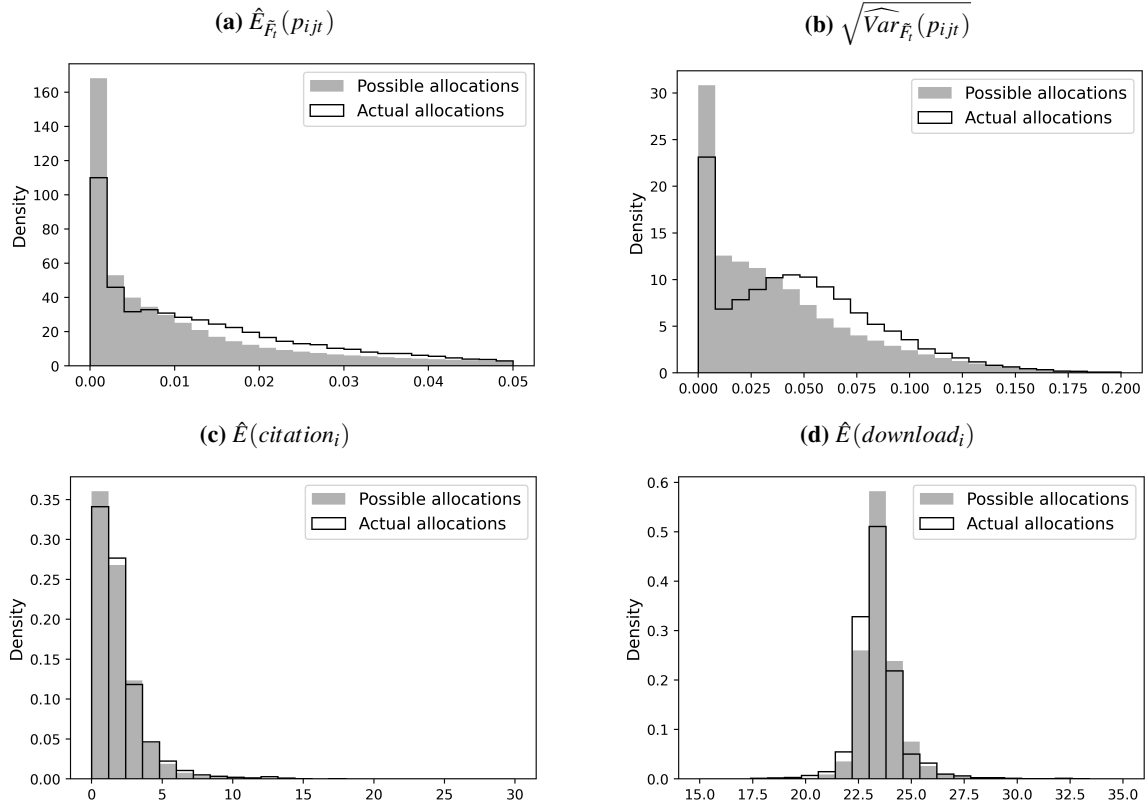
The mean of this distribution is the posterior expectation $\widehat{E}_{\widehat{F}_t}(p_{ijt})$, and the variance is the posterior variance $\widehat{Var}_{\widehat{F}_t}(p_{ijt})$.

My best-effort replication may deviate from the labs' actual posterior beliefs, but I do not expect the deviations to bias estimation results of the allocation model. I note a list of possible deviations in Appendix A.2.1. One example is that the labs did not always predict $\widehat{Var}_{\widehat{F}_t}(p_{ijt})$. When they did, the variable took the form of comparing predictions from multiple models side by side (Slabinski et al. (2007a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)). It is reasonable to believe the labs knew the value of analyzing the variation in predictions from different models, though they did not formally use a variable to represent that variation. $\widehat{E}_{\widehat{F}_t}(p_{ijt})$ and $\widehat{Var}_{\widehat{F}_t}(p_{ijt})$ therefore contain errors in the sense they may differ from the labs' actual posterior beliefs. But as long as the errors are random or at least not correlated with the actual allocation decisions, they should not bias the estimates of the allocation model. The errors may be correlated with the actual allocations if I fail to include in the training sample variables the labs actually used to form beliefs or make allocations. To minimize this risk, I include in my training samples all variables the labs ever mentioned using and all NIH evaluation metrics.

For the posterior beliefs about citations and downloads, I make the simplifying assumption that the labs' posterior beliefs stayed the same in all periods and were the same as the ground truth. This is equivalent to assuming that the labs only had incomplete information about the possibility of production and had perfect information about the kind of output produced if production does happen. Given that the projects were very well-defined, this assumption is reasonable. Another evidence in support of this assumption is that, unlike the probability of success of a trial, the labs did not form and update their posterior beliefs about the number of citations and downloads of a publication in a systematic way.

I use ridge regressions to model the posterior expectations about citations $\widehat{E}(citation_{iy})$ and about downloads $\widehat{E}(download_{iy})$. To align well with the ground truth, these models have to be able to make small prediction errors out of the training sample. Appendices A.2.2 and A.2.3 show details of model fitting and visual evidence of out-of-sample model fit in cross validation.

Figure 1.7: Estimated posterior beliefs about output



Note: Each panel shows two distributions. The outlined bars show the distribution of the variable for trials actually allocated at the four largest labs. The size of this data is 714,736. The grey unoutlined bars show the distribution of the variable for a random sample of all possible trials in those labs' choice sets. The size of this data is 105,533,539. See Appendix A.3.2 about sampling from choice sets. Panel 1.7a shows the distributions of the posterior expected probability of success of a trial. Panel 1.7b shows the posterior standard deviation of the probability of success of a trial. Panel 1.7c shows the expected five-year citations and mentions of a trial upon publication. Panel 1.7d shows the expected five-year downloads (in thousands) of a trial upon publication. An observation from a lab is inversely weighted by the total number of observations from this lab so each lab contributes equally to the aggregate distribution. The number of bins is coarse and equal to 25 to avoid depicting very tall bars near zero. The distributions are truncated to their respective ranges of x-axis values.

Figure 1.7 shows the estimated posterior beliefs and provides descriptive evidence that the labs explored high-variance projects in resource allocation and, in doing so, forwent some opportunities to exploit projects with high posterior expected output.

Figure 1.7a shows the distribution of $\hat{E}_{\tilde{F}_i}(p_{ijt})$ for trials the four largest labs actually allocated versus that for a random sample of trials in those labs' choice sets. While the distribution of $\hat{E}_{\tilde{F}_i}(p_{ijt})$ for the actual allocations is less right-skewed than that of the possible allocations, a considerable proportion of actual allocations had extremely low posterior expected probabilities of success. This is unlikely due to exhaustion

of good possible allocations with high posterior expected probabilities of success. The actual allocations accounted for only 0.7% of the random sample of the possible allocations.

Figure 1.7b shows the distribution of $\sqrt{\widehat{Var}_{\hat{F}_i}(p_{ijt})}$. The figure shows that on average the actual allocations had considerably more posterior variance in the probability of success than the possible allocations, providing evidence that the labs allocated a considerable amount of resources to high-variance projects.

Figures 1.7c and 1.7d show that the distributions of $\hat{E}(citation_i)$ and $\hat{E}(download_i)$. Both show the trials the labs allocated to were not very different from the trials the labs passed in terms of expected citations and downloads they would generate upon publication. This suggests choosing high-variance projects was associated with not being chiefly motivated by citations and downloads, providing evidence that the labs allocated to high-variance projects for the sake of exploration.

1.4 Model

I present an estimable model of the labs' allocation decision-making whose key feature is index approximation of value function. Under this model, in each period a lab first analyzes past information to form posterior beliefs about output from potential allocations. The lab then uses this posterior to compute a simple index to approximate the value function associated with allocating input to a project. The lab then allocates input to projects with the highest index values. This model has the advantage of being computationally tractable and of being empirically plausible, in the sense that the labs could have easily applied this model in actual decision-making.

I build intuition for the model as follows. First, I formalize the labs' input allocation problem into an objective function. I then discuss the intractability of a standard model of decision-making based on backward induction. I explain how index approximation differs from the standard model and why it is more tractable. I wrap up with a discussion of the theoretical foundation for index approximation.

I define the main model as well as several alternative models for robustness checks at the end of this section.

1.4.1 Objective Function

I formalize the labs' input allocation problem into an objective function that maximizes welfare from output over a finite time horizon. This objective function frames the allocation problem in line with the theoretical

literature.

I use C_{lt} to denote a lab l 's choice set on day t . This choice set consists of further trials of older projects already in the lab's portfolio and initial trials of new projects. For example, consider a day t where the lab has capacity to allocate $n_{lt} = 2$ new trials. Suppose the lab has had three trials of project i up to day $t - 1$; then the fourth and fifth trials of project i are in day t 's choice set. For another project i' the lab has received from the NIH and has not tried up to day $t - 1$; the first and second trials of project i' are in day t 's choice set.

I use a_{ijt} to denote input allocation decisions. $a_{ijt} \in \{0, 1\}$ where 1 represents allocating trial j_i to project i on day t and 0 represents not allocating j_i on day t . The vector $\mathbf{a}_{lt}$ has a length equal to the cardinality of the choice set C_{lt} and represents the action on each of the project-trials in the choice set.

I use Y_{ijt} to denote the outcome of trial j_i on day t . $Y_{ijt} = 1$ if the trial succeeded and produced a publication. I use Y_{ijk} to denote the outcome of stage k of trial j_i on day t .

I use Ω_t to denote the labs' information set on day t . Because the labs could observe each others' allocations and output at all times, all labs share the same information set on a day. Ω_t includes the actions $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{t-1}$ and outcomes and stage-specific outcomes $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$ observed before t from all labs.

I use p_{ijt} to denote the probability of success of trial j_i on day t . The labs and the economist have incomplete information about p_{ijt} . Let the prior distribution of p_{ijt} be $F_t(p_{ijt})$. The labs formed and updated posterior $\tilde{F}_t(p_{ijt}|\Omega_t)$ with supervised machine learning, and I made best-effort replication of the labs' posterior offline in Section 1.3.3.

I use $\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl})$ to denote the payoff (in welfare units) from project-trial j_i . It depends on three variables. $\pi_{ijt}(\cdot)$ depends on the actions $\mathbf{a}_{lt}$. Whenever $a_{ijt} = 0$, $\pi_{ijt}(\cdot) = 0$ because a trial not allocated does not pay off. When $a_{ijt} = 1$, $\pi_{ijt}(\cdot)$ also depends on p_{ijt} , the probability of success of the trial. $\pi_{ijt}(\cdot)$ also depends on $\boldsymbol{\theta}_{Xl}$, which are welfare weights predetermined by the lab.

$\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl})$ is not the same as output. One can think of it as a utility function that translates output into welfare according to the welfare weights $\boldsymbol{\theta}_{Xl}$. For example, consider a trial j_i that would produce a novel and biomedically important structure upon its success. The structure is a unit of output. Suppose the lab has a welfare weight of 5 on a novel structure and a welfare weight of 3 on a biomedically important structure; the welfare the lab receives from this trial upon its success is 8. The lab knew its welfare weights but the economist does not, so I will estimate $\boldsymbol{\theta}_{Xl}$ based on the lab's observed allocations. I describe the estimation procedure in Section 1.5.1.

A lab's objective is to maximize the posterior expected payoff over the horizon by choosing a sequence

of actions. The objective function is as follows:

$$\begin{aligned}
& \max_{\mathbf{a}_{l1}(\Omega_1), \dots, \mathbf{a}_{lT}(\Omega_T)} \sum_{t=1}^T \sum_{j_i \in C_{lt}(\Omega_t)} \int \pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t), \\
& \text{subject to } \sum_{j_i \in C_{lt}(\Omega_t)} a_{ijt} = n_{lt}, \\
& \text{and for all } j_i < j'_i \in C_{lt}(\Omega_t), \text{ if } a_{ijt} = 0 \text{ then } a_{ij'_t} = 0.
\end{aligned} \tag{1.1}$$

The first constraint is the capacity constraint. The second constraint rules out allocating a fifth trial to a project when the fourth trial has not been allocated.

1.4.2 Intractability of Backward Induction

The most standard way of modeling how agents choose actions in a problem like equation (1.1) is to assume they specify the value function of different actions as a Bellman equation and solve the Bellman equation by backward induction. The value function of a state is the value attained by equation (1.1) at this state at the sequence of optimal actions $\mathbf{a}_{l1}^*, \mathbf{a}_{l2}^*, \dots, \mathbf{a}_{lT}^*$. The state variable in our setting is the information set Ω_t .

The Bellman equation has two additively separable components, one component representing the posterior expected payoff in the current period and the other representing the continuation value. We can define the action-specific value function $V_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl})$ of project-trial j_i on day t as follows

$$V_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}) = \underbrace{\int \pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t)}_{\text{posterior expected payoff}} + \underbrace{E_{\Omega'_{t+1}} \left[\max_{\mathbf{a}_{l,t+1}} V_{i,j',t+1}(\Omega'_{t+1}, \mathbf{a}_{l,t+1}) \mid \Omega_t, \mathbf{a}_{lt} \right]}_{\text{continuation value}}. \tag{1.2}$$

The value function $V_{ijt}(\Omega_t; \boldsymbol{\theta}_{Xl})$ of project-trial j_i on day t is equal to $V_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl})$ evaluated at the optimal $\mathbf{a}_{l1}^*, \mathbf{a}_{l2}^*, \dots, \mathbf{a}_{lT}^*$.

As the continuation value integrates over the future evolutions of the state, solving the Bellman equation with backward induction and obtaining the optimal sequence of actions $\mathbf{a}_{l1}^*, \mathbf{a}_{l2}^*, \dots, \mathbf{a}_{lT}^*$ is computationally intractable in our setting due to the curse of dimensionality. The state variable Ω_t consists of all previous actions and outcomes. The number of possible actions on a day are sometimes on the order of millions. The time horizon has thousands of days. The number of possible states at the end of the horizon is astronomical.

1.4.3 Index Approximation

I present an alternative way of modeling how the labs chose actions, that is, assuming they used a simple index to approximate the value function.

Index approximation overcomes the curse of dimensionality because the approximated continuation value does not integrate over the future evolutions of the state. Let $V_{ijt}^A(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_l)$ represent the lab's approximation to the action-specific value function in equation (1.2),

$$V_{ijt}^A(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_l) = \underbrace{\int \pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t)}_{\text{as before}} + \underbrace{B_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Bl})}_{\text{approximation to continuation value}}. \quad (1.3)$$

It has two additively separable components in direct parallel to the Bellman equation. The first component represents the posterior expected payoff in the current period and is identical to that in equation (1.2). The second component is an unknown function that the lab is assumed to use to approximate the continuation value. The key difference between this component and the continuation value in equation (1.2) is that the former is not a function of the future state Ω'_{t+1} . The lab is assumed to use a heuristic $B_{ijt}(\cdot)$, such as the posterior variance of the payoff, to approximate the value of continuing exploration of a project, and this heuristic is based only on information available to the lab at time t . Theoretical literature would call $B_{ijt}(\cdot)$ the “exploration bonus.”

Because the computation of the approximation term only uses information currently available, the lab can solve equation (1.3) directly, without the need to backward induct. Assume the lab chose actions based on the approximate value function, acting as if the approximate value function was the true value function, it would simply choose actions $\mathbf{a}_{lt}^{A*}$ on each day t to maximize the sum of the approximate values on that day, as follows:

$$\begin{aligned} \mathbf{a}_{lt}^{A*} &= \underset{\mathbf{a}_{lt}}{\operatorname{argmax}} \sum_{j_i \in C_{lt}(\Omega_t)} V_{ijt}^A(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_l), \\ \text{subject to } &\sum_{j_i \in C_{lt}(\Omega_t)} a_{ijt} = n_{lt}, \\ &\text{and for all } j_i < j'_i \in C_{lt}(\Omega_t), \text{ if } a_{ijt} = 0 \text{ then } a_{ij'_t} = 0. \end{aligned} \quad (1.4)$$

With some very intuitive functional form assumptions about the payoff function $\pi_{ijt}(\cdot)$ (see Appendix A.3.1 for details), we can make $V_{ijt}^A(\cdot)$ only a function of a_{ijt} , not the full $\mathbf{a}_{lt}$ vector, and make the second constraint always holds at the solution. With those assumptions, $\mathbf{a}_{lt}^{A*}$ is equivalent to an index rule: the lab computes index $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l)$ for each project-trial in C_{lt} , and $\mathbf{a}_{lt}^{A*}$ are just allocating to the n_{lt} trials with the

highest index values.

Unlike actions based on backward induction, actions made using the index approximated value functions are not optimal. Actions made in this way just represent what might produce a reasonably good payoff. Extensive literature on dynamic allocation provides theoretical grounding for approximating the value function using an index in our setting.

Equation (1.1) is a finite-horizon multi-armed bandit (MAB). Allocation problems characterized by objective functions like equation (1.1) are called multi-armed bandits because of a classic example of this type of problem. Imagine you are playing a five-armed bandit slot machine and you have ten opportunities to pull. In what way should you allocate your pulls across arms to maximize the expected total payoff if you have little prior information about the payoff of each arm? Intuitively, you want to use a few pulls to sample payoffs from different arms to learn which arms are promising then use the remaining opportunities to pull those arms. In our setting, each project is an arm and each trial of a project is a pull of an arm.

Theoretical literature has found optimal or nearly optimal indices for some MABs. In the case of the standard infinite-horizon discounted MAB,²¹ the optimal index is the Gittins index (Gittins & Jones, 1979; Gittins, 1979). In the case of the standard finite-horizon MAB defined by the objective function

$$\begin{aligned} \max_{\mathbf{a}_{11}(\Omega_1), \dots, \mathbf{a}_{1T}(\Omega_T)} \sum_{t=1}^T \sum_{j_t \in C_{jt}(\Omega_t)} \int \pi_{ijt}(\mathbf{a}_{jt}, p_{ijt}; \boldsymbol{\theta}_{Xt}) d\tilde{F}^{(i)}(p_{ijt} | \Omega_t^{(i)}), \\ \text{subject to } \sum_{j_t \in C_{jt}(\Omega_t)} a_{ijt} = 1 \text{ for all } t, \end{aligned} \quad (1.6)$$

Lai & Robbins (1985) and Lai (1987) show that index rules do not provide exact solutions but are asymptotically optimal as the number of periods T goes to infinity, and have nearly optimal performance from both the Bayesian and frequentist viewpoints for moderate and small T . These nearly optimal index rules can be interpreted as the upper confidence bounds (UCB) for the posterior expected payoffs. Their results inspired a literature on indices that approximate the UCB (Agrawal, 1995; Bubeck & Cesa-Bianchi, 2012; Cappé et al., 2013). The simplest UCB index is just a function of the number of previous pulls of the arm (Auer et al., 2002).

Equation (1.1) differs from the standard finite-horizon MAB in a few ways, and, as a result, theoretical

²¹Defined by the objective function

$$\max_{\mathbf{a}_{11}(\Omega_1), \mathbf{a}_{12}(\Omega_2), \dots} \sum_{t=1}^{\infty} \beta^t \sum_{j_t \in C_{jt}(\Omega_t)} \int \pi_{ijt}(\mathbf{a}_{jt}, p_{ijt}; \boldsymbol{\theta}_{Xt}) d\tilde{F}^{(i)}(p_{ijt} | \Omega_t^{(i)}), \text{subject to } \sum_{j_t \in C_{jt}(\Omega_t)} a_{ijt} = 1 \text{ for all } t. \quad (1.5)$$

literature has yet to find an optimal index for it, but the literature has also shown indices may be good heuristics in those cases.

The first difference is that equation (1.1) does not assume stationarity of payoffs. $\tilde{F}_t(\cdot)$ in equation (1.1) is indexed with the t subscript while $\tilde{F}(\cdot)$ in equation (1.6) is not. As science advances continuously, the payoff of a trial might evolve over time independent of the states and actions. For example, a new technology that becomes available on day t could increase the probabilities of success of all trials in the future. Non-stationary MABs are called “restless bandits” (Whittle, 1988). Theoretical literature has found indices to be suboptimal for restless bandits in the general case (Ortner et al., 2012; Lattimore & Szepesvári, 2020). However, if one is willing to assume the nature of change in $\tilde{F}_t(\cdot)$, such as abrupt changes in unknown periods, Garivier & Moulines (2011) show that UCB-like indices can match the lower bound on regret up to a logarithmic factor. Such indices are based on the idea of “forgetting” and would discount information learned in early periods or use sliding windows to exclude that information.

Another difference is that equation (1.1) has a much more complex action space. Equation (1.1) allows for allocating multiple trials to the same project or to different projects in a period if $n_{it} > 1$ and n_{it} can change from period to period. All allocated trials reveal their payoffs. In contrast, equation (1.6) only allows for allocating one trial to one project in each period. MABs that allow for pulling multiple arms in each period, where all pulled arms reveal their payoffs, are called “combinatorial semi-bandits.” Works have shown that indices could work well for stochastic combinatorial semi-bandits (Kveton et al., 2015; Chen et al., 2016; Wang & Chen, 2018), albeit under quite strong assumptions on the action space and/or the functional form of the payoff function. Recent works push towards relaxing those assumptions (Lattimore & Szepesvári, 2020; Chen et al., 2021).

A further difference is that equation (1.1) does not assume independence between arms. Equation (1.1) integrates the payoff over the posterior distribution $\tilde{F}_t(p_{ijt}|\Omega_t)$, which is conditioned on the full information set at t . In contrast, equation (1.6) integrates over $\tilde{F}^{(i)}(p_{ijt}|\Omega_t^{(i)})$, where the information set is specific to each project i and learning about one project from its trial outcomes does not spill over to another project. Many research projects share similar characteristics, so their payoffs could be correlated. For example, 24% of projects in my data are membrane proteins, whose trials in general had low probabilities of success.²² When the agent has this kind of contextual information to help predict the payoff, the MABs are called “contextual

²²Only 1.2% of trials on membrane proteins succeeded while 1.8% of trials on non-membrane proteins succeeded in my data.

bandits” (Woodroffe, 1979; Langford & Zhang, 2007). Recent works have shown that UCB-like indices can achieve a nearly optimal regret guarantee on the order of $\tilde{O}(\sqrt{T})$ (Guan & Jiang, 2199; Zhou et al., 2020). These indices build models to correlate the contextual characteristics with the observed outcomes and then use the models to predict the UCB of the payoff of each pull.

Extensive practical applications of index heuristics in bandit-like problems also suggest using index approximation without tight theoretical justification might not be as alarming as it sounds. Algorithms based on index heuristics, especially the UCB, nowadays power applications ranging from product recommendations (Scale (2021)) to dynamic pricing (Yang et al. (2020)) to self-driving cars (Zhang (2021)). The real-world bandit problems are often too complex and messy to find an exact theoretical result so the uses of indices are often without tight theoretical justifications. This does not seem to concern practitioners. The practitioners’ goal is not to find the optimal decision model, but to find what works reasonably well and better than the alternatives. Given that the current theoretical literature often fails to find optimal solutions for complex MABs and needs to rely on index heuristics, it is implausible to think that the labs actually took the optimal actions. I therefore argue that index approximation is not just for computational convenience. Index approximation is a plausible empirical model of the labs’ decision-making because the labs could have applied this simple model in decision-making.

1.4.4 Models of Index Approximation

The main model modifies the simple, well-used UCB index from Auer et al. (2002). Under Auer et al. (2002), $B_{ijt}(\cdot)$ is a convex decreasing function of the amount of input previously allocated to the project.²³ Intuitively, UCB approximates the upper confidence bound of an allocation’s posterior expected payoff. Allocating input to a project that the lab has poor information about has a high UCB. As the lab allocates more input to the project and has better information about its productivity, the UCB of further allocations decreases. My modification to Auer et al. (2002) is an additional term that captures time-discounting of the value of older projects that the lab has attempted in the distant past. This term proxies for the unobserved changes in the

²³Under Auer et al. (2002),

$$B_{ijt}(\cdot, a_{ijt} = 1) = \begin{cases} \infty & \text{if } j = 1 \\ \sqrt{\frac{2\ln(N_{it})}{j-1}} & \text{if } j = 2, 3, 4, \dots \end{cases} \quad (1.7)$$

where N_{it} is the total number of pulls the agent has done so far. More recent implementation uses a fixed value θ_{B1} rather than $2\ln(N_{it})$ (Lattimore & Szepesvári, 2020). For my main model, I do not use an infinite value for $B_{ijt}(\cdot)$ when $j = 1$. If I do, V_{ijt}^A would be infinite and cause problems in estimation via maximum likelihood and in identifying θ_{B1} .

value of exploring a project over time from the lab's perspective, for example due to project-specific learning staying with individual researchers and personnel changes in the lab. The main model is as follows.

Main Model (UCB+D). *The labs set $B_{ijt}(\Omega_t, \mathbf{a}_{it}; \boldsymbol{\theta}_{Bl}) = a_{ijt}[\sqrt{\frac{\theta_{B1,l}}{j}} + \theta_{B2,l} \cdot (t - t'_{i,t})]$.*²⁴

The free parameter $\theta_{B1,l}$ captures the amount of exploration. The first few trials of a project have larger $B_{ijt}(\cdot)$ compared to later trials of the same project under this model and $\theta_{B1,l}$ determines how much larger. If a lab did not explore at all, the first few trials of its projects should not have larger $B_{ijt}(\cdot)$ relative to later trials. In that case, $\theta_{B1,l} = 0$. The free parameter $\theta_{B2,l}$ captures the amount of discounting of older projects. The variable $(t - t'_{i,t})$ measures the duration between the current period t and the period $t'_{i,t}$ in which the last previous trial on project i happened. A negative coefficient estimate of $\theta_{B2,l}$ would suggest the lab discounted projects it tried a while ago when it decided which projects to explore further. If the lab did not discount older projects, $\theta_{B2,l} \geq 0$.

I also define several alternative models for robustness checks. Those models are obvious candidates but one can surely add additional models. Theoretical literature and practical applications have proposed a wide range of other indices for MABs and we can in principle test each one of them. Ultimately, the models tested should be reasonably close to the allocation model the labs actually used. My conversations with NIH program officers and a project coordinator at one of the four largest labs reveal that the trial allocation process was quite heuristic. The labs used a “high-throughput” approach to trial allocation, where they allocated one initial trial to a lot of projects. If the projects were important or if the projects showed promise of success in previous trials, then they allocated more trials to those projects. As such, simple indices could well capture the actual allocation model. I therefore only chose to test the most parsimonious indices well-studied in theoretical literature.

The first of those alternative models is the static model. Under this model, the labs set the exploration bonus $B_{ijt}(\cdot)$ to zero and completely ignore exploration. This model does not have any free parameters in $B_{ijt}(\cdot)$ and is equivalent to restricting $\theta_{B1,l}$ and $\theta_{B2,l}$ in the main model to zeros.

Alternative Model 1 (static). *The labs set $B_{ijt}(\Omega_t, \mathbf{a}_{it}) = 0$.*

The second alternative model is the Gittins index, a seminal index in theoretical literature (Gittins & Jones, 1979; Gittins, 1979). The Gittins index prescribes optimal actions for the standard infinite-horizon

²⁴If $j = 1$, $t'_{i,t} = t$.

discounted MAB discussed in Section 1.4.2. Equation (1.1) was different, so under this model the labs would be using the Gittins index without its optimality guarantee. Computing the exact Gittins index is difficult. I therefore use Brezzi & Lai (2002)'s approximation to the index. This approximation explicitly includes the posterior variance of payoff $Var(\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}))$ in $B_{ijt}(\cdot)$.²⁵ As the first alternative model, this model does not have any free parameters, other than $\boldsymbol{\theta}_{Xl}$, in $B_{ijt}(\cdot)$.

Alternative Model 2 (Gittins). *The labs set $B_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}) = \psi(\cdot) \cdot \sqrt{Var(\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}))}$.*²⁶

The third alternative model is a simple UCB without discounting that closely resembles Auer et al. (2002). Compared to alternative models 1 and 2 which have no additional free parameters in $B_{ijt}(\cdot)$, this model has one free parameter $\theta_{B1,l}$ that needs to be estimated. The model is equivalent to restricting $\theta_{B2,l}$ of the main model to zero.

Alternative Model 3 (UCB). *The labs set $B_{ijt}(\Omega_t, \mathbf{a}_{lt}; \theta_{B1,l}) = a_{ijt} \sqrt{\frac{\theta_{B1,l}}{j}}$.*

The fourth alternative model relaxes how posterior variance enters $B_{ijt}(\cdot)$ compared to the Gittins index. It also has one free parameter in $B_{ijt}(\cdot)$.

Alternative Model 4 (FlexGittins). *The labs set*

$$B_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}) = \theta_{B1,l} \cdot \psi(\cdot) \cdot \sqrt{Var(\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}))}.$$

The last alternative model adds the term for time-discounting to the FlexGittins model. It has two free parameters, the same as the main model.

Alternative Model 5 (FlexGittins+D). *The labs set*

$$B_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}) = \theta_{B1,l} \cdot \psi(\cdot) \cdot \sqrt{Var(\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl}))} + a_{ijt} \cdot \theta_{B2,l} \cdot (t - t'_{i,t}).$$

²⁵

$$Var(\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl})) = \int (\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) - \int \pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t))^2 d\tilde{F}_t(p_{ijt}|\Omega_t). \quad (1.8)$$

²⁶The function $\psi(\cdot)$ is defined as

$$\psi(s) = \begin{cases} \sqrt{s/2} & \text{if } s \leq 0.2 \\ 0.49 - 0.11s^{-1/2} & \text{if } 0.2 < s \leq 1 \\ 0.63 - 0.26s^{-1/2} & \text{if } 1 < s \leq 5 \\ 0.77 - 0.58s^{-1/2} & \text{if } 5 < s \leq 15 \\ \{2\log(s) - \log(\log(s)) - \log(16\pi)\}^{-1/2} & \text{if } s > 15, \end{cases} \quad (1.9)$$

where $s = \frac{Var(p_{ijt}|\Omega_t)}{-\ln(\beta)E(p_{ijt}|\Omega_t)(1-E(p_{ijt}|\Omega_t))}$. I set the discount factor $\beta = 0.95$.

1.5 Model Fitting

I developed a two-step method to estimate and validate the model. In the first step, I estimate the free parameters in each model by maximizing the likelihood of the observed allocation decisions. In the second step, I use the estimated parameters to forward simulate the labs' entire history of input allocation and output. I compare the patterns of input allocated and output from the simulated data to those from the actual data to determine model fit.

1.5.1 Estimation

In this first step, I estimate the free parameters θ_l in each model by maximizing the likelihood of the observed allocation decisions. Our model assumes that the labs chose actions based on the approximate value function; therefore the observed actions $\mathbf{a}_{l1}^o, \mathbf{a}_{l2}^o, \dots, \mathbf{a}_{lT}^o$ are the solutions $\mathbf{a}_{l1}^{A*}, \mathbf{a}_{l2}^{A*}, \dots, \mathbf{a}_{lT}^{A*}$ to equation (1.4). We can use the likelihood for the observed allocations $P(\mathbf{a}_{lt}^o; \theta_l)$ to estimate θ_l .

To form $P(\mathbf{a}_{lt}^o; \theta_l)$, I start by rewriting the allocation problem in equation (1.4) as follows

$$\begin{aligned} \mathbf{a}_{lt}^o = \operatorname{argmax}_{\mathbf{a}_{lt}} \quad & \sum_{j_t \in C_{lt}} V_{ijt}^A(\Omega_t, a_{ijt}; \theta_l) + \varepsilon_{it}, \\ \text{subject to} \quad & \sum_{j_t \in C_{lt}} a_{ijt} = n_{lt}. \end{aligned} \tag{1.10}$$

I replaced $\mathbf{a}_{lt}^{A*}$ with $\mathbf{a}_{lt}^o$ in equation (1.10) to reflect that the observed allocations are the optimal actions based on the approximate value function. I introduced an additive error term ε_{it} . I assume all $\varepsilon \stackrel{iid}{\sim}$ Type I Extreme Value and capture the unobservables in decision-making. This will help me to write $P(\mathbf{a}_{lt}^o; \theta_l)$ as a closed-form function of θ_l . I also made some very intuitive functional form assumptions about the payoff function $\pi_{ijt}(\cdot)$, so that $V_{ijt}^A(\cdot)$ is only a function of a_{ijt} , and not a function of the full $\mathbf{a}_{lt}$ vector. These assumptions allow me to replace $V_{ijt}^A(\Omega_t, \mathbf{a}_{ijt}; \theta_l)$ with $V_{ijt}^A(\Omega_t, a_{ijt}; \theta_l)$. The functional form assumptions also guarantee that the second constraint of equation (1.4) holds at the solution, allowing me to drop the second constraint for equation (1.10). See Appendix A.3.1 for details of the functional form assumptions and derivations.

The solution $\mathbf{a}_{lt}^o$ to equation (1.10) is equivalent to the following index rule: one first computes index $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l) + \varepsilon_{it}$ for each project-trial in C_{lt} and then allocates to the n_{lt} trials with the highest index values.

For computational tractability, I assume the lab made iid decisions whether to allocate each trial, using a threshold rule. The lab allocated a trial whenever the trial's index value was greater than the threshold on that day. Doing so allows me to express the likelihood $P(a_{ijt}^o; \theta_l)$ with a simple closed form. Let $V_{lt}^{n_{lt}}(\theta_l)$ denote the n_{lt} th highest value of $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l)$ on day t . Define the threshold value which is equal to $V_{lt}^{n_{lt}}(\theta_l)$ plus an error term ε_{lt} .²⁷

$$threshold_{lt} = V_{lt}^{n_{lt}}(\theta_l) + \varepsilon_{lt}. \quad (1.11)$$

The likelihood of observing $a_{ijt}^o = 1$ is equal to the likelihood of $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l) + \varepsilon_{it}$ being greater than this threshold, so the likelihood $P(a_{ijt}^o = 1; \theta_l)$ is

$$\begin{aligned} P(a_{ijt}^o = 1; \theta_l) &= P(V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l) + \varepsilon_{it} > V_{lt}^{n_{lt}}(\theta_l) + \varepsilon_{lt}) \\ &= \frac{\exp(V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l))}{\exp(V_{ijt}^A(\Omega_t, a_{ijt} = 1; \theta_l)) + \exp(V_{lt}^{n_{lt}}(\theta_l))}. \end{aligned} \quad (1.12)$$

The total likelihood function sums over the log likelihood of the observed action for each project-trial in choice sets $C_{l1}, C_{l2}, \dots, C_{lT}$. These likelihoods include $P(a_{ijt}^o = 1; \theta_l)$ for trials actually allocated and $1 - P(a_{ijt}^o = 1; \theta_l)$ for trials in those choice sets but which were *not* allocated:

$$\theta_l^* = \underset{\theta_l}{\operatorname{argmax}} \sum_t^T \left(\underbrace{\sum_{j_i \in C_{lt}, a_{ijt}^o = 1} \log(P(a_{ijt}^o = 1; \theta_l))}_{\text{actual allocations}} + \underbrace{\sum_{j_i \in C_{lt}, a_{ijt}^o = 0} \log(1 - P(a_{ijt}^o = 1; \theta_l))}_{\text{actual nonallocations}} \right). \quad (1.13)$$

One can then estimate θ_l by maximizing the above likelihood function. Due to the simplicity of computing index approximations, estimation is feasible even though the choice sets contain millions of possible actions over thousands of days. A further trick to reduce computational burden is to compute in each iteration the log likelihoods for a random sample of the possible allocations in the choice sets, rather than for the full choice sets. The number of possible allocations in a full choice set could be large because n_{lt} could be large. Recall that the mean of n_{lt} for the four largest labs is 35. When $n_{lt} = 35$, the $(j+1)$ th, $(j+2)$ th, ..., $(j+35)$ th trials of every project in the lab's portfolio are in the full choice set. See Appendix A.3.2 for how I sample from the choice sets.

The estimated $\hat{\theta}_l$ includes two kinds of parameters. $\hat{\theta}_{Bl}$ are the free parameters in the exploration bonus

²⁷Note that this value is not necessarily the n_{lt} th highest index value on day t due to the iid error terms. In principle, one can obtain the n_{lt} th highest index value on day t with simulation draws of ε 's and use that as the threshold. Doing so would be computationally more challenging than my approach.

term $B_{ijt}(\cdot)$, if there are any. $\hat{\theta}_{Xl}$ are welfare weights that enter into the payoff function $\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \theta_{Xl})$. I let θ_{Xl} have eight parameters. Seven of these parameters correspond to the NIH evaluation metrics discussed in Section 1.3.1. For example, $\theta_{biomed,l}$ corresponds to $biomed_i$ and captures the amount of welfare lab l receives when it publishes a biomedically important structure. One additional parameter $\theta_{quant,l}$ captures the baseline amount of welfare lab l receives per unique structure. See Appendix A.3.1 for the specification of the payoff function $\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \theta_{Xl})$.

Identification of model parameters is based on revealed preferences and is very intuitive. Let us consider $\theta_{biomed,l}$. When we have a large positive $\theta_{biomed,l}$, it would increase the posterior expected payoff of biomedically important projects and increase V_{ijt}^A of those projects relative to $V_{lt}^{n_{lt}}$, which is based on the distribution of V_{ijt}^A of all project-trials in the lab's choice sets. In contrast, a large positive $\theta_{biomed,l}$ would decrease V_{ijt}^A for non-biomedically important projects relative to $V_{lt}^{n_{lt}}$. A large positive $\theta_{biomed,l}$ therefore increases the likelihood of allocating to biomedically important projects and the likelihood of not allocating to non-biomedically important projects. Correspondingly, if the lab actually allocated a high proportion of trials to biomedically important projects, such a data pattern will lead to a large positive estimate of $\hat{\theta}_{biomed,l}$. Variation in the binary variable $biomed_i$ in the choice sets therefore identifies $\theta_{biomed,l}$. Identification of other parameters follows the same intuition.

1.5.2 Simulation

In this second step, I use the estimated parameters of each model to forward simulate the labs' entire history of input allocation and output. Doing so would allow me to compare the patterns of input allocated and the output from the simulated data to that from the actual data to determine model fit.

I initialize each simulation with the following conditions.

- **Data generating process (DGP)**

We need to have a model of the “true” probability of success of a trial to simulate the outcomes of counterfactually allocated trials. I denote this model by F^* . This model is different from the posterior $\tilde{F}_t$ in Section 1.3.3. $\tilde{F}_t$ captures how the labs formed posterior beliefs. It does not have to produce an unbiased estimate of the true probability of success of a trial, while F^* has to. As such, F^* deviates from $\tilde{F}_t$ in several ways to correct the potential bias of and improve upon the machine learning models the labs described in their published journal articles. See Appendix A.2.1 for these implementation details.

F^* is trained on Ω_{T+1} , the full information set consisting of all trial outcomes and characteristics in the entire data.

- **Project space**

F^* may do a poor job imputing the probabilities of success of trials on projects the labs never attempted and are far out in the project space, so I restrict the project space to the projects the labs actually attempted. Applying this restriction still leaves the labs ample room to explore and/or exploit projects as only a tiny fraction of projects ever attempted successfully produced publications.

- **Allocation model**

This sets the allocation model and its corresponding parameters θ'_{Xl} and θ'_{Bl} during the simulation. To validate a model, θ'_{Xl} and θ'_{Bl} are equal to the estimated parameters $\hat{\theta}_{Xl}$ and $\hat{\theta}_{Bl}$ under the model.

- **Prior allocations and output**

Allocations and outcomes before 2005 are used as prior data and not simulated.

I then do the following to forward simulate the labs' entire history of input allocation and output since 2005.

Algorithm 1: Forward simulation of trial allocations and outcomes

For day t in 2005–2015,

- **Update posterior beliefs** Counterfactual information set Ω'_t consists of the actual and simulated actions and outcomes observed before t . Refit model of the lab's posterior belief $\tilde{F}'_t(\Omega'_t)$.
- **Make allocations** Form choice set C'_{lt} . Compute $\widehat{V}^A'_{ijt}$ of each choice in the choice set based on $\tilde{F}'_t(p'_{ijt}|\Omega'_t)$, θ'_{Xl} and θ'_{Bl} . Draw iid ε'_{it} from a Type-I Extreme Value distribution. Allocate to n_{lt} trials with the largest $\widehat{V}^A'_{ijt} + \varepsilon'_{it}$ values.
- **Simulate outcomes** Generate the true trial success probability $p^*_{ijt}|F^*(\Omega_{T+1})$ for the allocated trials. Draw outcomes $Y'_{ijt} \sim \text{Bernoulli}(\hat{p}^*_{ijt})$ for these trials.
- **Update information set** Add to the information set actions and outcomes observed in this period. Go to the next period.

End

1.6 Model Fitting Results

I find the main model fits the data extremely well and captures the labs' decision-making. During the maximum likelihood step, the main model by far has the smallest magnitude of log likelihood at convergence among the many alternative models I tested. It fits the data better with no or few additional parameters as compared to the alternatives. During the simulation step, the main model generates input allocation patterns and output that are very similar to those in the data. For all labs, the deviations of the simulated output from the actual output are within 10%. The alternative models fail to generate patterns matching the data as closely.

Based on the estimated parameters of this well-fitting model, I am able to reject the hypothesis that the labs did not explore at 95% confidence for all labs.

1.6.1 Estimation Results

Table 1.2 shows that the main model fits the data much better with no or few additional parameters than the alternatives through a comparison of log likelihoods. With the same number of parameters, the main model UCB+D has a magnitude of log likelihood that is only half the size of that of the FlexGittins+D model. In fact, the fit of the simple UCB model without discounting is already remarkable. With fewer parameters, the simple UCB model has a magnitude of log likelihood that is similar to that of the FlexGittins+D model. This table shows results for one large lab. Appendix Tables A.4.1–A.4.3 show results for the other large labs which are qualitatively similar. Among those labs, the main model has a magnitude of log likelihood that is 28% to 41% smaller than the FlexGittins+D model.

Table 1.2 also shows that the main model fits better through a comparison of the predicted likelihoods of the observed allocation decisions. It shows that, among the project-trials the lab actually allocated to, the main model on average predicts a 83.7% likelihood of allocating to each project-trial. Among the project-trials the lab did *not* actually allocate to, the main model on average predicts a 99.9% likelihood of *not* allocating to each project-trial. These predicted likelihoods are by far the highest among the models fitted. Appendix Tables A.4.1–A.4.3 show results for the other large labs which are qualitatively similar.

As a further robustness check, I examine the out-of-sample fit of the main model. To do so, I first fit the main model using only observed allocation decisions in odd years. I then use the estimated parameters to compute the average predicted likelihoods and log likelihoods in-sample and out-of-sample. I compute the in-sample results using decisions in odd years which I used to fit the model. I compute the out-of-sample results

Table 1.2: Comparison of likelihoods of different models

Model	Free parameters in $B_{ijt}(\cdot)$	Log likelihood	Avg $\hat{P}(a_{ijt}^o = 1; \theta_l)$ actual allocation	Avg $\hat{P}(a_{ijt}^o = 0; \theta_l)$ actual nonallocation
Static	0	-613,630	0.578	0.993
Gittins	0	-430,746	0.608	0.996
UCB	1	-204,228	0.763	0.998
FlexGittins	1	-412,112	0.614	0.996
FlexGittins+D	2	-230,048	0.710	0.998
UCB+D	2	-119,861	0.837	0.999

Note: I separately estimate each model during 2005–2008 and during 2009–2015 due to the change in the NIH’s preferences in 2009. Estimates are in Appendix Table A.4.5. The total log likelihood of each model adds up the log likelihoods from the two periods. Average $\hat{P}(a_{ijt}^o = 1; \theta_l)$ for actual allocation sums over the predicted likelihoods of all actually allocated trials in the two periods and divides the sum by the number of actually allocated trials. Average $\hat{P}(a_{ijt}^o = 0; \theta_l)$ for actual nonallocation sums over the predicted likelihoods of all actually not allocated trials in the choice sets and divides the sum by the number of actually not allocated trials. Estimation uses data between 2005 and 2015 from NESG, one of the four large labs. The number of trials actually allocated was 109,738. I reduced the sizes of the choice sets with random sampling (see Appendix A.3.2). The number of trials in the choice sets after random sampling is 32,410,947. Estimation results from the three other large labs are qualitatively similar (see Appendix Tables A.4.1–A.4.3).

Table 1.3: Estimates of main parameters of interest in UCB+D model

Parameter	2005–2008	2009–2015
	(1)	(2)
θ_{B1}	158.28 [156.35, 160.13]	119.52 [118.15, 121.41]
θ_{B2}	-2.28 [-2.23, -2.30]	-4.71 [-4.66, -4.73]
θ_{biomed}	21.50 [21.32, 21.72]	52.86 [52.69, 52.93]

Note: Table displays the estimates of the main parameters of interest in the UCB+D model for NESG, one of the four large labs. See Appendix Table A.4.5 for the full estimates in different models for the lab. See Appendix Table A.4.6 for the estimates of the main parameters of interest in the UCB+D model for other labs. The column with the header “2005–2008” uses data between 2005 and 2008. The number of trials actually allocated was 59,261. After random sampling (see Appendix A.3.2), the number of trials in the choice sets is 5,628,158. The columns with the header “2009–2015” uses data between 2009 and 2015. The number of trials actually allocated was 50,477. After random sampling, the number of trials in the choice sets is 26,782,789. 95% confidence intervals are computed using the MCMC approach in Chernozhukov & Hong (2003) and are shown in brackets. These confidence intervals are almost identical to those computed using Procedure 1 of Chen et al. (2018).

using decisions in even years. Appendix Table A.4.4 shows a comparison of the in-sample and out-of-sample results and suggests that the out-of-sample fit of the main model is quite similar to the in-sample fit for all labs.

Table 1.3 shows estimates of main parameters of interest in the main model. As these parameters reflect the lab’s preferences, the estimates are not in dollar values and are therefore hard to interpret on their own. Still, one can make interpretations by comparing the estimates with zero or with each other in different periods. The table shows results for one large lab. Appendix Table A.4.6 shows results for the other large labs which are qualitatively similar.

Estimates in Table 1.3 reject the hypothesis that the lab did not explore. Recall that θ_{B1} captures the amount of exploration under the main model. If the lab did not explore, $\theta_{B1} = 0$. Estimates of θ_{B1} in Table 1.3 are large and significantly different from zero at 95% confidence level, allowing us to reject the hypothesis that the lab did not explore. For all other large labs, we can similarly reject no exploration at 95% confidence.

Estimates in Table 1.3 also show that the lab discounted older projects in making allocations. Discounting

Table 1.4: Comparison of simulated outcomes of different models

Model	Projects attempted	Unique publications	Citations	Downloads (millions)
Static	14,175	597	1,787	14.1
Gittins	14,980	631	1,898	14.9
UCB	59,164	1,052	3,236	24.5
FlexGittins	15,005	621	1,830	14.6
FlexGittins+D	17,382	638	1,892	15.0
UCB+D	59,947	1,097	3,376	25.6
Actual	59,953	1,053	3,502	24.5 [†]

Note: Each simulation uses $\hat{\theta}_l$ from parameter estimates of the corresponding model. See Table A.4.5 for those estimates. Results are averaged from three simulations of each model. Actual outcomes are shown in the last row. Simulation results are for NESG, one of the four large labs. Simulation results from the three other large labs are qualitatively similar (see Appendix Tables A.4.7). [†]The download data on actually published projects are between 2007 and 2013, so actual five-year downloads may not be available for some projects. I predict five-year downloads for the actually published projects using the model described in Appendix A.2.3.

of older projects proxies for many unobserved factors in the allocation process. For example, if there were personnel changes in the lab over time and there were project-specific learning that stayed with individual researchers that frequently left, the lab would be less inclined to attempt projects that it had attempted in the distant past. Recall that a negative θ_{B2} captures discounting under the main model. If the lab did not discount older projects, $\theta_{B2} \geq 0$. Estimates of θ_{B2} are negative and significantly different from zero at 95%, allowing us to reject no discounting. For all other large labs, we can similarly reject no discounting at 95% confidence.

As another robustness check, I compare the estimates for θ_{biomed} during 2005–2008 and during 2009–2015. Recall that the NIH boosted its preference for biomedically important projects in 2009. As expected, the estimate during 2009–2015 is significantly larger than the estimate during 2005–2008 as their 95% confidence intervals do not overlap. This relationship holds for estimates from all other large labs.

We have shown the main model fits better than many alternatives. However, it is difficult to jump to the conclusion that the model correctly captures the actual allocation decisions. To investigate whether this model could generate patterns similar to those in the actual data, I use simulations.

1.6.2 Simulation Results

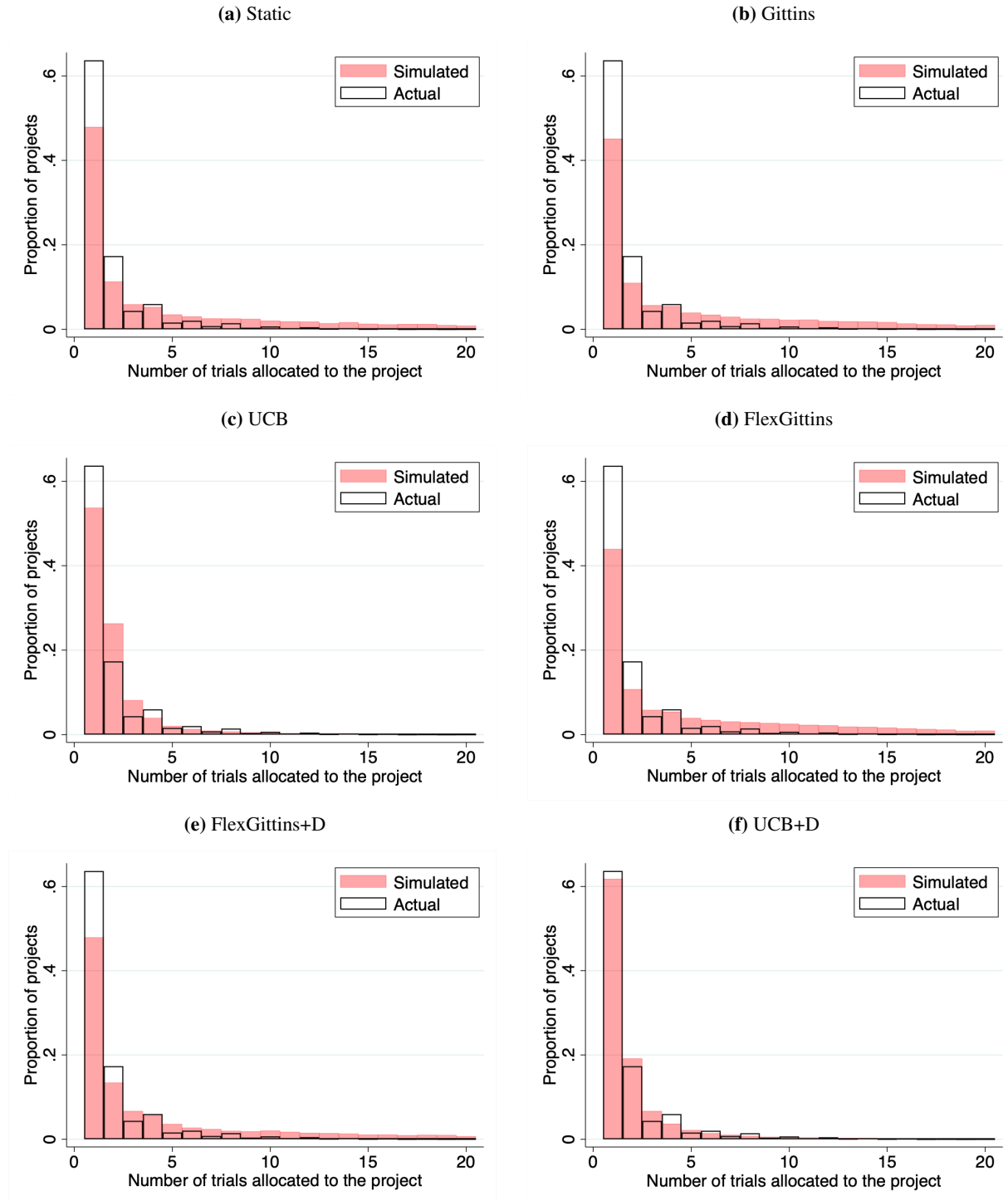
Table 1.4 shows simulated outcomes using estimated model parameters for one large lab. Results for other labs are in Appendix Table A.4.7 and are qualitatively similar.

The first column of Table 1.4 shows that the simulated number of projects attempted under the main model matches the actual number. The other models, in contrast, fall short by various degrees. Consistent with our intuition, the lab does not explore under the static model so the number of projects attempted is small at 14,175. The models based on the Gittins index encourage exploration by adding a positive exploration bonus so the numbers of projects attempted are slightly larger. The two UCB models explore extensively and the numbers of projects attempted under the two models are much closer to the actual. Figure 1.8 shows a further breakdown of the simulated input allocation across projects, plotting the distribution of the number of trials per attempted project under each model. One immediately sees that the simulated distribution of the main model fits the actual distribution the best.

The second column of Table 1.4 shows that the simulated number of unique publications under the main model matches the actual number. Figure 1.9 shows a further breakdown, plotting the simulated number of publications by year under each model.²⁸ The first panel of Figure 1.9 shows that output quantity under the static model peaks in 2006, the year immediately after the start of the simulation, then gradually declines. This is consistent with static maximization that the lab immediately allocates input to projects that it believes to be the most likely to pay off. In later years the lab exhausts such projects as it does not explore many new projects. The models based on the Gittins index produce slightly more output than the static model, but the output patterns are qualitatively similar. The two UCB models in contrast maintain steady output over the years because extensive exploration of new projects under the two models replenishes the lab's portfolio. The two models generate output patterns that are very similar to the actual output patterns.

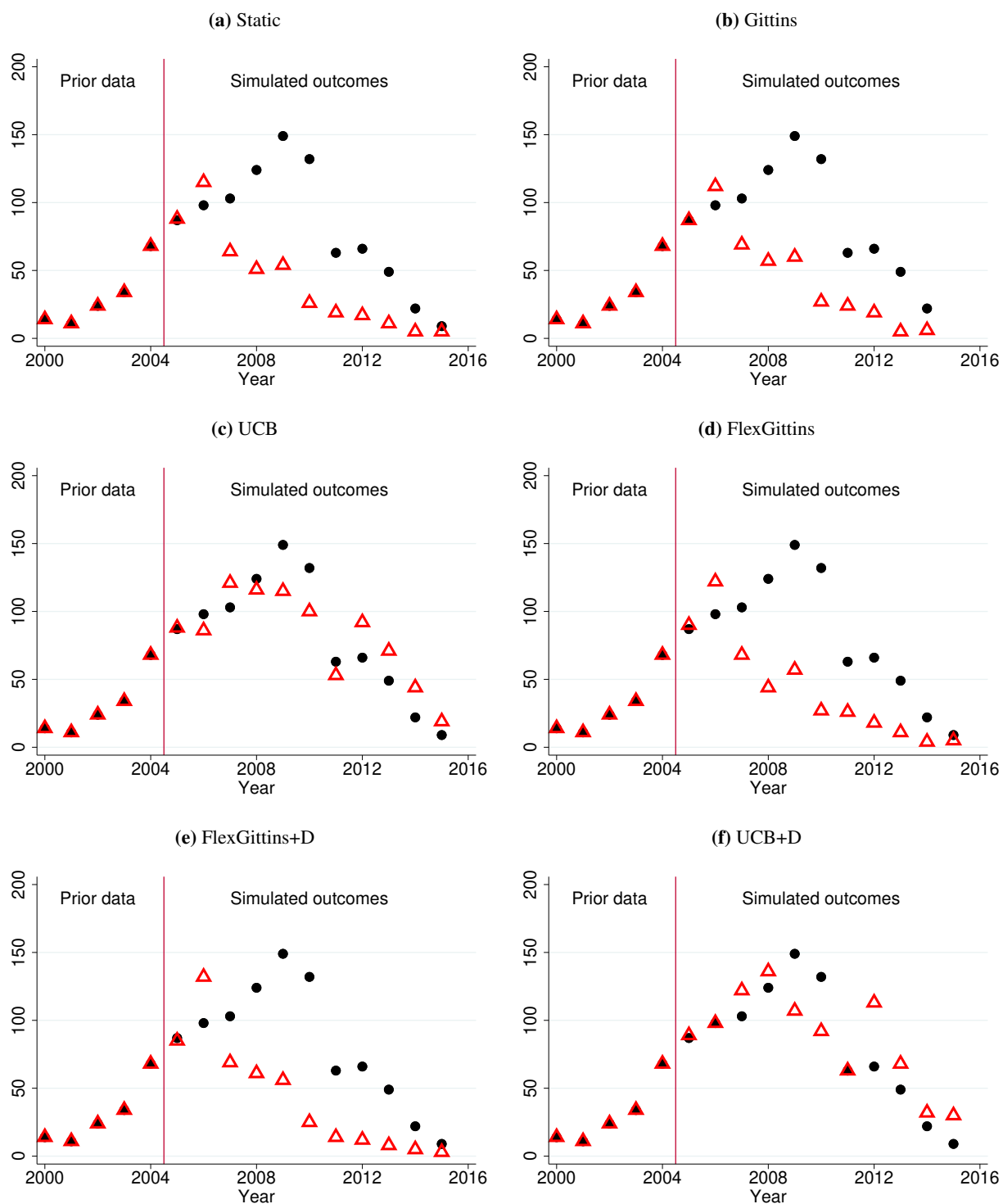
²⁸The actual series has a hump shape. Before 2009, the lab slowly built up capacity. In 2009, the lab became aware that the grant program would end in 2015 and started to slowly dismantle capacity.

Figure 1.8: Simulated distributions of input allocation across projects



Note: Each panel shows two distributions. The outlined bars show the actual distribution of the number of trials across projects at NESG. The pink unoutlined bars show the simulated distribution under the model. Each distribution is based on a single simulation of the corresponding model.

Figure 1.9: Simulated number of unique publications under different models



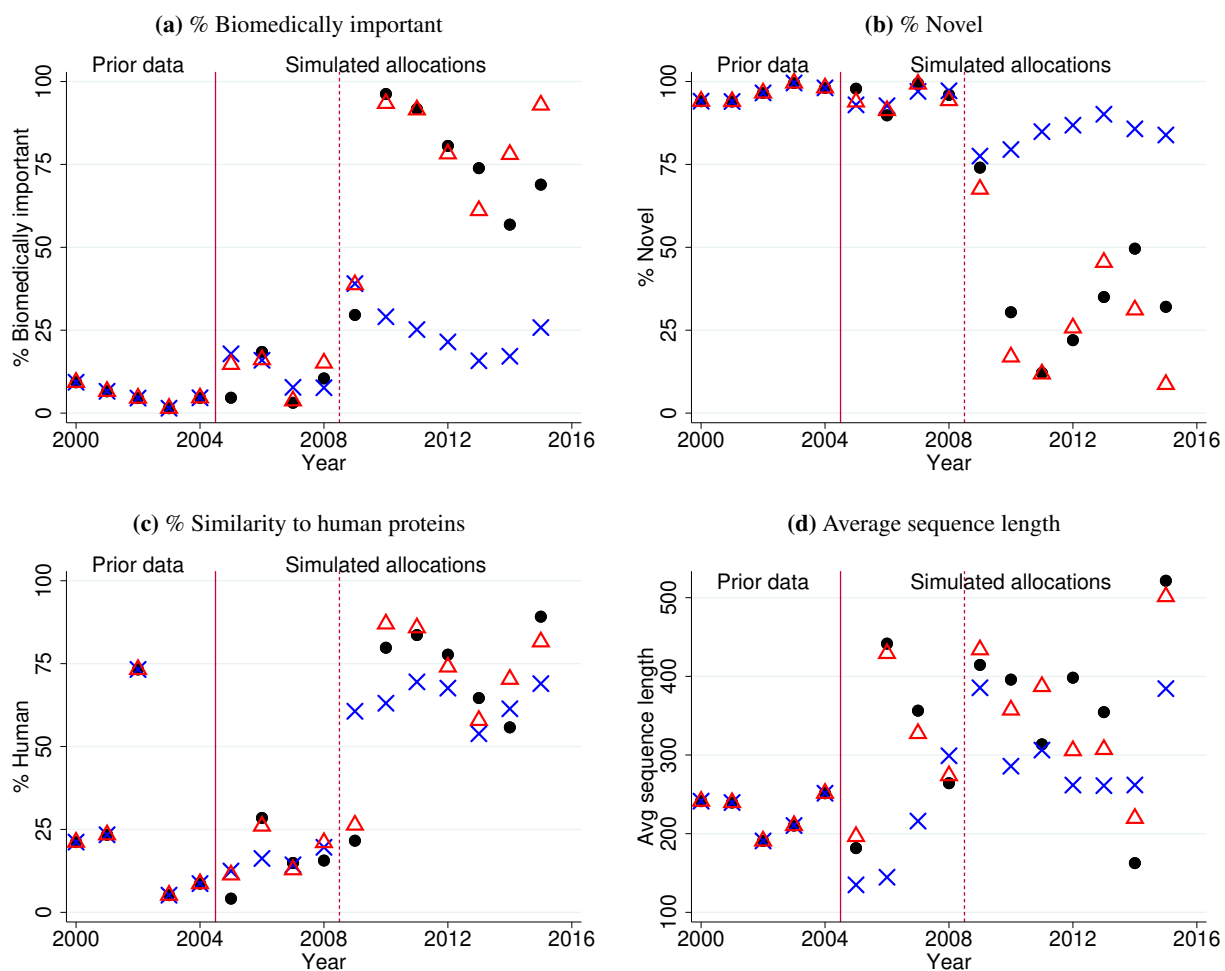
Note: Black dots represent actual numbers of publications on unique molecules at NESG. Red triangles represent simulated numbers of publications on unique molecules. Each series is based on a single simulation of the corresponding model.

The third and fourth columns of Table 1.4 show that the simulated numbers of citations and downloads under the main model match the actual numbers. These results suggest that the kinds of projects the lab allocates input to under the model are similar to the kinds of projects the lab actually allocated input to. To investigate further, Figure 1.10 compares the characteristics of trials actually allocated with those of trials allocated in simulations. The first panel displays the proportions of trials on projects that the lab cited as biomedically important. The main model produces allocation patterns very similar to the actual allocation patterns. Starting in 2009, the lab drastically shifted towards attempting more biomedically important projects, consistent with the change in NIH preferences in that year. The model captures this drastic shift almost perfectly, indicating that the lab changed the mix of project characteristics in its portfolio mostly through introducing new projects with the desired characteristics. In contrast, the increase is small for the static model around 2009 as the lab would rarely draw new projects under this model and only would change the mix of project characteristics by reallocating input within its existing portfolio. The second panel and the third panel show two more project characteristics that the NIH evaluated the lab's progress on. The fourth panel shows one project characteristic that the NIH did not evaluate on. Very similar to the first panel, these panels show that the main model produces allocation patterns very similar to the actual allocation patterns.

Results from other labs in Appendix Table A.4.7 are very similar to those in Table 1.4. For all four large labs, the deviations of the simulated outcomes under the main model from the actual outcomes are within 10%.

As a further robustness check, I examine the out-of-sample simulation results of the main model. To do so, I first fit the main model using only observed allocation decisions in odd years. I then use the estimated parameters to simulate the lab's entire input allocation history and output, in both odd and even years. As I do not use the observed decisions in even years to fit the model for this exercise, I deem these simulated results out-of-sample. Appendix Table A.4.8 shows a comparison of the in-sample and out-of-sample simulation results and suggests that the out-of-sample results are quite similar to the in-sample results for all labs.

Figure 1.10: Characteristics of trials allocated in simulations, static model versus UCB+D model



Note: Black dots represent characteristics of actually allocated trials at NESG. Blue crosses represent those of trials allocated under the static model. Red triangles represent those of trials allocated under the UCB+D model. Each series is based on a single simulation of the corresponding model.

I therefore claim with reasonable confidence that the main model captures the labs' actual allocation decision-making. Although we can never prove a model to be true because the scientific method only allows us to falsify untrue ones, we cannot reject the main model based on simulation results. Simulation results have shown that the main model could produce allocation patterns and outcomes that align well with the actual allocation patterns and outcomes.

1.7 Counterfactual—The Impact of Exploration on Productivity

With a well-fitting model of how the labs made the exploitation-exploration tradeoff, I study the effect of exploration on the labs' productivity through counterfactual simulations. Doing so involves simulating a counterfactual where the labs do not explore and comparing the outcomes with the baseline outcomes from the main model.

The simulation procedure for this counterfactual is largely identical to that for model validation in Section 1.5.2, with only differences in setting θ'_{Xl} and θ'_{Bl} . To simulate this counterfactual, I set the welfare weights θ'_{Xl} to $\hat{\theta}_{Xl,UCB+D}$, the ones recovered from estimating the main model. Intuitively, doing so maintains the labs' estimated preferences for project characteristics and studies how changing only the labs' model of exploration impacts the output. I set θ'_{Bl} to zeros to mimic no exploration. Intuitively, setting θ'_{Bl} to zeros makes the exploration bonus term $B_{ijt}(\cdot)$ zero for all trials in the labs' choice sets. The labs thus make allocations as if they follow the static model of decision-making—allocating only trials that they believe would have the highest posterior expected payoff.

Table 1.5 displays the counterfactual results for one large lab. Appendix Table A.4.9 shows results for other labs. Those results are qualitatively similar.

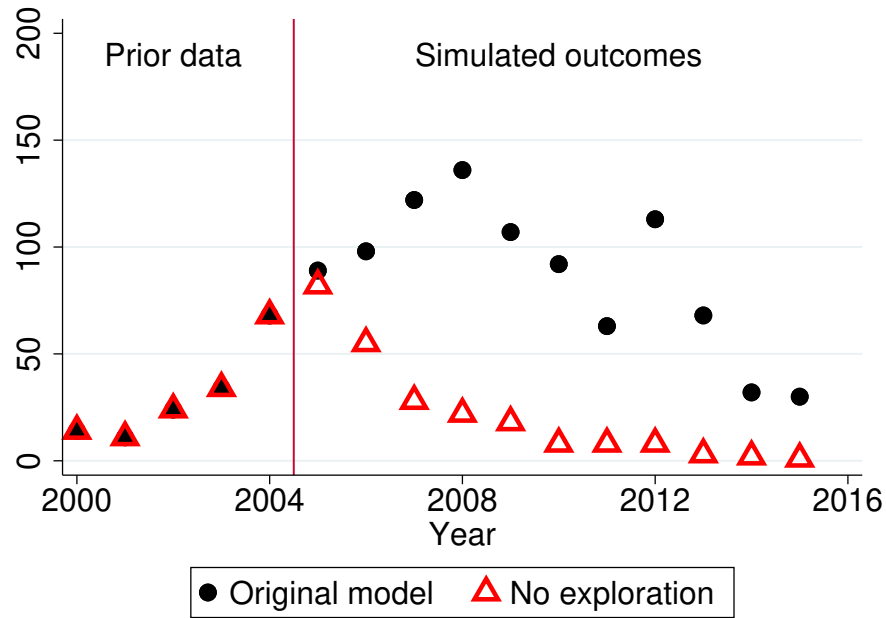
Based on results in the above tables, I find exploration was extensive and had a large impact on the labs' productivity. For the lab displayed in Table 1.5, no exploration would result in the lab attempting 79% fewer projects, producing 64% fewer unique publications and 65% fewer citations. Summing up results from all four large labs, no exploration would result in the labs attempting 82% fewer projects, producing 51% fewer unique publications and 57% fewer citations.

Applying back-of-the-envelope calculations with some simple assumptions, no exploration would result in a loss of at least \$650 to \$720 million of economic value. Suppose that the other smaller labs under the \$1.3 billion NIH program would have the same percentage decrease in output quantity had they not explored, and that the economic value of each unit of output is uniform; in that case the 51% decrease in output quantity would be equivalent to forgoing at least \$650 million of economic value, if the NIH had a nonnegative economic return to the program. Suppose instead that the other smaller labs would have the same percentage decrease in citations had they not explored, and that the economic value of each citation is uniform; in that case the 57% decrease in citations would be equivalent to forgoing at least \$720 million of economic value.

Table 1.5: Counterfactual outcomes, no exploration

Counterfactual model	Projects attempted	Unique publications	Citations	Downloads (millions)
Static	12,421	394	1,179	9.2
	(−79%)	(−64%)	(−65%)	(−64%)
Baseline model	59,947	1,097	3,376	25.6

Note: Each simulation uses $\hat{\theta}_{XI}$ from parameter estimates of the UCB+D model in columns (1) and (2) of Table A.4.5. Results are averaged from three simulations of the model. Output from the baseline model are identical to those in the second last row of Table 1.4. Parentheses show percentage differences as compared to the baseline model. Table shows results from NESG, one of the four large labs. See Appendix Table A.4.9 for results from other labs.

Figure 1.11: Simulated number of unique publications, no exploration

Note: Each series is based on a single simulation of the corresponding model for NESG. Black dots represent simulated output under the original UCB+D model. For this simulation, one sets welfare weights θ'_{XI} to the estimated welfare weights $\hat{\theta}_{XI,UCB+D}$ under the UCB+D model, and θ'_{BI} to $\hat{\theta}_{BI,UCB+D}$. This series is identical to the series represented by red triangles in Figure 1.9f. Red triangles represent simulated output under no exploration. For this simulation, one also sets welfare weights θ'_{XI} to the estimated welfare weights $\hat{\theta}_{XI,UCB+D}$ under the UCB+D model, but sets θ'_{BI} to zeros.

Figure 1.11 plots the simulated number of unique publications per year for one large lab under no exploration, compared to that under the original UCB+D model. The figure offers further evidence of how exploration boosted the labs' productivity. As the figure shows, output under no exploration is as high as that under the UCB+D model in 2005, the initial year simulated. In the following years, however, output under no exploration declines steadily as compared to that under the UCB+D model. Exploration enables the lab to find low-hanging fruits, which are in parts of the project space that the lab is not aware of at the beginning of 2005. In contrast, purely exploiting projects that the lab believes to be the most productive misses those low-hanging fruit. Over the years, no exploration results in substantial misallocation of resources to less productive projects.

1.8 Conclusion

Extensive theoretical literature has recognized the importance of making the exploitation-exploration tradeoff in resource allocation for innovation, yet little research has studied empirically how innovators make this tradeoff. This paper overcomes the empirical challenges with novel data and a new estimation method, and it studies how a group of large scientific labs made the exploitation-exploration tradeoff.

I find that a simple model captures the labs' decision-making extremely well. This model embodies extensive exploration and strongly resembles the simple UCB index of Auer et al. (2002). The only modification is that I added an additional variable to capture time-discounting of older projects. During the maximum likelihood step, I find this model fits the data the best with no or few additional parameters compared to the many alternative models I tested. During the simulation step, I find this model generates input allocation patterns and output that are very similar to those in the actual data. For all labs, the deviations of the simulated output from the actual output are within 10%.

Based on this well-fitting model, I find that the labs explored extensively. For all labs, I am able to reject no exploration at 95% confidence based on estimates of the model's free parameters. Counterfactual simulations show that exploration was extensive and that it had a large impact on the labs' productivity. Had the labs not explored, their output quantity would have decreased by 51%, and their citations would have decreased by 57%. The decrease is equivalent to a loss of least \$650 to \$720 million of economic value.

These results are valuable for several reasons. First, they confirm intuition from the theoretical literature about the value of exploration for resource allocation under incomplete information. Moreover, they suggest

sophistication in real-world decision-making. The labs were forward-looking and took into account the dynamic incentives to explore and acquire information, rather than being myopic and only performing static maximization. These results corroborate conventional wisdom related to resource allocation for innovation. For example, funding agencies often designate small early career grants to many young researchers who have scant track records. As researchers mature, track records play a more significant role in follow-on funding. Moreover, these answers have policy implications for the multi-billion-dollar innovation industry. If these results extend to the whole industry by any degree, exploration has a nontrivial contribution to innovation productivity through the channel of improving the resource allocation process. This research offers a new lens to look at the value of exploration of ideas.

This paper also opens up many possibilities for future research, one of them being further studies of how resource allocation policies could improve innovation productivity. Many open questions remain. For example, which policy features of funding programs make it easier for innovators to choose good allocation models? Besides the allocation model, does the design of the allocation process, such as the duration and scale, affect productivity? The answers to these questions can inform future policies. For example, when a funding agency designs early career grants, the level and duration of such grants are important considerations.

Another possibility is to apply the method developed in this paper to many other economic problems that embody the exploitation-exploration tradeoff. In principle, whenever economic agents need to make that tradeoff in dynamic resource allocation, we can use the index approximation approach to model their decisions. The approach overcomes the computational challenges in the workhorse estimation methods due to the curse of dimensionality. It allows one to recover an empirical model of decision-making in a problem with thousands of periods and millions of choices. The potential areas of application are numerous, e.g., the study of purchase decisions of experience goods, targeted marketing and product recommendations, investment decisions in new markets, and so on.

Chapter 2

The Impact of the General Data Protection Regulation on Internet Interconnection¹

2.1 Introduction

The Internet comprises thousands of independently owned, managed, and operated networks where network operators voluntarily exchange data via bilaterally-negotiated agreements (The Internet Society (2015)). The success of the Internet in creating economic surplus depends on these interconnection agreements. Hundreds of billions of dollars in transactions depend on the Internet's operation in the US alone, and these revenues have been growing rapidly.² The European Union (EU)'s General Data Protection Regulation (GDPR) serves as a landmark privacy law, regulating the collection, processing, and transfers of consumers' personal data that occur along with these transactions. Since the GDPR's approval in April 2016 and implementation in May 2018, it has inspired a wave of privacy regulation in countries such as Brazil, India, Japan, and the US (Goldberg et al., 2019). The unprecedented scale and scope of the GDPR makes it the most important privacy policy since the commercialization of the Internet in the 1990s and many hypothesize it would fundamentally change the Internet's operation and the digital economy.

¹Co-authored with Bradley Huffaker, kc claffy, and Shane Greenstein.

²From 2012 to 2017, payments for access to wireline forms of Internet access reached \$88.7 billion, growing more than 30% in those five years. Payments for access fees to wireless service reached over \$90.0 billion, an increase of 57%. In 2017, online advertising contributed \$105.9 billion in revenue to the GDP (Gross Domestic Product) among Internet Publishing and Broadcasting and Web Search Portals. That has grown 250% since 2012. The Census Bureau estimates electronic retailing at over \$545 billion for just electronic shopping and mail order houses (NAICS 4541), a growth of 65% over the same period.

This paper investigates whether the GDPR affected networks' decisions to interconnect at the internet layer. Consider this layer as analogous to the postal network. When consumers and providers of online content and services send each other "letters" containing digital data, networks ("post offices") deliver the mail.³ The GDPR restricts how and where content and service providers can collect, store, share, and monetize personal information contained in the "mail," bringing increased cost and complexity for these application firms, which may in turn impact the demand for "mail". We investigate whether the networks ("post offices") made fewer interconnections, or changed the types of interconnection agreements, post-GDPR in response to the decline in the demand for "mail" services. We will provide a precise definition of "the internet layer" in Section 2.2 and note features of the internet layer where the "postal network" analogy works less well.

Our study arrives against a backdrop of growing literature assessing the impacts of privacy regulation such as the GDPR on investment in applications. To date there is little research on the impact of these policies at the transport or internet layer of the Internet due to a lack of high quality data at these layers. Comprehensive data on the volume and types of traffic across the Internet does not exist and our various measures capture only part of the networks' interconnection activities. Still, our paper represents a rigorous first step towards an objective, data-driven analysis of the effects of the GDPR at the internet layer and adds to the body of research on the impact of online privacy regulation along this important margin.

We hypothesize that the GDPR shapes traffic and investment in connectivity at the internet layer through three different channels. First, we hypothesize that the GDPR operates as a tax on application firms and lowers the investment in applications. The GDPR raises the operational costs of online businesses that collect personal information as firms need to comply with a stringent set of obligations. The prospective enforcement of rules and the uncertainty about how enforcement operates also raise the expectations of fines and ongoing negotiation. The GDPR's restrictions on collecting and processing personal data could also hamper application firms' ability to monetize user participation, which may in turn reduce these firms' investment in content production or service provision supported by such monetization. This could lead to lower valuations for entrepreneurial startups in online commerce,⁴ a lower supply of applications, and lower

³We thank Dennis Carlton for suggesting this simple and insightful analogy as an accessible introduction to the internet layer.

⁴Jia et al. (2020, 2021) found a reduction in entrepreneurial ventures and market share of smaller firms after the implementation of the GDPR.

traffic.⁵

Second, the GDPR influences user participation.⁶ Participation reacts to rules limiting the collection, use, storage, and disposal of personal data, and limiting the resale and (re)disclosure of user data. If visitors value these privacy protections, then the GDPR may generate traffic from visits to online sites complying with the GDPR, and those visitors may engage more with the sites.⁷ The GDPR simultaneously reduces the value of those visits because it lowers the effectiveness of targeted advertising and targeted sales.⁸ The GDPR's restrictions on monetizing user participation and its impact on investment in applications supported by such monetization could also result in less user participation over time. These effects operate in opposite directions.

Moreover, we expect some of the GDPR's provisions to directly impact application firms' ability to transmit data. The GDPR's requirements on encryption, pseudonymization, and data minimization could impact the size of data being collected and transmitted.⁹ To prevent firms from simply moving personal data to a "data haven" with fewer restrictions, the GDPR restricts transfers of personal data outside the EEA (European Economic Area). This could reduce data traffic directly.

We expect negative changes in traffic generated by the application layer, if there is any, to lead to a decline in connectivity at the internet layer. This effect operates through changing the bargaining incentives of the networks when the derived demand for data exchange between networks falls. A simple bilateral bargaining model between networks, such as one in Besen et al. (2001), formalizes this intuition.¹⁰

We then dive into the empirical analysis. Our data comes from various data sources collected by the Center of Applied Internet Data Analysis (CAIDA) at the University of California, San Diego, and represents

⁵This is similar to Shiller et al. (2018) which found websites with larger proportions of visitors using ad blockers produced less content and had less traffic over time. Goldberg et al. (2019) found evidence of a decline in traffic at some existing firms, while Johnson et al. (2023) and Peukert et al. (2020) found evidence of a shift in traffic to the largest firms after the implementation of the GDPR.

⁶Miller & Tucker (2011, 2018) postulated a similar trade-off between privacy, participation and the costs of supplying services in the context of medical services.

⁷Empirical evidence on this hypothesis, however, is extremely lacking. Moreover, many provisions of the GDPR are motivated by views that these are intrinsic rights, and do not account for their consequences for online commerce. See Hoofnagle et al. (2019).

⁸This is similar to Goldfarb & Tucker (2011). Goldberg et al. (2019); Aridor et al. (2020) also hypothesized this effect.

⁹We thank an anonymous referee for suggesting mechanisms through which the GDPR impacts data flow and interconnection more directly.

¹⁰We present the derivations and comparative statics in Appendix B.1 for interested readers.

the state-of-the-art in inferring the presence of interconnection agreements and their types between networks on the world-wide scale, based on large collections of raw data on global network and IP address level topology of the Internet. Our data includes ownership information of all operating networks around the world, the number of observed agreements per network and the inferred type of each agreement. Using this network level data we can estimate the number of networks that are customers to a given networks. By combining the topology we can infer the number of interconnection points between pairs of networks with interconnection agreements on the level of IP addresses, the numerical labels assigned to unique devices connected to the Internet. We collect the datasets used in this paper quarterly, monthly or even daily. Most of the datasets go as far back as to early 2000s and are publicly accessible through CAIDA's website.¹¹

Our data, however, is not without limitations, and we explain those limitations thoroughly in Section 2.4, where we outline the collection of the data and the construction of the variables.¹² Among the limitations, we highlight that our data only captures part of the networks' activities – the formation and termination of interconnection agreements, and the types of agreements – and does not include important variables such as prices, capacity and actual data flow. Comprehensive data on those variables on the scale that we have for agreements has not been available for any academic research.

We begin by presenting descriptives which show persistent and similar growth in Internet interconnection of EEA countries versus non-EEA OECD (Organization of Economic Cooperation and Development) countries,¹³ though the levels of interconnectedness differ. We treat the GDPR's April 2016 approval and May 2018 enforcement as two cutoff dates for periods post policy treatment. We offer several reasons for this assumption and discuss them in detail in Section 2.5.

We then use a difference-in-differences approach, contrasting interconnection activities by networks owned by organizations headquartered in the EEA (treatment group) and networks owned by organizations headquartered in other countries (control group) before and after the approval and implementation of the policy. Given the wide territorial scope of the GDPR, we find it important to discuss whether it is ever possible to have a reasonable control group. We motivate our choice of control group in several different ways and discuss them thoroughly in Section 2.5. We also acknowledge the limitations of our empirical

¹¹<http://www.caida.org/data/overview/>. For more information about the data sources used in this paper, please see the data appendix (Appendix B.2).

¹²We provide additional details about data in Appendix B.2.

¹³Please see Appendix Table B.2.1 for lists of EEA countries and non-EEA OECD countries.

approach in the same section.

Contrasting changes in EEA networks' interconnection behavior before and after April 2016 and May 2018 relative to non-EEA OECD networks', we estimate zero effects across multiple measures. Networks in the EEA are similar to networks in non-EEA OECD countries in terms of the growth in the number of interconnecting parties and types of agreements reached. Networks' affiliation with the EEA also does not affect the observed numbers of IP-address-level interconnection points between each pair of interconnecting networks. We also find economically small effects of the GDPR on the entry and the number of networks that are customers of networks in EEA countries relative to non-EEA OECD countries. Overall, we discover no discernible change in EEA networks' interconnecting behavior across the measures we have. In Section 2.6, we present these results. We discuss several robustness checks to our main results in Section 6.7.

Our paper has an obvious policy implication: even stringent Internet privacy regulation that has evident negative impact at the application layer does not impact the incentive of network operators to interconnect and the short-run growth of interconnectivity. In the conclusion section of the paper, we discuss a number of possible reasons for this result.

Our paper also contributes by presenting data of unprecedented scale and scope.¹⁴ While a theoretical literature tackles questions on network operators and interconnection agreements,¹⁵ empirical research in Economics has been scant. Across the academic and policy arena, the lack of well-measured data describing the interconnectivity and traffic flow in the Internet has brought great attention, especially in issues such as net neutrality, international trade in digitally delivered goods, market power of big technology firms, and privacy regulations.¹⁶ Our data may represent a small step towards filling the data gap. We think future works should keep tackling the issue of unmet data needs. Specifically on the question of how privacy regulations impact the internet layer, if suitable data become available, future works may add additional results with

¹⁴To the best of our knowledge, the type of data used in our paper has only been used once in prior Economic literature, where D'Ignazio & Giovannetti (2009) obtained data from the London Internet Exchange (LINX) of its member networks and one type of agreement (peer-to-peer) between the members. Our data represents a significant improvement from their data as it covers virtually all operating networks in the world, a large number of agreements of both peer-to-peer and provider-to-customer types, and is publicly accessible.

¹⁵See for examples, Besen et al. (2001); Laffont et al. (2001); Choi et al. (2015).

¹⁶See discussions in Weller & Woodcock (2013); US International Trade Commission (2014); Meltzer (2014); Nicholson & McHenry (2016) for a few examples.

respect to variables such as prices, capacity and the actual levels of data flow.¹⁷

We organize the rest of the paper as follows. Section 2.2 provides the background in network interconnection. Section 2.3 provides the background in the GDPR. Section 2.4 describes the data, variable construction and limitations. Section 2.5 presents the main regression specification and explains justifications for our empirical strategy. Section 2.6 presents results across a number of measures of the impact of the GDPR on interconnection. Section 2.7 concludes.

2.2 Internet Interconnection

In Section 2.1, we use an analogy to the postal network to introduce the internet layer. We note this analogy, though useful, is not perfect. We offer a more precise definition of the internet layer in this section. The section explains the technicalities associated with the four layers of the Internet, and describes the demanders and suppliers associated with each layer and the flow of payment. We also discuss the contractual and institutional foundations behind interconnection, and argue networks can respond to policy changes by quickly changing the number or specifications of interconnection agreements.

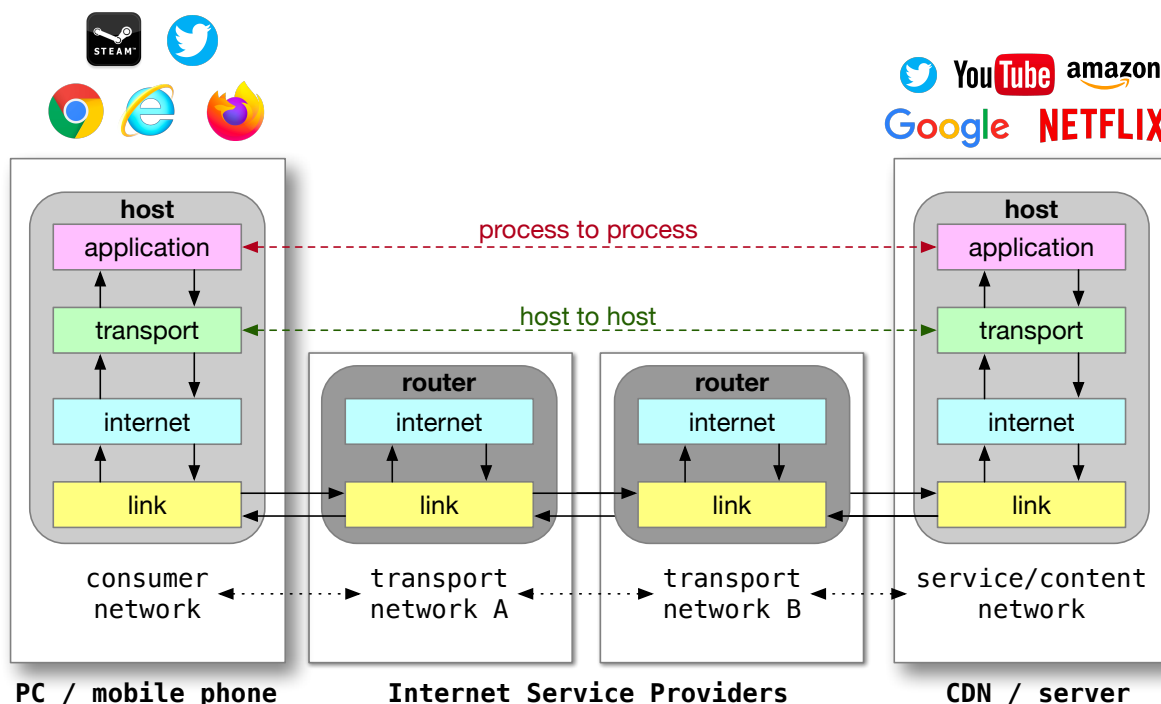
The Internet was designed with four layers of data exchange in mind: application, transport, internet, and link.¹⁸ Each layer uses a specific set of protocols, shared state, and provides a connection for higher layers. Processes in each layer communicate both with the layer directly above and below, but also across the same layer through connections provided by lower layers. Figure 2.1 provides a visual illustration of the four layers and how data exchange takes place between and across each layer.

As shown in Figure 2.1, a consumer's personal computer (PC) or smart phone has applications like web browsers and gaming platforms working at the application layer, and an operating system handling the transport, internet, and link layers. Consumers use their applications to connect, using lower layers, to other applications hosted on other devices remotely. The application layer is the layer where personal data

¹⁷We note two data sources that offer partial pictures of network capacity and the actual levels of data flow. First, a number of networks self-report capacity associated with their peering agreements at public peering points (Internet exchange points) on PeeringDB (<https://www.peeringdb.com/>). Second, Packet Clearing House (PCH) collects traffic statistics (peak, average, trough) of a number of Internet exchange points and makes it available at <https://www.pch.net/ixp/data>. The two data sources do not cover capacity or data flows associated with peering agreements at private peering facilities, which handle large volumes of traffic, nor do they cover statistics associated with the other type of agreement, the provider-to-customer agreement. As the scale and scope of the two pieces of data are quite different from those of the measures currently in this paper, we do not include them in our present analysis. We thank an anonymous referee for pointing us to the PCH data for future research.

¹⁸In an official specification document for the Internet regarding requirements for Internet hosts, the Internet Engineering Task Force (Braden, 1989) describes the four layers and specifies protocols associated with each layer.

Figure 2.1: Four layers of the Internet



is most relevant. A significant share of Internet traffic generated by applications may contain or depend on personal data due to, for examples, user authentication, third-party trackers,¹⁹ product recommendations, bots, improved search results, and spam.²⁰ Application categories may also have significant heterogeneity regarding the degrees to which they monetize personal data based on their business models. Consumers may pay service and content providers directly, and/or provide their engagement and personal information to these application firms who resell it on to advertisers targeting content, services, and ads. Metrics of engagement include ad views, ad clicks, and purchases resulting from referrals.

When applications connect, data exchange happens between the consumer and content/service providers. Application layer communication relies on lower layers of Internet infrastructure and communication protocols. The transport layer makes sure data from applications arrives correctly and reliably between end point

¹⁹A number of papers have found extensive third-party tracking activities associated with websites. Libert (2015) found nearly 90% of top one million websites by Alexa ranking leaked user data to third-parties of which the user was unlikely unaware; more than 60% websites spawned third-party cookies; and more than 80% websites loaded Javascript code from external parties onto users' computers. Englehardt & Narayanan (2016) found over 80,000 third-party trackers on the top one million websites. Karaj et al. (2018) found 71% of traffic to 1,330 highly visited websites in their data contained tracking.

²⁰By a number of estimates (Symantec (2010); Messaging Anti-Abuse Working Group (2011); Cisco Talos Intelligence Group (2020)), more than 80% of worldwide email traffic is spam.

devices. Protocols at this layer break data into packets before handing them off to the internet layer. The internet layer maintains global routing state, routing data packets to their destination address by selecting the next closest router. At this layer, the Internet can be conceptualized as a collection of different networks, each with its own set of routers and routing policies. Routers connect multiple networks and forward data packets destined either for their own networks or other networks. In Figure 2.1, the internet layer is visualized to facilitate moving data from the consumer's network to intermediary ISPs (transport networks A and B) then to the service/content network, and the service/content network may send data back the same route. Below the internet layer layer, the link layer forwards data packets to immediately adjacent (the "next hop") routers.

Some descriptive statistics at the internet layer may help the readers to contextualize this layer in relation to the application layer.²¹ By one estimate (Sandvine, 2018), the shares of Internet traffic of different application categories in 2018 were video streaming (58% downstream, 22% upstream), web (17%, 21%), gaming (8%, 3%), social (5%, 4%), marketplace (5%, 2%), file sharing (3%, 22%), messaging (2%, 8%), security (1%, 7%), storage (1%, 9%) and audio streaming (1%, 0%),²² where video streaming has experienced particularly strong growth in recent years.

In order to reach other networks, individual networks make direct connections with each other, as well as indirect connections through other networks that transport data traffic on their behalf. Consumers and service/content firms pay ISPs to connect their networks to each other. ISPs in turn pay each other where necessary to complete or enhance reachability to the rest of the Internet.

We note an important difference between the internet layer and the postal network.²³ In the postal network, there is a complete separation between the postal service and its user base: the post offices deliver the mail but do not create them. At the internet layer, that is not the case. As some application firms have grown, they began to self-supply network services. Google, Apple and Netflix have followed this expansion path. Conversely, some network firms, such as Comcast and AT&T, have expanded into applications and content. Networks can therefore be operated by consumers' and service/content firms' Internet Service Providers (ISPs), or service/content firms themselves. Though outside of the scope of this paper, it is an interesting open question how this vertical integration affects the integrated firms' responses to privacy regulations.

²¹ We thank an anonymous referee for the suggestion of incorporating these high level statistics.

²² The numbers do not sum up to 100 due to rounding.

²³ We thank an anonymous referee for pointing out this important difference and for offering an insightful discussion from which this paragraph draws on.

Network operators typically use a mix of agreements with different interconnection counterparties. As described by the Internet Society (The Internet Society, 2015), we can broadly classify these agreements as one of two types:

- Provider-to-customer (p2c) or customer-to-provider (c2p) is an agreement by which the provider network agrees to provide its customer network with connectivity to the rest of the Internet for a fee.
- Peer-to-peer (p2p) is an agreement by which two networks agree to a mutual exchange of traffic to and from their customer networks. Peering arrangements reduce the amount of traffic a network must send through its upstream transit provider network, lowering the average cost of traffic delivery. If the peers have similar negotiating power, they form a settlement-free agreement. Under an imbalance, the weaker network pays the other under a paid peering agreement.

Transition provider networks typically price p2c and c2p agreements as a metered service outside of the residential market on a per-megabit-per-second (Mbps) basis. Transit providers compete vigorously, resulting in a strong declining trend for the prices for transition from 1998 until present.²⁴ The duration of p2c and c2p agreements can be as short as one month or as long as multiple years. Due to the strong declining trend in prices, networks usually renegotiate even multiple-year agreements yearly. Customers have strong incentives to renegotiate or switch to a different provider to bring down the unit transit prices to the current market price. Networks also commonly use extremely short-term agreements, typically with no volume commits and with a duration of just one month, to fully capture the ever-decreasing market prices for transit (Norton, 2014).

p2p agreements may reduce the cost of traffic exchange even further when the volume of traffic is high. Potential peers typically negotiate p2p agreements on a case-by-case basis. Traffic volume often represents “a key determinant” of whether a peering agreement is reached as “the decision hinges upon whether or not there is sufficient value from peering to justify spending time and money” (Norton, 2014). A portion of the cost of peering involves purchasing circuits of fixed capacities between the peers at the peering point and this cost scales with the capacities of the circuits. When a network does not have a Point of Presence (POP) at the agreed peering location, it incurs additional cost to bring its traffic to the peering point. Networks incur additional cost associated with colocation, equipment, and peering ports. The split of the cost is specific to

²⁴Estimates based on a sample of US transit providers show that per Mbps transit prices averaged \$12.00 in 2008 and averaged \$0.63 in 2015 and yearly decreases between 2008 and 2015 ranged from 28% to 52% (see Norton (2014) Table 2-2).

the agreement and net payment between networks may occur, resulting in a paid peering agreement.²⁵

In many cases, setting up an interconnection does not require the deployment of additional hardware (Norton, 2014), and can be done very quickly. The two parties may simply utilize existing assets, such as configuring an existing port or purchasing circuits between their existing POPs. The process to interconnect can take as little as minutes. When the physical assets for interconnection, such as optical fibers and undersea cables, are not present, it can take substantially longer to install the hardware to interconnect, often in years. As we will discuss in more detail in later sections, our empirical analysis contrasts interconnection activities of networks in the EEA versus networks in non-EEA OECD countries and the vast majority of interconnections that we study involve both parties in developed countries. We expect the availability of physical hardware to have little constraint on the incentives to interconnect, at least in the short run that we study, and therefore networks can establish or terminate interconnections reasonably quickly in response to policy changes.

2.3 The GDPR

In this section, we provide an overview of the GDPR. We closely examine the text of the law, discuss where networks fit into the regulatory framework, and the mechanisms through which the law may impact them, presenting existing empirical evidence whenever appropriate. Combining the legal provisions and the empirical evidence with a discussion on the reaction of the popular media to the regulation, we find it easy to hypothesize significant negative impact of the regulation on networks' investment, with EEA networks harder hit. We need sound empirical evidence to support or refute this hypothesis.

Approved on April 14, 2016 and effective on May 25, 2018, the GDPR applies to most application firms and networks with EEA end customers because it applies to any organization that “processes personal data” of EEA consumers.²⁶ The GDPR defines “personal data” broadly, as any information that might identify a consumer (“data subject”).²⁷ It also defines “processing” broadly, as any operation that is performed on

²⁵A very rough estimate of the total cost of a p2p interconnection with a 10Gbps capacity at a European peering point using cross-continent transport stands at \$11,000 per month in 2014 (see Norton (2014) Table 5-1).

²⁶The GDPR was incorporated into the EEA Agreement on July 6, 2018, so its scope covers both EU member states and non-EU EEA member states (Iceland, Luxembourg and Norway). The GDPR was enforced in Iceland on July 15, 2018, in Norway on July 20, 2018, and in Luxembourg on August 20, 2018, all within three months of its enforcement in the EU. Iceland, Luxembourg and Norway collectively accounted for 1.4% of EEA population. The nationality of the consumer is not relevant, the relevant criterion is whether the person is located in the EEA (Hoofnagle et al., 2019)).

²⁷GDPR Art. 4(1). Under this definition, not only a person's name and physical addresses, but also IP addresses, cookies, and similar data are personal data.

personal data, whether or not by automated means.²⁸ The GDPR places the burden of responsibilities on organizations that determine the purposes and the means of processing of personal data (data “controllers”), while organizations that process personal data on behalf of controllers (data “processors”) also have to comply with a considerable portion of the GDPR.²⁹ Many application firms fall within the meaning of the GDPR as both data controllers and processors, while networks, routing data on behalf of application firms, fall within the meaning of the GDPR as data processors. The GDPR has a wide territorial scope. Even when a firm has no physical presence in the EEA, the GDPR applies if it is apparent that the firm envisages offering services to consumers located in the EEA.³⁰

We hypothesize that the GDPR shapes traffic and investment in connectivity at the internet layer through three different channels. First, the GDPR may operate as a tax on application firms due to compliance costs, regulatory uncertainty and threat of a fine. This may have negative effects on firms’ ability to invest in applications that generate traffic. Under the GDPR, firms need to fulfill major obligations such as keeping detailed, account-like records of their processing activities,³¹ incorporating protection into the technical design for the services with data protection “by design” and “by default,”³² developing “Data Protection Impact Assessments” (DPIA) for high-risk processing activities,³³ and so on (Hoofnagle et al., 2019). If firms fail to comply, serious violations can trigger administrative fines of up to 20 million euros or up to 4% of the total worldwide annual turnover.³⁴ Regulatory uncertainties further add to the cost. The rules, especially those in the recitals, were written with open-ended features to provide regulators with the flexibility to respond to unexpected and unanticipated issues. Firms therefore need to operate under the presumption of ongoing communications with the Data Protection Authorities over the unresolved features of the rules

²⁸GDPR Art. 4(2).

²⁹GDPR Art. 4(7), (8). In principle, if data processors violate the GDPR, the data controller will be considered responsible and liable (Hoofnagle et al., 2019)).

³⁰GDPR Recital 23.

³¹GDPR Art. 30.

³²GDPR Art. 25(1), (2). Data protection “by design” refers to measures such as pseudonymisation, which are designed to implement data-protection principles. Data protection “by default” means only personal data which are necessary for each specific purpose of the processing are processed and personal data are not made accessible without the individual’s intervention to an indefinite number of natural persons.

³³GDPR Art. 35(1). High-risk processing activities would include automated processing or profiling that leads to decisions that significantly affect people and sensitive data are processed on a large scale (Hoofnagle et al., 2019).

³⁴GDPR Art. 83(4), (5).

(Hoofnagle et al., 2019). Ernest & Young predicts the world's 500 largest companies would spend \$ 7.8 billion to comply with the GDPR.³⁵ While a report by DataGrail, a privacy management platform, estimates 74% of small- and mid-sized organizations would spend more than \$100,000 and 20% of them would spend more than \$1 million.³⁶

The GDPR may impose additional cost on application firms by hampering their ability to monetize user participation. This could further reduce application firms' investment in applications supported by such monetization and reduce traffic. Under the GDPR, consent must be freely given, specific, informed, unambiguous, and revocable³⁷ (Hoofnagle et al., 2019), preventing firms from using long and inaccessible consent processes to obtain personal data. The purpose limitation principle specifies that personal data shall be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."³⁸ This would limit application firms' ability to repurpose data in unanticipated ways. The data minimization principle specifies that personal data should be "limited to what is necessary in relation to the purposes for which they are processed" and shall be "kept in a form which permits identification of data subjects for no longer than is necessary" for those purposes.³⁹ This accounting comes with increased cost and complexity, and by intent reduces the window for monetization.

Available evidence from the empirical literature, though limited, largely supports the hypothesis that the GDPR is costly for application firms. Godinho de Matos & Adjerid (2022) studied the effectiveness of a campaign for obtaining GDPR-compliant consent for personal marketing and found such practices effective, though at additional cost to the firm to elicit such consent. Goldberg et al. (2019) found a 10% decrease in recorded e-commerce sales for a sample of EU firms after the GDPR's enforcement. Johnson et al. (2023) found that the week after the GDPR's enforcement, website use of web technology vendor fell by 15%. They also found websites were more likely to drop smaller vendors, which increased the relative concentration of the vendor market by 17%. Peukert et al. (2020) found similar effects and the magnitude of change was

³⁵Kahn, Jeremy, Stephanie Bodoni & Stefan Nicola. 2018. "It'll Cost Billions for Companies to Comply With Europe's New Data Law." Bloomberg Businessweek.

³⁶Lindsey, Nicole. 2019. "Understanding the GDPR Cost of Continuous Compliance." CPO Magazine.

³⁷GDPR Art 7(3).

³⁸GDPR Art. 5(1)(b).

³⁹GDPR Art. 5(1)(c), (e).

particularly large for websites with EU-specific top-level domains.⁴⁰ Aridor et al. (2020) found a 12.5% drop in observable consumers to a data analytics intermediary after the GDPR's enforcement and that resulted in declines in revenue from targeted ads for European travel platforms compared to their non-European counterparts. Lefrere et al. (2020) found the GDPR reduced the number of third-party cookies and tracking. However, they found no evidence that EU websites reduced content production relative to US website. This is in contrast to Shiller et al. (2018), which found websites with larger proportions of visitors using ad blockers produced less content and had less traffic over time.⁴¹

We also hypothesize that the GPDR influences traffic by changing user participation. Consumers receive many data-related rights under the GDPR, which may boost their participation and incentivize application firms to invest more. The GDPR specifies seven rights for consumers: the right 1) to access, 2) to data portability, 3) to rectify data, 4) to stop processing, 5) to object, 6) to erase data, and 7) to resist profiling and computerized decision making processes⁴² (Hoofnagle et al., 2019). If consumers value these privacy protections, then the GDPR may generate more use of the content and services that comply with the law. Empirical evidence on this, however, is lacking. To the best of our knowledge, we are not aware of academic works that have found consumers actually increase their demand for online content or services in response to better privacy protection. In contrast, Lefrere et al. (2020) found no effect of the GDPR on the amount of content that EU websites were able to publish, or the degree of average social media engagement and interaction with such content. Goldberg et al. (2019) found a large and significant 10% decline of recorded page views, visits and orders for a set of EU e-commerce firms after the GDPR became effective, suggesting the possibility of less user engagement with less personalized ad targeting and recommendations.⁴³

Given that the GDPR imposes various costs on application firms and the apparent lack of demand response from consumers to better privacy protection, one may expect the negative effects on investment to outweigh

⁴⁰A top-level domain is the last segment of a domain name. Common top-level domains include .com, .org and .us.

⁴¹Additional empirical evidence suggests content and service providers alter their behavior significantly following the implementation of the GDPR. Libert et al. (2018) found the GDPR has led to a 22% decrease in third-party cookies on a set of EU news sites (third-party cookies are information stored in browsers used for tracking and advertising, sent from sites other than the one the user is currently visiting). Degeling et al. (2019); Mohan et al. (2019) found extensive updates to websites' and cloud services' privacy policies. An exception is Iordanou et al. (2018), which found few changes in the amount of data flow associated with web tracking and in the percentage of this data flow attributed to tracking servers hosted in EU around the GDPR implementation window.

⁴²GDPR Art. 4(3), 8, 16, 17, 20, 21(1), 22.

⁴³Additional anecdotal evidence suggests consumers feel no better off under the GDPR. See for examples: Olenick, Doug. 2019. "Consumers Feel Privacy is No Safer under GDPR." SC Media. Tesserars, Lucy. 2018. "GDPR Three Months On: Most Consumers Feel no Better Off." MarketingWeek.

the positive, leading to lower overall investment in applications and lower traffic. The effect on investment may be especially pronounced for application firms located in the EEA.⁴⁴ Empirical evidence, though limited, supports this notion. Jia et al. (2020, 2021) show that the implementation of the GDPR strongly reduced venture capital investment in technology start-ups in Europe compared to their US counterparts and far away investors were more likely to respond negatively.

In addition to the GDPR's effect on investment in applications and user participation, we further expect some of the GDPR's provisions to directly impact the volume of data application firms are able to transmit. The GDPR requires that firms use encryption or pseudonymization to process personal data (data protection "by design"), which may increase data size. In contrast, the GDPR's data minimization requirement could reduce the amount of data being collected and transmitted. Moreover, to prevent data controllers from simply moving personal data to a "data haven" with fewer or no restrictions, the GDPR only allows transfers of personal data outside the EEA when the destination country or organization upholds privacy protection to a comparable level of that specified in the GDPR.⁴⁵ Application firms either bear the significant cost of achieving GDPR-level personal data protection even outside the EEA or choose to reduce the amount of data they transfer outside the EEA.

We expect that a decline in traffic generated by the application layer, if there is any, would lead to a decline in connectivity at the internet layer. This effect operates through changing the networks' bargaining incentives and gains from trade when the derived demand for data exchange between networks falls. A simple bilateral bargaining model between networks, such as one in Besen et al. (2001), formalizes this intuition.⁴⁶ The hypothesis that the GDPR's effects on the application layer may propagate to the internet layer coincides with the alarmist and negative discussion the popular media and opinion pieces have on the broad impact of the regulation.⁴⁷ As of this writing, these views continue to be the consensus. In extensive

⁴⁴The GDPR makes all EEA consumers more costly to serve, but a non-EEA application firm faces different costs for non-EEA consumers, and may not comply with the GDPR for their non-EEA consumers to lower their cost outside of the EEA. Non-EEA application firms may also choose not to comply, cutting out EEA consumers all together. We discuss in more detail the various evidence for noncompliance of non-EEA application firms and for differential impact of the GDPR at the application layer in Section 2.5, as support for our empirical strategy.

⁴⁵GDPR Art. 45-47. Adequacy status are evaluated by the European Commission. The US in general does not achieve adequacy. US-based firms may choose to participate the EU-US Privacy Shield, which requires firms to commit to a GDPR-like level of protection.

⁴⁶Appendix B.1 shows the derivations and comparative statics for interested readers.

⁴⁷A sampling from (the most credible) news sources gives a good sense of the range of concerns voiced as GDPR became binding. See for examples:

online search of news articles and editorials since the implementation of the GDPR, we have found no opinion or report to suggest any other impact on business than a costly impact, though views expressed in the news articles and editorials are often neither supported by systematic data collection, nor informed by a census of experience, and most of them stress the costs in unspecific terms. We need sound empirical works to support or refute the hypothesized impact and the uninformed discussions in the media.

2.4 Data

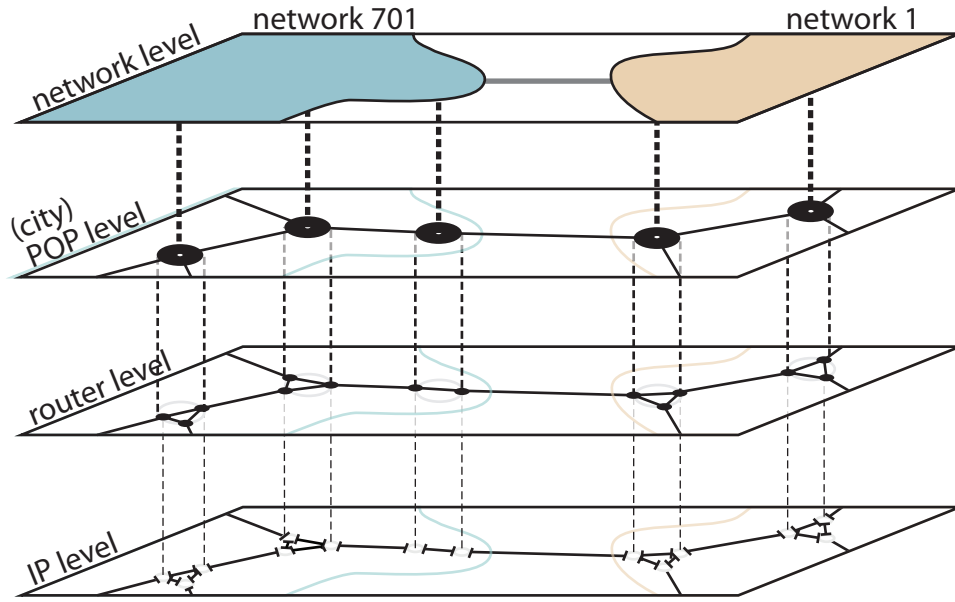
In this section, we describe our data, variable constructions and limitations. Our data comes from various data sources collected and compiled by the Center of Applied Internet Data Analysis (CAIDA) at the University of California, San Diego. Since 1998, CAIDA has been studying interconnectivity of the Internet by actively probing the Internet using its many monitors placed at various vantage points around the world. Its current flagship active measurement infrastructure, Archipelago, collects interconnectivity data on the IP-address-level from more than 200 monitors located on 6 continents in over 60 countries. CAIDA also collaborates with many organizations and compiles data collected from their monitors. Most notably, it collaborates with the Route Views Project at the University of Oregon and The Réseaux IP Européens Network Coordination Centre (RIPE NCC) in Europe to collect BGP routing tables that contain network-level interconnection paths announced across the Internet. Our main data on the network-level interconnection agreements comes from the routing tables, while our lower IP-address-level interconnection points for each agreement come from the active probes (Figure 2.2 visualizes the different levels at which we collect data and their relationships). CAIDA also gathers records of network registration information from the world's five regional Internet registries (RIRs), allowing us to identify countries or territories of organizations owning individual networks.⁴⁸

Table 2.1 provides a summary of the variables used in this paper, describing their units of observations,

-
- Bershidsky, Leonid. 2018. "Europe's Privacy Rules Are Having Unintended Consequences." Bloomberg.
Cool, Alison. 2018. "Europe's Data Protection Law Is a Big, Confusing Mess." New York Times.
Downes, Larry. 2018. "GDPR and the End of the Internet's Grand Bargain." Harvard Business Review.
Eiss, Robert. 2020. "Confusion over Data-Privacy Law Stalls Scientific Progress." Nature.
Hern, Alex. 2018. "Facebook Moves 1.5bn Users out of Reach of New European Privacy Law." Guardian.
Kostov, Nick & Sam Schechner. 2019. "GDPR Has Been a Boon for Google and Facebook." Wall Street Journal.
Satariano, Adam. 2018. "G.D.P.R., a New Privacy Law, Makes Europe World's Leading Tech Watchdog." New York Times.
Trentmann, Nina. 2018. "Companies Worry That Spending on GDPR May Not Be Over." Wall Street Journal.

⁴⁸We present a complete list of countries and territories in our sample in Appendix Table B.2.1.

Figure 2.2: Data collection at the internet layer



Notes: Our network-level interconnection agreements extracted from routing tables correspond to the topmost level in this figure. Our IP-address-level interconnection points for each agreement extracted from active probes correspond to the bottom level in this figure. Geolocating points of presence (PoP) and mapping routers to networks are challenging and open questions, therefore we do not use data on the middle levels.

frequency, sources and definitions. Table 2.2 presents summary statistics of variables described in Table 2.1. In the remainder of this section we discuss the data collection process and the limitations of the data. For additional information, please refer to the data appendix (Appendix B.2).

As shown in Table 2.1, a number of our key variables come from a dataset called *AS Relationships*. The dataset contains network-to-network level interconnection agreements extracted from routing tables contributed by Route Views and RIPE NCC. To correctly route data across the Internet, networks exchange routing and reachability information through a protocol called the Border Gateway Protocol (BGP). Each network router using the BGP protocol maintains a routing table. The table contains the connectivity information of the network and its immediate neighbors in the Internet and lists paths to particular network destinations. By placing monitors that peer directly with large networks, we can extract the full set of agreements used between the collecting networks and all visible destinations.

We then annotate the extracted agreements with algorithmically-inferred agreement types, as network operators consider the details of their business relationships as proprietary information and do not generally make them public. Our inference algorithm (Luckie et al., 2013) draws from a long literature of this type of inference including Gao (2001); Subramanian et al. (2002); Erlebach et al. (2002); Battista et al. (2003); Xia & Gao (2004); Dimitropoulos et al. (2005, 2007). It achieved over 98% accuracy of agreement type inference via direct validation with a set of network operators (Luckie et al., 2013). The algorithm succeeded in inferring 96% of the agreement types in our sample.

We compute the AS Relationships dataset monthly. We use data from January 2015 to June 2019 for our analysis. We first count the number of observed agreements each network k has in this data. The variable $numAgNtwrk_{kt}$ represents this count. We then aggregate individual agreements to the number of agreements between networks owned by each pair of countries (or territories) i and j . The variable $numAg_{ijt}$ represents this aggregate count of the number of agreements between country pairs ij . Breaking down the number of agreements between each country (or territory) pair by their agreement types, we make three variables $numProvAg_{ijt}$, $numPeerAg_{ijt}$, $numCustAg_{ijt}$ for when country (or territory) i 's networks are providers to, peers to, and customers of country (or territory) j 's networks respectively. We measure a network's centrality in the Internet by its customer cone, a commonly used measure of the number of networks that pay it directly or indirectly for transit. A network's customer cone is defined as itself and all the networks it was observed reaching following provider-to-customer agreements. Networks with larger customer cones have an especially important role in interconnecting the global Internet. The variable $NtwrkCustCone_{kt}$ represents the customer cone size of network k .

Our IP-address-level interconnection points within each agreement come from a dataset called *IPv4 Prefix-Probing*. The dataset consists of daily traceroutes from a subset of our Archipelago monitors to every announced BGP routing prefix (a prefix is a block of IP addresses) in the Internet. Each traceroute tries to reach each destination prefix and records the entire IP address-by-IP address path it takes. We then map each IP address to its network with the help of Route Views Prefix-to-AS mappings dataset (CAIDA, 2013) and bdrmapIT tool (Marder et al., 2018), identify IP pairs that form inter-network links and label the observed interconnection links by their IP addresses and network identifiers. The IPv4 Prefix-Probing dataset is available since December 2015 on a daily basis from multiple monitors, so we use data from December 2015 to June 2019. We do two aggregations. First we aggregate daily captures from multiple monitors to weekly captures of unique IP-address-to-IP-address connections. Then we aggregate individual connections

Table 2.1: Description of variables

Variable	Unit of Observation	Frequency	Description	Source	Additional Notes
numAg _{<i>ijt</i>}	ctry _{<i>t</i>} –ctry _{<i>j</i>}	monthly	The number of interconnection agreements between pairs of networks owned by the countries <i>i</i> and <i>j</i> .	AS Relationships	Agreements are available on network–network level and are aggregated to country–country level.
numProvAg _{<i>ijt</i>}	ctry _{<i>t</i>} –ctry _{<i>j</i>}	monthly	The number of interconnection agreements where country <i>i</i> ’s network is a provider to country <i>j</i> ’s network.	AS Relationships	Same as above.
numPeerAg _{<i>ijt</i>}	ctry _{<i>t</i>} –ctry _{<i>j</i>}	monthly	The number of interconnection agreements where country <i>i</i> ’s network and country <i>j</i> ’s network are peers.	AS Relationships	Same as above.
numCustAg _{<i>ijt</i>}	ctry _{<i>t</i>} –ctry _{<i>j</i>}	monthly	The number of interconnection agreements where country <i>i</i> ’s network is a customer of country <i>j</i> ’s network.	AS Relationships	Same as above. Value is identical to numProvAg _{<i>jit</i>} for ctry _{<i>j</i>} –ctry _{<i>i</i>} .
numAgIP _{<i>klt</i>}	ntwrk _{<i>k</i>} –ntwrk _{<i>l</i>}	weekly	The number of IP-address-level interconnection points between network <i>k</i> and network <i>l</i> , given <i>k</i> and <i>l</i> have an agreement.	IPv4 Prefix-Probing	Data is available daily and is aggregated to weekly.
numAgNtwrk _{<i>kt</i>}	ntwrk _{<i>k</i>}	monthly	The number of interconnection agreements network <i>k</i> has.	AS Relationships	
numNtwrk _{<i>it</i>}	ctry _{<i>i</i>}	quarterly	The number of networks country <i>i</i> owns.	AS Organizations	
NtwrkCustCone _{<i>kt</i>}	ntwrk _{<i>k</i>}	monthly	The number of networks network <i>k</i> can reach through its customer connections alone.	AS Relationships	A measure of a network’s importance in Internet routing.

Notes: In the computer science field, a network is referred to as an *Autonomous System* (AS). All data sources listed here (CAIDA (2015-2019a), CAIDA (2015-2019b), and CAIDA (2015-2019c)) are available through CAIDA’s webpage <http://www.caida.org/data/overview/>.

Table 2.2: Summary statistics

Variable	Observations	Mean	SD	Min	Max
<i>Panel A: unrectangularized variables</i>					
numAg _{ijt}	119,071	64.4	759.0	1	33,497
numProvAg _{ijt}	121,369	44.8	681.8	1	31,485
numPeerAg _{ijt}	62,241	30.8	140.1	1	4,155
numCustAg _{ijt}	121,369	44.8	681.8	1	31,485
numAgIP _{klt}	19,413,597	9.8	144.8	1	172,481
numAgNtwrk _{kt}	2,909,695	5.4	55.9	1	8,391
numNtwrk _{it}	3,597	357.3	1754.3	1	24,887
NtwrkCustCone _{kt}	2,909,695	7.8	263.3	1	37,061
<i>Panel B: rectangularized variables</i>					
numAg _{ijt}	1,085,400	7.1	252.2	0	33,497
numProvAg _{ijt}	2,160,000	2.5	161.9	0	31,485
numPeerAg _{ijt}	1,085,400	1.8	34.3	0	4,155
numCustAg _{ijt}	2,160,000	2.5	161.9	0	31,485

Notes: Panel A presents the variables with the appropriate levels of aggregation from the raw data. For numAg_{ijt}, numProvAg_{ijt}, numPeerAg_{ijt}, numCustAg_{ijt}, we also rectangularize the variables by filling in zero values for country pairs and dates with no observed agreements from our raw data and present the rectangularized variables in Panel B.

to the number of connections between each pair of networks k and l . The variable $numAgIP_{kl}$ represents the number of IP-address-level connections between networks k and l .

Although we know of no more rigorous data collection efforts of interconnection on the internet layer, we recognize that our data has limitations. First, we note that we are able to capture only part of networks' activities – the formation and termination of interconnection agreements, and the types of agreements. It is important to note that connectivity is not traffic, though there is evidence that IP address space advertised by BGP tables are strongly positively correlated with networks' self-reported traffic volume for a large set of peer-to-peer interconnections (Lodhi et al., 2014). We do not know how much traffic exchange happens across an interconnection or how that traffic has changed over time. If major changes in traffic occurred purely through existing interconnections, causing increased or decreased investment in Internet infrastructure, it would be invisible in our data.⁴⁹

Second, networks owned by organizations headquartered in a particular country or territory can have multiple points of presence (PoP) in many countries and locations within a country and a single Internet interconnection can represent multiple geographically distinct physical connections. Geolocating points of presence is a hard and an open question, so it is important to note the country subscripts of our variables indicate network ownership by organizations headquartered in those countries or territories instead of the exact physical locations of the networks. This measure is especially problematic for large global transit providers and content providers which have PoPs both within and outside the EEA. However, we note that though the relatively few large networks account for a substantial portion of global Internet traffic, the typical network is small and has limited geographic reach beyond its country of origin.⁵⁰ Throughout this paper, we use unweighted measures of the number of networks and the number of interconnections. This to some extent alleviates the concern that the imperfect measurement of locations of a few large networks drives the results.

Moreover, the number of agreements we capture, though extremely large, is a subset of all agreements. Individual routers do not maintain a full set of Internet paths, but rather a set of “best” paths for each destination based on local preferences. Networks also do not announce their peer-to-peer paths to their

⁴⁹This can happen when networks add or limit new capacity at existing interconnection points, utilizing the framework of existing agreements and the agreed terms of those agreements, or replacing old agreements by new ones with different terms.

⁵⁰For reference, if we measure the combined value of an organization's users and content purely in terms of the number of IP addresses in its customer cone, an organization at the 95% percentile only accounts for 0.01% of the full routed IP address space, an organization at the 99% percentile accounts for 0.2%, while Amazon.com, Inc. accounts for 1.21%. The distribution of actual traffic across agreements may be even more skewed than the size distribution of the networks measured by customer cones.

providers so many peer-to-peer agreements are not observable in the data we use. A truly complete set of agreements would require collecting BGP tables and traceroute data from vantage points in the majority of Internet networks, while our data collection is limited to vantage points where we have our own or partner monitors. Over time, monitors were added at new vantage points, resulting in more visibility in parts of the Internet and hence a greater number of discoverable agreements. To keep visibility consistent throughout our sample periods, we extracted agreements only from a set of monitors that operated throughout our sample periods, January 2015–June 2019 for AS Relationships and December 2015–June 2019 for IPv4 Prefix-Probing.

Similarly, the number of interconnection points we capture is a subset of all interconnection points associated with each agreement. We do not have the ability to target an interconnection directly, but must instead target destinations and infer interconnections from paths that our monitors cross to reach those destinations. We thus miss interconnections not observed by our monitors.

We also note interconnection agreements are more complex than allowed for in our approach. The types of agreements between the same two networks can differ by peering location or even by prefix. Our inference algorithm oversimplifies these cases by assigning a single agreement type to each pair of networks (CAIDA, 2015-2019a).

Finally, sometimes technical problems occur with monitors, resulting in changes in visibility of some paths. In October 2018, configuration changes in three RIPENCC partner monitors placed in Amsterdam, Barcelona and Zurich caused permanent disappearance of around 2450 network-to-network interconnections from our sample. We dropped all of the affected interconnections throughout our sample.

2.5 Empirical Strategy

In this section, we present our main regression specification and provide justifications for key assumptions in our empirical strategy.

Our empirical strategy, in short, constitutes using a difference-in-differences approach to compare interconnection activities of networks owned by organizations headquartered in the EEA (treatment group) versus networks owned by organizations headquartered in non-EEA OECD countries (control group) before and after the GDPR approval date (April 2016) and implementation date (May 2018). Our main regression

specification is as follows:

$$\log(outcome_{mt}) = \beta_0 + \beta_1 POST_{e,mt} \times EEA_{mt} + \beta_2 POST_{a,mt} \times EEA_{mt} + \gamma_m D_m + \gamma_t D_t + \varepsilon_{mt}, \quad (2.1)$$

where m is the unit of observation of the outcome variable of interest. We take the log of the outcome variable to reflect estimated effects in percentage changes. m can take country pair subscript ij , network pair subscript kl , country subscript i or network subscript k . $POST_{e,mt}$ is an indicator variable equal to 1 if time t is after the GDPR effective date. $POST_{a,mt}$ is an indicator variable equal to 1 if time t is after the GDPR approval date. EEA_{mt} is an indicator variable equal to 1 if the observation m is in the treatment group, and equal to 0 if the observation m is in the control group. A dummy D_m for each unit of observation m and a dummy D_t for each time period t are included. The difference-in-differences effect is identified by the coefficients on the interaction terms $POST_{e,mt} \times EEA_{mt}$ and $POST_{a,mt} \times EEA_{mt}$.

The validity of our approach hinges on the validity of our assumptions that any potential policy effect did not set in before the approval of the GDPR and that our control group is reasonable. We therefore focus our discussion on the justifications for these assumptions. We also acknowledge the limitations of our empirical approach.

We offer several justifications for our assumption about the timing of the effect: 1) a robustness check using an alternative cutoff date, 2) conversations with network operators on their decision horizon, and 3) empirical evidence at the application layer supporting stark cutoff dates.

We first discuss a robustness check that uses December 2015 as an alternative cutoff date to study whether networks responded to the law prior to its approval. Examining the timeline of the creation of the law, we think December 2015 is the earliest possible date for firms to respond to the future law. Consultation for the law began as early as 2009 and the European Commission published a proposal text in 2012. In 2013, the European Parliament adopted a compromised text, based on almost 4,000 proposed amendments. In 2015, the Council of the European Union published its proposal for the GDPR and started negotiations with the European Parliament. The Parliament and Council reached agreement on the text of the GDPR in December 2015 (Hoofnagle et al., 2019). Given the intensity of negotiation and the amount of changes the proposal went through, we think it was unlikely for firms to respond before the text of the law was fixed. Most of our variables are available well before December 2015, allowing us to use December 2015 as an alternative cutoff date and test whether networks responded in anticipation of the law. We discuss results from this robustness check in our results section.

Moreover, through conversations with network operators, we learn network operators respond to real-time changes in actual data flows at the application layer, rather than respond to potential changes on the longer time horizon, due to the fact that networks can establish and terminate interconnection agreements relatively quickly. As such, we think they were unlikely to respond before changes happened at the application layer.

Empirical evidence at the application layer supports the notion that changes at the application layer happened after the GDPR effective date of May 2018. Jia et al. (2021) discuss that, within the two years between the GDPR's approval and effective dates, many organizations chose to roll out their compliance strategy only days and weeks before the effective date. Goldberg et al. (2019) show large declines in page views, visits, orders and revenue from EU consumers at a set of e-commerce sites relative to a control group within four weeks after the policy implementation date.⁵¹ Johnson et al. (2023) show a cliff-like decline in websites' use of web technology within thirty days after the policy implementation date.⁵² Aridor et al. (2020) similarly show an immediate effect on consumer opt-out behavior and firm revenue.⁵³ We think such evidence of large and quick responses at the application layer helps to justify our choice of stark cutoff dates as well.

Now we move to discuss our choice of control group. Our treatment group consists of networks owned by organizations headquartered in EEA countries. We choose networks owned by organizations headquartered in non-EEA OECD countries as the control group. We think this is a relevant comparison because networks in developed countries have similar growth rates of interconnection prior to the GDPR. As we will show in a series of graphs later in the results section, the parallel pre-trends needed for the difference-in-differences approach are visually apparent for the treatment and control groups across all outcome variables of interest. We exclude networks in non-EEA non-OECD countries and territories from the control group for worries that networks in developing countries might behave differently from networks in more developed countries prior to the GDPR. Appendix Table B.2.1 presents complete lists of countries and territories in treatment group, in the control group), and are excluded.

We are well aware of the concern that, given the GDPR's global ambition and wide territorial scope, application firms in non-EEA OECD countries also need to incur substantial cost to comply with the law

⁵¹See their Figure 2 for details.

⁵²See their Figure 1 for details.

⁵³See their Figures 3–6 for details.

if they want to serve EEA consumers. This would in turn change the derived demand for data exchange associated with non-EEA OECD networks and bias our results towards zero given our choice of control group. We discuss thoroughly how we address this concern in five ways: 1) a robustness check using a first differences approach, 2) a discussion of the extent of compliance of non-EEA application firms, 3) an organization of our results based on our confidence of the validity of the control group across our eight outcome measures and various subsample breakdowns, 4) interpretation of our results as differential impact and empirical evidence of differential impact at the application layer, and 5) a straightforward acknowledgement of problems with our control group for some of our outcome measures and subsamples.

First, we note that we can perform a robustness check to our main difference-in-differences approach by simply first differencing our outcome variables within EEA subsamples and studying whether the approval or the implementation of the GDPR impacted the rate of interconnection growth within EEA countries or networks. This test helps to rule out the scenario under which the GDPR had significant and identical effect on EEA and non-EEA OECD networks. We discuss further the rationale and results of this robustness check in our results section.

We then discuss the extent of compliance to the GDPR among non-EEA application firms. We hypothesize that non-EEA consumers are unlikely to enjoy similar protection as EEA consumers following the GDPR, even if these firms choose to comply. These firms may also choose not to comply, cutting out EEA consumers all together. As compliance to the GDPR can be extremely costly, there is incentive for non-EEA application firms to limit compliance to EEA consumers. While some non-EEA application firms allegedly improved privacy protection for non-EEA consumers following the GDPR, the degree of protection non-EEA consumers enjoyed fell far short from their EEA counterparts.⁵⁴ In fact, the California Consumer Privacy Act (CCPA) was strongly motivated by the goal to bring privacy protection of Californian residents on par with that of EEA residents under the GDPR as firms did not voluntarily do so. Peukert et al. (2020) suggest websites catering to non-EU audiences decreased their use of third-party web technology vendors following the GDPR. However, they found the magnitude of change for those websites at 2.2–3.6 percentage points in their most reliable specifications to be far smaller than the magnitude of change for websites catering to EU audience at

⁵⁴The Associated Press investigated Facebook's claim on global GDPR compliance and found its implementation of many GDPR provisions were vague for non-EEA consumers. While Facebook did not publicize the fact, the Associated Press also found users in six Asian countries did not get the protection through manual checks. See Jesdanun, Anick. 2018. "How Google, Facebook will adapt to Europe's New Privacy Law." The Associated Press.

7.1–9.3 percentage points.⁵⁵ A lot of anecdotal evidence also suggests many content and service providers located outside the EEA simply stopped serving EEA consumers.⁵⁶ When EEA consumers were not blocked, they could be offered a very stripped down version of the content.⁵⁷

Our above discussion helps us to identify the outcomes and subsamples that are less likely to be plagued by the bias towards zero across our eight different outcome measures and various subsample breakdowns. We present our results in Section 2.6 with this consideration, showing first the measures we are the most confident about. We discuss the rationale for our confidence in those cases in Section 2.6.

Moreover, we are confident that our estimates, interpreted as the differential effects between EEA networks and non-EEA OECD networks, will be nonzero if the policy has any effect on decisions to interconnect. A number of empirical works have found significant and often large differential effects of privacy regulation at the application layer using a difference-in-differences approach and control groups similar to ours. For example, Goldfarb & Tucker (2011) found the EU’s 2002 e-Privacy Directive reduced online display ad effectiveness in the EU relative to other countries. Jia et al. (2021) used EU-based technology ventures as their treatment group and US-based ventures as their primary control group and found EU venture deals declined by as much as 26.1% after the implementation of the GDPR. Jia et al. (2020) found the impact of the GDPR on EU venture investment relative to their US counterparts was larger when ventures and lead investors were not in the same state or union. Aridor et al. (2020) found the GDPR resulted in a 12.5% drop in trackable consumers on European travel platforms as compared to their non-European counterparts and resulted in declines in revenue for the European firms. These differential changes in derived demand for data exchange motivate our expectation of differential changes at the internet layer, if the policy does affect the internet layer.

In addition to all of above, We acknowledge the limitation of our difference-in-differences research design

⁵⁵We believe their estimates for websites catering to non-EU audience are overestimates. They explored four different definitions for “websites that cater to EU audience”: 1) the website has a top-level domain that is specific to a country in the EU (for examples, .de or .fr); 2) the website appears on Alexa’s rank for any country in the EU; 3) the website returns content in any of the official languages of member countries of the EU, except English; 4) the website is visited by users in Germany but not users in the US in Nielsen clickstream data. In each of the four cases, “websites that cater to non-EU audience” are defined as the websites that did not meet the criterion. We note that all four definitions would misclassify a large number of European-based English language sites with common top-level domains such as .com, .org and .net as catering to non-EU audience.

⁵⁶For example, Joseph O’Connor, a web developer, compiled an list of 1,361 websites (mostly US-based news sites) that blocked visitors from the EU after the GDPR effective date. See O’Connor, Joseph. 2018-2019. “Websites Not Available in the European Union after GDPR.” <https://data.verifiedjoseph.com/dataset/websites-not-available-eu-gdpr>. As of March 2019, the last time the list was updated, 1,129 websites on the list remained blocked.

⁵⁷Sentance, Rebecca. 2018. “GDPR: Which Websites are Blocking Visitors from the EU?” Econsultancy.

that it is not able to estimate the absolute effect of the GDPR on interconnection decisions at the internet layer. We also acknowledge that the differential impact for some of our subsamples and outcome measures are more likely to tend towards zero while others less so. We discuss this in more detail when we present results by outcome measure and subsample in our next section.

2.6 Results

In this section, we present the regression results specific to each outcome variable. We also discuss results from various robustness checks at the end of this section.

2.6.1 The Number of Agreements between Countries

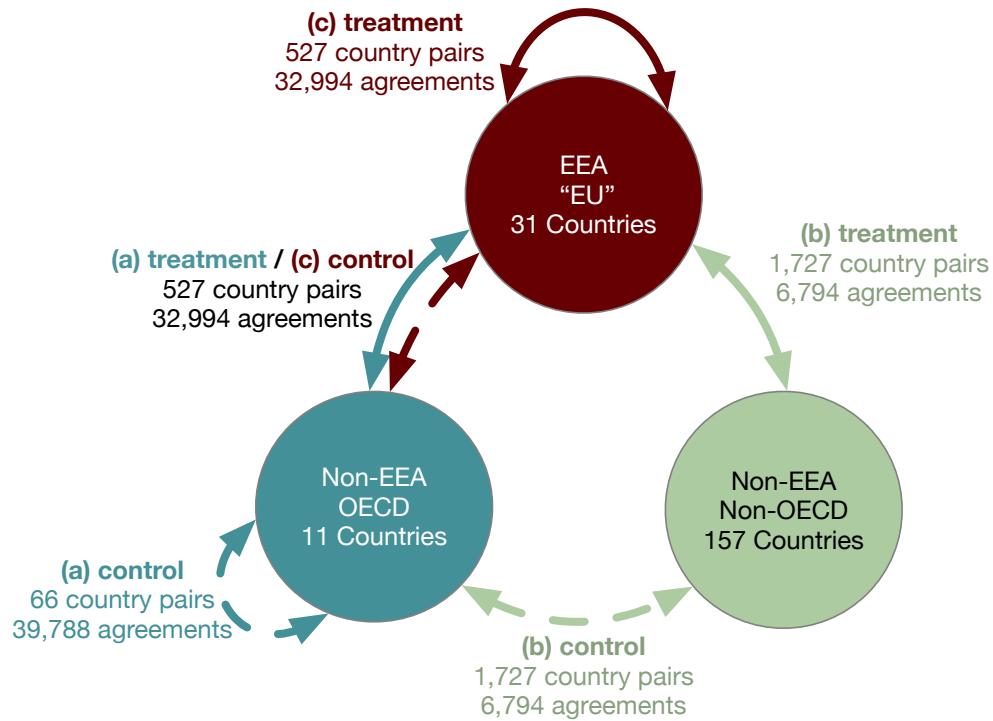
In this subsection, we study the outcome variable $numAg_{ijt}$, the number of interconnection agreements between pairs of networks owned by the countries i and j . As the unit of observations is a country pair, we need to hold fixed the EEA membership status (or OECD status) of the counterparty of interconnection while we compare the outcomes for EEA countries (treatment group) and for countries in the OECD but not in the EEA (control group).

We therefore construct three subsamples based on counterparties: (a) the counterparties are non-EEA OECD countries, (b) the counterparties are non-EEA non-OECD countries, and (c) the counterparties are EEA countries. Within each subsample, we then keep only observations where networks or countries are in the EEA (treatment group) or are in the OECD but not in the EEA (control group) and compare their outcomes. Figure 2.3 illustrates visually the construction of the three subsamples.

We note that a bias towards zero is less likely to impact regression results for subsamples (a) and (b) than results for subsample (c). The control group of either subsample (a) or subsample (b) does not involve EEA countries. As we discussed in the previous section, we believe non-EEA application layer firms are far less likely to comply with the GDPR in markets outside the EEA or change their behavior in those markets due to the regulation. Their derived demand for data exchange at the internet layer from and to those markets therefore should change little. We are concerned that results for subsample (c) may be biased towards zero as non-EEA application firms need to comply with the GDPR in EEA markets or they may exit those markets, whichever would reduce derived demand for data exchange at the internet layer.

Figure 2.4 shows a comparison of the total number of agreements in the EEA countries and in the

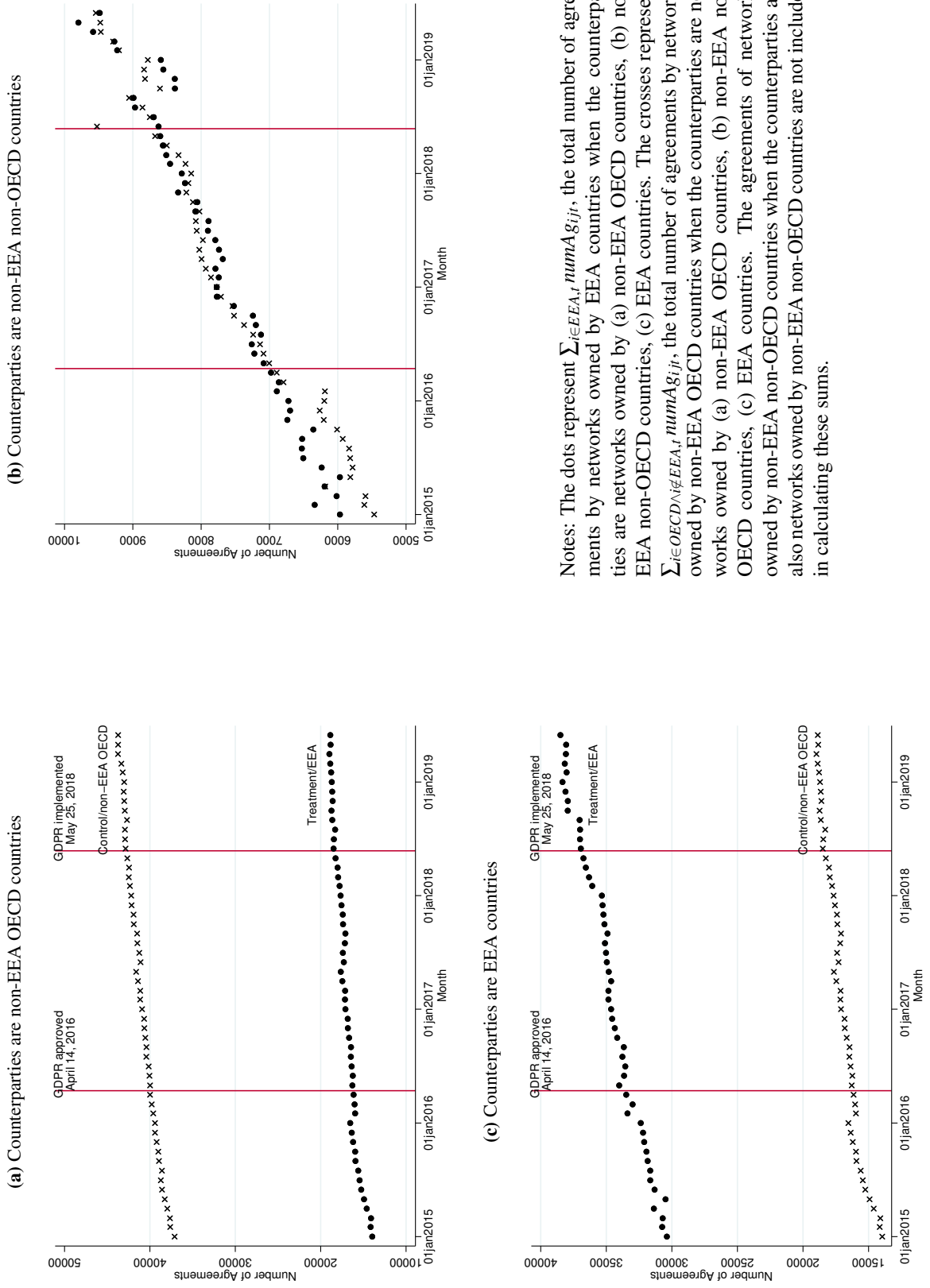
Figure 2.3: Three subsamples for the analysis on the country pair level



Notes: Interconnections are bidirectional, as represented by the double-headed arrows. We present the total number of agreements in each treatment/control group in March 2016, the month before the GDPR's approval. Subsample (a) fixes non-EEA OECD countries (or territories) as interconnection counterparties. Subsample (b) fixes non-EEA non-OECD countries (or territories) as interconnection counterparties. Subsample (c) fixes EEA countries (or territories) as interconnection counterparties. Interconnections between EEA countries and non-EEA OECD countries contribute to both subsample (a) and subsample (c). For example, a country pair France–US contributes to both the treatment group in (a) and the control group in (c).

non-EEA OECD countries, holding fixed the counterparties. We make a few observations. First, despite the differences in levels, EEA countries and non-EEA OECD countries exhibit remarkable parallel trends in setting up agreements with counterparties that are non-EEA OECD countries, non-EEA non-OECD countries, and EEA countries throughout the sample period. Second, agreements with developing countries or territories have a lot more noise in measurement compared to agreements within OECD countries or EEA countries.

Figure 2.4: Number of interconnection agreements by EEA and non-EEA OECD countries by counterparty



We then run the regression specification in Equation 2.1 on each of the three subsamples. The outcome variable is $(numAg_{ijt} + 1)$, where we add one to $numAg_{ijt}$, the number of agreements between countries i and j in month t , to account for zero values. The unit of observation m is country pair ij .

The results are shown in Table 2.3. The main effect, based on the coefficient on $POST_{e,ijt} \times EEA_{ijt}$, is not significantly different from zero across the three subsamples. The only significant result in this table comes from the coefficient on $POST_{a,ijt} \times EEA_{ijt}$ for the non-EEA non-OECD counterparty subsample and we test the robustness of this result. Table 2.3 clusters standard error by country pair. Alternatively, one might expect the interconnection decisions of one particular country to other countries to have correlated errors. This may be especially true for interconnection decisions from an EEA or OECD country to developing countries based on the EEA/OECD networks' global interconnection strategy to remote and low demand areas. Therefore, we cluster standard error by EEA and OECD countries in the country pairs for the non-EEA non-OECD counterparty subsample as a robustness test, resulting in 43 clusters as compared to 6,751 clusters in Column 2 of Table 2.3. The coefficient on $POST_{a,ijt} \times EEA_{ijt}$ is no longer significant and is therefore likely a spurious result.

2.6.2 The Number of Agreements between Countries by Agreement Type

In this subsection, we further break down the number of agreements between country pairs to provider-to-customer, peer-to-peer, and customer-to-provider types. As before, we prioritize results for subsamples where interconnection counterparties are non-EEA OECD countries or non-EEA non-OECD countries.

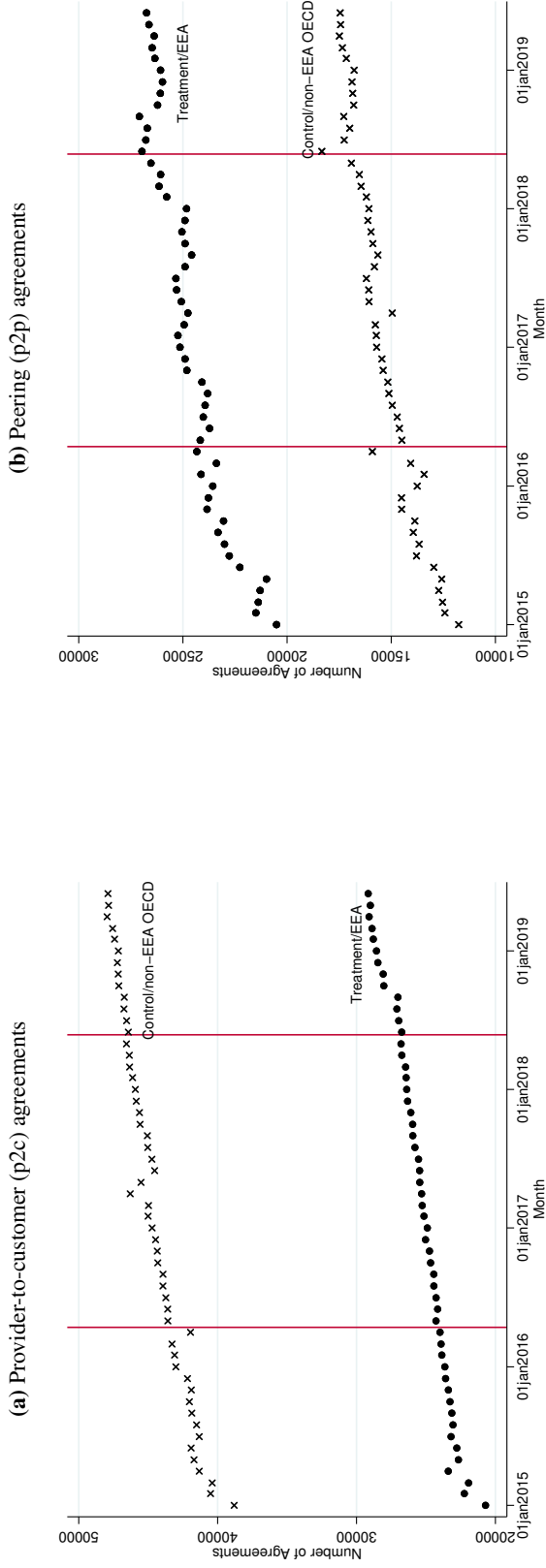
Figure 2.5 shows a comparison of the total number of agreements in the EEA countries and in the non-EEA OECD countries, by agreement type. We still observe EEA countries and non-EEA OECD countries have remarkable parallel trends by agreement type throughout the sample period. Based on visual evidence, the GDPR does not have heterogeneous effects on different types of agreements.

Table 2.3: The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e \times EEA$	−0.009 (0.029)	0.003 (0.005)	−0.003 (0.016)
$POST_a \times EEA$	−0.007 (0.024)	0.017*** (0.005)	0.011 (0.017)
Group dummies	country pairs	country pairs	country pairs
Time dummies	months	months	months
Clusters	418	6,751	880
R^2	0.991	0.948	0.987
Observations	22,572	364,554	47,520

Notes: The dependent variable is $\log(numAg_{ijt} + 1)$. The variable $numAg_{ijt}$ is rectangularized as described in Table 2.2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Column (1) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by an EEA country. All regressions include month dummies and country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Figure 2.5: Number of agreements by EEA and non-EEA OECD countries by agreement type



Notes: (a) The dots represent $\sum_{i \in EEA, j} numProvAg_{ij}$, the total number of agreements by networks owned by EEA countries where these networks are providers. The crosses represent $\sum_{i \in OECD \setminus EEA, j} numProvAg_{ij}$, the total number of agreements by networks owned by non-EEA OECD countries where these networks are providers. (b) The dots represent $\sum_{i \in EEA, j} numPeerAg_{ij}$. The crosses represent $\sum_{i \in OECD \setminus EEA, j} numPeerAg_{ij}$. Peering agreements between EEA countries and non-EEA OECD countries contribute to counts in both series. (c) The dots represent $\sum_{i \in EEA, j} numCustAg_{ij}$. The crosses represent $\sum_{i \in OECD \setminus EEA, j} numCustAg_{ij}$. Agreements between networks owned by non-EEA non-OECD countries are not included in calculating these sums.

We then run the regression specification in Equation 2.1 on each agreement type for each of the three counterparty subsamples. The outcome variables are $(numProvAg_{ijt} + 1)$, $(numPeerAg_{ijt} + 1)$ and $(numCustAg_{ijt} + 1)$. We add one to $numProvAg_{ijt}$, $numPeerAg_{ijt}$ and $numCustAg_{ijt}$, the number of provider-to-customer, peer-to-peer and customer-to-provider agreements between countries i and j in month t , to account for zero values. The unit of observation m is country pair ij .

We show the results in Table 2.4. We see a few significant results in the non-EEA non-OECD counterparty subsample. As previously, once we cluster standard error by EEA and OECD countries in the country pairs for the non-EEA non-OECD counterparty subsample as a robustness test, the significance of these results disappear. We also note these results, though sometimes significant, lack systematic patterns and are economically small in magnitude.⁵⁸

2.6.3 The Number of IP-Address-Level Interconnection Points per Agreement

Our earlier results suggest the GDPR did not change whether agreements were made and what types of agreements were made between networks. One hypothesis for the absence of behavior change is that setting up an agreement is such a substantial decision that changes in usage and bargaining friction due to the GDPR are small in comparison. Networks may only change the capacity associated with each interconnection in response to lower usage instead of cancelling an agreement altogether. If that is the case, we are unlikely to observe effects of the GDPR on the extensive margin. The GDPR's impact may be on how dense the two networks' interconnection is. Motivated by this consideration, we examine how the GDPR affected the number of IP-address-level interconnection points two networks had, conditional on them having an agreement.

The outcome variable we study in this section is $numAgIP_{klt}$, the number of IP-address-level interconnection points between network k and network l , given k and l have an agreement. As the unit of observations is a network pair, we hold fixed the EEA membership status (or OECD status) of the counterparty of interconnection while we compare the outcomes for EEA countries (treatment group) and for countries in the OECD but not in the EEA (control group).

As previously, we construct three subsamples based on counterparties: (a) the counterparties are in non-

⁵⁸To illustrate how economically small the implied effect based on the coefficients is, we take for example the coefficient -0.038 on $POST_e \times EEA$ from column (8) of Table 2.4, the largest significant result in the table. The dependent variable for the regression in column (8) is $\log(numPeerAg_{ijt} + 1)$. It has a mean of 0.109 and an SD of 0.544. Therefore, being in the treatment group post GDPR effective date has an effect which is a tiny fraction of one standard deviation of the outcome.

Table 2.4: The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty and agreement type

	Counterparty is non-EEA OECD			Counterparty is non-EEA non-OECD			Counterparty is EEA		
	Provider (1)	Peer (2)	Customer (3)	Provider (4)	Peer (5)	Customer (6)	Provider (7)	Peer (8)	Customer (9)
$POST_e \times EEA$	-0.032 (0.023)	-0.040 (0.042)	-0.020 (0.022)	0.002 (0.003)	-0.007 (0.004)	-0.013*** (0.004)	-0.007 (0.012)	-0.038** (0.016)	0.005 (0.013)
$POST_a \times EEA$	0.021 (0.026)	0.031 (0.035)	-0.025 (0.027)	0.011*** (0.004)	0.007* (0.004)	-0.006** (0.003)	0.027* (0.016)	0.002 (0.017)	-0.021 (0.014)
Group dummies	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs
Time dummies	months	months	months	months	months	months	months	months	months
Clusters	473	418	473	6,751	6,751	6,751	1,376	880	1,376
R^2	0.984	0.984	0.985	0.941	0.930	0.925	0.978	0.980	0.977
Observations	25,542	22,572	25,542	364,554	364,554	364,554	74,304	47,520	74,304

Notes: The dependent variable is $\log(numProvAg_{ijt} + 1)$ for columns (1), (4), (7), $\log(numPeerAg_{ijt} + 1)$ for columns (2), (5), (8), and $\log(numCustAg_{ijt} + 1)$ for columns (3), (6), (9). The dependent variables are rectangularized as described in Table 2.2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Columns (1), (2), (3) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by a non-EEA OECD country. Columns (4), (5), (6) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by a non-EEA non-OECD country. Columns (7), (8), (9) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by an EEA country. All regressions include month dummies and country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

EEA OECD countries, (b) the counterparties are in non-EEA non-OECD countries, and (c) the counterparties are in EEA countries. Within each subsample, we then keep only observations where networks or countries are in the EEA (treatment group) or are in the OECD but not in the EEA (control group) and compare their outcomes.

The control group of either subsample (a) or subsample (b) does not involve networks in EEA countries. As before, we are less concerned about the results for these subsamples than those for subsample (c). Results from subsample (c) may be biased towards zero as non-EEA application firms need to comply with the GDPR in EEA markets or they may exit those markets, whichever would reduce derived demand for data exchange at the internet layer.

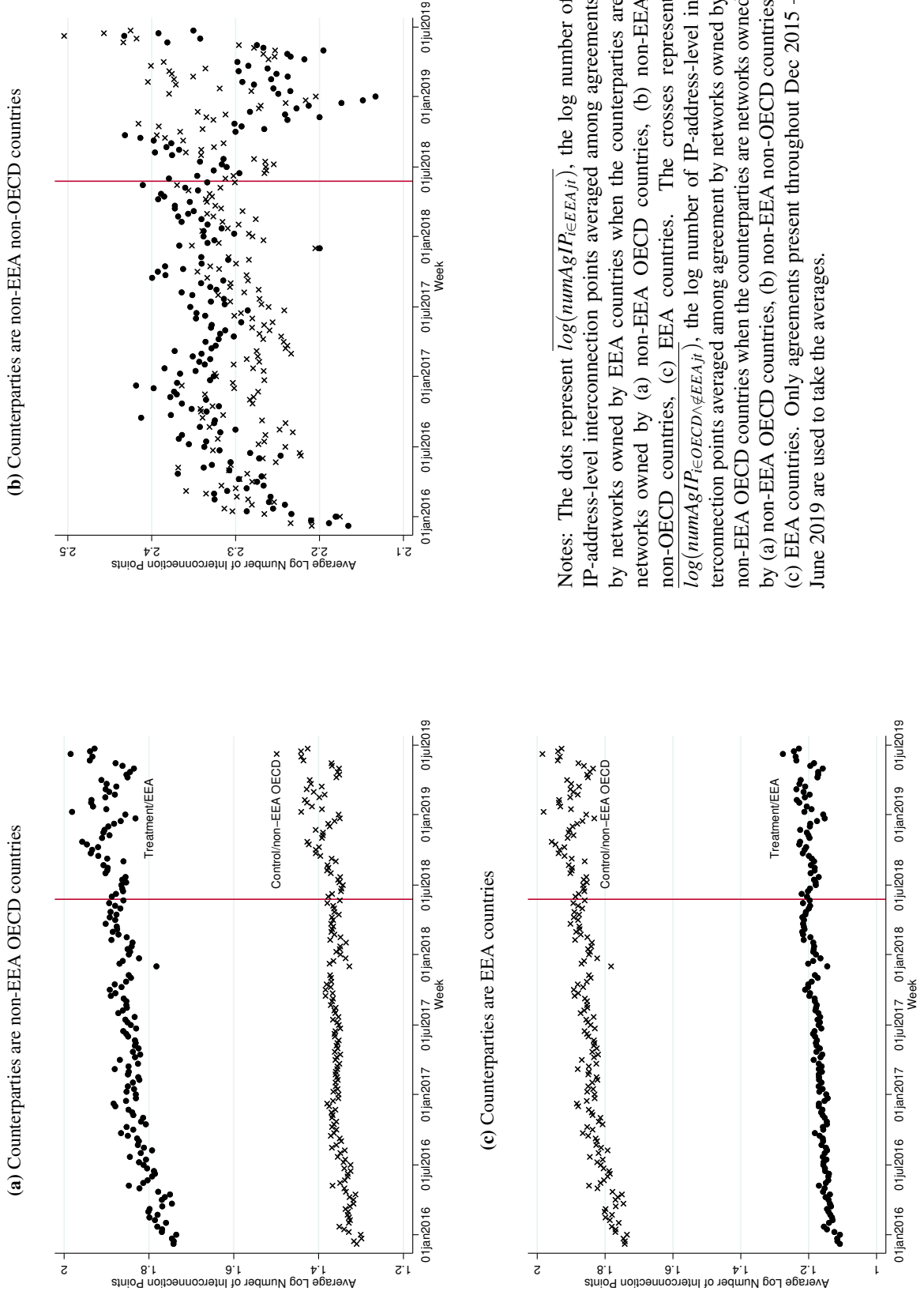
Figure 2.6 shows a comparison of the average log number of interconnection points per agreement in the EEA countries and in the non-EEA OECD countries, holding fixed the interconnection counterparties. We first note that observed interconnection points with developing countries have a lot of noise in our measurement while observed interconnection points among EEA and OECD countries are quite precisely measured, reflecting the large number of vantage points inside developed countries. When interconnection points are well-measured, we observe that, despite the differences in levels, EEA countries and non-EEA OECD countries still exhibit remarkable parallel trends in terms of the number of interconnection points per agreement throughout the sample period.

Given this particular data source only started in December 2015, close to the GDPR approval date, we do not include the interaction term $POST_{a,kl,t} \times EEA_{kl,t}$ in our regression. Therefore, instead of Equation 2.1, we run the following regression on each of the three subsamples,

$$\log(numAgIP_{kl,t}) = \beta_0 + \beta_1 POST_{e,kl,t} \times EEA_{kl,t} + \gamma_{kl} D_{kl} + \lambda_t D_t + \varepsilon_{kl,t}. \quad (2.2)$$

We take the log of the outcome variable $numAgIP_{kl,t}$ to reflect estimated effects in percentage changes. $POST_{e,kl,t}$ is an indicator variable equal to 1 if time t is after the GDPR effective date. $POST_{a,kl,t}$ is an indicator variable equal to 1 if time t is after the GDPR approval date. $EEA_{kl,t}$ is an indicator variable equal to 1 if the network pair kl is in the treatment group for the subsample, and equal to 0 if the network pair kl is in the control group for the subsample. A dummy D_{kl} for each network pair kl and a dummy D_t for each week t are included. The difference-in-differences effect is identified by the coefficient on the interaction term $POST_{e,kl,t} \times EEA_{kl,t}$.

Figure 2.6: Average log number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries by counterparty

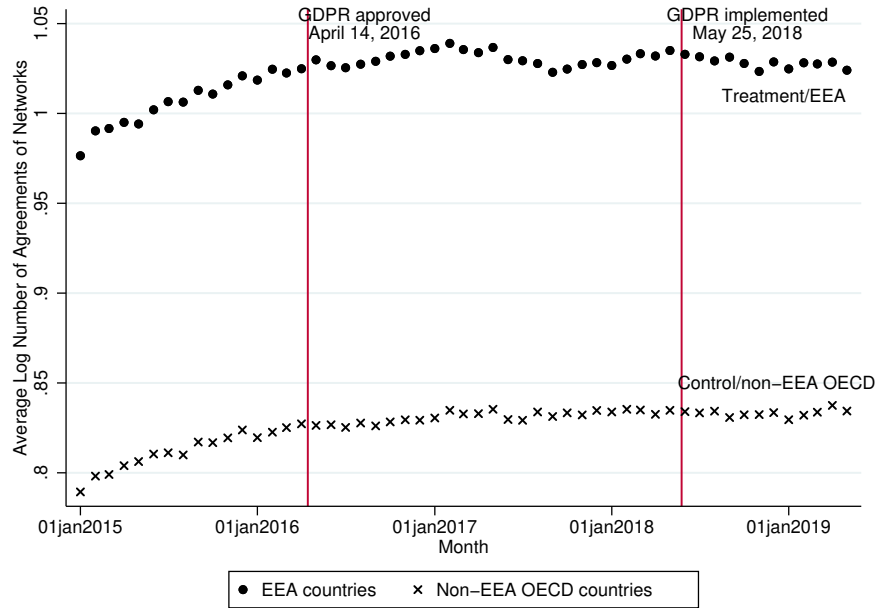


The results are shown in Table 2.5 and are in no case significantly different from zero. We include agreements present for at least 150 weeks for our regressions in Table 2.5. Given we study the intensive margin, alternatively we keep only agreements present for all of 169 weeks between December 2015 and June 2019. Doing so substantially reduces the sample size and the results are similar to those in Table 2.5.

2.6.4 The Number of Agreements by Networks

In addition to interconnection behavior between pairs of countries or networks, we study how the GDPR might have impacted the number of agreements per network, the number of networks per country and the sizes of the customer cones of each networks. We stress that these estimates are differential impact between EEA networks and non-EEA OECD networks.

Figure 2.7: Average log number of interconnection agreements by networks in EEA and non-EEA OECD countries



Notes: The dots represent $\overline{\log(\text{numAgNtwrk}_{k \in \text{EEA}, t})}$, the log number of agreements averaged among networks owned by EEA countries. The crosses represent $\overline{\log(\text{numAgNtwrk}_{k \in \text{OECD} \setminus k \notin \text{EEA}, t})}$, the log number of agreements averaged among networks owned by non-EEA OECD countries. Non-EEA and non-OECD countries' networks are not included in taking the averages. Only networks present throughout Jan 2015 – June 2019 are used to take the averages. The first red vertical line represents 14 April 2016, the approval date of the GDPR. The second red vertical line represents 25 May 2018, the implementation date of the GDPR. Regression including month and network fixed effects has the coefficient on $\text{POST}_t \times \text{EEA} = -.004$ ($se = 0.007$, clustered by country) and the coefficient on $\text{POST}_t \times \text{EEA} = 0.006$ ($se = 0.007$, clustered by country). Both are insignificant at conventional levels of significance.

Table 2.5: The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e \times EEA$	0.039 (0.023)	0.003 (0.049)	-0.032 (0.024)
Group dummies	network pairs	network pairs	network pairs
Time dummies	weeks	weeks	weeks
Clusters	128	522	307
R^2	0.871	0.827	0.867
Observations	2,593,805	494,374	1,886,031

Notes: The dependent variable is $\log(numAgIP_{klt})$. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Column (1) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by an EEA country. Only agreements present for at least 150 weeks are used. The GDPR approval date Apr 2016 is close to the sample starting date Dec 2015, so $POST_a \times EEA$ is not included in the regressions. All regressions include week dummies and network pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Figure 2.7 shows a comparison of the average log number of agreements by networks in the EEA countries and in the non-EEA OECD countries. We observe visually apparent parallel trends between the two groups prior to the approval of the GDPR, between the approval and implementation of the GDPR, as well as after the implementation of the GDPR.

We then run the regression specification in Equation 2.1. The outcome variable is $numAgNtwrk_{kt}$, the number of agreements network k has in month t . The unit of observation m is network k .

We show the results below Figure 2.7. The coefficient on $POST_e \times EEA$ and the coefficient on $POST_a \times EEA$ are both insignificant at conventional levels of significance. This result suggests the GDPR does not have differential impact on the number of interconnection agreements on EEA networks relative to their non-EEA OECD counterparts.

2.6.5 The Number of Networks

Figure 2.8 shows a comparison of the average log number of networks per country in the EEA countries and in the non-EEA OECD countries. Again, we observe visually apparent parallel trends between the two groups prior to the approval of the GDPR, between the approval and implementation of the GDPR, as well as after the implementation of the GDPR.

We then run the regression specification in Equation 2.1. The outcome variable is $numNtwrk_{it}$, the number of networks country i has in quarter t . The unit of observation m is country i .

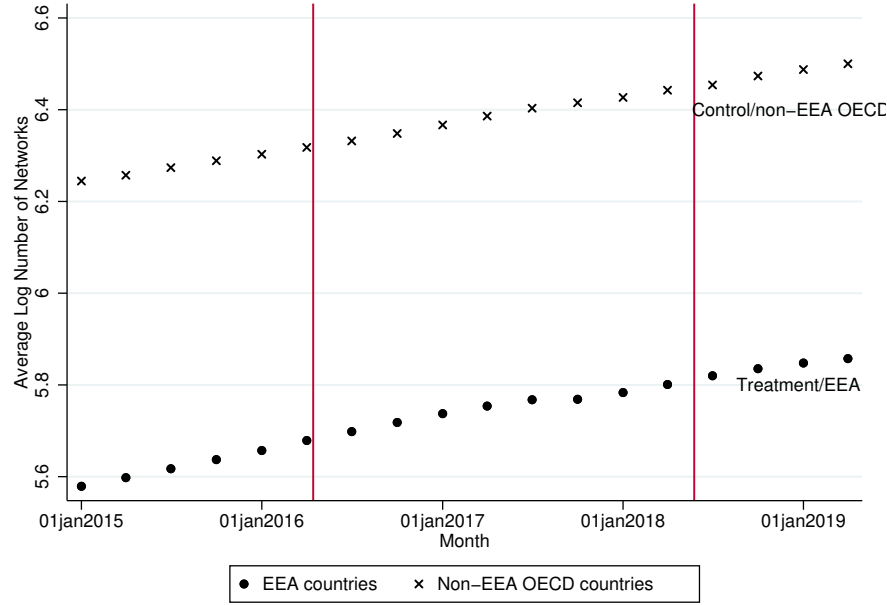
We show the results below Figure 2.8. The coefficient on $POST_e \times EEA$ and the coefficient on $POST_a \times EEA$ are both insignificant at conventional levels of significance. This result suggests the GDPR does not differentially impact the number of networks in EEA countries compared to non-EEA OECD countries.

2.6.6 Customer Cone of Networks

Figure 2.9 shows a comparison of the average log customer cone of networks in the EEA countries and in the non-EEA OECD countries. We observe visually apparent parallel trends between the two groups prior to the approval of the GDPR, between the approval and implementation of the GDPR, as well as after the implementation of the GDPR.

We then run the regression specification in Equation 2.1. The outcome variable is $NtwrkCustCone_{kt}$, the size of network k 's customer cone in month t . The unit of observation m is network k .

Figure 2.8: Average log number of networks in EEA and non-EEA OECD countries



Notes: The dots represent $\overline{\log(numNtwrk_{i \in EEA,t})}$, the log number of networks averaged among EEA countries. The crosses represent $\overline{\log(numNtwrk_{i \in OECD \setminus EEA,t})}$, the log number of networks averaged among non-EEA OECD countries. Non-EEA and non-OECD countries' networks are not included in taking the averages. Regression including quarter and country fixed effects has the coefficient on $POST_e \times EEA = -.002$ ($se = 0.017$, clustered by country) and the coefficient on $POST_a \times EEA = 0.016$ ($se = 0.024$, clustered by country). Both are insignificant at conventional levels of significance.

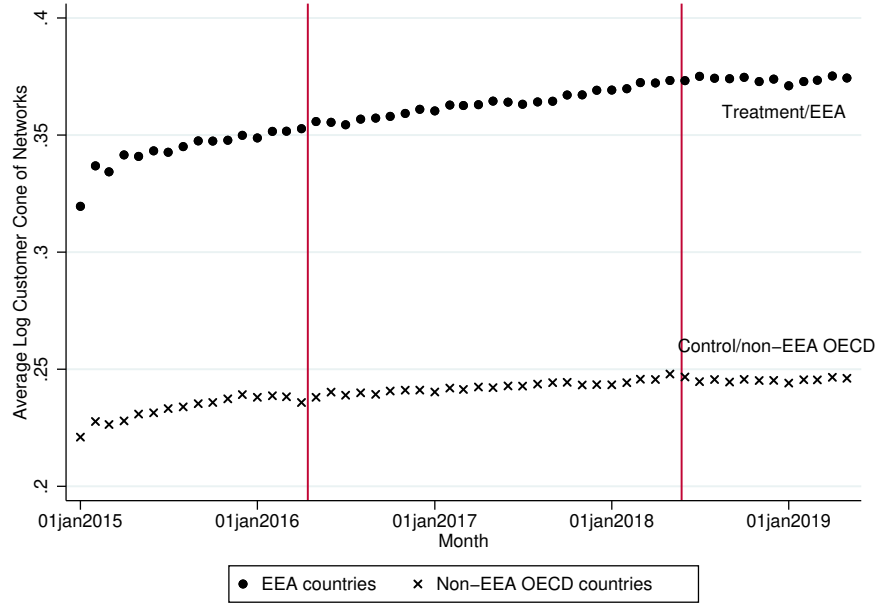
We show the results below Figure 2.9. Though both the coefficient on $POST_e \times EEA$ and the coefficient on $POST_a \times EEA$ are significantly different from zero, their magnitudes are economically very small, suggesting the GDPR has little impact on the centrality of networks in EEA countries compared to non-EEA OECD countries.

2.6.7 Robustness Checks

In this subsection, we discuss our robustness checks. We first replace the logged outcome variables in all of our specifications with their original unlogged values. We present the results in Appendix B.3.1. As shown, all of the results are qualitatively similar to results with logged outcome variables. This alleviates the concern that our zero estimates are driven by taking the log of the outcome variables.

We then perform a robustness check by redefining $POST_a$ to equal to 1 if the observation is made after December 2015. As discussed in our empirical strategy section, it is possible that networks invested in

Figure 2.9: Average log customer cone of networks in EEA and non-EEA OECD countries



Notes: The dots represent $\overline{\log(NtwrkCustCone_{k \in EEA,t})}$, the log customer cone averaged among networks owned by EEA countries. The crosses represent $\overline{\log(NtwrkCustCone_{k \in OECD \wedge i \notin EEA,t})}$, the log customer cone averaged among networks owned by non-EEA OECD countries. Non-EEA and non-OECD countries' networks are not included in taking the averages. Only networks present throughout Jan 2015 – June 2019 are used to take the averages. Regression including month and network fixed effects has the coefficient on $POST_e \times EEA = -.007^*$ ($se = 0.004$, clustered by country) and the coefficient on $POST_a \times EEA = 0.011^{***}$ ($se = 0.004$, clustered by country). Though both are significantly different from zero, their magnitudes are economically small.

interconnection decisions in anticipation of the enforcement of the GDPR even before the law was approved. We think the earliest possible date for the firms to respond in anticipation is December 2015, the time when the text of the law was fixed. The results from this robustness check for our various outcomes are almost identical to our main results presented in Tables 2.3–2.5 and Figures 2.7–2.9. This alleviates the concern that our main specifications did not capture possible effects due to anticipation.

Lastly, we perform a first differences regression for each outcome using only the subsample consisting of the treatment group in our main regression. The rationale for this robustness check is as follows. Consider a scenario in which the GDPR did have a nonzero effect on networks. This would mean the trends before and after the policy cutoff dates were different for the treatment group (and possibly also for the control group). Also suppose that the effects of the policy were identical for the treatment and control groups, or in other words, the trends for the treatment group and control group changed for the exact same amount post policy

cutoffs. Then our difference-in-differences approach would not be able to detect the effect of the policy. We therefore use a first differences approach to test whether there were differential trends before and after the GDPR for the treatment group alone. We will be able to rule out the above scenario if the first differences estimates are zero.

Specifically, we run the following regression:

$$\log(outcome_{mt}) - \log(outcome_{m,t-1}) = \beta_0 + \beta_1 POST_{e,mt} + \beta_2 POST_{a,mt} + \gamma_m D_m + \varepsilon_{mt}, \quad (2.3)$$

where m is the unit of observation of the outcome variable of interest. m can take country pair subscript ij , network pair subscript kl , network subscript k , or country subscript i . $POST_{e,mt}$ is an indicator variable equal to 1 if time t is after the GDPR effective date. $POST_{a,mt}$ is an indicator variable equal to 1 if time t is after the GDPR approval date. A dummy D_m for each unit of observation m is included. We do not include a time dummy as such a dummy would absorb any effect of the GDPR. The effect on first differences is identified by the coefficients on the terms $POST_{e,mt}$ and $POST_{a,mt}$. We show these results in Appendix B.3.2. As shown, the results are very precisely estimated and in all cases do not exceed 0.6 percentage points. These results suggest that the rate of growth in interconnection in the EEA after the GDPR did not change when compared to that before the GDPR. We can therefore rule out the scenario under which the GDPR had significant and identical impact on EEA and non-EEA OECD networks.

2.7. Conclusion

The effectiveness of the Internet in creating economic surplus depends on efficient interconnections bilaterally negotiated by independently operated networks. In this paper, we investigate whether the approval and implementation of the GDPR affects the growth in interconnection of the Internet in Europe. Despite evidence that the GDPR so far had significant effects at the application layer on European firms, we find no visible consequences at the infrastructure layer across the multiple measures we have. Occasionally we estimate statistically significant effects, which prove to be not robust. Our robustness checks suggest that our results are not driven by taking the log of outcome variables or our choice of April 2016 as the first policy cutoff date. Using the first differences approach as an additional robustness check, we show that EEA networks had similar growth rates in interconnection agreements before and after the GDPR approval or implementation, therefore our main results are not driven by our choice of control group.

A number of possible reasons could have contributed to this finding. First, the lack of discernible short-run effect on interconnection could have arisen from slow investment and behavioral changes at the internet layer. This seems unlikely because renegotiations of interconnection agreements happen frequently and we observe continued growth across all network connections.

It is also possible that despite the evident behavioral changes at the application layer due to the GDPR, the effect is small compared to other considerations in negotiating interconnection agreements. That could happen if, for example, the regular growth in data due to growth in many applications overwhelms any short-run impact of the GDPR. In that case, network operators may rationally expect the long run effect of the GDPR to be small even at the application layer.

Another possibility is that the GDPR may change behaviors at the internet layer only along the margins that our measures do not capture. For example, if changes in derived demand for data flow only prompt network operators to change capacity at existing interconnection points, then these changes are not visible to us.

In addition, the GDPR has seen limited and heterogeneous enforcement until the end of our sample period and it may not materially reduce Internet interconnections. While regulators have issued large fines during the period,⁵⁹ they have issued relatively few of them⁶⁰ and in many cases have stopped short of enforcement.⁶¹ Enforcement actions may directly and materially affect data flow and interconnections.⁶²

Finally we only observe the short run, so we cannot rule out that more gradual changes due to the GDPR may surface in the longer run, which is an open question. If we are able to observe a much longer period of time, we will be able to use the data from additional periods and the same methodology to study the effect of the GDPR in the longer run.

Our results have immediate policy implications. As many countries are contemplating implementing their own versions of privacy and data protection regulations, there are concerns about whether such regulations

⁵⁹French Data Protection Authority (CNIL) issued a €50 million fine, the largest GDPR fine to date, to Google Inc. on January 21, 2019 due to issues with Google's consent practices when users configure their Android phones (CNIL, 2019).

⁶⁰Regulators have issued a total of 57 GDPR fines up to June 2019.

⁶¹The Irish Data Protection Commission conducted a sweep on 38 data controllers between August 2019 and December 2019. They found a range of violations but allowed for a six-month period for compliance before taking enforcement actions (Data Protection Commission, 2020)).

⁶²The UK Information Commissioner's Office (ICO) found that implementing its own best practices regarding opt-in consent on its own website reduced data flows to Google Analytics by 90%. (Cross, Tim. 2019. "The ICO's Cookie Consent Rate Dropped 90 Percent After Implementing its Own Best Practices." Video Ad News.)

may negatively impact the growth of the Internet, reduce technology firms' incentives in operating and innovating, reduce the use of the Internet in productivity enhancing activities, and reduce the economic surplus generated through the use of the Internet in the country and beyond. Our results suggest limited effects of such regulations at the internet layer for the measures we are able to capture.

Our results also speak to the debate on the allocation of rents generated through the successful commercialization of the Internet. The enormous rents associated with the exploitation of Web 2.0 and mobile web represent a large portion of the private returns to innovation in the 21st century. These rents have been overwhelmingly captured by players at the application layer, notably the "big tech" companies, while firms at the internet layer captures little of the rents. Our study is consistent with the view that the cost of the GDPR has been a shock to rents, and the costs have been borne by the application layer, paid out of the rents from innovation.

Our results also mask the potential heterogeneity in the GDPR's impact on Internet data flow across application categories. The GDPR affects application firms to different degrees, depending on their business models and how much those models depend on monetization of personal data. Similarly, applications differ in the amount of data traffic generated. The interaction of the two affects the extent to which the GDPR affects the derived demand for data exchange at the internet layer. Video traffic includes Netflix and Youtube, and may be less affected by GDPR due to subscription-based business models, and could grow even as other traffic drops. This combination may explain the visible effects of the GDPR to certain application firms but little evidence of effects on interconnection decisions at the internet layer.

We also note that current empirical works, including this paper, study the impact of the GDPR on suppliers of Internet services and content at various layers of the Internet. Empirical evidence on consumers' responses to privacy regulation is extremely lacking. As policy makers strive to enhance consumer welfare through better privacy protection while trying to minimize such laws' impact of the digital economy, evaluating the laws' effect on consumers is an important direction for future research to allow for the overall welfare analysis.

In addition to policy implications, our paper presents data consisting of virtually all operating networks in the world and a large number of interconnection agreements among them across many years, which opens the possibility of investigating a range of economic- and policy-relevant questions. We acknowledge that important pieces of data are still missing, notably the actual data flows between networks on a similar scale. Future works should keep bridging the unmet data needs.

Chapter 3

Upgraded Software and Embedded Improvements: Tracking Vulnerabilities and Bugs on the Web¹

3.1 Introduction

The growth of networking changed the support and evolution of enterprise software. Networks enabled users to download software instead of requiring physical distribution. That enabled quick improvement to existing software with updated versions and patches. Unfortunately, an ever-changing set of security concerns accompanied those gains – arising from potentially exploiting software vulnerabilities (Arora et al., 2008; Cavusoglu et al., 2008). Cybersecurity hacking incidents have led to ransomware attacks, and data breaches are estimated to have cost trillions of stolen data (Steinberg, 2019; Accenture, 2019; IBM, 2020).

With networked software, users and vendors share responsibility for cybersecurity. Unfortunately, many cybersecurity incidents are due to users' servers running outdated software and IT departments failing to install software updates promptly (Ranger, 2019). For example, in 2017 UK's National Health Service fell victim to ransomware that exploited a vulnerability in their server software for which a security update had been available for over a month – but not installed (Acronis, 2020; Palmer, 2017). At suppliers, discovering vulnerabilities induces a scramble to compose software updates that address the vulnerabilities or “bugs,”

¹Co-authored with Raviv Murciano-Goroff and Shane Greenstein.

two labels we will use interchangeably in this paper. Moreover, suppliers must decide whether to report and disclose security bugs before, during, or after the release of updates (Arora et al., 2008).

Vendor decisions depend on both previous and expected user actions, which frames several challenging trade-offs (Liu et al., 2017; Li et al., 2021; Dey et al., 2015; Dissanayake et al., 2022; Arora et al., 2008; Mookerjee et al., 2011; August et al., 2014; Mitra & Ransbotham, 2015). For each software vulnerability, vendors must estimate both the fraction of users who are currently at risk because of that vulnerability as well as the rate at which users who are risk will install a software update eliminating that vulnerability. While research has identified many factors that shape patching, a critical trade-off informs virtually all analyses. If users respond to new disclosures quickly, as when users associate patching with low friction and high benefits, the supplier accelerates their exposure. On the other hand, suppose users face high administrative costs adjusting their routines. In that case, their reactions may be slow, and vendors will want to disclose vulnerabilities later to reduce the flow of information to hackers.²

Despite the importance of the topic, analysis remains uninformed by the actual empirical distribution of vulnerabilities and the determinants of this distribution, one of the core primitives of these tradeoffs. This study addresses this gap by characterizing nearly two decades of upgrades and the resulting distribution of users for one popular piece of enterprise software, namely, the Apache web server. We also analyze the determinants of user actions and develop implications for prioritizing disclosures of security vulnerabilities and software updates.

Web server software offers a valuable lens for understanding updating and cybersecurity. First, web servers are ubiquitous and critical to the modern web-based commercial internet. Tens of millions of firms in the United States and hundreds of millions across the globe use web servers to support billions of web pages. Second, though hidden from view due to its open-source status, the value is enormous. For example, the value of open-source customer-facing US web server software in 2018 was between \$4.5 and \$11.2 billion, and the latest software correlates with productivity (Murciano-Goroff et al., 2021).

Additionally, due to technical interdependence and complementarities with browsers, server improvements relate directly to progress in the internet user experience. Server improvement correlates with, and sometimes directly causes, the rate of progress in electronic commerce, which supports hundreds of billions of dollars in transactions a year. As we will show, installing improved versions also reduces security vulnerabilities.

²Arora et al. (2008) present a theory that speaks to this trade-off. At the extreme of the theory presented in that paper, if firms are inattentive to software updates, a vendor might decide to keep the presence of vulnerabilities a secret forever.

The Apache Web Server, the focus of this study, was among the earliest mass-market platforms to pioneer upgrades over networks. In the 1990s, Apache grew into the world's second most popular open-source software, and today millions of businesses use it to support their web presence. Like other web servers, Apache's experience is a valuable proxy for the shared responsibility between user and supplier for cybersecurity. For example, as discussed in the paper, Apache has reported 28 severe vulnerabilities and 130 less severe ones since the turn of the millennium and until 2018, the end of our sample period. The average time for managing a significant vulnerability in Apache is 3.6 months, and 4.8 months for a less substantial bug. This study fills in the missing piece: how fast do users respond to the availability of secure versions, and what determines the variance in the installation distribution?

No study has ever constructed such an extensive census of security vulnerabilities from the installations of servers. We do this by building on previous research methods (Murciano-Goroff et al., 2021), which collected the most extensive data project using the Internet Archive to document user upgrades to homepage software at medium and large U.S. business establishments from 2000 to 2018. This paper focuses on over 150,000 Apache users in which we can assemble profiles of software vintages. Then, we match this with data on severe security vulnerabilities reported to Apache. This serves as quasi-natural experiments because discovering and disclosing security bugs in server software are exogenous to users. Once matched, we can observe whether and when organizations eliminate the vulnerability with an upgrade or patch. We can also measure the distribution of user propensity to respond to publicized security threats.

The baseline patterns are striking. We first show that the response has a robust firm-specific behavioral component, which we characterize by the establishment's level of attentiveness. Speaking loosely for brevity, users belong to one of three camps. Two camps are opposites, with one group persistently attentive and another persistently not. Attentive users respond quickly to every vulnerability, while inattentive users never respond. A third camp, arguably more than half the sample, displays a hybrid of behavior, with inattentive actions in the short run and attentive actions over several years.

These results frame the question: why do users accelerate their rate of upgrading when a severe bug arises? We collect data on a subset of users, assembling a complete dataset of over eleven thousand attentive and inattentive users. However, a reverse causality issue is created in an extended panel due to a correlation between the (low) propensity to upgrade and a (high) prevalence of software containing known security vulnerabilities. This motivates using a hazard model approach that accounts for firm-specific proclivities to upgrade.

We find that many attentive users adopt a regular cycle of upgrading. They display a significant propensity to upgrade merely because a new version has appeared, suggesting many users gain protection against vulnerabilities as a byproduct of routine administrative processes that support maintaining frontier software. We also find that many users accelerate upgrading when their web server supports electronic commerce instead of merely an informational or coordinating role, suggesting commercial concerns play a role. We also find deceleration when users must upgrade in a complex IT operation. Finally, attentive security users respond to releasing new software versions by installing minor updates. In contrast, we find that disclosures of bugs—even severe security bugs—alone do not accelerate software patching as much.

In summary, disclosing bugs has limited effects on instigating software updating in attentive and inattentive users. Instead, users primarily respond to the release of patches rather than information about the bugs themselves. The implications of these findings are counterintuitive. The results are inconsistent with updating models where users perceive updating activities as low friction or high benefit. Instead, these findings align with models in which users face high administrative costs to adjust operational routines, and users adopt hybrid processes only after accumulating information about extreme dangers (Dey et al., 2015; Dissanayake et al., 2022).

These findings imply that disclosing security bugs might be detrimental to inattentive users, who effectively behave as if they ignore available updates. On the one hand, revealing the presence of dangerous security bugs could motivate users to patch their software. But on the other hand, posting information about security holes provides a roadmap for hackers to exploit and increases the attack space and vulnerability among inattentive websites. Moreover, releasing the information increases the externalities across organizations. Software vendors and governments may wish to publicize available software updates while providing less information about the vulnerabilities. We discuss these tradeoffs in conclusion.

The following two sections discuss the literature and setting. Section 3.4 describes the data. Section 3.5 presents the empirical strategy. Section 3.6 analyzes updates and presents statistical evidence about the causes behind the distribution of behavior. Finally, the paper concludes with a discussion of the implications.

3.2 Contributions

A vast literature measures the quantity and quality of IT and its impact on the productivity of organizations (Jorgenson et al., 2006, 2016; Brynjolfsson & Hitt, 2003; Tambe & Hitt, 2012). Researchers have found

persistent differences across organizations in the investments in and returns from IT (Foster et al., 2001; Aral et al., 2006).³ Our prior work (Murciano-Goroff et al., 2021) made a direct link with productivity analysis. It showed that high productivity correlates with high-quality web software, which arises at firms that upgrade more frequently and maintain software close to the frontier. In contrast, in the present study, we link upgrading behavior to a new valuable outcome, cybersecurity vulnerabilities, which has yet to be a focus of prior productivity analysis. Similar to patterns identified in previous research, we find persistent differences across firms in IT investment behavior, which motivates analyzing the determinants of firm heterogeneity in investment in software quality. Unlike that work, we focus analysis on user responsiveness to announcements, the availability of patches, and the causal determinants of the variance in upgrade activity.

Only two prior empirical research studies have examined longitudinal investment in cybersecurity and linked it to outcomes. Li et al. (2021) examined hospital adoption of security software and investment in related activities, while Liu et al. (2017) examined higher education and governance and associated actions. Both link these factors to the propensity to suffer a security incident.⁴ Furthermore, both focus on exogenous organizational features and processes, using cross-sectional variance to infer causal determinants. Our study takes a different approach, gaining causal inference from measuring software upgrades and vulnerabilities directly. In addition, our study looks at a more extended period, a more significant number, and a wider variety of organizations.

We contribute to previous research that examined how the breadth of security vulnerability disclosure affects the rate and volume of hacking attempts (Mitra & Ransbotham, 2015). Analyzing data on patch releases, prior research considered how much time software vendors keep security vulnerabilities secret and whether this could accelerate the development of software updates and security patches. In contrast, our research gains visibility into user responses to releases, analyzing heterogeneity in the timing and extent of a user's updating.

We overlap in addressing a crucial question: Should reports of security vulnerabilities in software be made immediately, or should reports be given only to software vendors to prepare for patches and updates

³Some of this variance can be explained by the differences in returns related to the size and scale of implementations and strategic investments that enable leadership persistence (McElheran, 2015; Bessen & Righi, 2019; Tambe et al., 2020; Zolas et al., 2021).

⁴Li et al. (2021) stresses the returns at organizations that invest in on-premises processes, such as anti-virus, intrusion detection, and authentication. Liu et al. (2017) found behavior consistent with a tradeoff between granting autonomy and flexibility in using information systems and enforcing standardized, organization-wide security protocols—the more complex the computing environment, the higher the returns on centralized governance that limits vulnerabilities.

before being released to the general public? Immediate release is known as “full disclosure,” while the latter is “limited disclosure.” With full disclosure, security professionals can close the security hole (Mookerjee et al., 2011), while limited disclosure provides time for vendors to develop software updates and for users to install those updates. Empirical evidence suggests that full disclosures of security vulnerabilities accelerate attacks on vulnerable software. However, there is no significant difference in the total volume of attacks because of a fully or limited disclosed bug (Mitra & Ransbotham, 2015).

That frames a related question: If vendors receive information about bugs before disclosing the vulnerability to software users, how long should suppliers hold that information before releasing it? Addressing this question requires understanding how delays shape the attacks that exploit the vulnerability. Arora et al. (2008) find that software vendors release software updates at a slower rate than socially optimal. In this analysis, users are portrayed as attentive to security vulnerabilities and compliant in installing software updates when available. However, Arora et al. (2008) allow for some degree of heterogeneity in the uptake of software updates.⁵ While the authors conceptualize that the quality of the updates shapes variance in uptake and endogeneity in updating, they do not estimate the distribution of such conflict from actual investments in upgrades.

We fill this gap with our empirical analysis. Our findings suggest that the number of inattentive organizations should become an essential feature of models analyzing trade-offs. Using the Arora et al. (2010), for example, the presence of a large number of inattentive organizations pushes toward policies with less disclosure. In addition, it suggests software suppliers should care not only about the quality of their software but also about the prevalence of that vulnerability. The number of users of a software version with a bug is an additional moderator that can also accelerate the release of new software updates and decelerate disclosure. Finally, models must account for the spread of information about vulnerabilities among hackers if disclosure does not occur.

Our empirical approach builds on research about optimal patching policies (Dey et al., 2015; Dissanayake et al., 2022). In these analyses, users patch on a regular cycle or in reaction to a cumulative signal of threat levels. A hybrid model uses a regular cycle but varies it with additional information. When the costs of disrupting operations are high, users adopt the regular process and only differ from it in the presence of

⁵For example, suppose users are only responsive to software updates but not to reports of vulnerabilities. In that case, vulnerability disclosures may primarily focus hackers’ attention without accelerating the hardening of defenses. In addition, if users are only responsive to software updates rather than bug disclosures, the optimal disclosure policy may forego revealing any information about security vulnerabilities.

other information about threats if these too are high. Conversely, a hybrid approach can be chosen by a user with sufficiently low disruption costs, who then faces low costs when responding to new information about vulnerabilities. Our finding informs this literature: We find a distribution of user activity inconsistent with models that presume all users perceive updating activities as predominantly low friction or high benefit.

Our study also contributes to the literature investigating software upgrading through the release of incremental versions. Much of it focuses on the incentives of for-profit suppliers that can monetize the upgrades (Leyden, 2018; Comino et al., 2019).⁶ A related line analyzes trade-offs in incentives for upgrading enterprise software in the presence of security breaches (August & Tunca, 2006; August et al., 2014; Kang, 2022).⁷ In contrast, we focus on Apache, a not-for-profit enterprise software supplier. While this supplier is not incentivized to monetize upgrades, it does aspire to satisfy users' requests for secure software and compatibility with other IT processes. Our finding about the distribution of attentive users and slow upgrades alters has implications for future research into how fast a supplier releases a new version.

3.3 Setting

After software is released, users of that software often discover bugs or mistakes in the program. Some of those bugs inhibit the quality of the software's functionality, while other bugs create security vulnerabilities that can be exploited by malicious actors. Therefore, upon learning of a bug, vendors of software typically create and release updates to their software that users can install to fix the bug and mitigate any security vulnerabilities.

Server software is a particular type of computer program that enables users to connect over the Internet and retrieve website content. When an individual visits an organization's website, the individual's computer sends a request to that organization's server. The server processes that request using server software, which

⁶Leyden (2018) analyzes upgrades in the context of iPhone app updates, focusing on incentives to increase the rate of upgrades to monetize higher-quality apps with more frequent sales. Comino et al. (2019) show that these incentives operate differently on iPhones and Android distribution due to the composition of apps and quality control processes in each platform.

⁷Kang (2022) emphasizes user incentives for upgrading enterprise software with many complements and the costs of accounting for such operational complexity. In a for-profit setting, August et al. (2014) investigate optimal trade-offs between the cloud-supported provision of upgrades or on-premise upgrades in the face of heterogeneous user valuation of quality. For-profit firms target their promotions to segments that demand low, medium, or high security, depending on the risks and costs of alternatives. Relatedly, August & Tunca (2006) analyze incentives to patch in both a for-profit and free setting, assuming that upgrade behavior reflects forward-looking incentives but ignores externalities on others. In the for-profit environment, incentives to fix are too low, requiring vendor subsidies to induce optimal behavior. With freeware, the incentives are too low (high) when the risks and costs are minor (significant).

determines which content to send back to the individual. For example, after an individual connects to Amazon.com, the Amazon server software determines which products and prices to display to that individual. After an individual connects to their bank's website, the server software determines who the individual is, what information they should or should not have access to from the bank's database, and what transactions that user should be allowed or not allowed to make.

Because server software plays these important roles in both guarding and transmitting sensitive data, server software vendors have developed procedures for finding and fixing security vulnerabilities. Software vendors typically accept reports from users about bugs and potential security vulnerabilities. These bugs are vetted by teams of security experts. These reported bugs, however, are kept secret from the public, so malicious actors are not tipped off to their existence. When software vendors believe it is prudent to do so, vendors will publicly acknowledge the existence of a particular bug or security vulnerability. Vendors make these acknowledgements when they feel that it is important to warn their users about security risks, either to encourage the users to more carefully monitor their systems or to encourage their users to take mitigating actions such as updating their software. Finally, once software vendors have developed a fix for a bug, the vendor will release a software updated that includes that fix.⁸

This procedure for handling reports of bugs and releasing software updates for handling security vulnerabilities reflects the open-source development process and communities that created the original web servers. Apache is the most popular open-source web server used by websites around the world today. Apache descended from the first web server, built on the new diffusing World Wide Web, released by Tim Berners-Lee in 1991. In 1993, the National Center for Supercomputing Applications (NCSA) at the University of Illinois developed a computer program called the NCSA HTTPd server. The HTTPd server software supported sharing content on the web through the Hypertext Transfer Protocol (HTTP). NCSA made HTTPd available as shareware within academic and research settings, along with the underlying code.

Apache's adoption spread quickly, in part, because the servers did not restrict the usage or modification of the software. Many web admins took advantage by adding improvements as needed or communicating with the lead programmer, Robert McCool, who coordinated adding and releasing of new extensions. When, in

⁸While this is the process that most software vendors hope will occur, in reality, some bugs are discovered and handled outside of this procedure. In particular, some bugs are discovered when a user notices and discusses problems with program without being aware that the problem is indicative of an underlying vulnerability. In those cases, the date at which the bug is discovered and the date at which the vendor publicly acknowledges the vulnerability may be the same. In addition, the vendor may acknowledge the bug before a software update is ready to be released.

the spring of 1994, McCool left the University to become one of the first ten employees of the newly founded Netscape, the development of the web server software fragmented with eight different teams working on eight distinct versions of the software. In 1995, the eight teams decided to coordinate their efforts into one server known as Apache (ostensibly because it was “a patchy web server”). The University of Illinois then transferred the server software development to the Apache organization without licensing or restrictions. Apache grew in popularity as the commercial internet grew, becoming widely used.

Apache continues to be developed by volunteers. Apache’s growth and deployment rely upon donations and a community of technically skilled users who provide new features motivated by intrinsic and extrinsic rewards. Yet, Apache is affiliated with many revenue-generating economic activities. For example, there is a significant labor market for Apache programmers, administrators, and third-party consultants.

Apache, and the organization that supports it, the Apache Foundation, follow the common process for handling reports of bugs and security vulnerabilities: When users of Apache software discover a bug, they report that to the Apache Foundation. We refer to these bugs as *reported* bugs. These bugs are analyzed by Apache’s core team as well as the National Institute of Standards and Technology (NIST) and then scored based on the potential of that bug to create harm for users.⁹ The bugs that score highest for their impact on the security of the Apache software we call *severe security* bugs. Eventually, the team at Apache releases information about those bugs. We call those *publicly acknowledged bugs*. Finally, Apache releases software updates that fix the reported bugs. In order to decrease the possibility that malicious actors will exploit a bug, Apache typically publicly acknowledges a bug at the same time that they release a new software update that fixes that bug.¹⁰

The setting informs our empirical strategy. We envision server administrators supporting Apache server software as part of their job to maintain IT infrastructure within their respective organizations.¹¹ Drawing on prior research, we posit that each organization has a baseline regarding security and software updating

⁹The National Institute of Standards and Technology (NIST) maintains the National Vulnerabilities Database (NVD). When a bug is reported then it is entered in the NVD and a score is computed based on the Common Vulnerability Scoring System (CVSS). <https://nvd.nist.gov/vuln-metrics/cvss#>

¹⁰The statement of this process is available on the website of the Apache Software Foundation Security Team at <https://www.apache.org/security/>. To the best of our knowledge, the overview of this process has stayed the same since the early days of the Apache Software Foundation.

¹¹IT decisions are made by individual server administrators within organizational contexts. Decisions about installing, updating, and changing server software is an interplay of both individual and organizational-level factors. In this paper, we discuss server administrators and organizations interchangeably as decision-makers regarding server software.

practices (Dey et al., 2015; Dissanayake et al., 2022). For some firms, this baseline may be to keep their software at the technological and security frontier by immediately installing any updates and security patches as they become available. These firms we dub the *frontier chasers*. For other firms, this baseline may be to largely ignore software updates and rarely install software updates. These firms we call the *do-nothing* firms. Whether a firm is one of these extreme categories may be a function of many complex factors, including organizational capital, management practices, and cybersecurity human capital. Furthermore, the organization's geographic location may influence local labor market conditions and the costs of hiring labor to operate a process.

In between the extremes of the frontier chasers and do-nothing firms, there are firms which intermittently install software updates and security patches. In addition to decision not to chase the frontier or forgo updates entirely, these *do-something* firms must decide under what circumstances they will respond to released software updates or new information about the presence of security bugs in their software. This responsiveness to information and announcements about cybersecurity vulnerabilities is a function of similar factors to those that influence the initial decision regarding the baseline software updating cycle.¹²

Both costs and benefits influence the baseline probability that an organization updates their server software as well as the responsiveness of organizations to the release of software updates and severe security vulnerability announcements.

The costs affiliated with software updates may derive from frictions that the organizations encounter when updating. For example, if administrators must multitask across a large set of operational needs, then updating the Apache server may involve time and high opportunity costs. If an organization has been operating for a longer time, administrators may inherit configurations that accumulate myopic decisions from the past. If, however, a server is hosted on a cloud provider, such as Amazon Web Services (AWS), administrators may find it less costly to manage the transition from one software version to another. In addition, the technical complexity of a website is likely to be associated with organizations being less responsive to changes in known vulnerabilities in software. Websites that depend on large numbers of interconnected or interdependent technologies are more challenging to update and often require more forethought and planning to avoid introducing incompatibilities when transitioning software versions. Because of that friction, websites that are

¹²This conceptualization parallels the rational inattention framework applied to organizations (Matějka & McKay, 2015). Much like that literature, organizations have a baseline propensity towards an action, updating. In addition, each organization has factors that influence their attentiveness to new information, namely responsiveness to the presence of bugs in their software and the release of new versions of software.

more technically complex are also less likely to respond quickly to the discovery of security vulnerabilities.

In addition to the costs, different users may experience various perceived benefits from installing software updates. These may include direct benefits, such as faster loading times for their site or the ability to display and deploy frontier web apps that engage users. For example, deploying servers to support Web 2.0 applications may motivate upgrades. These updates may also decrease the probability of security breaches. The prominence of an organization's homepage is also likely to increase the benefits from being responsive to the discovery of cybersecurity vulnerabilities. Highly trafficked websites are also more likely to be targeted by hackers. Finally, the ability of a website to conduct e-commerce and financial transactions also correlates with the benefits from updating after the announcement of cybersecurity vulnerabilities in software. Cybercriminals are likely to target sites that collect and store financial records, so server administrators at organizations with such sensitive data are also more likely to carefully watch for updates patching security vulnerabilities.

3.4 Data

The centerpiece of our data is the largest dataset ever collected on business web server use between 2000 and 2018, which we compiled for hundreds of thousands of medium to large organizations in the United States. We keep the observations associated with using Apache server software and supplement this data with organization and website characteristics from various sources. We summarize all data sources in Table 3.1 . In this section, we discuss our sample construction in more detail. Table 3.1 organizes the data sources in two panels, corresponding with our empirical analysis on two related aspects of organizations' software use and updating decisions. Panel (a) describes the data sources contributing to the analysis of the extent of security vulnerability. Panel (b) illustrates the data source contributing to analyzing the determinants of organizations' updating decisions.

Row (a)i of Table 3.1 shows our full data on almost five million months of observed usage of Apache server software for 150,854 U.S. organizations between 2000 and 2018. The data on the use of Apache server software and the version of Apache used comes from analyzing the server headers of U.S. organization websites collected by the Internet Archive (IA). When a browser visits a website, the visitor's computer connects with the hosting website. Upon connecting, the web server responds with the web page's content and data, known as the server header. By default, server headers contain the name of the responding web

server's vendor (e.g., Apache, Microsoft IIS, Nginx) and the version used (e.g., Apache 2.2.3). The IA is a nonprofit organization that stores snapshots of websites. Their computers regularly connect to large numbers of websites and record both the content and the server header. Therefore, the observation for this data is organization-month. Still, AI does not capture every website monthly, so server software usage in uncaptured months is unobserved.

Rows (a)ii and (a)iii show our data on 158 known security vulnerabilities of Apache server software which we merge with our Apache usage data. The data are publicly available from the Apache Software Foundation and the National Vulnerability Database (NVD). Combining data, we cover the degree of severity and the version(s) affected by each security vulnerability, the date the vulnerability was reported to the Apache Foundation, the date it was made public, and the date of release of new versions that fix the vulnerability. We also observed 94 vulnerabilities where someone was credited with reporting the vulnerability.

Row (a)iv shows a broad cross-sectional sample of organization characteristics which we also merge with our usage data. This data comes from Mint Global by Bureau van Dijk and includes characteristics such as an organization's location, industry, number of employees, and revenue. We examine only organizations based in the U.S. who have at least 50 employees. Merging this with our data on usage results in an overlap of 145,515 organizations.

Row (b)i shows a restricted sample of around three million months of observed usage of Apache server software of 70,092 U.S. organizations between 2000 and 2018. We use this sample to analyze the determinants of those organizations' Apache server software updating decisions. For several reasons, this sample is smaller than our full sample shown in Row (a)i. First, we include in this sample only organizations that were, on average, captured by the IA every three months or more frequently. This ensures that when an organization updates its Apache server software, we can precisely measure the time-to-update. Second, we include only organizations with at least ten months of observed Apache usage. As Apache makes a new minor release every 4.5 months on average, the restriction ensures that we observe organizations' updating decisions over at least two average software release cycles. We also drop an observation if the time gap between the observations' captures and the following observation of the organization is more than 4.5 months. Moreover, we drop some observations associated with each organization's first observed Apache updating cycle, and adjust for censoring.¹³

¹³For example, the time-to-update corresponding to the first observed updating cycle is left-censored, creating problems in our empirical analysis. Similarly, we also drop the last observation of Apache usage of each organization due to right-censoring.

Table 3.1: Summary of sources of data

Data		Description	Frequency	Source
<i>(a) Security vulnerability of Apache users</i>				
i.	Organizations' Apache web server usage	A version of Apache web server of 150,854 homepages of US organizations	Monthly between Jan 1, 2000, and August 31, 2018; 4,863,383 observations in total; on average, each organization has 32 observations ¹⁴	The Internet Archive
ii.	Apache web server security vulnerabilities	Security vulnerability reported to the Apache Software Foundation; variables include the version(s) affected, date of reporting, the date the vulnerability was made public, date of release and version of the fix, and (when available) who was credited with the discovery/reporting of the vulnerability	158 unique vulnerabilities; 5,192 vulnerability-version pairs; for vulnerabilities reported before or on August 31, 2018	The Apache Foundation
iii.	Severity of security vulnerabilities	Severity rating of "high," "medium," or "low" for each security vulnerability	28 vulnerabilities were rated "high," 123 were rated "medium," and 7 were rated "low"	National Vulnerability Database (NVD)
iv.	Organization characteristics	214,199 organizations in the US with at least 50 employees; organizations are identified by their homepages; variables include location, NAICS code, number of employees and revenue	A cross-sectional snapshot of the database on August 28, 2018; merging this data with (a) results in an overlap of 145,515 organizations	Mint Global by Bureau van Dijk
<i>(b) Upgrading behavior of Apache users</i>				
i.	Organizations' Apache web server usage (frequent captures, uncensored)	A version of the Apache web server of 70,092 homepages of US organizations that were frequently captured by the Internet Archive; ¹⁵ each observation is not from a left-censored updating cycle and is not the last observation of Apache usage of the organization in the raw data ¹⁶	Monthly between Jan 1, 2000, and August 31, 2018; 3,104,455 observations in total; on average, each organization has 44 observations. ¹⁷	The Internet Archive

¹⁴ If an organization's website only became active after Jan 1, 2000, or went offline before Dec 31, 2018, the months before the website went online and the months after the website went offline do not have observations. For some organizations, the Internet Archive did not capture the active websites every month. Therefore, those months do not have observations. In addition, some organizations have used other web servers (such as Microsoft's IIS or Nginx) during the period. Those observations are not included.

¹⁵We only include organizations that were, on average, captured by the Internet Archive every three months or more frequently. Moreover, we only have organizations with at least ten monthly observations of Apache web server usage. Doing so allows us to observe organizations' upgrading decisions over a duration equal to at least two average release cycles of the Apache web server software, as the average release cycle of an Apache web server is 4.5 months. We then drop an observation if the time gap between the observation and the following observation of the organization is more than 4.5 months apart.

¹⁶Our hazard models account for time-dependent covariates that vary monthly. Left-censoring due to the first data capture is problematic for all observations associated with the left-censored updating cycles. Right-censoring due to the last data capture is only difficult for the final observation of the organization's Apache usage. We, therefore, drop the problematic observations.

¹⁷Footnote 14 applies.

Table 3.1 (Continued): Summary of sources of data

ii.	Apache web server security vulnerabilities	Identical to (a)ii	Similar to (a)ii	The Apache Foundation
iii.	Severity of security vulnerabilities	Identical to (a)iii	Similar to (a)iii	National Vulnerability Database (NVD)
iv.	Apache web server version release dates	The release date of each minor version (e.g., 2.4.1) of the Apache web server software	132 releases of minor versions between 1996 and 2019; those minor versions belong to eight different major versions (e.g., 2.4)	Authors' compilation
v.	Websites' Amazon Web Service (AWS) use	Whether an organization's web server using an IP address associated with AWS	A snapshot of all current IP addresses associated with AWS taken on March 25, 2020	AWS
vi.	Alexa web traffic ranking	Ranking of top one million websites by traffic	Yearly between 2010 and 2018	Alexa
vii.	Websites' technology use	Technologies used in building 24,263 organizations' homepages; 45 technology categories, including analytics, e-commerce, and web frameworks; 531 technologies including jQuery, Google Tag Manager, and WordPress	Captured between 2016 and 2018, merging the organizations in this data with (b)i results in an overlap of 11,824 organizations	HTTP Archive
viii.	Organization characteristics	Identical to (a)iv	Merging the organizations in this data with the intersection of (b)i and (b)vii results in an overlap of 11,824 organizations	Mint Global by Bureau van Dijk
ix.	Organization characteristics (public firms)	US public firms' characteristics; variables include total assets, capital expenditure, depreciation and amortization, employees, cash flow, and net income	Yearly panel between 2000 and 2018; merging the organizations in this data with the intersection of (b)i and (b)vii results in an overlap of 1,216 organizations	Compustat

The table shows the emergence of our final data set. Rows (b)ii and (b)iii show our data on the known security vulnerabilities of Apache server software which we merge with the Apache usage data in Row (b)i. These data are identical to those shown in Rows (a)ii and (a)iii. Row (b)iv shows our data on the release dates of Apache server software versions we compiled between 1996 and 2019. This data allows us to examine whether the releases of new versions affect organizations' updating decisions. For example, the Apache Foundation often simultaneously maintains multiple significant versions (e.g., Apache 1.3, Apache 2.2) of the Apache server software. As a result, we observe releases of new minor versions (e.g., Apache 1.3.27, Apache 2.2.3) corresponding to each major version on average every 4.5 months.

To further enable the analysis of the determinants of updating decisions, we have collected data on websites' characteristics. First, we took a snapshot of all IP addresses associated with Amazon Web Services (AWS), a significant provider of cloud computing services, in 2020. This data is shown in Row (b)v and allows us to examine whether an organization's website ever used an IP address associated with AWS. This serves as a proxy for whether the organization likely used the cloud to host its web server.¹⁸ We obtained the Alexa traffic rankings of the top one million websites between 2010 and 2018, shown in Row (b)vi. This allows us to measure the intensity of usage of organizations' websites.

We also obtained data for 24,263 organizations and the technologies used in building those organizations' websites between 2016 and 2018. This data is shown in Row (b)vii. We collected this data from the HTTP Archive, an organization that takes snapshots of websites and analyzes them for their technical attributes. Specifically, the HTTP Archive runs an open-source tool developed by the company Wappalyzer on each website, which flags the technologies being used. The data capture both technology categories that contribute to the basic architecture of websites, such as web frameworks and databases, and technology categories that support monetization and e-commerce, such as marketing automation and payment processors. This data helps us construct proxies for websites' technical complexity and the organizations' intent to monetize their websites. We conceptualize the number of web technologies as being correlated with the technical complexity of the website and, thus, the technical complexity when attempting to update the site's web server. We think monetization tools indicate that security vulnerabilities would be particularly damaging and that significant benefits exist to keep the website secure. Merging this data with the usage data in Row (b)i results in an overlap of 11,824 organizations.

¹⁸It has some measurement error because the IP addresses associated with the AWS could have changed over time.

We use two distinct data sources of organization characteristics to analyze determinants for updating. The first data source in Row (b)viii is identical to Row (a)iv, a broad cross-sectional sample of organizations with few organization characteristics. However, as it does not have a time dimension, it does not allow us to study how changes in an organization's financials affect its software updating decisions across time. Merging this data with the intersection of our data on Apache web server usage in Row (b)i and our data on website technologies in Row (b)vii results in an overlap of 11,824 organizations.

Our second data source of organization characteristics in Row (b)ix is from Compustat and covers the full panel of U.S. public firms every year. This data contains a wide range of organization characteristics such as the total assets, capital expenditure, cash flow, and income, allowing us to examine organization characteristics that might affect updating decisions deeply. The data's time dimension also will enable us to study the effects of changes in financials within organizations. A drawback to using this data source is the reduction of sample size. Merging this data with server usage in Row (b)i and our data on website technologies in Row (b)vii results in only an overlap of 1,216 organizations. Given that the two sources of organization characteristics each have its strength, we keep both samples in our analysis.

Tables 3.2 and 3.3 show summary statistics for the final samples. Table 3.2 shows summary statistics for our broad sample of organizations, while Table 3.3 shows those for public firms. Appendix Table C.1 shows detailed variable definitions. Each observation in the two tables represents one month in an updating cycle of an organization that used Apache server software. During that month, the server software's and the organization's characteristics could vary. For example, when a severe security vulnerability of the minor version of Apache (e.g., Apache 2.2.3) organization i used became publicly known in month t , the binary variable $hasSevereBugs_{it}$ switches on in that month and stays on in the following months until organization i makes an update to another version that fixes the vulnerability.

Table 3.2: Summary statistics of the broad sample of organizations

Variable	Obs.	Mean	SD	Min	Max
<i>timeToUpdateMonthEnd_{it}</i>	665,360	18,978	18.508	1	195
<i>updated_{it}</i>	665,360	0.052	0.221	0	1
<i>hasSevereBugs_{it}</i>	665,360	0.702	0.457	0	1
<i>hasAlexaRank2010_i</i>	665,360	0.714	0.452	0	1
<i>aws_{it}</i>	665,360	0.036	0.187	0	1
<i>techCategories_i</i>	665,360	10.047	3.849	1	22
<i>techs_i</i>	665,360	13.520	6.104	1	38
<i>monetizationTechs_i</i>	665,360	2.842	2.240	0	15
<i>newVersionAvailable_{it}</i>	665,360	0.965	0.184	0	1
<i>newMinorVersionAvailable_{it}</i>	665,360	0.879	0.325	0	1
<i>infrequentUpdater_i</i>	594,004	0.386	0.487	0	1
<i>frequentUpdater_i</i>	594,004	0.240	0.427	0	1
<i>logEmployment_i</i>	665,360	6.003	1.568	4.007	14.715
<i>logRevenue_i</i>	665,360	9.941	3.136	0	19.169

Notes: Each observation represents one month in an updating cycle of an organization that used Apache server software. The number of organizations in this data is 11,824. The event is making an update of the version of Apache server software an organization used. The number of events is 34,364. The number of updating cycles is 51,312. $51,312 - 34,364 = 16,948$ updating cycles are right-censored.

Table 3.3: Summary statistics of the sample of public firms

Variable	Obs.	Mean	SD	Min	Max
<i>timeToUpdateMonthEnd_{it}</i>	64,204	18.760	17,827	1	155
<i>updated_{it}</i>	64,204	0.048	0.213	0	1
<i>hasSevereBugs_{it}</i>	64,204	0.728	0.445	0	1
<i>hasAlexaRank2010_i</i>	64,204	0.768	0.422	0	1
<i>aws_{it}</i>	64,204	0.057	0.231	0	1
<i>techCategories_i</i>	64,204	10.412	3.767	1	22
<i>techs_i</i>	64,204	14.271	6.290	1	36
<i>monetizationTechs_i</i>	64,204	3.194	2.260	0	13
<i>newVersionAvailable_{it}</i>	64,204	0.971	0.169	0	1
<i>newMinorVersionAvailable_{it}</i>	64,204	0.896	0.305	0	1
<i>infrequentUpdater_i</i>	56,504	0.405	0.491	0	1
<i>frequentUpdater_i</i>	56,504	0.220	0.414	0	1
<i>logEmployment_{iy}</i>	44,575	7.176	2.940	0	13.090
<i>logCapx_{iy}</i>	44,575	9.448	3.232	0	17.015
<i>logTotalAssets_{iy}</i>	44,575	20.002	4.112	0	28.424
<i>logDepreciation_{iy}</i>	44,575	16.207	4.269	0	24.029
<i>logIncome_{iy}</i>	44,575	7.176	15.908	-23.779	24.455
<i>logCashflow_{iy}</i>	43,210	0.035	0.171	-2.534	1.652

Notes: The number of organizations in this data is 1,216. These organizations have been public for at least one year between 2000 and 2018. The number of events is 3,055. The number of updating cycles is 4,716. $4,716 - 3,055 = 1,661$ updating cycles are right-censored. The rest of the notes of Table 3.2 applies.

3.5 Empirical Strategy

Our empirical analysis has three components. The first component is to document the extent of the security vulnerabilities in the Apache server software being used by organizations over time. We use the data with almost five million observations from 150,854 organizations, described in Panel (a) of Table 3.1, to perform this analysis. We examine the proportion of organizations using Apache server software with severe security vulnerabilities and illustrate changes over time and heterogeneity across the industry, geography, and firm size. We present the visual evidence in the Results section.

The second component seeks to understand what factors are associated with organizations' baseline server software updating decisions. We perform this analysis on our broad sample of 11,824 organizations (shown in Table 3.2) and our sample of 1,216 public firms (shown in Table 3.3). The start time is when an organization first adopts a particular version of the Apache server software, and the event of interest is when the organization updates from that version. The month t subscript is associated with each observation. We construct the time-to-update as the outcome variable. The time-to-update variable is always positive and highly left-skewed for both samples. Ordinary least squares (OLS) and probit regression models are inappropriate for analyzing an outcome variable with this distribution. Our preferred empirical strategy is based on survival models, a standard approach for analyzing time-to-event outcome variables.

Our empirical strategy is based on the Cox proportional hazards analysis. The standard Cox proportional hazards regression model is as follows:

$$h(t) = h_0(t) \exp(b_1 X_1 + b_2 X_2 + \dots + b_j X_j). \quad (3.1)$$

$h(t)$ is the hazard function, representing the expected monthly updates given that the version has continued usage for t months. $h_0(t)$ is the baseline hazard and represents the hazard when all the independent variables $X_1, \dots, X_j$ are equal to zero.

The standard Cox model relies on several assumptions our setting does not satisfy. Importantly, the standard model assumes that there are no time-varying covariates. In our setting, several covariates, such as the binary variable *hasSevereBugs_{it}* vary over time. The modification to the standard model is simple conceptually, where we include a time subscript:

$$h(t) = h_0(t) \exp(b_1 X_{1t} + b_2 X_{2t} + \dots + b_j X_{jt}). \quad (3.2)$$

Our main independent variables of interest include *hasSevereBugs_{it}* (whether the version in use had known severe security vulnerabilities), *hasAlexaRank2010_i* (whether the website had high traffic), and *newVersionAvailable_{it}* (whether newer versions of the Apache server software was available). The independent variables could also include organization characteristics, such as location, industry, and the number of employees.¹⁹

While the standard model allows us to explain differences in the rates that organization update their server software using observable organization and environmental attributes, there is the possibility that unobserved organizational routines around updating drive these differences. Specifically, the model in Equation 3.2 assumes that the baseline hazard $h_0(t)$ is identical across organizations. We discover considerable heterogeneity across the organization in updating decisions that cannot be explained by observable covariates. This raises a concern, since not accounting for the differences in the organizations' baseline hazard, or propensity in updating, would give rise to a reverse causality issue. For example, organizations that lag in updating their Apache server software are more likely to use older versions with more known severe security vulnerabilities. Suppose that we have assumed that all organizations have the same baseline hazard and estimated a Cox model. The estimated coefficient on *hasSevereBugs_{it}* would be biased downward, meaning that we are more likely to find a negative effect of severe security vulnerabilities on the organizations' decisions to update.

Therefore, in order to control for different organizations' different baseline hazards, we also use estimate a stratified Cox model:

$$h_i(t) = h_{0i}(t) \exp(b_1 X_{1t} + b_2 X_{2t} + \cdots + b_j X_{jt}), \quad (3.3)$$

where the i subscript on the hazard functions denotes stratum organization i .

The third component of our analysis is to understand what factors are associated with organizations being more or less responsive to discovery of security vulnerabilities and availability of new software versions. We do this by modifying the stratified Cox model in Equation 3.3 to include interactions between organization and

¹⁹The implementation involves considerable manipulation of the data. To accommodate time-varying covariates, we need to shape the data so that each observation corresponds to a different combination of covariate values that change over time. For example, suppose an updating cycle consists of three months. If there are no time-varying covariates, this updating cycle will make one Cox analysis observation where the outcome variable time-to-update equals 3. If there are covariates that vary at the monthly frequency, then this updating cycle would make three separate observations. We also need to use two distinct variables to keep track of the time-to-update. The first observation has a starting time-to-update at month 0 and an ending time-to-update at month 1. It is tied to the combination of covariate values in the first month. The second observation has a starting time-to-update at month 1 and an ending time-to-update at month 2. It is tied to the combination of covariate values in the second month. The third observation has a starting time-to-update at month 2 ending time-to-update at month 3. It is tied to the combination of covariate values in the third month. The event equals 1 ("updated") only for the third observation. As the time-varying covariates in our data vary at a monthly frequency, our data has been shaped so that each observation represents one month in an updating cycle of an organization.

environmental attributes and the discovery of security vulnerabilities. Since the stratified Cox model absorbs the differences in the baseline rate of updating, the interaction terms isolate the factors most associated with differences in responsiveness to software updates and severe vulnerabilities.

We think of the discovery and reporting of severe security vulnerabilities as quasi-random and exogenous to the updating decisions of any individual organization in our data. Though testing exogeneity assumptions formally is generally difficult, our assumption in this setting is intuitive. Developers and IT professionals discover bugs in the software when they interact with it; for example, when they use the software to build applications or develop new features of the software itself—the chances of finding bugs increase when the total interaction with the software increases. However, as many individuals and organizations around the work worldwide on the Apache web server, changes in activities of any particular organization (other than the Apache Foundation) are tiny compared to the total amount of interaction. To confirm that individual organizations’ actions do not have an outsized impact on discovering security vulnerabilities in the Apache server software, we inspect our data on who was credited with discovering or reporting each vulnerability to the Apache Software Foundation. Other than the staff at the Apache Software Foundation Security Team, we find almost no overlap in the names and organizational affiliations of the reporters.

Our main independent variable of interest *hasSevereBugs_{it}* captures whether the version of the software in use had publicly known severe security vulnerabilities, whose timing is not exogenous. We do not use a variable constructed from the reporting date as our main independent variable because most reporting occurred privately. It would be implausible that many organizations in our data would become aware of the vulnerabilities and respond. Still, it was highly correlated when a security vulnerability was discovered and publicized. The two dates are sometimes identical if the discovery started from a conversation on a public online forum. There is a remaining concern about endogeneity because the Apache Foundation often had a choice about when to disclose the vulnerability publicly. The date is usually close to the release of new versions that offer fixes. To overcome the endogeneity concern, we include two additional variables in our preferred specifications of the Cox model. *newVersionAvailable_{it}* captures whether newer versions of Apache server software to what organization *i* was using in month *t*, major or minor, was available. *newMinorVersionAvailable_{it}* captures whether a newer minor version of Apache server software was available. Including these variables also helps us to investigate how much organizations responded to vulnerabilities versus releases of new versions.

With all of the Cox models used in our analysis, we have data censoring due to the start and end of our

sample. The first updating cycle of each organization is left-censored because the usage started before the first capture of use in the data. The observed time-to-update for each monthly observation in the left-censored updating cycles is a lower bound of the actual time-to-update. Because this creates problems in measuring the time-to-update, we have dropped all observations belonging to the first updating cycle of each organization. The last updating cycle of each organization is right-censored because the usage might have ended after the final capture of use in the data. The time-to-update for each monthly observation is measured without problem because we know when the previous update happened. We also know that there was no update until the last month that we observed. However, we cannot determine whether the organization made an update right after the last month observed, so the variable $updated_{it}$ has a missing value for the last observation of the last updating cycle of the organization. Therefore, we also drop each organization's last observation in our survival analysis.

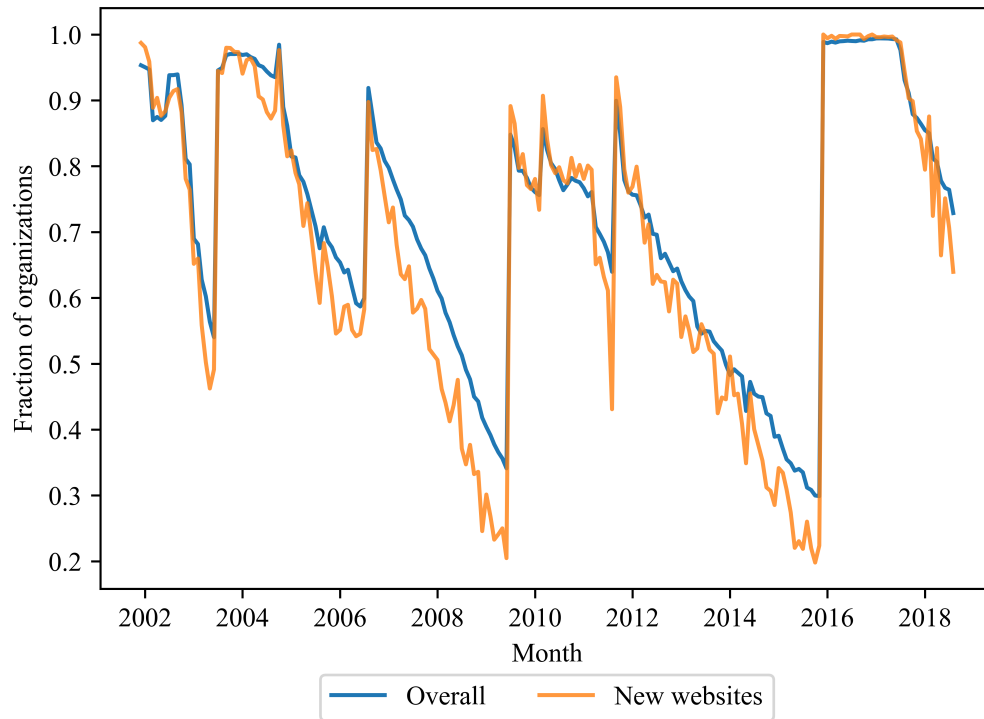
3.6 Results

3.6.1 Usage of Server Software with Security Vulnerabilities

Figure 3.1 shows the fraction of organizations using Apache versions with known severe security vulnerabilities. In this figure, we date the discovery of each severe security vulnerability by the date that the vulnerability was *reported to* the Apache Foundation. While many of these vulnerabilities were unknown beyond a small group of professionals at the moment of their discoveries, we note that at any moment during our sample period, a substantial fraction of organizations used versions that had known severe security vulnerabilities that could have been exploited. The fraction was the smallest in November 2015 at around 30% and shot to the highest at almost 100% in December 2015 with the discovery of a critical security vulnerability in a module that assigns content metadata to the content selected for an HTTP response. New websites that just went online did slightly better than websites in the full sample. This is intuitive because new websites likely had a smaller technology burden than older websites which had to remain compatible with the other technologies they already used. As a result, new websites could adopt more recent versions of the Apache software, which usually had fewer known security vulnerabilities.

Figure 3.2 shows the fraction of organizations using Apache versions with severe security vulnerabilities that were publicly known. This figure represents a more alarming picture of the extent of the security threat of using compromised versions of Apache than that depicted in Figure 1. For this figure, we only determine

Figure 3.1: Fraction of organizations using Apache versions with severe security vulnerabilities

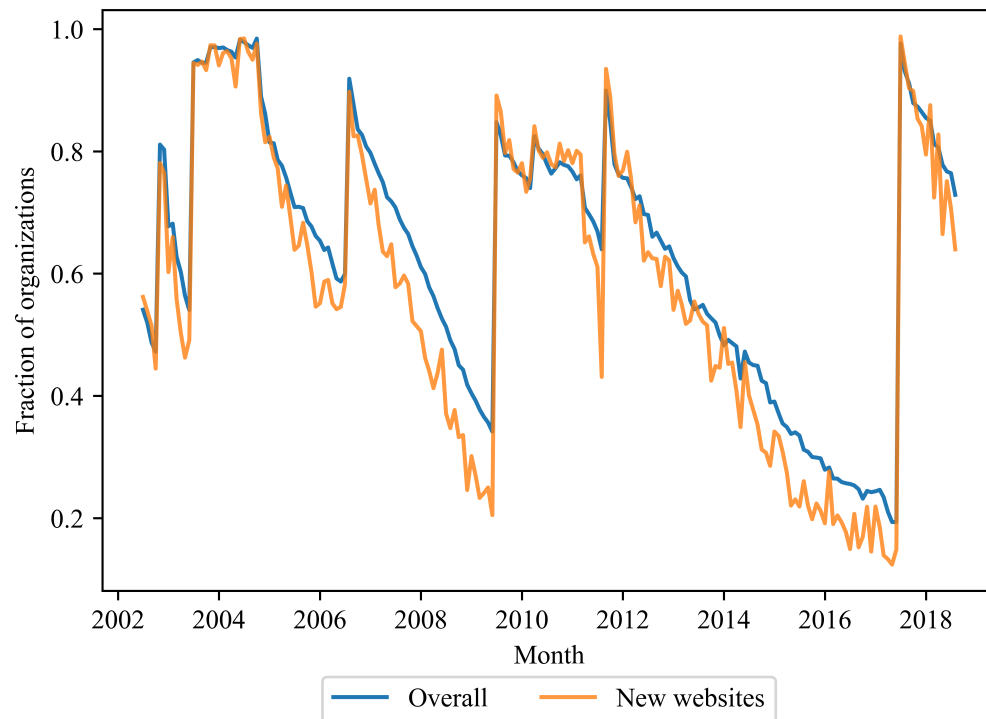


Notes: This figure plots the number of organizations in our sample between December 2001 and August 2018. The determination of when a severe security vulnerability was discovered is by the time the vulnerability was *reported to* the Apache Software Foundation. The first severe vulnerability report in our data is from November 2001. A new website is the first capture of a website by the Internet Archive between Jan 1, 2000, and August 31, 2018. Only vulnerabilities rated to have “high” severity by the National Vulnerability Database (NVD) are included.

the organizations’ use of Apache to be problematic when the vulnerabilities were publicly known and easily exploitable by malicious parties. Nevertheless, the fraction of organizations using Apache versions with publicly known severe security vulnerabilities was above 19% in all months we observed. The fraction peaked in October 2004 at 98% when several severe security vulnerabilities affecting Apache 2.0.x became public in the latter half of 2004. Again, new websites were better at adopting the more up-to-date, secure versions, but not by much. Besides a smaller technology burden, new users could be more discerning and aware of recently announced security vulnerabilities as they had to select a software product actively.

Figure 3.3 shows the fraction of organizations using Apache versions with severe security vulnerabilities that were publicly known and already fixed in newer versions of the Apache server software. While the previous figure illustrates the extent of the exploitability of vulnerable Apache web servers, this figure shows the extent of inaction, either due to the high costs of updating to newer version fixes due to inattentiveness.

Figure 3.2: Fraction of organizations using Apache versions with severe security vulnerabilities that were publicly known

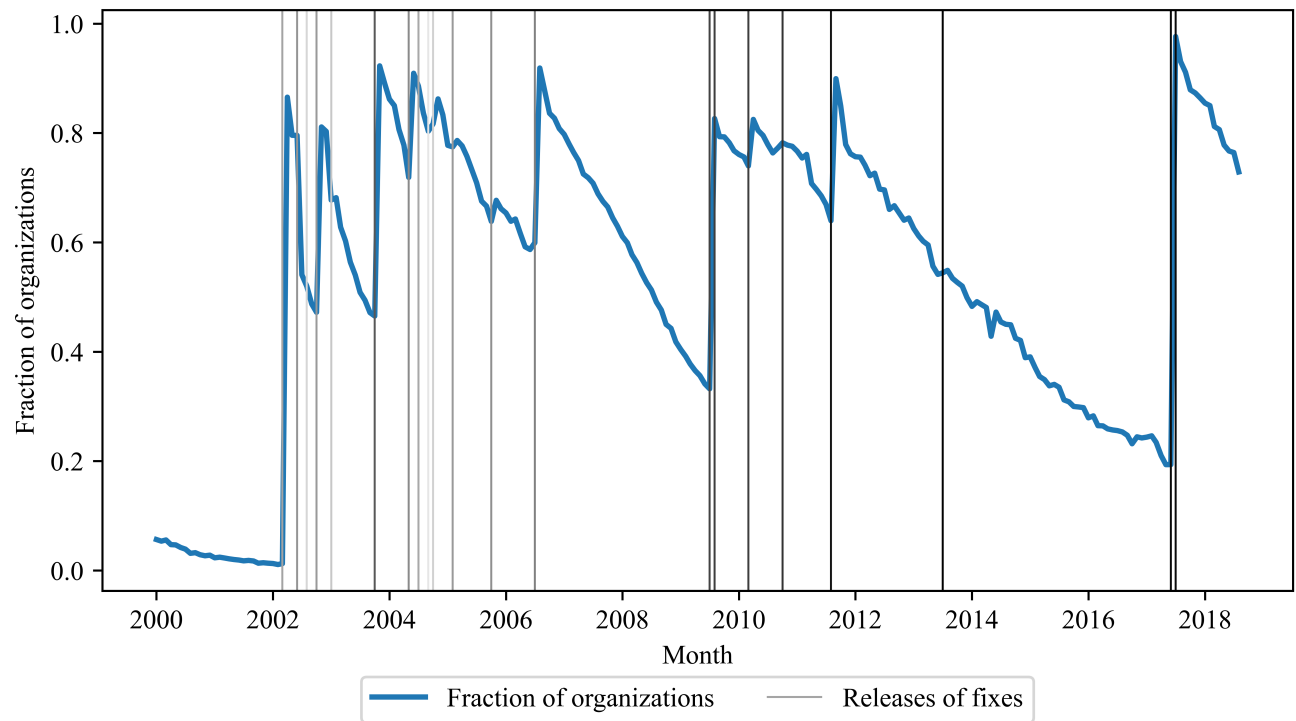


Notes: This figure plots the fraction of organizations in our entire sample between July 2002 and August 2018. The determination of when a severe security vulnerability became public is by the Apache Software Foundation. The first severe vulnerability that became public in our data was in June 2002. The rest of the notes in Figure 3.1 apply.

Except for the early period until March 2002, in all months that we observed, more than 19% of organizations used Apache versions with severe security vulnerabilities already fixed in newer versions. Those 19% represent an approximate estimate of the most *inattentive* users. In this figure, we also plot the release dates of the fixes, represented by the vertical grey lines. This allows us to gauge organizations' level and speed of response to those fixes.

One immediately recognizes that the response was slow and unthorough. Quite a significant fraction of organizations did not respond to releases simultaneously. For example, in May 2006, the Apache Foundation released Apache 1.3.35, Apache 2.0.58, and Apache 2.2.2, which fixed a vulnerability that allowed remote attackers to inject arbitrary web scripts. However, the updating of those releases was slow. The fraction of organizations using problematic versions of Apache peaked at 92% after the release, but the rate declined to only 1.7% per month for the next three years. By mid-2009, over 30% of organizations operated the

Figure 3.3: Fraction of organizations using Apache versions with severe security vulnerabilities that were already fixed in new releases



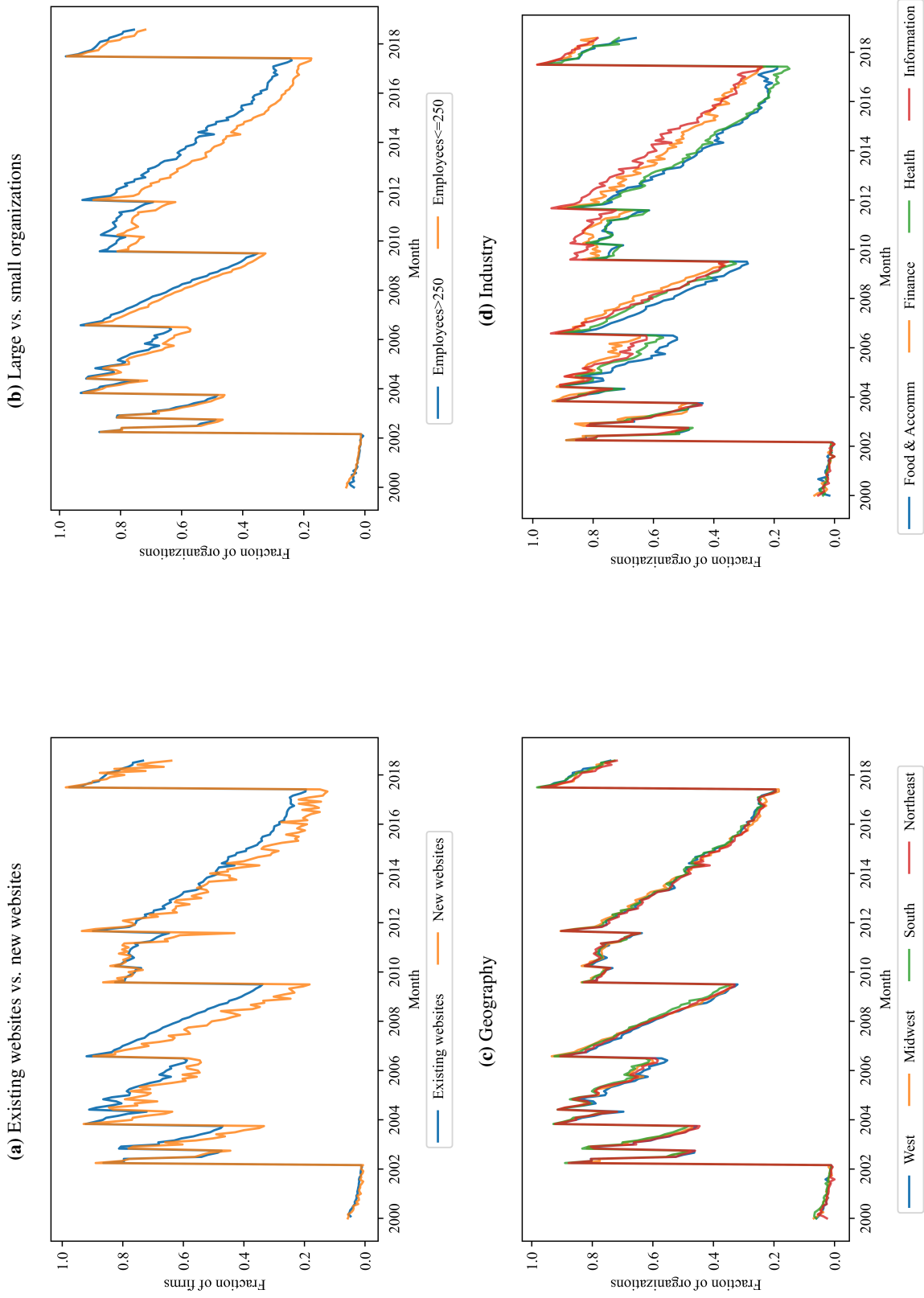
Notes: This figure plots the fraction of organizations in our entire sample between Jan 2000 and August 2018. The determination of when a severe security vulnerability was fixed is by the Apache Software Foundation. The first severe vulnerability fix release date in our data is September 1998. The rest of the notes in Figure 3.1 apply.

affected versions of Apache server software.

Figure 3.4 explores the heterogeneity in these patterns across organizations and websites with different characteristics. We discover some heterogeneity. However, the overall finding from these plots is that compromised Apache was prevalent across different stages of the development of websites and organizations, large or small, in different states and industries.

We split our sample in Panel (a) on new and existing websites. If new users were more discerning when actively selecting a software product, they might select newer versions that fixed well-known severe vulnerabilities and avoid versions with those vulnerabilities. Indeed, in the plot, we show that a lower fraction of new websites installed the problematic versions of the server software relative to existing sites.

Figure 3.4: Heterogeneity in the use of Apache versions with severe security vulnerabilities that were already fixed in new releases



Notes: Panels (b), (c), and (d) include observations from organizations that have their information in the Mint Global database. The number of such organizations is 145,515 (96% of the total sample). The classification of industries is by two-digit NAICS code. The rest of the notes of Figure 3.3 apply.

Panel (b) examines the differences based on the organization's size. For example, smaller organizations might be more agile and capable of updating their server software faster due to a smaller coordination cost across their IT-related departments. Indeed, the plot shows that larger organizations more frequently maintained vulnerable software even when fixes were available, whereas smaller organizations were more inclined to switch to newer versions with fixes.

Panel (c) breaks down the organizations by the geographic region of the organization's headquarters. If different geographic regions have different levels of attentiveness regarding security issues in software—for example, if Silicon Valley-based firms more frequently hear about security bugs through informal networks—we might expect geographic variation in using vulnerable versus secure software. Furthermore, if the local labor market for security and engineering professions differs across regions, this could influence software updating decisions. In the plot, however, we find a relatively slight variation in software usage with and without severe security vulnerabilities across geography.

Panel (d) displays the variation across industries. We find some substantive differences in the software updating decisions. For example, organizations operating in the Health and Food & accommodation industries less frequently install server software with severe vulnerabilities than those operating in the Finance and the broadly defined Information industries.

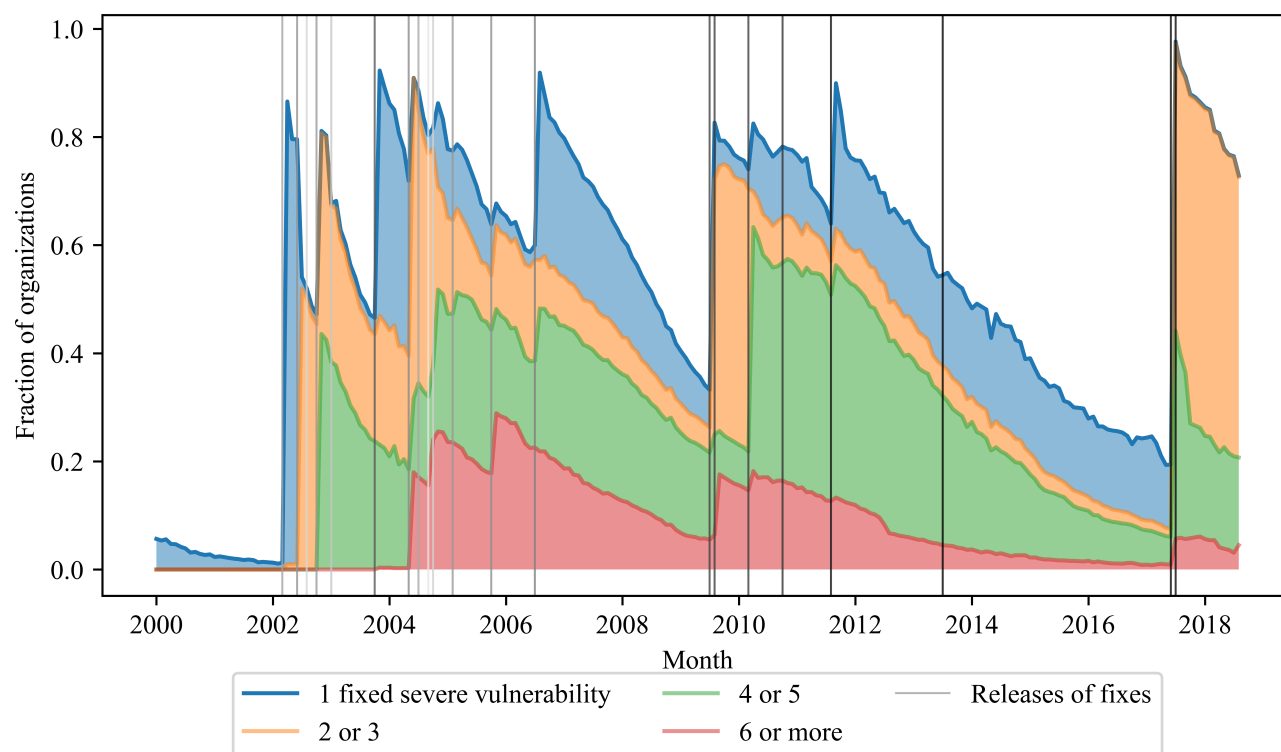
Finally, in Figure 3.5, we show the fraction of organizations with one, two, or more acknowledged and fixed security vulnerabilities in the software they used at the time, while previous figures show the presence or absence of any severe security vulnerability. The plot documents that a large fraction of organizations operate with multiple severe security vulnerabilities in their server software. This implies that malicious actors have multiple avenues for attacking these organizations' websites.

3.6.2 Do Organizations Update Their Software in Response to Vulnerabilities?

The second component of our analysis is a survival analysis of the determinations of organizations' software updating decisions using the Cox proportional hazards model. A central question of interest is whether organizations speed up updating when severe security vulnerabilities arise and if they do, by how much? For this estimate we use the sample of 11,824 websites in which we have measures of many possible determinants.

Table 3.4 shows the Cox proportional hazards analysis. Column (1) establishes the baseline by correlating an organization's updating decision and whether its server software had publicly known severe security vulnerabilities. The specification used in Equation 3.2, where we account for the time-varying independent

Figure 3.5: Fraction of organizations using Apache versions with severe security vulnerabilities that were already fixed in new releases (breakdown by the number of vulnerabilities)



Notes: The notes of Figure 3.3 apply.

variable $hasSevereBugs_{it}$. The coefficient on $hasSevereBugs_{it}$ is -0.065 and significant at 1% level, which implies that the presence of severe security vulnerabilities is correlated with a $1 - \exp(-0.065) = 6.3\%$ decrease in the probability of updating per month.

Figure 3.6 Panel (a) visualizes this relationship over a long-time horizon by plotting the Kaplan-Meier survival curves. The horizontal axis shows the number of months an organization has been using a software version. The vertical axis shows the probability that the organization is still using that software version at each time interval. We split the observations into those impacted by publicly known severe security vulnerabilities (dashed line) and those in line. For example, this figure shows that 26.4% of observations affected by known severe security vulnerabilities continued using the same version after 30 months, whereas 25.1% of observations not impacted by known severe security vulnerabilities continued using the same version.

Table 3.4: Cox proportional hazards analysis of updating decisions (broad sample)

	(1)	(2)	(3)	(4)	(5)	(6)
<i>hasSevereBugs_{it}</i>	−0.065*** (0.016)	−0.058*** (0.017)	−0.111*** (0.018)	0.278*** (0.018)	0.124*** (0.019)	0.232*** (0.059)
<i>newVersionAvailable_{it}</i>			−0.072* (0.042)		0.266*** (0.039)	0.267*** (0.038)
<i>newMinorVersionAvailable_{it}</i>			0.258*** (0.029)		0.679*** (0.026)	0.588*** (0.071)
<i>hasSevere_{it} × hasAlexaRank2010_i</i>						−0.051 (0.042)
<i>hasSevere_{it} × aws_{it}</i>						−0.115 (0.102)
<i>hasSevere_{it} × techCategories_i</i>						−0.007 (0.005)
<i>hasSevere_{it} × monetizationTechs_i</i>						−0.000 (0.010)
<i>newMinor_{it} × hasAlexaRank2010_i</i>						0.156*** (0.051)
<i>newMinor_{it} × aws_{it}</i>						0.348* (0.143)
<i>newMinor_{it} × techCategories_i</i>						−0.008 (0.007)
<i>newMinor_{it} × monetizationTechs_i</i>						0.019 (0.012)
<i>infrequentUpdater_i</i>		−0.134*** (0.021)	−0.135*** (0.021)			
<i>frequentUpdater_i</i>		0.161*** (0.023)	0.166*** (0.023)			
<i>hasAlexaRank2010_i</i>			0.069*** (0.020)			
<i>aws_{it}</i>			−0.061 (0.047)			−0.405*** (0.142)
<i>techCategories_i</i>			−0.017*** (0.005)			
<i>techs_i</i>			0.003 (0.004)			
<i>monetizationTechs_i</i>			0.016*** (0.006)			
<i>logEmployment_i</i>			−0.009 (0.008)			
<i>logRevenue_i</i>			0.001 (0.004)			
State dummies	N	N	Y	N	N	N
Industry dummies	N	N	Y	N	N	N
Stratification	N	N	N	Firm	Firm	Firm
Observations	665,360	594,004	594,004	665,360	665,360	665,360
No. of events	34,364	29,584	29,584	34,364	34,364	34,364

Notes: Each observation represents one month in an updating cycle of an organization that used Apache server software. The event is the decision to update the version of Apache server software. Standard errors are clustered at the organization level.

Both results suggest that severe security vulnerabilities are associated with less updating. However, as discussed in the Empirical Strategy section, these estimates are based on a cross-sectional variation of organizations and suffer from reverse causality. Organizations that update software less frequently are more likely to accumulate bugs over time. This would make the decision to update and the presence of security vulnerabilities appear negatively correlated. We, therefore, need to control for organizations' baseline propensity for updating.

One possibility to control for organizations' baseline propensity for updating is to include variables that capture that propensity. We construct two binary variables *infrequentUpdater_i* and *frequentUpdater_i* based on the first nine monthly observations of Apache usage of each organization. If the organization did not update its Apache severance during the first nine months, we let *infrequentUpdater_i* = 1. If the organization updated its Apache server more than once during the first nine months, we let *frequentUpdater_i* = 1. We do not include the first nine observations for each organization in estimating the Cox model. Column (2) of Table 3.4 shows the results. The coefficient on *hasSevereBugs_{it}* becomes slightly less negative, indicating the two binary variables were able to account for some differences in the baseline across organizations. Column (3) of Table 3.4 controls for additional covariates, including the full sets of state and industry dummies and numerous variables across organization and website characteristics. Still, there is sizable unexplained heterogeneity in organizations' updating decisions, apparently not explained by organizations' geography, industry, size, and use of websites (see Murciano-Goroff et al. (2021) and Section 3.6.1). As a result, these variables are unlikely to be able to capture the very rich heterogeneity in the baseline propensity to update across organizations.

In Column (4) of Table 3.4, we estimate the Cox model stratified at the organization level as in Equation 3.3. Doing so allows each organization to have a completely different baseline hazard and overcomes the reverse causality issue. The stratification approach is similar to the organization fixed-effects in a linear regression model. We use the within-organization variations of the independent variables to study their effects on the updating decisions. In this column, we run the model with no additional controls. The stratification of organizations changes the sign of their security vulnerabilities. The coefficient is 0.278, both positive and significant at 1%. This implies that if an organization's Apache server is disclosed to have severe security vulnerabilities, holding all else fixed, it is associated with a $\exp(0.278) - 1 = 32.0\%$ increase in the organization's probability of updating per month.

The fact that the addition of controls in Column (1)—(3) does not change the sign or significance of

the coefficient on $hasSevereBugs_{it}$ while stratification by organization in Column (4) changes the sign of the coefficient is telling. Organizations' software updating behavior is more associated with unobservable organization characteristics than observed ones. Updating decisions largely cannot be explained by obvious factors, such as industry, geography, website traffic, usage of cloud services, or even technologies on site. Instead, organizations likely have operational differences in software updating habits for various idiosyncratic reasons. Stated succinctly, some organizations are attentive and others are inattentive, and both extremes arise in the data with some prevalence.

Before we can claim that the estimate from our stratified Cox model is causal, we need to address the additional concern that public disclosures of vulnerabilities are correlated with new releases. The Apache Foundation often discloses vulnerabilities close to the releases of new versions that fix the vulnerabilities, so the two may be positively correlated. Moreover, disclosure follows the discovery, and discovering a security vulnerability might instigate the software developers to speed up releasing new versions to offer fixes. Organizations might then update because of the latest software release rather than the security vulnerability. In Column (5) of Table 3.4, we add variables to control for new releases. We find the coefficient on $hasSevereBugs_{it}$ of 0.124 large, positive, and highly significant. This coefficient is much smaller than the coefficient in Column (4), which suggests that the estimate in Column (4) is biased upward due to the omission of variables that capture new releases. Given that we have controlled for new releases, we are confident to interpret this result as causal, meaning that the disclosure of severe security vulnerabilities of an organization's Apache server software in use increases the organization's probability of updating per month by $\exp(0.124) - 1 = 13.2\%$, holding all else fixed.

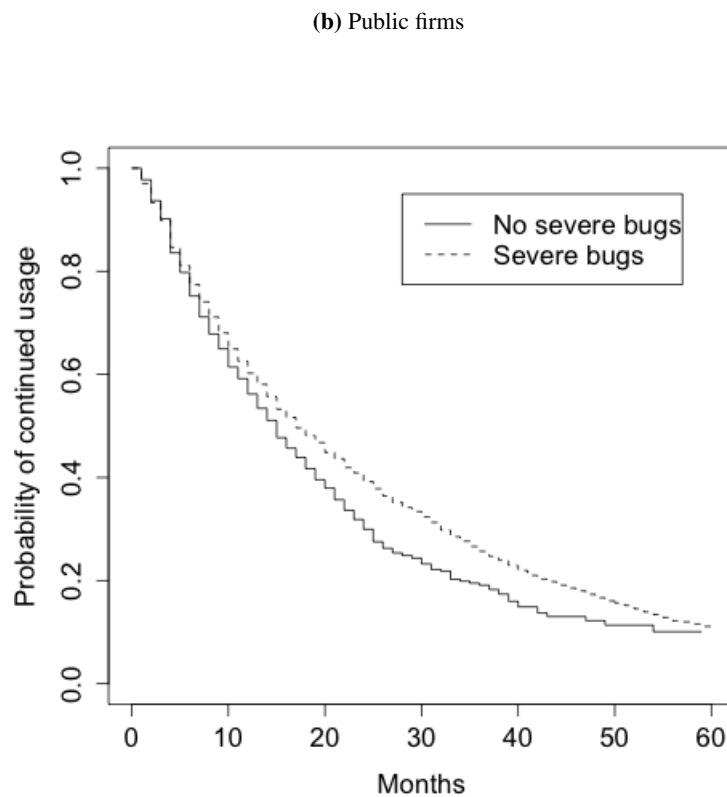
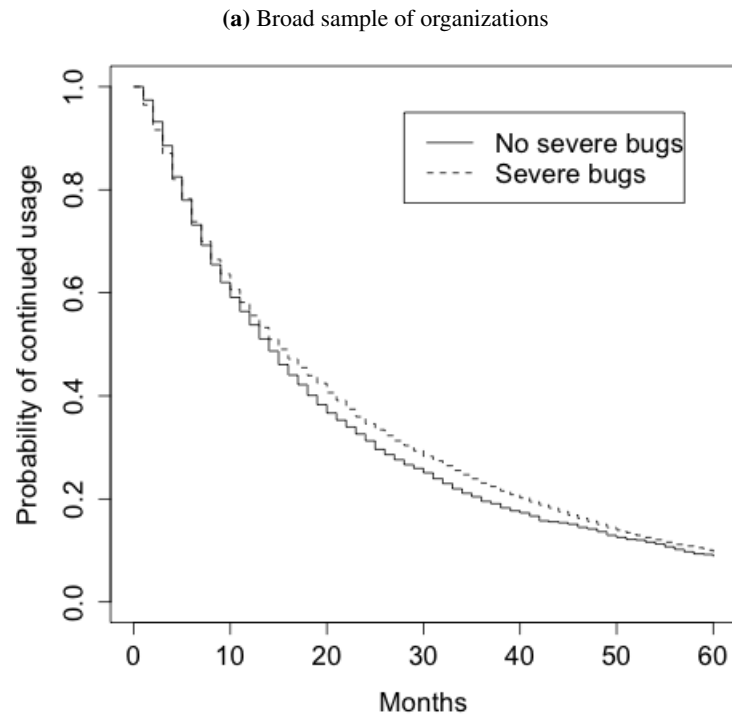
Based on the results in Column (5) of Table 3.4, we also find that organizations respond to vulnerabilities and new releases. We find that the coefficient on $newVersionAvailable_{it}$ of 0.266 to be positive, significant, and larger than the coefficient on $hasSevereBugs_{it}$. This implies that the fact that the Apache server software just became available, major or minor, is associated with speeding up updating by $\exp(0.266) - 1 = 30.5\%$, holding all else fixed. The effect is even stronger when the release is a new minor version of the version the organization was using on $newMinorVersionAvailable_{it}$ is 0.679, meaning that it is associated with increasing the probability of updating per month by $\exp(0.679) - 1 = 197\%$, holding all else fixed. We delay our discussion of Column (6) of Table 3.4 until the following subsection.

In Table 3.5 and Figure 3.6 Panel (b), we repeat the analysis we have done for the broad sample of 11,824 organizations on the sample of 1,216 public firms. The results from public firms are qualitatively very similar

to those from the broad sample of organizations. However, the smaller sample size means the coefficients are often imprecisely estimated. Column (1) of Table 3.5 shows a negative coefficient on *hasSevereBugs_{it}* in the unstratified Cox model without controls. The Kaplan-Meier survival curves in Figure 3.6 Panel (b) offer visual evidence to confirm the direction of correlation. Columns (2) and (3) show that the negative coefficient persists when controlling for baseline updater types, geography, industry, and numerous other organization and website characteristics. The coefficient turns positive, large, and significant at 1% once we stratify the Cox model at the firm level in Column (4). Controlling for new releases, Column (5) shows that the coefficient on *hasSevereBugs_{it}* becomes much smaller. Still, the coefficient on *newMinorVersionAvailable_{it}* is huge at 0.817 and highly significant at 1%. This implies that releasing new minor versions increases the probability of monthly updating for public firms by $\exp(0.817) - 1 = 226.4\%$, holding all else fixed. Controlling for firms' financials in Column (6) changes the results only slightly. We delay our discussion of Column (7) of Table 3.5 until the following subsection, when we discuss firms' rationales for updating.

Our results so far demonstrate that organizations are most responsive to the availability of new software than the announcement of a severe security vulnerability in the software they use. In addition, the comparison of stratified versus unstratified specifications of the Cox model demonstrates that much of the variation in organizations' decisions around software updating is due to idiosyncratic reasons rather than being driven by observable attributes.

Figure 3.6: Continued usage of software versions with and without severe security vulnerabilities



Notes: Each plot shows the Kaplan-Meier survival curves of the version of the software in usage. Panel (a) shows the sample of 11,824 organizations based on Mint Global organization characteristics. Panel (b) shows the sample of 1,216 public firms based on Compustat.

Table 3.5: Cox proportional hazards analysis of updating decisions (public firms)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>hasSevereBugs_{it}</i>	−0.092 (0.057)	−0.088 (0.065)	−0.108 (0.077)	0.192*** (0.069)	0.032 (0.072)	0.102 (0.089)	0.414 (0.274)
<i>newVersionAvailable_{it}</i>			−0.183 (0.231)		0.099 (0.151)	0.039 (0.204)	0.080 (0.204)
<i>newMinorVersionAvailable_{it}</i>			0.519*** (0.142)		0.817*** (0.100)	0.844*** (0.133)	0.636* (0.339)
<i>hasSevere_{it} × hasAlexaRank2010_i</i>							0.104 (0.195)
<i>hasSevere_{it} × aws_{it}</i>							0.376 (0.388)
<i>hasSevere_{it} × techCategories_i</i>							−0.018 (0.028)
<i>hasSevere_{it} × monetizationTechs_i</i>							−0.076 (0.050)
<i>newMinor_{it} × hasAlexaRank2010_i</i>							0.748*** (0.231)
<i>newMinor_{it} × aws_{it}</i>							0.255 (0.393)
<i>newMinor_{it} × techCategories_i</i>							−0.075** (0.035)
<i>newMinor_{it} × monetizationTech_i</i>							0.154** (0.064)
<i>infrequentUpdater_i</i>		−0.117 (0.072)	−0.540 (0.090)				
<i>frequentUpdater_i</i>		0.107 (0.078)	0.215** (0.100)				
<i>hasAlexaRank2010_i</i>			0.064 (0.106)				
<i>aws_{it}</i>			0.313 (0.182)				−0.352 (0.598)
<i>techCategories_i</i>			−0.051* (0.026)				.
<i>techs_i</i>			0.010 (0.018)				
<i>monetizationTechs_i</i>			0.024 (0.028)				
<i>logEmployment_{iy}</i>			0.003 (0.016)			0.030 (0.030)	0.030 (0.029)
<i>logCapx_{iy}</i>			−0.031 (0.016)			−0.028 (0.042)	−0.029 (0.042)
<i>logTotalAssets_{iy}</i>			−0.017 (0.029)			−0.231** (0.101)	−0.227** (0.102)
<i>logDepreciation_{iy}</i>			−0.002			−0.030	−0.032

Table 3.5 (Continued): Cox proportional hazards analysis of updating decisions (public firms)

			(0.016)			(0.041)	(0.043)
<i>logIncome_{iy}</i>			0.001			−0.001	−0.001
			(0.002)			(0.004)	(0.004)
<i>logCashflow_{iy}</i>			−0.130			0.356	0.326
			(0.212)			(0.478)	(0.462)
State dummies	N	N	Y	N	N	N	N
Industry dummies	N	N	Y	N	N	N	N
Stratification	N	N	N	Firm	Firm	Firm	Firm
Observations	64,204	56,504	38,016	64,204	64,204	43,210	43,210
No. of events	3,055	2,588	1,683	3,005	3,005	1,970	1,970
Notes: The notes of Table 3.4 apply.							

3.6.3 Do Costs or Benefits Influence Updating Decisions?

Though most variations in organizations' decisions to update are driven by the organization-specific unobservable, we are interested in investigating whether the findings are influenced by the perceived benefits of the update or the estimated costs of forgoing timely updates. For example, organizations that operate e-commerce websites may see monetary benefits in avoiding security vulnerabilities. On the other hand, organizations with more technically complex websites may find it more challenging to install updates because mistakes may result in downtime.

Again, we use the stratified Cox model analysis. The results are shown in Column (6) of Table 3.4 and Column (7) of Table 3.5. We estimate the stratified Cox model in both columns, including interaction terms indicating that an organization may see benefits or costs from maintaining up-to-date software. We first include interaction terms with a measure of the traffic to the website based on Alexa ranking. For example, high-traffic websites may see considerable benefits of being secure and significant costs of being attacked because they have many users.

Next, we include interaction terms with a binary dummy indicating whether AWS likely hosted the website. Delegating to a third-party hosting the website may make it costly for the organization to perform maintenance regularly. The third party may also behave differently in handling software updates. We also include interaction terms with a variable that captures the number of web technologies embedded in the website. More embedded technologies are associated with high technical complexity and higher costs of updating. Moreover, we include interaction terms with embedded monetization technologies, such as

marketing automation apps and payment processing features—the number of monetization technologies proxies for the degree of monetization intention of the organization through its website. Though monetization technologies also increase the technical complexity of the website and therefore make updating costly, the organization may again see a big benefit of being secure and significant costs of being attacked due to disruption of e-commerce.

Both Column (6) of Table 3.4 and Column (7) of Table 3.5 shows that organizations have a higher probability of updating in the face of the release of new minor versions when they have high-traffic websites. The coefficient on the interaction term $newMinor_{it} \times hasAlexaRank2010_i$ is 0.156 in Column (6) of Table 3.4 and highly significant. This implies that organizations have an $\exp(0.156) - 1 = 16.9\%$ higher probability of updating per month when a new minor version is available if Alexa ranked the organization's website as among the top one million by traffic in 2010. The estimate for the public firms is even larger, implying high-traffic websites of public firms have a $\exp(0.748) - 1 = 211.3\%$ higher probability of updating per month when a new minor version is available than their low-traffic peers. Public firms have higher traffic websites than organizations in our broad sample. Among the public firms that reached the Alexa ranking in 2010, the median ranking was around 170,000. Among the organizations in our broad sample that reached the Alexa ranking in 2010, the median ranking was around 290,000. Our results suggest that high-traffic websites, especially those of public firms, have an exceptionally high incentive to update. This is consistent with the proposition that organizations are, to some extent, incentive-driven in their software updating decisions.

Column (6) of Table 3.4 also shows that organizations are more likely to update in the face of releases of new minor versions if they use AWS. However, the result from Column (7) of Table 3.5 is imprecisely estimated and insignificant. Nevertheless, we find two results regarding AWS in Table 3.4. First, the coefficient on the aws_{it} variable is large, negative, and significant, meaning that if a site is hosted on the AWS, it is associated with a smaller probability of updating per month at the baseline. However, the interaction term $newMinor_{it} \times aws_{it}$ is large, positive, and significant, meaning that when faced with new minor releases, being hosted on the AWS is associated with a higher probability of updating per month. This suggests that organizations may face different incentives in updating on the cloud. For example, monitoring day-to-day may become costly, but new releases may prompt IT professionals to perform checkups and maintenance.

In addition, Column (7) of Table 3.5 also suggests that public firms are less likely to update in the face of new releases when their websites have more embedded technologies but are more likely to respond when their websites have more monetization-related technologies. The coefficient on the interaction term

$newMinor_{it} \times techCategories_i$ is -0.075 and is significant at the 5% level. This implies that one additional category of technology embedded in the website is associated with a $1 - \exp(-0.075) = 7.2\%$ decrease in the probability to update per month when a new minor version is available. The coefficient on the interaction term $newMinor_{it} \times monetizationTechs_i$ is 0.154 and is significant at the 5% level. This implies that one additional monetization technology embedded in the website is associated with 16.6% increase in the probability to update per month when a new minor version is available. These results are also consistent with the proposition that organizations are incentive-driven in software updating decisions to some extent. Costs due to technical complexity and benefits due to undisrupted e-commerce influence organizations' decisions to update.

3.6.4 Do Discoveries of Severe Security Vulnerabilities Instigate Developments of New Versions of Software?

In the previous subsections, we analyzed how organizations responded to announcements of severe security vulnerabilities and releases of new software versions. In this subsection, we investigate if the software producers are responsive to reports of a vulnerability in their software by speeding up new releases.

Using our data on the reports of severe security vulnerabilities and releases of new versions, we perform a survival analysis of versions using the Cox proportional hazards model. We construct our data so that each observation represents one month in a release cycle of a major version. The event is the release of a newer minor version for that major version. For example, Apache has a major version, 2.2. In July 2006, the Apache Foundation released minor version 2.2.3. In January 2007, the Apache Foundation released minor version 2.2.4. One of updating cycles for Apache 2.2 is six months, lasting from August 2006 to January 2007, ending with the event ("updated") in January 2007. As we allow the variable capturing whether there were reports of severe security vulnerabilities to vary monthly, this updating cycle corresponds to six monthly observations. Our Cox model specification is identical to that in Equation 3.2, which allows for time-varying covariates.

Table 3.6 displays the result of our Cox analysis of releases in response to reports of vulnerabilities. Column (1) suggests that reports of severe security vulnerabilities increase the probability of releasing a new version per month by $\exp(0.596) - 1 = 81.5\%$. Even with a few observations, the coefficient is significant at the 5% level. As discussed in the Empirical Strategy section, the reporting of bugs is quasi-random. We, therefore, interpret this effect to be causal. Moreover, we investigate whether reporting

Table 3.6: Cox proportional hazards analysis of releasing of new minor versions

	(1)	(2)
<i>SevereBugsReported_t</i>	0.596** (0.257)	
<i>logOrganizationsImpacted_t</i>		0.067** (0.027)
Observations	486	486
No. of events	95	95

Notes: Each observation represents one month in a release cycle of a major version of Apache server software. The event is releasing a new minor version. Our data's first severe vulnerability report is November 2001 so this sample is between November 2001 and August 2018. *SevereBugsReported_t* has a mean of 0.103 and a standard deviation of 0.304. *organizationsImpacted_t* has a mean of 15,578 and a standard deviation of 12,953. Standard errors are clustered at the level of major version.

severe security vulnerabilities impacting many users will likely speed up new releases more. We proxy for the overall usage level of a version by the number of organizations in our sample using that version each month. *logOrganizationsImpacted_t* captures the logged number of organizations affected by the severe security vulnerability reported in month t . We again find a positive and significant effect. The coefficient on *logOrganizationsImpacted_t* is 0.067, meaning that a 10% increase in the number of organizations impacted by the vulnerability leads to approximately $\exp(0.1 \times 0.067) - 1 = 0.67\%$ increase in the probability of releasing a new version per month.

These results illustrate the relative importance of different channels of how organizations gain protection against severe security vulnerabilities. Our survival analysis of organizations' updating decisions shows that organizations respond to severe security vulnerabilities directly. Still, this channel plays a much less important role than regular, routine updating activities in response to releases of new versions. The timely incorporation of bug fixes in recent new releases by the Apache Foundation has played an essential role in protecting organizations against severe security vulnerabilities.

3.7 Conclusion

This study examined the largest assembled census of security vulnerabilities among a population of users. We first examine one hundred and fifty thousand Apache users and show that the response has a robust firm-specific component, which we characterize by the establishment's level of attentiveness. Many users are either attentive or not to cybersecurity, with the majority displaying a hybrid of behavior, with inattentive actions in the short run and attentive actions over several years. Next, we collect data on a subset of users, assembling a complete dataset of over eleven thousand attentive and inattentive users. We address a reverse causality issue with a hazard model approach that accounts for firm-specific proclivities to upgrade, and the model identifies a causal explanation for the action.

Many users display a significant propensity to upgrade merely because a new version has appeared, suggesting many users gain protection against vulnerabilities as a byproduct of routine administrative processes that support maintaining frontier software. We also find that many users accelerate upgrading when their web server supports electronic commerce instead of merely an informational or coordinating role, suggesting commercial concerns play a role. We also find deceleration when users must upgrade in a complex IT operation. Finally, attentive security users respond to releasing new software versions by installing minor updates. In contrast, we find that disclosures of bugs—even severe security bugs—alone do not accelerate software patching as much. In summary, disclosing bugs has limited effects on instigating software updating in attentive and inattentive users. In other words, users primarily respond to the release of patches rather than information about the bugs themselves.

The findings are inconsistent with models of updating with low friction or high benefit. Instead, these findings align with models in which users face high administrative costs to adjust operational routines, and users depart from those routines only after accumulating information about extreme dangers. These findings also imply that disclosing security bugs might face difficult trade-offs. On the one hand, revealing the presence of dangerous security bugs could motivate users to patch their software. But on the other hand, posting information about security holes provides a roadmap for hackers to exploit and increases the attack space and vulnerability among inattentive websites. If the externalities across organizations increase significantly when inattentive users become more vulnerable, in the presence of many inattentive users, software vendors and governments might wish to publicize available software updates while providing less information about the vulnerabilities. In addition, suppose the inattentive users are treated as already

vulnerable with or without the latest disclosure due to their installation of outdated software. In that case, the reasoning changes again, potentially disclosing bugs to benefit only attentive users, and at the cost of making inattentive users incrementally more vulnerable.

This study offered a novel approach to measuring the source of vulnerabilities in a population of users, but with that novelty came a bevy of new open questions. First, our results also mean that outsourcing the management of servers to cloud providers, such as AWS, may confer an additional benefit to users. This may be especially significant for websites that try to avoid data breaches and downtime for monetizing websites. The implications also require further analysis. Second, our findings imply that all full or limited disclosures are dangerous for some users and not others, and the dangers necessarily add vulnerabilities to already vulnerable websites. That may change the priorities for updates and bug disclosures, but these trade-offs require further analysis. Third, the results may imply that insurance companies can play a role in incentivizing better cybersecurity by making the opportunity cost of not updating larger. Finally, we have also identified a puzzle for future research. The distribution of attentiveness varies dramatically across the population of users, and many measures of heterogeneity explains only a small fraction of that variance. The determinants of variance in attentiveness deserve further attention.

References

- Accenture (2019) “Ninth Annual Cost of Cybercrime Study,” <https://www.accenture.com/us-en/insights/security/cost-cybercrime-study>.
- Acronis (2020) “The NHS Cyber Attack,” <https://www.acronis.com/en-us/blog/posts/nhs-cyber-attack/>.
- Agrawal, Rajeev (1995) “Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-Armed Bandit Problem,” *Advances in Applied Probability*, 27 (4), 1054–1078.
- Aguirregabiria, Victor and Pedro Mira (2010) “Dynamic Discrete Choice Structural Models: A Survey,” *Journal of Econometrics*, 156 (1), 38–67.
- Aral, Sinan, Erik Brynjolfsson, and D.J. Wu (2006) “Which Came First, IT or Productivity? Virtuous Cycle of Investment and Use in Enterprise Systems,” in *Twenty Seventh International Conference on Information Systems*.
- Aridor, Guy, Yeon-Koo Che, and Tobias Salz (2020) “The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR,” NBER Working Paper 26900.
- Arora, Ashish, Ramayya Krishnan, Rahul Telang, and Yubao Yang (2010) “An Empirical Analysis of Software Vendors’ Patch Release Behavior: Impact of Vulnerability Disclosure,” *Information Systems Research*, 21 (1), 115–132.
- Arora, Ashish, Rahul Telang, and Hao Xu (2008) “Optimal Policy for Software Vulnerability Disclosure,” *Management Science*, 54 (4), 642–656.
- Arrow, Kenneth (1962) “Economic Welfare and the Allocation of Resources for Invention,” in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, 609–625, Princeton: Princeton University Press.
- Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer (2002) “Finite-Time Analysis of the Multiarmed Bandit Problem,” *Machine Learning*, 47 (2), 235–256.
- August, Terrence, Marius Florin Niculescu, and Hyoduk Shin (2014) “Cloud Implications on Software Network Structure and Security Risks,” *Information Systems Research*, 25 (3), 489–510.
- August, Terrence and Tunay I Tunca (2006) “Network Software Security and User Incentives,” *Management Science*, 52 (11), 1703–1720.
- Azoulay, Pierre (2012) “Turn the Scientific Method on Ourselves: How Can We Know Whether Funding Models for Research Work? By Relentlessly Testing Them Using Randomized Controlled Trials,” *Nature*, 484 (7392), 31–33.

- Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat (2019) “Public R&D Investments and Private-Sector Patenting: Evidence from NIH Funding Rules,” *Review of Economic Studies*, 86 (1), 117–152.
- Azoulay, Pierre, Joshua S Graff Zivin, and Gustavo Manso (2011) “Incentives and Creativity: Evidence from the Academic Life Sciences,” *RAND Journal of Economics*, 42 (3), 527–554.
- Babnigg, György and Andrzej Joachimiak (2010) “Predicting Protein Crystallization Propensity from Protein Sequence,” *Journal of Structural and Functional Genomics*, 11 (1), 71–80.
- Battista, G Di, M Patrignani, and M Pizzonia (2003) “Computing the Types of the Relationships Between Autonomous Systems,” in *IEEE INFOCOM 2003*, 1, 156–165, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies.
- Bergemann, Dirk and Juuso Välimäki (1996) “Learning and Strategic Pricing,” *Econometrica*, 64 (5), 1125–1149.
- Berman, Helen M, Margaret J Gabanyi, Andrei Kouranov, David I Micallef, John Westbrook, and Protein Structure Initiative network of investigators (2017) “Protein Structure Initiative—TargetTrack 2000-2017—all data files,” <https://zenodo.org/record/821654>, accessed June 25, 2019.
- Besen, Stanley, Paul Milgrom, Bridger Mitchell, and Padmanabhan Srinagesh (2001) “Advances in Routing Technologies and Internet Peering Agreements,” *American Economic Review*, 91 (2), 292–296.
- Bessen, James and Cesare Righi (2019) “Shocking Technology: What Happens When Firms Make Large IT Investments?”, Boston University School of Law, Law and Economics Research Paper No. 19-6.
- Bolton, Patrick and Christopher Harris (1999) “Strategic Experimentation,” *Econometrica*, 67 (2), 349–374.
- Borouh, Mark A (2021) “New Data on U.S. R&D: Summary Statistics from the 2019–20 Edition of National Patterns of R&D Resources,” *National Center for Science and Engineering Statistics (NCSES)*, <https://ncses.nsf.gov/pubs/nsf22314>.
- Braden, Robert T. (1989) *Requirements for Internet Hosts – Communication Layers* (RFC 1122): Internet Engineering Task Force, <https://datatracker.ietf.org/doc/rfc1122>.
- Brezzi, Monica and Tze Leung Lai (2002) “Optimal Learning and Experimentation in Bandit Problems,” *Journal of Economic Dynamics and Control*, 27 (1), 87–108.
- Brynjolfsson, Erik and Lorin M Hitt (2003) “Computing Productivity: Firm-Level Evidence,” *Review of Economics and Statistics*, 85 (4), 793–808.
- Bubeck, Sébastien and Nicolo Cesa-Bianchi (2012) “Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems,” *arXiv preprint arXiv:1204.5721*.
- Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost (2021) “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND,” *Nature Methods*, 18 (4), 366–368.
- Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2015) “Fast and Sensitive Protein Alignment Using DIAMOND,” *Nature Methods*, 12 (1), 59–60.
- CAIDA (2013) “Routeviews Prefix-to-AS mappings (pfx2as) for IPv4 and IPv6,” <https://www.caida.org/catalog/datasets/routeviews-prefix2as/>.

- (2015-2019a) “AS Relationships,” <https://www.caida.org/catalog/datasets/as-relationships/>.
- (2015-2019b) “Inferred AS to Organization Mapping Dataset,” <https://www.caida.org/catalog/datasets/as-organizations/>.
- (2015-2019c) “IPv4 Prefix-Probing Traceroute Dataset,” https://www.caida.org/catalog/datasets/ipv4_prefix_probing_dataset/.
- Cantoni, Davide and Noam Yuchtman (2014) “Medieval Universities, Legal Institutions, and the Commercial Revolution,” *Quarterly Journal of Economics*, 129 (2), 823–887.
- Cappé, Olivier, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz et al. (2013) “Kullback–Leibler Upper Confidence Bounds for Optimal Sequential Allocation,” *Annals of Statistics*, 41 (3), 1516–1541.
- Caria, Stefano, Grant Gordon, Maximilian Kasy, Simon Quinn, Soha Shami, and Alex Teytelboym (2020) “An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan,” CESifo Working Paper.
- Cavusoglu, Hasan, Huseyin Cavusoglu, and Jun Zhang (2008) “Security Patch Management: Share the Burden or Share the Damage?” *Management Science*, 54 (4), 657–670.
- Chen, Wei, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu (2016) “Combinatorial Multi-Armed Bandit with General Reward Functions,” *Advances in Neural Information Processing Systems*, 29.
- Chen, Wei, Liwei Wang, Haoyu Zhao, and Kai Zheng (2021) “Combinatorial Semi-Bandit in the Non-Stationary Environment,” in *Uncertainty in Artificial Intelligence*, 865–875, Proceedings of Machine Learning Research.
- Chen, Xiaohong, Timothy M Christensen, and Elie Tamer (2018) “Monte Carlo Confidence Sets for Identified Sets,” *Econometrica*, 86 (6), 1965–2018.
- Chernozhukov, Victor and Han Hong (2003) “An MCMC Approach to Classical Estimation,” *Journal of Econometrics*, 115 (2), 293–346.
- Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim (2015) “Net Neutrality, Business Models, and Internet Interconnection,” *American Economic Journal: Microeconomics*, 7 (3), 104–141.
- Chruszcz, Maksymilian, Alexander Wlodawer, and Wladek Minor (2008) “Determination of Protein Structures—A Series of Fortunate Events,” *Biophysical Journal*, 95 (1), 1–9.
- Cisco Talos Intelligence Group (2020) “Email and Spam Data – Comprehensive Threat Intelligence,” https://talosintelligence.com/reputation_center/email_rep.
- CNIL (2019) “The CNIL’s Restricted Committee Imposes a Financial Penalty of 50 Million Euros Against GOOGLE LLC,” https://edpb.europa.eu/news/national-news/2019/cnils-restricted-committee-imposes-financial-penalty-50-million-euros_en.
- Cohen, Wesley M and Daniel A Levinthal (1989) “Innovation and Learning: the Two Faces of R&D,” *Economic Journal*, 99 (397), 569–596.
- Comino, Stefano, Fabio M Manenti, and Franco Mariuzzo (2019) “Updates Management in Mobile Applications: iTunes Versus Google Play,” *Journal of Economics & Management Strategy*, 28 (3), 392–419.

- Crawford, Gregory S and Matthew Shum (2005) “Uncertainty and Learning in Pharmaceutical Demand,” *Econometrica*, 73 (4), 1137–1173.
- Data Protection Commission, The (2020) “Report by the Data Protection Commission on the Use of Cookies and Other Tracking Technologies,” <https://dataprotection.ie/en/news-media/publications/report-dpc-use-cookies-and-other-tracking-technologies>.
- DeepMind (2020) “AlphaFold: Using AI for Scientific Discovery,” <https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery>.
- Degeling, Martin, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz (2019) “We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR’s Impact on Web Privacy,” in *Proceedings 2019 Network and Distributed System Security Symposium*, arXiv preprint arXiv:1808.05096 [cs].
- Dey, Debabrata, Atanu Lahiri, and Guoying Zhang (2015) “Optimal Policies for Security Patch Management,” *INFORMS Journal on Computing*, 27 (3), 462–477.
- Dickstein, Michael J (2018) “Efficient Provision of Experience Goods: Evidence from Antidepressant Choice,” Working Paper.
- Dimitropoulos, Xenofontas, Dmitri Krioukov, Marina Fomenkov, Bradley Huffaker, Young Hyun, kc claffy, and George Riley (2007) “AS relationships: inference and validation,” *ACM SIGCOMM Computer Communication Review*, 37 (1), 29–40.
- Dimitropoulos, Xenofontas, Dmitri Krioukov, Bradley Huffaker, kc claffy, and George Riley (2005) *Inferring AS Relationships: Dead End or Lively Beginning?*, 3503 of Lecture Notes in Computer Science, 113–125, Berlin, Heidelberg: Springer Berlin Heidelberg.
- Dissanayake, Nesara, Asangi Jayatilaka, Mansoor Zahedi, and M Ali Babar (2022) “Software Security Patch Management—A Systematic Literature Review of Challenges, Approaches, Tools and Practices,” *Information and Software Technology*, 144, 106771.
- D’Ignazio, Alessio and Emanuele Giovannetti (2009) “Asymmetry and Discrimination in Internet Peering: Evidence from the LINX,” *International Journal of Industrial Organization*, 27 (3), 441–448.
- Ederer, Florian and Gustavo Manso (2013) “Is Pay for Performance Detrimental to Innovation?” *Management Science*, 59 (7), 1496–1513.
- EMBL-EBI (2021) “EMBL-EBI Pfam to PDB Mapping,” <http://ftp.ebi.ac.uk/pub/databases/Pfam/mappings/>.
- Englehardt, Steven and Arvind Narayanan (2016) “Online Tracking: A 1-million-site Measurement and Analysis,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1388–1401.
- Erdem, Tülin and Michael P Keane (1996) “Decision-Making under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets,” *Marketing Science*, 15 (1), 1–20.
- Erlebach, Thomas, Alexander Hall, and Thomas Schank (2002) “Classifying Customer-Provider Relationships in the Internet,” TIK-Report 145.
- Foster, Lucia, John C Haltiwanger, and Cornell John Krizan (2001) “Aggregate Productivity Growth: Lessons from Microeconomic Evidence,” in *New Developments in Productivity Analysis*, 303–372: University of Chicago Press.

- Furman, Jeffrey L and Scott Stern (2011) “Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research,” *American Economic Review*, 101 (5), 1933–63.
- Ganglmair, Bernhard, Timothy Simcoe, and Emanuele Tarantino (2019) “Learning When to Quit: An Empirical Model of Experimentation in Standards Development,” Working Paper.
- Gao, Lixin (2001) “On Inferring Autonomous System Relationships in the Internet,” *IEEE/ACM Transactions on Networking*, 9 (6), 733–745.
- Garivier, Aurélien and Eric Moulines (2011) “On Upper-Confidence Bound Policies for Switching Bandit Problems,” in *International Conference on Algorithmic Learning Theory*, 174–188.
- Gittins, John C (1979) “Bandit Processes and Dynamic Allocation Indices,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 41 (2), 148–164.
- Gittins, John C and David M Jones (1979) “A Dynamic Allocation Index for the Discounted Multiarmed Bandit Problem,” *Biometrika*, 66 (3), 561–565.
- Goldberg, Samuel G, Garrett A Johnson, and Scott K Shriver (2019) “Regulating Privacy Online: An Economic Evaluation of the GDPR,” Law & Economics Center at George Mason University Scalia Law School Research Paper Series No. 22-025.
- Goldfarb, Avi and Catherine E. Tucker (2011) “Privacy Regulation and Online Advertising,” *Management Science*, 57 (1), 57–71.
- Guan, Melody and Heinrich Jiang, “Nonparametric Stochastic Contextual Bandits,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1).
- Guan, R., E. Marcos, P. O’Connell et al. (2014) “Crystal Structure of an engineered protein with denovo beta sheet design, Northeast Structural Genomics Consortium (NESG) Target OR486,” <https://www.rcsb.org/structure/4R80>.
- Henderson, Rebecca and Iain Cockburn (1996) “Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery,” *RAND Journal of Economics*, 27 (1), 32–59.
- Hill, Ryan and Carolyn Stein (2020) “Scooped! Estimating Rewards for Priority in Science,” Working Paper.
- (2021) “Race to the Bottom: Competition and Quality in Science,” Working Paper.
- Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero (2022) “The Streetlight Effect in Data-Driven Innovation,” Working Paper.
- Hoofnagle, Chris Jay, Bart van der Sloot, and Frederik Zuiderveen Borgesius (2019) “The European Union General Data Protection Regulation: What It Is and What It Means,” *Information & Communications Technology Law*, 28 (1), 65–98.
- Hotz, V Joseph and Robert A Miller (1993) “Conditional Choice Probabilities and the Estimation of Dynamic Models,” *The Review of Economic Studies*, 60 (3), 497–529.
- Huang, Hongzhan, Peter B McGarvey, Baris E Suzek, Raja Mazumder, Jian Zhang, Yongxing Chen, and Cathy H Wu (2011) “A Comprehensive Protein-Centric ID Mapping Service for Molecular Data Integration,” *Bioinformatics*, 27 (8), 1190–1191.

- IBM (2020) “IBM Report: Compromised Employee Accounts Led to Most Expensive Data Breaches Over Past Year,” <https://newsroom.ibm.com/2020-07-29-IBM-Report-Compromised-Employee-Accounts-Led-to-Most-Expensive-Data-Breaches-Over-Past-Year>.
- Iordanou, Costas, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris (2018) “Tracing Cross Border Web Tracking,” in *Proceedings of the Internet Measurement Conference 2018*, 329–342.
- Jaffe, Adam B (2002) “Building Programme Evaluation into the Design of Public Research-Support Programmes,” *Oxford Review of Economic Policy*, 18 (1), 22–34.
- Jahandideh, Samad, Lukasz Jaroszewski, and Adam Godzik (2014) “Improving the Chances of Successful Protein Structure Determination with a Random Forest Classifier,” *Acta Crystallographica Section D: Biological Crystallography*, 70 (3), 627–635.
- Jaroszewski, Lukasz, Lukasz Slabinski, John Wooley, Ashley M Deacon, Scott A Lesley, Ian A Wilson, and Adam Godzik (2008) “Genome Pool Strategy for Structural Coverage of Protein Families,” *Structure*, 16 (11), 1659–1667.
- Jia, Jian, Ginger Zhe Jin, and Liad Wagman (2020) “GDPR and the Localness of Venture Investment,” Working Paper.
- (2021) “The Short-Run Effects of the General Data Protection Regulation on Technology Venture Investment,” *Marketing Science*, 40 (4), 661–684.
- Johnson, Garrett A., Scott K. Shriver, and Samuel G. Goldberg (2023) “Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR,” *Management Science*, forthcoming.
- Jorgenson, Dale W, Mun S Ho, and Jon D Samuels (2016) “The Impact of Information Technology on Postwar US Economic Growth,” *Telecommunications Policy*, 40 (5), 398–411.
- Jorgenson, Dale Weldeau, Mun S Ho, and Kevin J Stiroh (2006) *Productivity: Information Technology and the American Growth Resurgence (Volume 3)*: The MIT Press.
- Kang, Hye Young (2022) “Too Much Can be as Bad as Too Little: Product Update Strategy for Online Digital Platform Complementors,” *Industrial and Corporate Change*, 31 (6), 1494–1516.
- Karaj, Arjaldo, S. Macbeth, Rémi Berson, and J. M. Pujol (2018) “WhoTracks .Me: Shedding Light on the Opaque World of Online Tracking,” *arXiv preprint arXiv:1804.08959*.
- Katehakis, Michael N and Herbert Robbins (1995) “Sequential Choice from Several Populations.,” *Proceedings of the National Academy of Sciences*, 92 (19), 8584–8585.
- Kawashima, Shuichi, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa (2007) “AAindex: Amino Acid Index Database, Progress Report 2008,” *Nucleic Acids Research*, 36 (suppl_1), D202–D205.
- Keller, Godfrey, Sven Rady, and Martin Cripps (2005) “Strategic Experimentation with Exponential Bandits,” *Econometrica*, 73 (1), 39–68.
- Khmelnitskaya, Ekaterina (2022) “Competition and Attrition in Drug Development,” Working Paper.
- Klausen, Michael Schantz, Martin Closter Jespersen, Henrik Nielsen et al. (2019) “NetSurfP-2.0: Improved Prediction of Protein Structural Features by Integrated Deep Learning,” *Proteins: Structure, Function, and Bioinformatics*, 87 (6), 520–527.

- Krieger, Joshua L (2021) “Trials and Terminations: Learning from Competitors’ R&D Failures,” *Management Science*, 67 (9), 5525–5548.
- Kveton, Branislav, Zheng Wen, Azin Ashkan, and Csaba Szepesvari (2015) “Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits,” in *Artificial Intelligence and Statistics*, 535–543, Proceedings of Machine Learning Research.
- Laffont, Jean-Jacques, Scott Marcus, Patrick Rey, and Jean Tirole (2001) “Internet Peering,” *American Economic Review*, 91 (2), 287–291.
- Lai, Tze Leung (1987) “Adaptive Treatment Allocation and the Multi-Armed Bandit Problem,” *Annals of Statistics*, 15 (1), 1091–1114.
- Lai, Tze Leung and Herbert Robbins (1985) “Asymptotically Efficient Adaptive Allocation Rules,” *Advances in Applied Mathematics*, 6 (1), 4–22.
- Lane, J.N., C. Ayoubi, K. Boudreau, E. Guinan, and Lakhani K. (2022) “Generating Innovation in the Lab: Experimental Evidence from the Life Sciences,” Working Paper.
- Lane, Julia I, Jason Owen-Smith, Rebecca F Rosen, and Bruce A Weinberg (2015) “New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value,” *Research Policy*, 44 (9), 1659–1671.
- Langford, John and Tong Zhang (2007) “The Epoch-Greedy Algorithm for Contextual Multi-Armed Bandits,” *Advances in Neural Information Processing Systems*, 20 (1), 96–1.
- Lattimore, Tor and Csaba Szepesvári (2020) *Bandit Algorithms*, Cambridge: Cambridge University Press.
- Lefrere, V, L Warberg, C Cheyre, V Marotta, and A Acquisti (2020) “The Impact of the GDPR on Content Providers,” in *The 2020 Workshop on the Economics of Information Security*.
- Leyden, Benjamin T (2018) *There’s an app (update) for that*, Working Paper.
- Li, Danielle, Lindsey R Raymond, and Peter Bergman (2020) “Hiring as Exploration,” NBER Working Paper 27736.
- Li, He, Sungjin Yoo, and William J Kettinger (2021) “The Roles of IT Strategies and Security Investments in Reducing Organizational Security Breaches,” *Journal of Management Information Systems*, 38 (1), 222–245.
- Libert, Timothy (2015) “Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites,” arXiv preprint arXiv:1511.00619.
- Libert, Timothy, Lucas Graves, and Rasmus Kleis Nielsen (2018) “Changes in Third-Party Content on European News Websites after GDPR,” University of Oxford Reuters Institute for the Study of Journalism Factsheet.
- Liu, Che-Wei, Peng Huang, and Henry Lucas (2017) “IT Centralization, Security Outsourcing, and Cybersecurity Breaches: Evidence from the US Higher Education,” in *ICIS 2017 Proceedings*.
- Lodhi, Aemen, Natalie Larson, Amogh Dhamdhere, Constantine Dovrolis, and kc claffy (2014) “Using PeeringDB to Understand the Peering Ecosystem,” *ACM SIGCOMM Computer Communication Review*, 44 (2), 20–27.

- Luckie, Matthew, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and kc claffy (2013) “AS Relationships, Customer Cones, and Validation,” in *Proceedings of the 2013 Conference on Internet Measurement Conference*, 243–256.
- Marder, Alexander, Matthew Luckie, Amogh Dhamdhere, and Bradley Huffaker (2018) “Pushing the Boundaries with bdrmapIT: Mapping Router Ownership at Internet Scale,” in *Proceedings of the 2018 Internet Measurement Conference*, 56–69.
- Godinho de Matos, Miguel and Idris Adjerid (2022) “Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider,” *Management Science*, 68 (5), 3330–3378.
- Matějka, Filip and Alisdair McKay (2015) “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model,” *American Economic Review*, 105 (1), 272–298.
- McElheran, Kristina (2015) “Do Market Leaders Lead in Business Process Innovation? The Case(s) of E-Business Adoption,” *Management Science*, 61 (6), 1197–1216.
- Meltzer, Joshua P (2014) “The Importance of the Internet and Transatlantic Data Flows for US And EU Trade and Investment,” Global Economy and Development Working Paper 79.
- Messaging Anti-Abuse Working Group (2011) *Email Metrics Program: The Network Operators’ Perspective*, Report #14: Third and Fourth Quarter 2010.
- Miller, Amalia R. and Catherine Tucker (2018) “Privacy Protection, Personalized Medicine, and Genetic Testing,” *Management Science*, 64 (10), 4648–4668.
- Miller, Amalia R. and Catherine E. Tucker (2011) “Can Health Care Information Technology Save Babies?” *Journal of Political Economy*, 119 (2), 289–324.
- Miller, Robert A (1984) “Job Matching and Occupational Choice,” *Journal of Political Economy*, 92 (6), 1086–1120.
- Miller, Susan, Joel Janin, Arthur M Lesk, and Cyrus Chothia (1987) “Interior and Surface of Monomeric Proteins,” *Journal of Molecular Biology*, 196 (3), 641–656.
- Mitra, Sabyasachi and Sam Ransbotham (2015) “Information Disclosure and the Diffusion of Information Security Attacks,” *Information Systems Research*, 26 (3), 565–584.
- Mohan, Jayashree, Melissa Wasserman, and Vijay Chidambaram (2019) “Analyzing GDPR Compliance Through the Lens of Privacy Policy,” in *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 11721 of Lecture Notes in Computer Science, 82–95: Springer International Publishing.
- Mookerjee, Vijay, Radha Mookerjee, Alain Bensoussan, and Wei T Yue (2011) “When Hackers Talk: Managing Information Security under Variable Attack Rates and Knowledge Dissemination,” *Information Systems Research*, 22 (3), 606–623.
- Moore, Peter B (2007) “Let’s Call the Whole Thing Off: Some Thoughts on the Protein Structure Initiative,” *Structure*, 15 (11), 1350–1352.
- Murciano-Goroff, Raviv, Ran Zhuo, and Shane Greenstein (2021) “Hidden Software and Veiled Value Creation: Illustrations from Server Software Usage,” *Research Policy*, 50 (9), 104333.
- Myers, Kyle (2020) “The Elasticity of Science,” *American Economic Journal: Applied Economics*, 12 (4), 103–34.

- Nicholson, Jessica R and Giulia McHenry (2016) “Measuring Cross-Border Data Flows: Data, Literature, and Considerations,” <https://www.ntia.doc.gov/other-publication/2016/measuring-cross-border-data-flows-data-literature-and-considerations>.
- NIGMS (2007a) “Frequently Asked Questions,” <http://www.nigms.nih.gov/Initiatives/PSI/Background/FAQs.htm>.
- (2007b) “Frequently Asked Questions for PSI-2,” <http://www.nigms.nih.gov/Initiatives/PSI/Background/FAQsPSI-2.htm>.
- (2007c) “Request for Information (RFI): To Solicit Input for the Assessment of the Protein Structure Initiative,” <http://grants.nih.gov/grants/guide/notice-files/NOT-GM-07-108.html>.
- (2008a) “Protein Structure Initiative (Pilot Phase) Fact Sheet,” <http://www.nigms.nih.gov/Initiatives/PSI/Background/PilotFacts.htm>.
- (2008b) “Report of the Future Structural Genomics Initiatives Meeting,” <http://www.nigms.nih.gov/News/Reports/FutureSGMeeting102008.htm>.
- (2009a) “Concept Clearance: High-Throughput Structural Biology,” https://www.nigms.nih.gov/News/Reports/council_concept_clearance_2009.
- (2009b) “Protein Structure Initiative Advisory Committee (PSIAC): Recommendations for the Future of the PSI,” http://www.nigms.nih.gov/News/Reports/PSIAC_Future_2009.htm.
- (2011a) “PSI Production Phase Fact Sheet,” <http://www.nigms.nih.gov/Research/FeaturedPrograms/PSI/Background/PSI2FactSheet.htm>.
- (2011b) “PSI:Biology FOA Frequently Asked Questions,” https://www.nigms.nih.gov/research/specificareas/PSI/PSI_biology/background/pages/PSIbiology_faqs.aspx.
- NIH (2019) “NIGMS Funding Opportunities Search,” <https://www.nigms.nih.gov/grants/Pages/Funding.aspx?expired=1>.
- (2021) “NIH RePORT Advanced Projects Search,” <https://reporter.nih.gov/>.
- Norton, William B (2014) *The 2014 Internet Peering Playbook: Connecting to the Core of the Internet*: DrPeering Press.
- Ortner, Ronald, Daniil Ryabko, Peter Auer, and Rémi Munos (2012) “Regret Bounds for Restless Markov Bandits,” in *International Conference on Algorithmic Learning Theory*, 214–228.
- Pakes, Ariel (1986) “Patents as Options: Some Estimates of the Value of Holding European Patent Stocks,” *Econometrica*, 755–784.
- Palmer, Danny (2017) “WannaCry Ransomware: Hospitals were Warned to Patch System to Protect against Cyber-Attack - But Didn’t,” <https://www.zdnet.com/article/wannacry-ransomware-hospitals-were-warned-to-patch-system-to-protect-against-cyber-attack-but-didnt/>.
- Petsko, Gregory A (2007) “An Idea Whose Time Has Gone,” *Genome Biology*, 8 (6), 1–3.
- Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer (2020) “European Privacy Law and Global Markets for Data,” CEPR Discussion Paper No. DP14475.

- Price, W Nicholson, Yang Chen, Samuel K Handelman et al. (2009a) “Supplementary Information for Understanding the Physical Properties that Control Protein Crystallization by Analysis of Large-Scale Experimental Data,” *Nature Biotechnology*, 27 (1), 51–57.
- (2009b) “Understanding the Physical Properties that Control Protein Crystallization by Analysis of Large-Scale Experimental Data,” *Nature Biotechnology*, 27 (1), 51–57.
- Ranger, Steve (2019) “Cybersecurity: One in Three Breaches are Caused by Unpatched Vulnerabilities,” <https://www.zdnet.com/article/cybersecurity-one-in-three-breaches-are-caused-by-unpatched-vulnerabilities/>.
- Roberts, Kevin and Martin L Weitzman (1981) “Funding Criteria for Research, Development, and Exploration Projects,” *Econometrica*, 49 (5), 1261–1288.
- Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen (2017) “A Tutorial on Thompson Sampling,” *arXiv preprint arXiv:1707.02038*.
- Rust, John (1987) “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher,” *Econometrica: Journal of the Econometric Society*, 999–1033.
- (1994) “Structural Estimation of Markov Decision Processes,” *Handbook of Econometrics*, 4, 3081–3143.
- Sandvine (2018) “2018 Global Internet Phenomena Report,” <https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf>.
- Scale (2021) “Netflix Explains Recommendations and Personalization,” <https://scale.com/blog/Netflix-Recommendation-Personalization-TransformX-Scale-AI-Insights>.
- Shiller, Benjamin, Joel Waldfogel, and Johnny Ryan (2018) “The Effect of Ad Blocking on Website Traffic and Quality,” *The RAND Journal of Economics*, 49 (1), 43–63.
- Slabinski, Lukasz, Lukasz Jaroszewski, Ana PC Rodrigues, Leszek Rychlewski, Ian A Wilson, Scott A Lesley, and Adam Godzik (2007a) “The Challenge of Protein Structure Determination—Lessons from Structural Genomics,” *Protein Science*, 16 (11), 2472–2482.
- Slabinski, Lukasz, Lukasz Jaroszewski, Leszek Rychlewski, Ian A Wilson, Scott A Lesley, and Adam Godzik (2007b) “XtalPred: A Web Server for Prediction of Protein Crystallizability,” *Bioinformatics*, 23 (24), 3403–3405.
- Steinberg, Scott (2019) “Cyberattacks Now Cost Companies \$200,000 on Average, Putting Many out of Business,” <https://www.cnbc.com/2019/10/13/cyberattacks-cost-small-companies-200k-putting-many-out-of-business.html>.
- Subramanian, L., S. Agarwal, J. Rexford, and R.H. Katz (2002) “Characterizing the Internet Hierarchy from Multiple Vantage Points,” in *Proceedings of Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, 2, 618–627.
- Symantec (2010) “Messagelabs Intelligence: 2010 Annual Security Report,” <https://retelur.files.wordpress.com/2007/10/symantec-messagelabsintelligencefinal2010-110111043129-phpapp02.pdf>.
- Tambe, Prasanna and Lorin M Hitt (2012) “The Productivity of Information Technology Investments: New Evidence from IT Labor Data,” *Information Systems Research*, 23 (3-part-1), 599–617.

- Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson (2020) *Digital Capital and Superstar Firms*, NBER Working Paper 28285.
- The Internet Society (2015) “Policy Brief: Internet Interconnection,” <https://www.internetsociety.org/policybriefs/internetinterconnection/>.
- Timmins, Christopher (2002) “Measuring the Dynamic Efficiency Costs of Regulators’ Preferences: Municipal Water Utilities in the Arid West,” *Econometrica*, 70 (2), 603–629.
- UniProt (2021a) “Programmatic Access — Retrieving Individual Entries,” https://www.uniprot.org/help/api_retrieve_entries.
- (2021b) “Programmatic Access—Mapping Database Identifiers,” https://www.uniprot.org/help/api_idmapping.
- (2021c) “Taxonomy—Eukaryota,” <https://www.uniprot.org/taxonomy/2759>.
- (2021d) “Taxonomy—Homo sapiens (Human),” <https://www.uniprot.org/taxonomy/9606>.
- UniProt Consortium, The (2021) “UniProt: The Universal Protein Knowledgebase in 2021,” *Nucleic Acids Research*, 49 (D1), D480–D489.
- US International Trade Commission (2014) “Digital Trade in the U.S. and Global Economies, Part 2,” https://www.strtrade.com/media/publication/7238_pub4485.pdf.
- Van Montfort, Rob LM and Paul Workman (2017) “Structure-Based Drug Design: Aiming for a Perfect Fit,” *Essays in Biochemistry*, 61 (5), 431–437.
- Varadi, Mihaly, John Berrisford, Mandar Deshpande et al. (2020) “PDBe-KB: a Community-Driven Resource for Structural and Functional Annotations,” *Nucleic Acids Research*, 48 (D1), D344–D353.
- Wang, Siwei and Wei Chen (2018) “Thompson Sampling for Combinatorial Semi-Bandits,” in *International Conference on Machine Learning*, 5114–5122, Proceedings of Machine Learning Research.
- Weitzman, Martin L (1979) “Optimal Search for the Best Alternative,” *Econometrica*, 641–654.
- Weller, Dennis and Bill Woodcock (2013) *Internet Traffic Exchange: Market Developments and Policy Challenges*, OECD Digital Economy Papers No. 207, https://www.oecd-ilibrary.org/science-and-technology/internet-traffic-exchange_5k918gpt130q-en.
- Whittle, Peter (1988) “Restless Bandits: Activity Allocation in a Changing World,” *Journal of Applied Probability*, 25 (A), 287–298.
- Woodroffe, Michael (1979) “A One-Armed Bandit Problem with a Concomitant Variable,” *Journal of the American Statistical Association*, 74 (368), 799–806.
- Wootton, John C (1994) “Non-Globular Domains in Protein Sequences: Automated Segmentation Using Complexity Measures,” *Computers & Chemistry*, 18 (3), 269–285.
- Wrapp, Daniel, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan (2020) “Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation,” *Science*, 367 (6483), 1260–1263.
- wwPDB (2013) “PDB Structure Download Statistics,” <http://www.wwpdb.org/downloadStats.php>.

- Xia, Jianhong and Lixin Gao (2004) “On the Evaluation of AS Relationship Inferences [Internet Reachability/Traffic Flow Applications],” in *IEEE Global Telecommunications Conference 2004*, 3, 1373–1377.
- Yang, Jeremy, Dean Eckles, Paramveer Dhillon, and Sinan Aral (2020) “Targeting for Long-Term Outcomes,” *arXiv preprint arXiv:2010.15835*.
- Zhang, Jason (2021) “Deep Understanding Tesla FSD Part 3: Planning & Control,” <https://saneryee-studio.medium.com/deep-understanding-tesla-fsd-part-3-planning-control-9a25cc6d04f0>.
- Zhou, Dongruo, Lihong Li, and Quanquan Gu (2020) “Neural Contextual Bandits with UCB-Based Exploration,” in *International Conference on Machine Learning*, 11492–11502, Proceedings of Machine Learning Research.
- Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson et al. (2021) *Advanced Technologies Adoption and Use by US Firms: Evidence from the Annual Business Survey*, NBER Working Paper 28290.

Appendix A

Appendix to Chapter 1

A.1 Data and Variable Construction

A.1.1 Project Rationale

I construct two binary variables, $novel_i$ and $biomedical_i$, based on textual descriptions in the TargetTrack information system. This section describes the variable construction process.

First, I use keywords to identify projects that were novel and/or biomedically important. TargetTrack contains a variable called *targetCategoryList* where labs give projects categorical labels such as “biomedical,” “structural coverage,”¹ and so on. It also contains a text field called *targetRationale* where labs give textual descriptions of projects’ rationales. Whenever *targetCategoryList* and *targetRationale* contain the following keywords, I set $novel_i$ equal to 1:

big,² coverage of protein universe, diversity, first structure of class, low sequence identity, mega,³ metagenomic, new fold, no structural information, no structure, number of homologs,⁴ pfam, remote homologs, structural coverage, structural template for unsolved, structure coverage, unsolved families, without any solved structures, without structure.

¹“Structural coverage” means the project is in part of the structure space with no or few published structures.

² BIG and MEGA domain families were defined by the PSI-2 Target Selection Committee as having high value for extensive coverage. These families contained hundreds to tens of thousands of members and many subfamilies which could not be modeled well due to a lack of structural coverage.

³Same as above.

⁴This typo occurs in the raw data.

Whenever *targetCategoryList* and *targetRationale* contain the following keywords, I set *biomedical_i* equal to 1:

activator, adhesion, antibiotic, binding, biochemistry, biological interest, biomedical, cascade, catalyze, cell development, community nominated, community-nominated,⁵ community-nominated, community request, conserved, disease, coronavirus, drug, drug development, drug target, effector, enzyme, essential, function, functional studies, functional, gpcr, high value, high-value,⁶ hiv, homeostasis, host, immune, immunity, infection, infectious, inhibitor, interaction, interact, legionella, medical school, metabolism, mitochondria, model system, operon, parkinsons, partnership, pathogen, pathology, pathway, phosphatase, pneumonia, protein family of high biological importance, reagent, receptor, resistance, resistant, salmonella, school of medicine, secret, sensor, shen lab, shen_lab, shen_selection, stem cell, substrate, syndrome, synthesis, t-cell, t cell, therapeutic, thorson lab, toxoplasma, transcription, transport, tuberculosis, tumor, university, vaccine, vibrio, virulence, virulent.

Second, I use labs' selection protocols of projects for additional information. TargetTrack contains a field where labs describe the protocols they used to conduct each stage of the trials. One type of protocol is the selection protocol. For example, 15 projects were selected because of the protocol "TSel_101," which states "These proteins are important for cell development." I read the descriptions associated with each selection protocol and manually classified whether each protocol was "novel" and/or "biomedical."⁷ Then I set *novel_i* equal to 1 if the project was selected due to a "novel" protocol. I set *biomedical_i* equal to 1 if the project was selected due to a "biomedical" protocol.

Lastly, TargetTrack has a field that contains a list of reference IDs of each molecule in large-scale bioinformatics databases.⁸ These reference ids may yield additional information. Whenever the list of reference ids contains BIG and MEGA reference ids,⁹ I set *novel_i* equal to 1.

A.1.2 Lab Funding

The NIGMS released the following FOAs directly tied to the PSI.

⁵Same as above.

⁶Same as above.

⁷The manual classification is available upon request.

⁸These reference ids include, but are not limited to, the molecule's id in the Protein Data Bank (PDB), UniProt, and the National Center for Biotechnology Information (NCBI) database.

⁹See footnote 2.

Table A.1.1: Funding opportunity announcements (FOA) tied to PSI

Id	Title	Year
RFA-GM-99-009	PILOT PROJECTS FOR THE PROTEIN STRUCTURE INITIATIVE (STRUCTURAL GENOMICS)	1999
PA-99-116	PROTEIN STRUCTURE INITIATIVE (STRUCTURAL GENOMICS)	1999
PA-99-117	PROTEIN STRUCTURE INITIATIVE (STRUCTURAL GENOMICS) – SBIR/STTR	1999
RFA-GM-00-006	PILOT PROJECTS FOR THE PROTEIN STRUCTURE INITIATIVE (STRUCTURAL GENOMICS)	2000
RFA-GM-05-001	LARGE-SCALE CENTERS FOR THE PROTEIN STRUCTURE INITIATIVE	2004
RFA-GM-05-002	SPECIALIZED CENTERS FOR THE PROTEIN STRUCTURE INITIATIVE	2004
RFA-GM-06-004	Structural Genomics Knowledgebase (U01)	2006
RFA-GM-10-004	PSI:Biology Knowledgebase (U01)	2009
RFA-GM-10-005	Centers for High-Throughput Structure Determination (U54)	2009
RFA-GM-10-006	Centers for Membrane Protein Structure Determination (U54)	2009
RFA-GM-10-007	Consortia for High-Throughput-Enabled Structural Biology Partnerships (U01)	2009
PAR-10-214	High-Throughput-Enabled Structural Biology Research (U01)	2010
PAR-11-176	High-Throughput-Enabled Structural Biology Partnerships (U01)	2011

These FOAs allowed me to search directly all grants associated with them in the NIH RePORT database. In addition to the FOAs, I performed a direct search of the labs' names and abbreviations using RePORT's advanced search functionality to obtain data on each labs' supplementary funding. The search term I used

was (quotation marks included):

“[lab full name]” OR “[lab abbreviation]”

I then aggregate each lab’s sum of research grants by year from the search results.

A.1.3 Matching Projects to UniProt Molecule Information

As a preliminary to using the UniProt data, I match projects from the TargetTrack information system to their molecule information on UniProt through two methods.

TargetTrack has a field containing a list of reference ids of each molecule i in large-scale bioinformatics databases. These reference ids include, but are not limited to, the molecule’s id in the Protein Data Bank (PDB), UniProt, and the National Center for Biotechnology Information (NCBI) database. When the UniProt id of the molecule is available in this field, the mapping is direct. I also use the following id types, which easily convert into UniProt molecule id through UniProt’s ID Mapping service (Huang et al., 2011; UniProt, 2021b):

- *PDB_ID*: a molecule’s id in the Protein Data Bank (PDB), a database for 3D structures.
- *P_REFSEQ_AC*: a molecule’s id in NCBI’s RefSeq protein database.
- *EMBL*: a molecule’s corresponding gene’s id in European Molecular Biology Laboratory (EMBL)/GenBank/DNA Data Bank of Japan (DDBJ) CDS database.
- *P_ENTREZGENEID*: a molecule’s corresponding gene’s id in GeneID (Entrez Gene) database.
- *P_GI*: a molecule’s GI number assigned by NCBI.

When the first method fails to find a match (usually due to an entirely missing reference id field or obsolete records in the relevant databases), I use a second method: directly searching the molecule’s sequence of amino acids against all protein sequences in UniProt.¹⁰ I perform this search using DIAMOND (Buchfink et al., 2015, 2021), a very fast algorithm for searching similar sequences. The diamond command I used was:

```
diamond blastp -d [database name] -q [input sequences in .fasta]
-o [output in .csv] -f 6 qseqid qlen sseqid slen evalue bitscore pident length
-b4.0 --top 5
```

¹⁰Downloadable in .fasta format at <https://www.uniprot.org/downloads>.

It produces search results with the following variables:

- *qseqid*: query sequence's identifier (the full sequence in this case).
- *qlen*: query sequence's length.
- *sseqid*: search result's UniProt id.
- *slen*: search result's length.
- *evalue*: the number of expected hits of similar quality that could be found just by chance in a random database of the same size. E-value is a commonly used measure for the degree of similarity between the query sequence and the search result.
- *bitscore*: the required size of a sequence database in which the current match could be found just by chance. Bit score does not depend on the size of the database and is a common alternative measure for the degree of similarity between the query sequence and the search result.
- *pident*: percentage of identical matches between the query sequence and the search result over the alignment length.
- *length*: the alignment length between the query sequence and the search result.

If the query sequence's best match search result, determined by the e-value, a standard metric for assessing sequence similarity, has at least 95% *pident* and the alignment length *length* is at least 67% of both *qlen* and *slen*, I map the query sequence to the result sequence's UniProt id.

I was able to match 262,984 (78.4%) of the 335,553 projects to their UniProt entries through the id mapping method and match an additional 58,593 (17.5%) projects through the direct search. Overall, I was able to map 321,577 (95.8%) projects to their UniProt entries. I then used UniProt's programmatic access for individual entries (UniProt (2021a)) to pull each molecule's information from UniProt. I successfully pulled this information for 319,986 (95.4%) projects.

A.1.4 Data Glossary

This paper uses hundreds of project characteristics extracted from a variety of sources. This data glossary offers a comprehensive view of these variables.

* Variable is included in the characteristics $\mathbf{X}_{ijt}$ in training $\tilde{F}_t(\Omega_t)$.

† Variable is included in the characteristics $\mathbf{X}_{ijt}$ in training $F^*(\Omega_T)$.

‡ Variable is included in the characteristics $\mathbf{X}_{ijt}$ in training $\text{ridge}((\mathbf{X}, \text{citation})_T)$.

§ Variable is included in the characteristics $\mathbf{X}_{ijt}$ in training $\text{ridge}((\mathbf{X}, \Delta \text{download})_T)$.

Please see Appendix A.2 for these models.

Table A.1.2: Data glossary

Variables	Description
$[4 \text{ cap letters then } 6 \text{ digits}]_i^{*\dagger}$	Amino acid attributes from the AAindex database (Kawashima et al. (2007)). Each attribute had an identifier that had four capital letters followed by six digits. I started with the 567 attributes in AAindex1, and then normalized and clustered them to a set of around 30 attribute classes as in Babnigg & Joachimiak (2010). I used scikit-learn’s implementation of affinity propagation clustering, which automatically picked 34 clusters. I then kept the cluster center of each class. For each cluster center attribute, I calculated the local average value, the local minimum, and the local maximum of the sum of the attribute in a seven-amino acid sliding window for molecule i as in Babnigg & Joachimiak (2010). This resulted in 102 variables.
$[\text{consortium abbreviation}]_{ijt}^{*\dagger\ddagger\text{§}}$	Binary variable = 1 if trial j_i was conducted by the given consortium at time t . Only consortia with more than 70 observations of projects in the TargetTrack database have their corresponding variables. 36 variables in total.
$[\text{gene}]_i^{\dagger\ddagger\text{§}}$	Binary variable = 1 if molecule i is coded for by the given gene. From UniProt. 48,548 variables in total. Most variables are very sparse. For † I only include genes that have occurred more than 200 times in my data. For ‡ and § I only include genes that are associated with at least one molecule whose structure was successfully published in my sample.

Table A.1.2 (Continued): Data glossary

$[\text{keyword}]_i$ † ‡ §	Binary variable = 1 if molecule i is associated with the given keyword in UniProt. Examples of keywords include “Alzheimer disease,” “Antioxidant,” “RNA-binding,” “Viral envelope protein.” 1,053 variables in total. Most variables are very sparse. For † I only include keywords that have occurred more than 200 times in my data and remove the keyword “3D-structure” because this is the outcome. For ‡ and § I only include keywords that are associated with at least one molecule whose structure was successfully published in my sample.
$[\text{superkingdom-phylum}]_i$ * † ‡ §	Binary variable = 1 if molecule i comes from an organism in the specific superkingdom and phylum. From UniProt. Due to the large number of species molecules in TargetTrack represent, I do not go down the UniProt taxonomy below phylum. 81 variables in total.
$\text{aminoAcid_}[X]_i$ * †	Counts the number of times amino acid “X” is in molecule i . 20 variables for each of amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Contents of certain amino acids are linked to more successes of trials (Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
$\text{aminoAcidPercent_}[X]_i$ * †	Calculate the amino acid “X” content in molecule i in percentages. 20 variables for each of amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Contents of certain amino acids are linked to more successes of trials (Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
biomedical_i * † ‡ §	Binary variable = 1 if project i was biomedically important. See Appendix A.1.1 for variable construction.

Table A.1.2 (Continued): Data glossary

$citation_{iy}$	five-year citations and mentions of project i published in year y , from PDBe (Varadi et al. (2020)). When multiple structures were published on molecule i , I take the mean values of the five-year citations and year of publication.
$download_{im}$	Number of downloads of published project i in month m across the three major structure databases in the world, from wwPDB (wwPDB (2013)). Available for Aug 2007–Nov 2013.
$\hat{E}(citation_{iy})$	Expected five-year citations and mentions of project i published in year y . See Appendix A.2.2 for construction.
$\hat{E}(download_{iy})$	Expected five-year downloads of project i published in year y . See Appendix A.2.3 for construction.
$\hat{E}_{\tilde{F}_t}(p_{ijt})$	Best-effort replication of the labs' posterior expectation of the probability of success of trial j_i on day t . See Appendix A.2.1 for construction.
$eukaryote_i^{*\dagger\ddagger\S}$	Maximal percentage identity of molecule i to any eukaryotic molecule. To construct this variable, I search each molecule i against all UniProt protein sequences in the Eukaryota superkingdom (UniProt (2021c)). From the search results, I take the maximal percentage identity of i to any eukaryotic molecule as the variable $eukaryote_i$. Due to potentially large number of search results, the search algorithm DIAMOND (Buchfink et al. (2015, 2021)) by default cuts off results at $evalue = 0.001$. $evalue$ is a well-understood metric for search quality in this field. If there are no search results meeting the cutoff, I let $eukaryote_i = 0$.
$exposedAminoAcid_ [X]_i^{*\dagger}$	Counts the number of times amino acid “X” is on the predicted exposed surface of molecule i . 20 variables for each of amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Exposed surface was predicted using the NetSurfP (Klausen et al. (2019)) program with the cutoff of relative solvent accessibility (rsa) > 0.25 . Contents of certain amino acids on the exposed surface of the molecule are linked to more successes of trials (Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).

Table A.1.2 (Continued): Data glossary

<i>exposedAminoAcidPercent</i> _ $[X]_i^{*\dagger}$	Calculate the amino acid “X” content on the predicted exposed surface of molecule i in percentages. 20 variables for each of amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Exposed surface was predicted using the NetSurfP (Klaussen et al. (2019)) program with the cutoff of relative solvent accessibility (rsa) > 0.25. Contents of certain amino acids on the exposed surface of the molecule are linked to more successes of trials (Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
<i>extinctCoeffReduced</i> $_i^{*\dagger}$	Molar extinction coefficient of molecule i with reduced cysteines. Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Slabinski et al. (2007b) used the extinction coefficient as a feature to predict project success.
<i>extinctCoeffOxidized</i> $_i^{*\dagger}$	Molar extinction coefficient of molecule i with disulfid bridges. Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Slabinski et al. (2007b) used the extinction coefficient as a feature to predict project success.
<i>funding</i> $_{ly}$	Total sum of research grants consortium l received from NIH in year y . See Appendix A.1.2 for variable construction.
<i>gaps</i> $_i^{*\dagger}$	The average number of insertions in molecule i ’s alignment compared to homologs in UniProt protein sequences. Computed by searching sequence i against UniProt protein sequences using DIAMOND (Buchfink et al. (2015, 2021)). The output variable <i>gaps</i> captures this value. Insertions were included as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Jahandideh et al. (2014).

Table A.1.2 (Continued): Data glossary

$gapOpen_i^{*\dagger}$	The average number of insertion openings in the alignment compared to homologs in UniProt protein sequences. Computed by searching sequence i against UniProt protein sequences using DIAMOND (Buchfink et al. (2015, 2021)). The output variable $gapOpen$ captures this value. Insertions were included as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Jahandideh et al. (2014).
$gravyIndex_i^{*\dagger}$	Grand average of hydropathicity index (GRAVY) of molecule i , used to represent the hydrophobicity value of a molecule. Calculated using Biopython's ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Hydrophobicity is a key determinant of success of trials (Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
$hasPrevSuccess_{ijkt}^*$	Binary variable = 1 if at least one previous trial on molecule i successfully completed stage k before date t .
$hasPrevFailure_{ijkt}^*$	Binary variable = 1 if at least one previous trial on molecule i failed at stage k before date t .
$human_i^{*\dagger\ddagger\text{§}}$	Maximal percentage identity of molecule i to any human molecule. To construct this variable, I search each molecule i against all UniProt protein sequences in the Homo sapiens (human) species (UniProt (2021d)). From the search results, I take the maximal percentage identity of i to any human molecule as the variable $human_i$. Due to potentially large number of search results, the search algorithm DIAMOND (Buchfink et al. (2015, 2021)) by default cuts off results at $evalue = 0.001$. $evalue$ is a well-understood metric for search quality in this field. If there are no search results meeting the cutoff, I let $human_i = 0$.

Table A.1.2 (Continued): Data glossary

$instabilityIndex_i^{*\dagger}$	Instability index of molecule i , which is an estimate of the stability of the protein in a test tube. Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Instability Index was included as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Jahandideh et al. (2014).
$isoelectricPoint_i^{*\dagger}$	Isoelectric point of molecule i . Calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module. Isoelectric point is a key determinant of success of trials (Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
$membrane_i^{*\ddagger\text{§}}$	Binary variable = 1 if project i ’s UniProt information contains the word “membrane.”
$molecularWeight_i^{*\ddagger\text{§}}$	Molecular weight of molecule i , calculated using Biopython’s ProteinAnalysis function from Bio.SeqUtils.ProtParam module.
$novel_i^{*\ddagger\text{§}}$	Binary variable = 1 if project i was novel. See Appendix A.1.1 for variable construction.
$p_{ij,k-1,t_{k-1}}^{*\dagger}$	The predicted probability of success of stage $k - 1$ of project-trial j_i that started in period t_{k-1} . If $k = 0$, this variable is set to 1.
$percentCoil_i^{*\dagger}$	Predicted percentage of coil secondary structure in molecule i . Predicted using the NetSurfP (Klausen et al. (2019)) program. Secondary structure features were used in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Jahandideh et al. (2014).
$percentCoiledCoil_i^{*\dagger}$	Percentage of coiled-coil regions in molecule i from UniProt. Coiled-coil regions were used in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014).

Table A.1.2 (Continued): Data glossary

<i>percentDisordered_i</i> ^{*†}	Predicted percentage of disordered region in molecule <i>i</i> . Predicted using the NetSurfP (Klaussen et al. (2019)) program. Disordered region was used as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014).
<i>percentDisorderedUniprot_i</i> ^{*†}	Percentage of disordered region in molecule <i>i</i> from UniProt. Disordered region was used as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014).
<i>percentExposed_i</i> ^{*†}	Predicted percentage of amino acids on the exposed surface of molecule <i>i</i> . Exposed surface was predicted using the NetSurfP (Klaussen et al. (2019)) program with the cutoff of relative solvent accessibility (rsa) > 0.25. Extent of the exposed surface of the molecule are linked to more successes of trials (Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014)).
<i>percentHelix_i</i> ^{*†}	Predicted percentage of helix secondary structure in molecule <i>i</i> . Predicted using the NetSurfP (Klaussen et al. (2019)) program. Secondary structure features were used in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014).
<i>percentLowComplexity_i</i> ^{*†}	Predicted percent low-complexity regions in molecule <i>i</i> . Computed using the SEG program (Wootton (1994)). Low-complexity regions were used as features in Slabinski et al. (2007a,b); Jaroszewski et al. (2008).
<i>percentSignalPeptide_i</i> ^{*†}	Percentage of signal peptide in molecule <i>i</i> . From UniProt. Slabinski et al. (2007a,b); Price et al. (2009a); Babnigg & Joachimiak (2010); Jahandideh et al. (2014) state molecules containing signal peptides have very low chances of success.

Table A.1.2 (Continued): Data glossary

$percentStrand_i^{*\dagger}$	Predicted percentage of strand secondary structure in molecule i . Predicted using the NetSurfP (Klaussen et al. (2019)) program. Secondary structure features were used in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014).
$percentTransmembraneHelices_i^{*\dagger}$	Percentage of transmembrane helices in molecule i . From UniProt. Transmembrane helices were used as a feature in Slabinski et al. (2007a,b); Jaroszewski et al. (2008); Price et al. (2009a,b); Babnigg & Joachimiak (2010); Jahandideh et al. (2014) .
$pfam_i$	A list of protein families associated with molecule i , from UniProt (UniProt (2021a)).
$phase1_{ijkt}^{\dagger}$	Binary variable = 1 if stage k of project-trial j_i started in phase 1 of PSI (pilot phase). I let this variable be 1 if the stage start year is before or in 2005. Phase 1 ended in 2004. However, based on Figures A.4.1 and 1.6, one can clearly see that 2004 and 2005 are transition periods: output quantity jumped up in 2004. I therefore let 2004 and 2005 be part of both Phase 1 and Phase 2.
$phase2_{ijkt}^{\dagger}$	Binary variable = 1 if stage k of project-trial j_i started in phase 2 of PSI (production phase). I let this variable be 1 if the stage start year is between 2004 and 2010. Phase 2 is between 2005 and 2008. However, based on Figures A.4.1 and 1.6, one can clearly see that 2009 and 2010 are transition periods: output quantity stayed high but citations reversed the trend. I therefore let 2009 and 2010 be part of both Phase 2 and Phase 3.
$phase3_{ijkt}^{\dagger}$	Binary variable = 1 if stage k of project-trial j_i started in phase 3 of PSI (biomedical phase). I let this variable be 1 if the stage start year is 2009 and beyond.
$prevPub_{iy}^{*\dagger\ddagger\text{\textcolor{teal}{\$}}}$	Number of publications on molecule i by the start of year y , from UniProt (UniProt (2021a)).

Table A.1.2 (Continued): Data glossary

$prevStruct_{iy}^{*\dagger\ddagger\S}$	Number of already published structures in the same protein families associated with molecule i by the start of year y . To construct this variable, I first pull from UniProt the list of protein families $pfam_i$ associated with molecule i . I then obtain a mapping of each protein family to its associated structures from EMBL-EBI (2021) and the structures' publication dates (I take the structure's deposition date to the PDB as the publication date) from Varadi et al. (2020). Merging the datasets results in $prevStruct_{iy}$. If i is associated with multiple protein families, I take the average of the number of already published structures in each protein family associated with i .
$prevSuccesses_{ijk}^{*}$	Number of previous trials on molecule i that have successfully completed stage k before date t .
$prevTrials_{ijk}^{*}$	Number of previous trials on molecule i that have reached stage k before date t .
$refId_i$	A list of reference ids of molecule i in TargetTrack, used to map i to its information in UniProt.
seq_i	Sequence representation of molecule i 's amino acids, unique identifier of project i .
$seqLength_i^{*\dagger\ddagger\S}$	The number of amino acids in molecule i .
$simPrevProj_{it}$	The maximal degree of similarity between project i and all previously attempted projects at time t , measured by the bit score (see Appendix A.1.3 for the definition of bit score). Computed by searching sequence i against all sequences attempted before time t using DIAMOND (Buchfink et al. (2015, 2021)). The maximum of the output variable <i>bitscore</i> among research results was used as $simPrevProj_{it}$.

Table A.1.2 (Continued): Data glossary

$surfaceRuggedness_i^{*\dagger}$	Surface ruggedness of molecule i , defined by the total accessible surface of molecule i divided by the accessible surface predicted based on molecular mass. The total accessible surface of the molecule i is calculated by summing the predicted absolute solvent accessibility of each amino acid from NetSurfP (Klausen et al. (2019)). The accessible surface predicted based on molecular mass is calculated using the formula $6.3(molecularMass)^{0.73}$ (Miller et al. (1987)). Jahandideh et al. (2014) used this variable as a feature.
$trialId_{ij}$	Trial id of project-trial j_i , unique identifier of trial j_i in TargetTrack.
$\widehat{Var}_{\hat{F}_i}(p_{ijt})$	Best-effort replication of the posterior variance of the labs' beliefs about the probability of success of trial j_i on day t . See Appendix A.2.1 for construction.
Y_{ijt}	Binary variable = 1 if trial j_i on date t was successful.
Y_{ijk}	Binary variable = 1 if intermediate stage k of trial j_i on date t was successful. $Y_{ij0t} = 1$ if DNA was successfully cloned. Y_{ij1t} is only defined when $Y_{ij0t} = 1$ and is equal to 1 if protein was successfully expressed. Y_{ij2t} is only defined when $Y_{ij0t} = 1$ and $Y_{ij1t} = 1$ and is equal to 1 if protein was successfully purified. Y_{ij3t} is only defined when $Y_{ij0t}, Y_{ij1t}, Y_{ij2t} = 1$ and is equal to 1 if protein was successfully crystalized for X-ray crystallography or prepared for NMR or cryo-EM. Y_{ij4t} is only defined when all previous stages were successful and is equal to 1 if the structure was successfully produced and deposited to the Protein Data Bank (PDB) for publication.

A.2 Constructing Posteriors

A.2.1 Random Forest for Trial Success Probabilities

There are two kinds of models of trial success probability in this paper. The first one is a model that captures how labs formed posterior beliefs. It does not have to produce an unbiased estimate of the true probability of success of a trial. However, it has to produce an unbiased estimate of the labs' perceived posterior beliefs

about the probability of success. My implementation of $\tilde{F}_t$ closely follows the machine learning approach the labs described in published journal articles.

The second one is a model of the true data generating process of trial success probability F^* , which is used in simulating counterfactual outcomes. Estimating F^* is different from estimating the posterior using $\tilde{F}_t$ because F^* needs to produce an unbiased estimate of the true probability of success of a trial. As such, my implementation of F^* deviates from $\tilde{F}_t$ in several ways to correct the potential bias of and improve upon the machine learning systems the labs described.

In this appendix, I first explain my implementation of $\tilde{F}_t$; then I move on to discuss how my implementation of F^* deviates from that of $\tilde{F}_t$.

Implementation of $\tilde{F}_t$

My implementation of $\tilde{F}_t$ fits stage-specific models to account for information embodied in outcomes of intermediate stages. Recall that each trial j_i has multiple sequential stages and the overall probability of success of j_i is equal to the product of the probabilities of success for all sequential stages $p_{ijt} = \prod_{k=0}^4 p_{ijk t}$. The intermediate outcomes Y_{ijk} for stages $k = 0, 1, \dots$ up to the point when the overall trial failed/succeeded provides information for future trials' potential.

For a given quarter $q(t)$ and each of the stages $k = 0, 1, 2, 3, 4$, I let the information set $\Omega_{k,q(t)}$ consist of project-trial outcomes realized before quarter $q(t)$ at stage k and these project-trials' characteristics. Following when the labs started to use machine learning to form posterior, I let $q(t)$ to be between 2005 and 2015. For $q(t) = 2005Q1$, I use the trial outcomes realized before 2005 and these trials' characteristics as the initial information set Ω_{2005Q1} . Project-trial characteristics $\mathbf{X}_{ijk t}$ I use for training $\tilde{F}_t$ and prediction of labs' posterior beliefs fall under three categories:¹¹

- Physicochemical properties of molecule i based on scientific reasoning. These variables were identified by the series of journal articles the labs published and were quite similar across labs and time.
- Other characteristics of project i , for examples, novelty, biomedical importance, and the number of prior publications on molecule i .
- Past successes and failures of project i at stage k .

¹¹Please see Appendix A.1.4 for the full list of variables used.

Then, for the given quarter $q(t)$ and each of the stages $k = 0, 1, 2, 3, 4$, I fit a random forest model $\tilde{F}_{k,q(t)}(\Omega_{k,q(t)})$ using `RandomForestClassifier` from python package `scikit-learn`. Random forest is an *ensemble*¹² machine learning method. The algorithm constructs a large number of decision trees at training time. Each decision tree is a learning model that aims to find the project-trial characteristics predictive of success/failure in the training set. When it comes to prediction, the trained random forest classifier $\tilde{F}_{k,q(t)}(\Omega_{k,q(t)})$ would pool individual trees and average predicted values of $\{\hat{p}_{ijkt}^{ntree}\}$ from individual trees as the final output. As Jahandideh et al. (2014) did, I set the number of trees in the random forest equal to 1000.

Decision trees and random forests are known for often overfitting without regularization. To avoid overfitting, I regularize by restricting the hyperparameters max_depth ,¹³ $min_samples_leaf$,¹⁴ $max_features$,¹⁵ and $min_samples_split$.¹⁶ I perform model selection with a grid search of the combinations of the four hyperparameters.¹⁷ For each hyperparameter combination, I evaluate the model with five-fold cross validation using `scikit-learn`'s `cross_validate` function. In each iteration of the cross-validation, the function fits a random forest on four out of five cross-validation folds and then computes the cross-validation score by comparing the model's predictions with the actual data from the remaining fold. I use the average log likelihood (`log_loss` scoring in `scikit-learn`) as the cross-validation scoring method. I choose the hyperparameter combination that maximizes the average log likelihood in cross-validation.

After training the models $\tilde{F}_{k,q(t)}$ for $k = 0, 1, 2, 3, 4$ for a given $q(t)$, I predict $\widehat{E}_{\tilde{F}_{q(t)}}(p_{ijt})$ and $\widehat{Var}_{\tilde{F}_{q(t)}}(p_{ijt})$ for each project-trial in the choice set C_{it} at decision time t as follows. I first collect the predictions $\{\hat{p}_{ijkt}^{ntree}\}$ from the 1000 individual decision trees in $\tilde{F}_{k,q(t)}(\Omega_{k,q(t)})$, and then compute $\hat{p}_{ijt}^{ntree} = \prod_{k=0}^4 \hat{p}_{ijkt}^{ntree}$. There are

¹²Ensemble methods use multiple learning models to obtain better predictive performance than could be obtained from any of the constituent learning models alone.

¹³This hyperparameter determines the maximum depth of each decision tree.

¹⁴This hyperparameter determines the minimum number of observations a node in the decision tree must have before it can be split.

¹⁵This hyperparameter determines the maximum number of features to consider when looking for the best split.

¹⁶This hyperparameter determines the minimum number of observations required to split a node.

¹⁷To reduce computational burden, I do not perform model selection for all $\tilde{F}_{k,q(t)}$. Rather, for each $k = 0, \dots, 4$, I construct $\Omega_{k,T}$ using all outcomes at stage k and only perform model selection for $\tilde{F}_{k,T}$ on this full training set. I then use the selected hyperparameters to train the models $\tilde{F}_{k,q(t)}$ where $q(t) = 2005Q1, 2005Q2, \dots, 2015Q4$. The set of max_depth used in grid search is $[int(\log(sample_size, 2)), 2 \cdot int(\log(sample_size, 2)), 3 \cdot int(\log(sample_size, 2)), 4 \cdot int(\log(sample_size, 2))]$. The set of $min_samples_leaf$ used in grid search is $[1, 2, 4]$. The set of $max_features$ used in grid search is $[0.1, 0.2, 0.3, 0.4]$ of the total number of features. The set of $min_samples_split$ used in grid search is $[8, 16, 32, 64, 128]$.

1000 values in the set $\{\hat{p}_{ijt}^{ntree}\}$. I let

$$\hat{E}_{\tilde{F}_{q(t)}}(p_{ijt}) = \bar{p}_{success,ijt}^{ntree}, \quad (\text{A.1})$$

$$\widehat{Var}_{\tilde{F}_{q(t)}}(p_{ijt}) = s^2(p_{success,ijt}^{ntree}) \quad (\text{A.2})$$

Although I extensively reference the labs' implementations of machine learning systems when I implement $\tilde{F}_t$, my estimate of the posterior is not a perfect replica of the labs' posteriors. I note why replicating perfectly the labs' posterior beliefs would be difficult and where my implementation corresponds to and deviates from the labs' learning and updating process below:

- I include in $\mathbf{X}_{ijt}$ the set of physicochemical properties of molecules identified in the labs' published journal articles (Slabinski et al., 2007a,b; Jaroszewski et al., 2008; Price et al., 2009a,b; Babnigg & Joachimiak, 2010; Jahandideh et al., 2014). This set is the union of the sets of such properties in different articles (to minimize the risk of selection on unobservables) and is fixed for all labs and time periods in my implementation. In contrast, though similar across lab and time, the set of physicochemical properties the labs used in training and prediction still varied. It is impossible to capture all of these potential variations during the labs' long operational history (some may not have been recorded by the published articles).
- The construction of some variables in $\mathbf{X}_{ijt}$ requires using software packages that are constantly being updated or have become obsolete. I make my best effort to construct variables using methods as close to the labs' original approach as possible (see Appendix A.1.4).
- The $\mathbf{X}_{ijt}$ in my implementation includes past trial outcomes of projects while the labs' implementations did not explicitly include those characteristics. Still, it is reasonable to believe that researchers working on a project would update their beliefs on the potential of the project upon seeing a trial success/failure.
- I use random forest as the model of posterior updating for all labs and time periods. In contrast, the machine learning models the labs used in training and prediction varied across lab and across time. It is impossible to capture all of these potential variations during the labs' long operational history (some may not have been recorded by the published articles).
- I set the frequency of "updating" and refitting models at the quarterly interval. In contrast, the labs' actual belief updating frequency is not clearly documented. I use the quarterly interval because training

models at a finer interval, such as at the daily frequency, places large computational and storage burden. The day-to-day change of the information set Ω was also relatively small. Therefore, to improve computational tractability, I coarsen the frequency of refitting new models to quarterly.

- My model predicts the overall potential of success of a trial while the labs' implementations focused on predicting the potential of success of bottleneck stages of a trial. That is, for stages where success rates were usually reasonable (for example cloning the DNA), the labs were often not explicitly reliant on something as rigorous as supervised machine learning systems to form and update beliefs, while they were explicitly reliant on such systems for predicting the potential of success in crystallizing a molecule and studying its structure through X-ray crystallography.
- The output the labs' systems produced may not exactly be $\widehat{E}_{\tilde{F}_{q(t)}}(p_{ijt})$ and $\widehat{Var}_{\tilde{F}_{q(t)}}(p_{ijt})$. For example, the model in Slabinski et al. (2007b) predicted the probability of success as an intermediate outcome. The final output was an integer score between 1 and 5, where 1 represents "optimal" and 5 represents "very difficult." The labs' systems did not always predict $\widehat{Var}_{\tilde{F}_{q(t)}}(p_{ijt})$. When they did, the measure took the form of comparing predictions from multiple models side by side (Slabinski et al., 2007a,b; Babnigg & Joachimiak, 2010; Jahandideh et al., 2014). It is reasonable to believe that labs had some understanding that predictions from different models (or submodels of an ensemble model) differed, and looking at how those predictions varied was valuable, though they did not percolate the idea down to form an additional metric just to measure that variation. This seems consistent with the notation that the labs used heuristics to guide their exploration of high-variance projects.

Implementation of F^*

The implementation of F^* is almost identical to that of $\tilde{F}_t$ except for a few deviations. First of all, a new model $\tilde{F}_{k,q(t)}$ (for stages $k = 0, \dots, 4$) is trained for every quarter $q(t)$ between 2005Q1 and 2015Q4, incorporating new trial outcomes realized in each quarter. In contrast, F_k^* (for stages $k = 0, \dots, 4$) is trained only on the full information set Ω_T . Ω_T covers the characteristics and outcomes of all trials in my trial allocations and outcomes dataset in the entire sample period.

Second, F^* uses additional covariates to correct the potential bias of $\tilde{F}_t$ in predicting trial success probabilities. The model $\tilde{F}_t$ may be biased in predicting trial success probabilities because it does not account for the propensity of observing a specific stage of a trial. To see this, think about the probability of success of

stage 1 of a trial. We observe stage 1 of a trial only if stage 0 of the trial was successful. If the probabilities of success of stages 0 and 1 are positively correlated, then we are more likely to observe stage 1 of trials that are more likely to succeed in stage 1. Therefore, models trained with the observed data on stage 1 would produce prediction results that are positively biased. Correcting this bias is simple if we assume that the selection into observing a given stage is only based on observable characteristics of trials: we can use the predicted probability of success of the previous stage as the propensity score of observing the given stage. As such, I include $p_{ij,k-1,t_{k-1}}^*$, the predicted probability of success of stage $k-1$ of trial j_i that started in period t_{k-1} , as a covariate when I train F_k^* . For stage $k=0$, I set this variable to 1. The labs’ published articles offer no discussion about this source of bias, so I do not include this variable in training $\tilde{F}_t$.

Another difference between F^* and $\tilde{F}_t$ is that F^* does not include variables on previous trial outcomes as covariates. In simulations, all previous trial outcomes of a project are simulated and should not shift the project’s true probability of success; therefore, the simulated counterfactual outcome of a trial should not be based on the simulated previous outcomes.

To further improve the predictive power of F^* , I include in F^* project-trial characteristics the labs did not use in their machine learning systems. I include keywords and genes associated with the molecule i . I also include three phase-specific dummy variables to capture the effects of different phases of the grant program on the probabilities of success, as I learned during my conversations with NIH program officers that the labs underwent retooling corresponding to the changes of phases.

A.2.2 Ridge Regression for Citations

Let the model be $\text{ridge}(\mathbf{X}, \text{citation})_T$, where the training set $(\mathbf{X}, \text{citation})_T$ represents the characteristics and citation_{iy} of all published projects in my data. The goal of this model is to predict $E(\text{citation}_{iy} | \mathbf{X}_{iy}, \text{ridge}((\mathbf{X}, \text{citation})_T))$, the expected number of five-year citations a project i published in year y would generate conditional on the project’s characteristics $\mathbf{X}_{iy}$.

The number of characteristics that could potentially predict higher citations is very large. Characteristics ranging from the organism the molecule i is from to the gene that expresses molecule i could all contribute to the biomedical significance and research interests on molecule i . The number of characteristics is on the order of hundreds, most of which are very sparse, while I only have 10,424 observations.¹⁸ This calls for

¹⁸The total number of published projects in my data is 10,501. 77 published projects in my data did not give the PDB ids of their publications so I was not able to map their citation information.

regularization to avoid overfitting.

I use a ridge regression from the python package `scikit-learn`. The project characteristics $\mathbf{X}_{iy}$ included for model fitting are shown in Appendix A.1.4. I choose the regularization hyperparameters using cross-validation with the `RidgeCV` function provided by the `scikit-learn` package.

I standardize the outcome variable $citation_{iy}$ by subtracting away the lab mean value of this variable and then dividing by the lab standard deviation as publications from different labs had large variations in citation numbers. When the model makes a prediction, I multiply the predicted value with the lab standard deviation and add the lab mean to get the predicted citations.

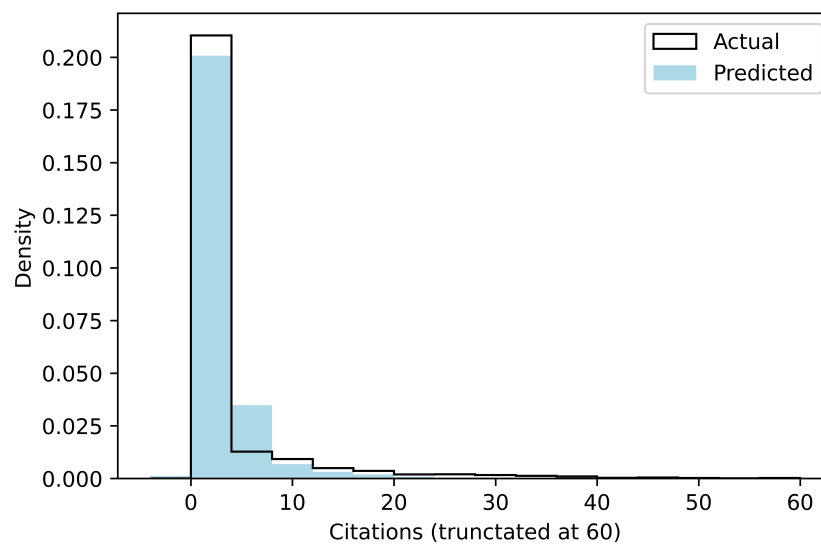
I assess model fit by comparing the actual citations with their out-of-sample predicted citations under five-fold cross validation. Figure A.2.1 shows the distribution of the predicted citations correctly captures a high proportion of zero values in the actual data. Figure A.2.2 shows a scatterplot of predicted citations against the actual citations. A linear regression of the predicted citations on the actual citations without constant shows an $R^2 = 0.580$.

The predicted citations $\hat{E}(citation_{iy}|\mathbf{X}_{iy}, \text{Ridge}((\mathbf{X}, citation)_T))$ has a y subscript because some of the important characteristics vary with time, for example, the number of publications on molecule i prior to the year of publication of the structure. In Section 1.3.3, since we do not know which exact year the structure of each trial in the choice set would have been published if the trial was allocated, I remove the y subscript by averaging the predicted citations for each molecule across years so that

$$\hat{E}(citation_i) = \frac{1}{16} \sum_{y=2000}^{2015} \hat{E}(citation_{iy}|\mathbf{X}_{iy}, \text{ridge}((\mathbf{X}, citation)_T)). \quad (\text{A.3})$$

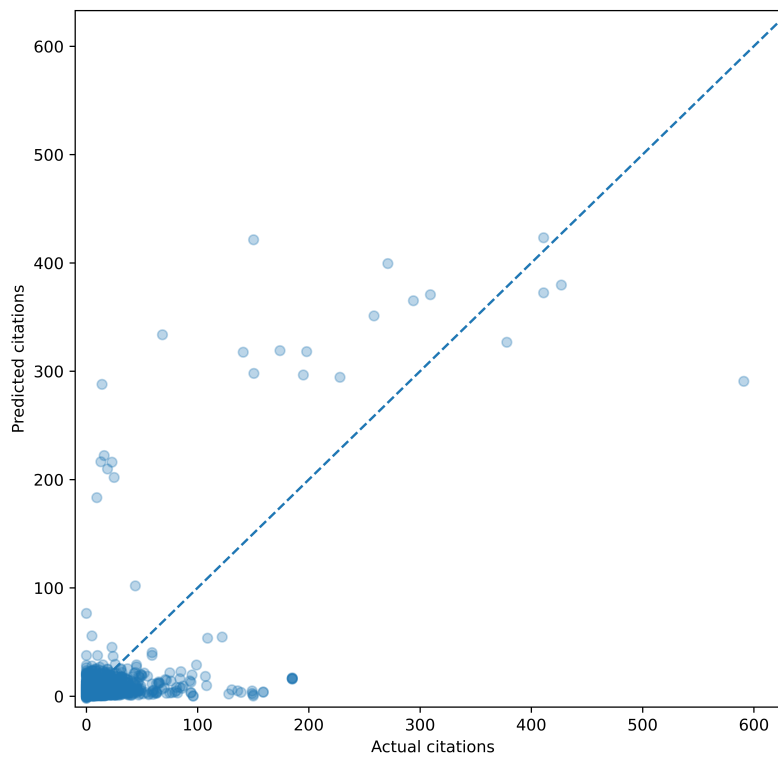
In simulations, I simulate both the outcome (success or failure) of an allocated trial and the date that outcome is realized. In that case, the y subscript is preserved for the predicted citations of the publication.

Figure A.2.1: Distributions of actual citations and predicted citations



Note: Distributions are truncated at 60. Bin width is 4.

Figure A.2.2: Scatterplot of predicted citations against actual citations



A.2.3 Ridge Regression for Downloads

Before we start predicting downloads, the raw download data needs to be transformed because the number of downloads shows a strong time trend over the lifecycle of a publication. The raw download data I obtain consists of monthly downloads on the structure level between Aug 2007 and Nov 2013 for all publicly available structures human beings know of. The number of structure-month observations is 5,484,800. As some structures in a given month observed were published a long time ago while some just got published, comparing these structures' raw download counts in a month would be misleading. As shown in Figure A.2.3, downloads peak within a month since the publication of a structure and then sharply decline over the following months until reaching some steady level in approximately two years.

I perform a transformation of the raw download data to detrend it as follows. Let $pubAge(i, m)$ be the age of publication i (in months) in month m . To detrend, I first compute $\overline{download}_{pubAge(i, m)}$, the mean downloads of structures that have been published for $pubAge(i, m)$ months.¹⁹ I then compute how much the number of downloads structure i had in month m deviates from this mean, $\widetilde{download}_{im} = download_{im} - \overline{download}_{pubAge(i, m)}$. I then define $\Delta download_i$, the average deviation of structure i 's monthly downloads from the mean download trend, by the average of $\widetilde{download}_{im}$, in other words $\Delta download_i = \overline{\widetilde{download}_{im}}$. I treat the variable $\Delta download_i$ as the outcome variable. If there are multiple structures on the same project i , I take the mean of their average deviations as $\Delta download_i$. I then match the download data with the project-trial characteristics of completed projects in my data.

Let the model be $ridge((\mathbf{X}, \Delta download)_T)$, where the training set $(\mathbf{X}, \Delta download)_T$ represents the characteristics and $\Delta download_i$ of all published projects in my data. The goal of this model is to predict $E(\Delta download_{iy} | \mathbf{X}_{iy}, ridge((\mathbf{X}, \Delta download)_T))$, the expected average deviation of the structure's monthly downloads from the mean download trend for a project i published in year y , conditional on the project's characteristics $\mathbf{X}_{iy}$. There is a y subscript because some of the important characteristics vary with time, for example, the number of publications on molecule i prior to the year of publication of the structure.

The number of characteristics that could potentially predict higher downloads is very large. Characteristics ranging from those of molecule i 's organism to those of molecule i 's gene could all contribute to the level of interest on molecule i . The number of characteristics is on the order of hundreds, most of which are very

¹⁹I pool observations with $pubAge > 25$ months in computing this mean as the mean number of downloads flattens by 25 months since publication.

sparse, while I only have 10,424 observations. This calls for regularization to avoid overfitting.

I use a ridge regression from the python package `scikit-learn`. The project characteristics $\mathbf{X}_{iy}$ included for model fitting are shown in Appendix A.1.4. I choose the regularization hyper-parameters using cross-validation with the `RidgeCV` function provided by the `scikit-learn` package.

I assess model fit by comparing the actual $\Delta\text{download}_{iy}$ with their out-of-sample predicted values under five-fold cross validation. Figure A.2.4 shows a comparison of the distributions. As ridge regression shrinks all regression coefficients towards zero, the distribution of the predicted values is narrower. Still, the predicted values capture the rank order of the actual data well. Figure A.2.5 shows a scatterplot. The plot shows a relationship quite close to the line $y = x$. Figure A.2.6 shows a binned scatterplot.

In Section 1.3.3 where we show descriptives, I perform additional transformations on the predicted value $\hat{E}(\Delta\text{download}_{iy}|\mathbf{X}_{iy}, \text{ridge}((\mathbf{X}, \Delta\text{download})_T))$. First, the predicted value has a y subscript because some of the important characteristics vary with time, for example, the number of publications on molecule i prior to the year of publication of the structure. Since we do not know which exact year the structure of each trial in the choice set would have been published if the trial was allocated, I remove the y subscript by averaging the predicted values for each molecule across years so that

$$\hat{E}(\Delta\text{download}_i) = \frac{1}{16} \sum_{y=2000}^{2015} \hat{E}(\Delta\text{download}_{iy}|\mathbf{X}_{iy}, \text{ridge}((\mathbf{X}, \Delta\text{download})_T)). \quad (\text{A.4})$$

Second, the variable $\hat{E}(\Delta\text{download}_i)$ predicts the average deviation of monthly downloads from a trend and is difficult to interpret. I therefore transform this variable to a prediction of five-year downloads $\hat{E}(\text{download}_i)$ by using the following formula:

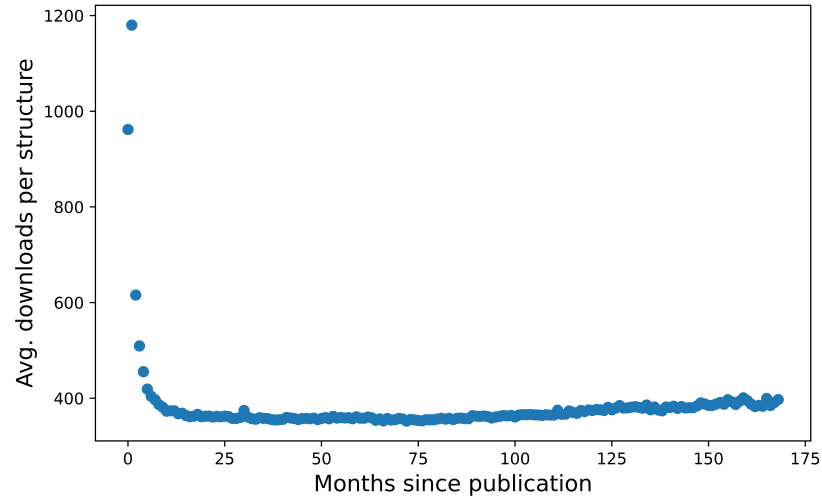
$$\hat{E}(\text{download}_i) = \sum_{\text{pubAge}=0}^{59} \overline{\text{download}_{\text{pubAge}}} + 60 \times \hat{E}(\Delta\text{download}_i), \quad (\text{A.5})$$

where $\sum_{\text{pubAge}=0}^{59} \overline{\text{download}_{\text{pubAge}}} = 23960.11$ is computed based on the full download data with 5,484,800 structure-month observations.

In simulations, I simulate both the outcome (success or failure) of an allocated trial and the date that outcome is realized. In that case, the y subscript is preserved for the predicted average deviation of monthly downloads of the published projects. Therefore the predicted five-year downloads becomes

$$\hat{E}(\text{download}_{iy}) = \sum_{\text{pubAge}=0}^{59} \overline{\text{download}_{\text{pubAge}}} + 60 \times \hat{E}(\Delta \text{download}_{iy} | \mathbf{X}_{iy}, \text{ridge}((\mathbf{X}, \Delta \text{download})_T)). \quad (\text{A.6})$$

Figure A.2.3: Average downloads per structure in months since publication



Note: The plot is based on 5,484,800 structure-month observations of download counts between Aug 2007 and Nov 2013. Each blue dot aggregates in this data the average monthly downloads for structures published m months ago.

Figure A.2.4: Distributions of predicted $\hat{E}(\Delta download_{iy})$ against actual $\Delta download_{iy}$

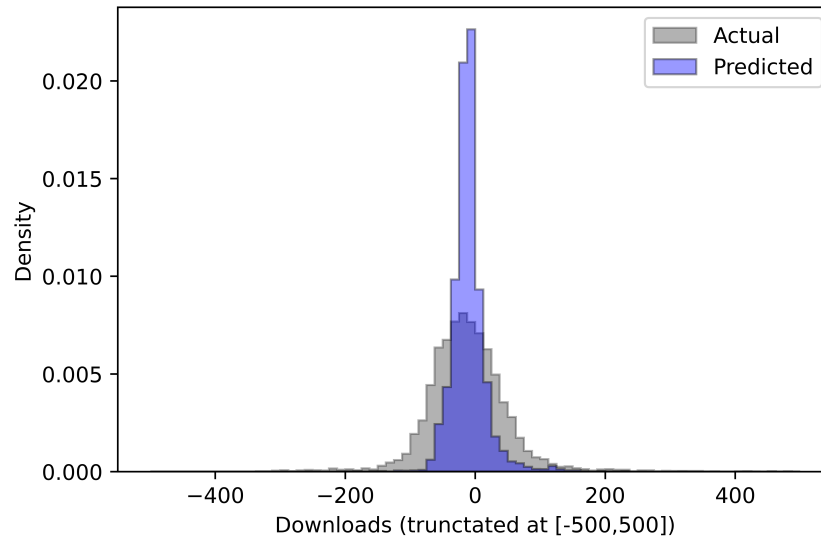


Figure A.2.5: Scatterplot of predicted $\hat{E}(\Delta download_{iy})$ against actual $\Delta download_{iy}$

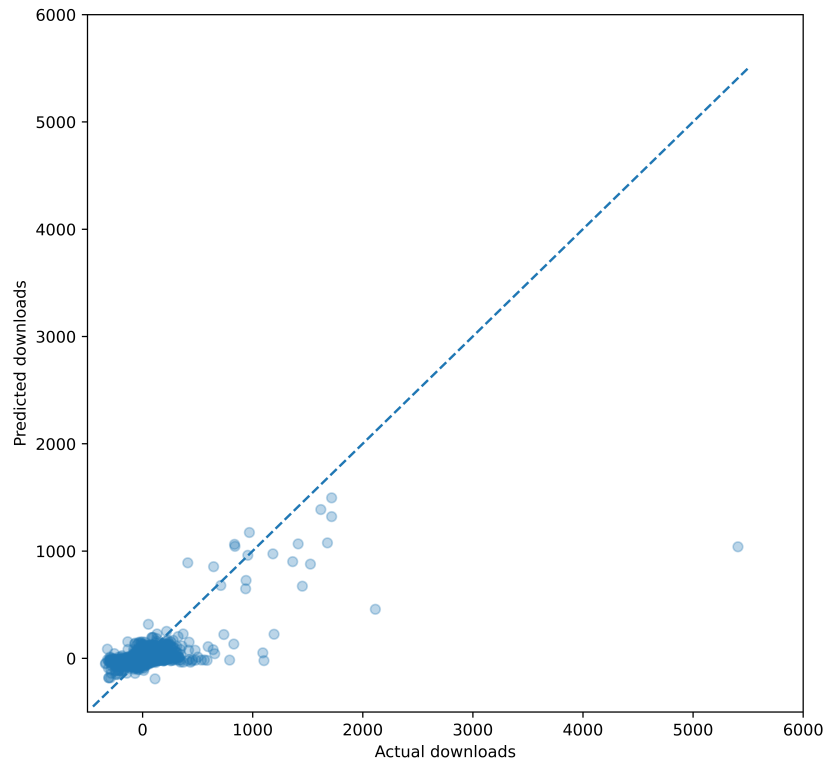
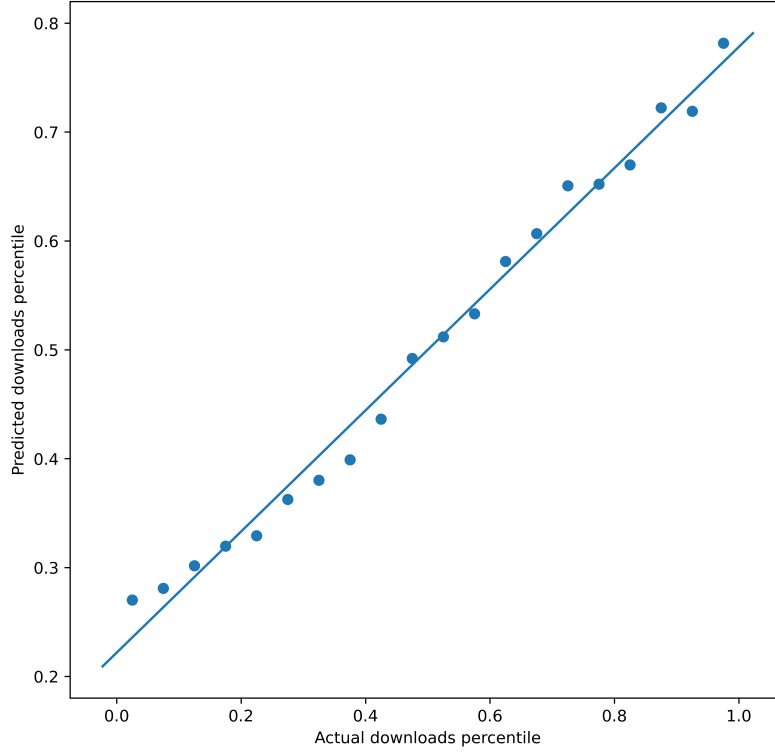


Figure A.2.6: Binned scatterplot of predicted $\hat{E}(\Delta\text{download}_{iy})$ against actual $\Delta\text{download}_{iy}$



A.3 Additional Details on Modeling and Estimation Procedure

A.3.1 Functional Form of Payoff Function $\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl})$

To begin, recall that the payoff $\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl})$ of project-trial j_i at t given action $\mathbf{a}_{lt}$ has a probability distribution depending on p_{ijt} . The lab or the economist does not perfectly know p_{ijt} , but previous outcomes of trials reveal information about it so one can form a posterior $\tilde{F}_t(p_{ijt}|\Omega_t)$. Integrating $\pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl})$ over the posterior, we obtain the posterior expected payoff $\pi_{ijt}(\Omega_t, \mathbf{a}_{lt}; \boldsymbol{\theta}_{Xl})$.

I define a function $q(\mathbf{a}_{lt}, p_{ijt})$ that maps the probability of success to the probability of *payoff* of project-trial j_i at t given actions. When a project-trial is not allocated on day t , it does not pay off even though it may have a nonzero probability of success. Moreover, the labs often simultaneously allocated multiple trials to the same project. A successful trial j_i of project i should only receive payoff if the simultaneous trials $(j-m)_i, (j-m+1)_i, \dots, (j-1)_i$ fail, because we have assumed only the first successful trial/publication on a project produces welfare. Each trial j_i is a Bernoulli trial with probability p_{ijt} . We can express $q(\mathbf{a}_{lt}, p_{ijt})$

as follows:

$$q(\mathbf{a}_{lt}, p_{ijt}) = a_{ijt}(1 - p_{ijt})^m p_{ijt}, \quad (\text{A.7})$$

where trials $(j - m)_i, (j - m + 1)_i, \dots, (j - 1)_i$ are in the choice set C_{lt} and trial $(j - m)_i$ is the smallest-numbered trial of project i in C_{lt} .²⁰ When $m = 0$, trial j_i is the smallest-numbered trial of project i in C_{lt} and its probability of payoff is simply $a_{ijt} p_{ijt}$.

I then specify a deterministic reward function, which captures the amount of payoff a lab will get when a trial pays off. I let the reward function $r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl})$ be a function of project i 's characteristics $\mathbf{X}_{it}$ on day t . $\boldsymbol{\theta}_{Xl}$ are the welfare weights on $\mathbf{X}_{it}$ and are to be estimated. To reduce the number of parameters, I restrict $\mathbf{X}_{it}$ to correspond to the set of NIH evaluation metrics. $\mathbf{X}_{it}$ includes a constant 1 to capture preference for quantity; *novel_i* and *prevStruct_{iy}* to capture preference for novelty; *biomed_i* and *prevPub_{iy}* to capture preference for biomedical importance; and *human_i*, *eukaryote_i*, and *membrane_i* to capture preferences for human, eukaryotic, and membrane proteins, respectively. I let $r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl})$ have a simple linear form

$$r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) = 1 \cdot \theta_{quant,l} + \text{biomed}_i \cdot \theta_{biomed,l} + \dots + \text{membrane}_i \cdot \theta_{membrane,l}. \quad (\text{A.8})$$

Whenever a trial pays off, the lab receives a baseline amount $\theta_{quant,l}$ plus additional amounts depending on the other characteristics of the project.

We can then break the posterior expected payoff into a few pieces:

$$\begin{aligned} \int \pi_{ijt}(\mathbf{a}_{lt}, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t) &= \int r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) \cdot q(\mathbf{a}_{lt}, p_{ijt}) d\tilde{F}_t(p_{ijt}|\Omega_t) \\ &= a_{ijt} \cdot r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) \underbrace{\int \overbrace{[(1 - p_{ijt})^m p_{ijt}]^{\text{let it be } M_{ijt}}} d\tilde{F}_t(p_{ijt}|\Omega_t)}_{\text{estimated offline}}. \end{aligned} \quad (\text{A.9})$$

Notice that M_{ijt} only depends on p_{ijt} . Since we have estimated $\tilde{F}_t(p_{ijt}|\Omega_t)$ offline (see Appendix A.2.1), we can estimate $E_{\tilde{F}_t}(M_{ijt})$ and $\text{Var}_{\tilde{F}_t}(M_{ijt})$ offline as well.²¹

Also notice that given the breakdown of the posterior expected payoff in equation (A.9), V_{ijt}^A does not depend on the full action vector $\mathbf{a}_{lt}$. It only depends on the action a_{ijt} . Plugging in equation (A.9) into V_{ijt}^A in

²⁰ $q(\mathbf{a}_{lt}, p_{ijt}) = a_{ijt}(1 - p_{i,j-m,t}) \dots (1 - p_{i,j-1,t}) p_{ijt}$. As all trials on day t share the same information set Ω_t , the posteriors for $p_{i,j-m,t}, \dots, p_{ijt}$ are the same.

²¹ Each trial j_i is a Bernoulli trial with probability of success p_{ijt} ,

$$E_{\tilde{F}_{q(t)}}(M_{ijt}) = E_{\tilde{F}_{q(t)}}((1 - p_{ijt})^m p_{ijt}) = [1 - E_{\tilde{F}_{q(t)}}(p_{ijt})]^m \cdot E_{\tilde{F}_{q(t)}}(p_{ijt}), \quad (\text{A.10})$$

equation (1.3) and evaluating it at $a_{ijt} = 1$, we obtain:

$$\begin{aligned} V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l) &= \int \pi_{ijt}(a_{ijt} = 1, p_{ijt}; \boldsymbol{\theta}_{Xl}) d\tilde{F}_t(p_{ijt}|\Omega_t) + B_{ijt}(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_{Bl}) \\ &= r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) E_{\tilde{F}_t}(M_{ijt}) + B_{ijt}(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_{Bl}). \end{aligned} \quad (\text{A.12})$$

For the main model, $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l) = r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) E_{\tilde{F}_t}(M_{ijt}) + \sqrt{\frac{\theta_{B1,l}}{j}} + \theta_{B2,l} \cdot (t - t'_{i,t})$. For alternative model 1, where $B_{ijt}(\cdot) = 0$, this reduces to $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l) = r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) E_{\tilde{F}_t}(M_{ijt})$. For alternative model 2, $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l) = r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) E_{\tilde{F}_t}(M_{ijt}) + \psi(\cdot) r(\mathbf{X}_{it}; \boldsymbol{\theta}_{Xl}) \text{Var}_{\tilde{F}_t}(M_{ijt})$. Evaluating V_{ijt}^A at $a_{ijt} = 0$, we obtain $V_{ijt}^A(\Omega_t, a_{ijt} = 0; \boldsymbol{\theta}_l) = 0$ for all models.

Moreover, with this functional form, the second constraint in equation (1.4) will always be guaranteed by the solution. Notice that in most models including the main model, for all $j_i < j'_i \in C_{lt}$, $V_{ijt}^A(\Omega_t, a_{ijt} = 1; \boldsymbol{\theta}_l) \geq V_{ij't}^A(\Omega_t, a_{ij't} = 1; \boldsymbol{\theta}_l)$ because the two terms only differ by $E_{\tilde{F}_t}(M_{ijt}) \geq E_{\tilde{F}_t}(M_{ij't})$. For models based on the Gittins index, the two terms also differ by $\text{Var}_{\tilde{F}_t}(M_{ijt}) \geq \text{Var}_{\tilde{F}_t}(M_{ij't})$ and the constraint continues to be satisfied. In equation (1.10), we add an ε_{it} to both terms and the constraint continues to be satisfied.

A.3.2 Specifying Choice Set C_{lt}

As discussed in Section 1.2, the major labs in my data received new projects through three ways with close NIH involvement: 1) a centralized planning committee periodically assigned families of novel molecules; 2) the biomedical research community nominated projects; and 3) the labs determined projects of their own interest, which they reported to the NIH well in advance. These processes placed limits on the new projects the labs could plausibly consider when they made trial allocations. These limits allow me to considerably reduce C_{lt} .

I restrict C_{lt} to include only the following project-trials. For an older project i that the lab has attempted up until trial j_i in period $t' < t$, I include trials $(j+1)_i, \dots, (j+n_t)_i$ in C_{lt} . For a new project i' that the lab has not attempted until t but attempts within the next six months, I include trials $1_{i'}, \dots, (n_t)_{i'}$ in C_{lt} . One can also consider using alternative windows for the new projects, such as projects attempted within the next three months or nine months. Doing so changes the magnitudes of the estimates, but all qualitative results are the

$$\begin{aligned} \text{Var}_{\tilde{F}_{q(t)}}(M_{ijt}) &= \text{Var}_{\tilde{F}_{q(t)}}((1 - p_{ijt})^m p_{ijt}) \\ &= \{\text{Var}_{\tilde{F}_{q(t)}}(p_{ijt}) + [E_{\tilde{F}_{q(t)}}(p_{ijt})]^2\} \times \{\text{Var}_{\tilde{F}_{q(t)}}(1 - p_{ijt}) + [E_{\tilde{F}_{q(t)}}(1 - p_{ijt})]^2\}^m \\ &\quad - [E_{\tilde{F}_{q(t)}}(p_{ijt})]^2 \times \{[E_{\tilde{F}_{q(t)}}(1 - p_{ijt})]^2\}^m. \end{aligned} \quad (\text{A.11})$$

Plugging in $\hat{E}_{\tilde{F}_{q(t)}}(p_{ijt})$ and $\widehat{\text{Var}}_{\tilde{F}_{q(t)}}(p_{ijt})$ into the above equations, one obtains $\hat{E}_{\tilde{F}_{q(t)}}(M_{ijt})$ and $\widehat{\text{Var}}_{\tilde{F}_{q(t)}}(M_{ijt})$.

Table A.3.1: Trials included in the reduced choice set for estimation

Project	Trial	Actually allocated on day t ?	Notes
i	$(j+1)_i$	Y	Trials $(j+1)_i$ through $(j+m-1)_i$ were actually allocated on day t , include all in C_{lt}^R . When no trial was allocated on day t , m equals 1.
	$\vdots$	$\vdots$	
	$(j+m-1)_i$	Y	
	$(j+m)_i$	N	Trials $(j+m)_i$ through $(j+n_{lt})_i$ were not actually allocated on day t , include one random trial $(j+r)_i$ in C_{lt}^R .
	$\vdots$	$\vdots$	
	$(j+r)_i$	N	
	$\vdots$	$\vdots$	
	$(j+n_{lt})_i$	N	

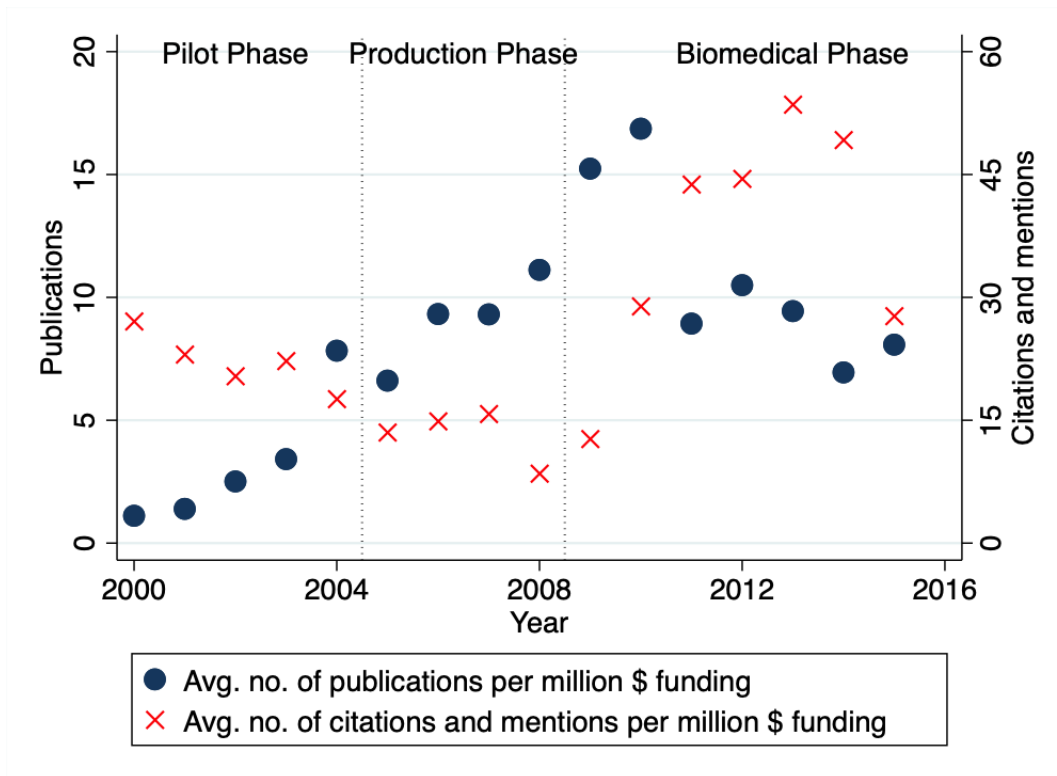
Notes: For an older project that has been attempted before t , j equals the number of trials allocated to the project before t . For a new project, $j = 0$. Trials in black are included in the reduced choice set C_{lt}^R . Trials in grey are in the choice set C_{lt} but are excluded from C_{lt}^R to reduce computational burden.

same as when using six months as the window. Likewise, adding some new projects that the lab could have considered but never actually attempted could change the estimates, but the qualitative results should stay the same.

I further reduce the sizes of the choice sets used in estimation by taking random subsamples of C_{lt} . The reason is related to computation. Labs at times allocate hundreds of project-trials on a day and they usually had tens of thousands of projects in their portfolios. The sizes of some choice sets C_{lt} could be on the order of millions. Given that we have thousands of periods, if we use the full choice sets C_{lt} , we need to compute log likelihood for billions of choices in each iteration of maximum likelihood. This would result in a very large memory burden and slow computation. Moreover, it is not necessary to include every possible choice in C_{lt} to consistently estimate θ_l . A random subsample of the choices on each side of the threshold value would be sufficient. Due to the sheer number of projects in each lab's portfolio, the set of actual trials is much smaller than the set of not-allocated trials. I therefore reduce the sizes of the choice sets for estimation by taking random subsamples of the latter on the level of project and date. Let the reduced choice sets be $C_{lt}^R \subset C_{lt}$. Table A.3.1 shows the project-trials in C_{lt}^R after random sampling from C_{lt} .

A.4 Additional Results

Figure A.4.1: Observed output: number of publications and citations



Note: The blue dots show the number of published structures on unique molecules in a given year divided by the lab consortium's funding in millions in that year. The red crosses show the number of 5-year citations and mentions the published structures in that year generated, divided by the lab consortium's funding in millions in that year. Each plot shows the average value across the four large lab in each year. The disaggregated values are in Figure 1.6.

Table A.4.1: Comparison of likelihoods of different models, JCSG

Model	Free parameters		Avg $\hat{P}(a_{ijt}^o = 1; \theta_l)$	Avg $\hat{P}(a_{ijt}^o = 0; \theta_l)$
	in $B_{ijt}(\cdot)$	Log likelihood	actual allocation	actual nonallocation
Static	0	-1,308,021	0.691	0.891
Gittins	0	-1,104,171	0.640	0.910
UCB	1	-544,118	0.854	0.981
FlexGittins	1	-1,056,772	0.640	0.917
FlexGittins+D	2	-349,531	0.644	0.981
UCB+D	2	-206,951	0.909	0.995

Note: Estimation uses data between 2005 and 2015 from JCSG, one of the four large labs in the data. The number of trials actually allocated was 320,295. The number of trials in the choice sets after random sampling is 5,807,902. The rest of the notes of Table 1.2 apply.

Table A.4.2: Comparison of likelihoods of different models, MCSG

Model	Free parameters	Log likelihood	Avg $\hat{P}(a_{ijt}^o = 1; \theta_l)$	Avg $\hat{P}(a_{ijt}^o = 0; \theta_l)$
	in $B_{ijt}(\cdot)$		actual allocation	actual nonallocation
Static	0	-636,089	0.644	0.994
Gittins	0	-473,481	0.669	0.996
UCB	1	-274,230	0.769	0.998
FlexGittins	1	-402,754	0.681	0.997
FlexGittins+D	2	-225,201	0.780	0.998
UCB+D	2	-157,987	0.850	0.999

Note: Estimation uses data between 2005 and 2015 from MCSG, one of the four large labs in the data. The number of trials actually allocated was 141,059. The number of trials in the choice sets after random sampling is 39,136,035. The rest of the notes of Table 1.2 apply.

Table A.4.3: Comparison of likelihoods of different models, NYSGRC

Model	Free parameters in $B_{ijt}(\cdot)$	Log likelihood	Avg $\hat{P}(a_{ijt}^o = 1; \theta_l)$ actual allocation	Avg $\hat{P}(a_{ijt}^o = 0; \theta_l)$ actual nonallocation
Static	0	-677,382	0.538	0.990
Gittins	0	-550,827	0.561	0.992
UCB	1	-339,498	0.682	0.996
FlexGittins	1	-529,851	0.565	0.993
FlexGittins+D	2	-400,671	0.626	0.994
UCB+D	2	-288,162	0.700	0.996

Note: Estimation uses data between 2005 and 2015 from NYSGRC, one of the four large labs in the data. The number of trials actually allocated was 139,276. The number of trials in the choice sets after random sampling is 23,883,552. The rest of the notes of Table 1.2 apply.

Table A.4.4: Out-of-sample fit of UCB+D model

Lab	Sample	Avg Log likelihood	Avg $\hat{P}(a_{ijt}^o = 1; \hat{\theta}_l)$ actual allocations	Avg $\hat{P}(a_{ijt}^o = 0; \hat{\theta}_l)$ actual nonallocations
JCSG	in	-0.033	0.909	0.994
	out	-0.036	0.914	0.996
MCSG	in	-0.004	0.826	0.999
	out	-0.004	0.871	0.999
NESG	in	-0.004	0.817	0.999
	out	-0.003	0.850	0.999
NYSGRC	in	-0.013	0.656	0.996
	out	-0.011	0.734	0.997

Note: To compute these results, I first estimate the UCB+D model using only observed allocation decisions in odd years. I then compute the in-sample results using the estimates and the odd years' decisions which I used to fit the model. I compute the out-of-sample results using even years' decisions. The rest of the notes of Table 1.2 apply.

Table A.4.5: Estimates of parameters in different models, NESG

	UCB+D		Static		Gittins		UCB		FlexGittins		FlexGittins+D	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	2005–2008	2009–2015	2005–2008	2009–2015
θ_{B1}	158.3 [156.4, 160.1] -2.3	119.5 [118.1, 121.4] -4.7	-	-	-	-	210.5 [208.8, 212.5]	188.8 [188.5, 189.2]	1.6 [1.6, 1.6]	1.4 [1.4, 1.4]	1.4 [1.4, 1.4]	0.8 [0.8, 0.8]
θ_{B2}	-2.3, -2.2	[-4.7, -4.7]	-	-	-	-	-	-	-	-	-6.5 [-6.4, -6.6]	-8.4 [-8.4, -8.4]
θ_{quant}	105.6 [105.5, 105.8]	95.9 [95.8, 95.9]	288.9 [286.3, 291.4]	232.6 [232.4, 232.9]	167.5 [167.1, 167.8]	162.3 [162.1, 162.5]	106.1 [106.0, 106.3]	146.1 [146.1, 146.2]	150.0 [149.3, 151.0]	149.5 [149.4, 149.8]	138.4 [137.9, 139.1]	143.5 [143.3, 143.7]
θ_{novel}	28.4 [27.8, 28.9]	-23.3 [-23.4, -23.1]	202.6 [200.5, 204.3]	20.7 [20.5, 20.8]	67.6 [67.3, 67.8]	-8.5 [-8.6, -8.4]	28.7 [28.6, 28.9]	-39.8 [-39.8, -39.8]	48.4 [47.9, 49.1]	-14.4 [-14.5, -14.3]	38.7 [38.1, 39.4]	-7.7 [-7.7, -7.7]
$\theta_{prevStructZ}$	18.1 [17.6, 18.6]	20.1 [20.1, 20.3]	9.8 [9.4, 10.0]	56.5 [56.4, 56.6]	3.8 [3.8, 3.9]	7.9 [7.8, 8.0]	22.3 [22.1, 22.5]	47.6 [47.5, 47.7]	3.0 [2.7, 3.2]	5.2 [5.1, 5.2]	-0.4 [-0.7, -0.1]	4.9 [4.8, 5.0]
θ_{hiomed}	21.5 [21.3, 21.7]	52.9 [52.7, 52.9]	45.3 [43.1, 47.0]	158.5 [158.3, 158.8]	7.3 [7.1, 7.5]	44.1 [44.0, 44.2]	21.5 [21.4, 21.7]	102.5 [102.4, 102.5]	8.0 [7.8, 8.3]	38.0 [37.8, 38.4]	3.4 [2.8, 3.9]	29.3 [29.1, 29.3]
$\theta_{prevPubZ}$	-9.2 [-9.3, -9.0]	-3.0 [-3.1, -3.0]	-14.5 [-14.7, -14.2]	-1.7 [-1.7, -1.7]	-6.9 [-7.0, -6.8]	-0.5 [-0.5, -0.4]	-9.4 [-9.6, -9.2]	-6.8 [-6.8, -6.7]	-6.9 [-7.2, -6.7]	-4.5 [-4.5, -4.4]	-3.8 [-3.9, -3.6]	-3.1 [-3.1, -3.1]
θ_{human}	67.1 [66.9, 67.3]	124.0 [123.9, 124.0]	105.9 [104.8, 106.9]	191.0 [190.7, 191.3]	51.4 [51.2, 51.6]	90.5 [90.4, 90.6]	69.8 [69.7, 70.0]	160.6 [160.6, 160.7]	40.2 [40.1, 40.5]	83.0 [82.9, 83.1]	36.3 [36.0, 36.7]	74.5 [74.5, 74.6]
$\theta_{eukaryote}$	-29.8 [-29.9, -29.7]	-35.9 [-36.0, -35.8]	67.7 [67.1, 69.1]	61.6 [61.5, 61.8]	-8.3 [-8.7, -8.0]	-26.9 [-27.0, -26.8]	-30.0 [-30.4, -29.8]	-0.3 [-0.3, -0.2]	-14.3 [-14.4, -13.9]	-33.7 [-33.7, -33.6]	-11.7 [-12.0, -11.6]	-33.4 [-33.4, -33.4]
$\theta_{membrane}$	58.3 [58.0, 58.6]	13.1 [13.1, 13.3]	96.5 [96.1, 97.2]	52.3 [52.1, 52.4]	35.2 [35.1, 35.4]	28.4 [28.3, 28.5]	58.1 [57.9, 58.5]	22.7 [22.7, 22.8]	31.2 [31.1, 31.4]	24.8 [24.8, 24.9]	23.5 [23.1, 23.9]	22.1 [22.1, 22.2]

Note: Table displays the full estimates of parameters in different models for NESG, one of the four large labs. Results from other labs are available upon request. $prevStructZ_{iy}$ represents the number of standard deviations by which $prevStruct_{iy}$ differs from the yearly mean $prevStruct_y$. $prevPubZ_{iy}$ represents the number of standard deviations by which $prevPub_{iy}$ differs from the yearly mean $prevPub_y$. The rest of the notes of Table 1.3 apply.

Table A.4.6: Estimates of main parameters of interest in UCB+D model, other labs

Parameter	JCSG		MCSG		NYSGRC	
	2005–2008 (1)	2009–2015 (2)	2005–2008 (3)	2009–2015 (4)	2005–2008 (5)	2009–2015 (6)
θ_{B1}	558.2 [551.1, 573.4]	1001.9 [977.6, 1028.0]	273.3 [266.2, 284.8]	127.1 [126.3, 127.8]	61.6 [61.2, 61.9]	115.6 [114.2, 116.9]
θ_{B2}	-247.9 [-247.6, -248.5]	-104.7 [-102.9, -106.6]	-3.9 [-3.9, -3.9]	-3.8 [-3.8, -3.8]	-2.9 [-2.9, -3.0]	-3.9 [-3.9, -3.9]
θ_{biomed}	-34.3 [-34.6, -34.1]	105.1 [104.7, 105.8]	12.7 [12.6, 12.7]	84.8 [84.8, 84.8]	33.0 [32.9, 33.1]	89.4 [89.1, 89.8]

Note: Table displays the estimates of the main parameters of interest in the UCB+D model for the other three large labs. Full estimates of parameters in different models are available upon request. 95% confidence intervals are computed using the MCMC approach in Chernozhukov & Hong (2003) and are shown in brackets. These confidence intervals are almost identical to those computed using Procedure 1 of Chen et al. (2018).

Table A.4.7: Simulated outcomes of UCB+D model, other labs

Lab	Model	Projects attempted	Unique publications	Citations	Downloads (millions)
JCSG	UCB+D	40,881	1,607	1,495	38.4
	Actual	40,881	1,512	1,463	36.3 [†]
MCSG	UCB+D	77,503	2,040	2,524	46.3
	Actual	78,740	2,276	3,145	50.0 [†]
NYSGRC*	UCB+D	59,734	626	2,579	14.7
	Actual	59,734	617	2,575	14.4 [†]

Note: Each simulation uses $\hat{\theta}_l$ from parameter estimates of the corresponding model. Results are averaged from three simulations of each model. [†]The download data on actually published projects are between 2007 and 2013, so actual five-year downloads may not be available for some projects. I predict five-year downloads for the actually published projects using the predictive model described in Appendix A.2.3. *I note data problems related to NYSGRC. For this particular lab, more than half of the trials that produced structures either miss key stage dates or have those dates in wrong orders (for example, publication is at an earlier date than previous stages). As a result, I am not able to correctly simulate the dates and outcomes of different stages of trials. Though this set of simulation results looks quite nice, please take them with a grain of salt.

Table A.4.8: Out-of-sample simulation results for UCB+D model

Lab	Model	Projects attempted	Unique publications	Citations	Downloads (millions)
JCSG	in	40,881	1,607	1,495	38.4
	out	40,881	1,621	1,481	38.6
	actual	40,881	1,512	1,463	36.3 [†]
MCSG	in	77,503	2,040	2,524	46.3
	out	77,504	2,051	2,553	46.4
	actual	78,740	2,276	3,145	50.0 [†]
NESG	in	59,947	1,097	3,376	25.6
	out	59,913	1,085	3,336	25.3
	actual	59,953	1,053	3,502	24.5 [†]
NYSGRC*	in	59,734	626	2,579	14.7
	out	59,734	628	2,594	14.7
	actual	59,734	617	2,575	14.4 [†]

Note: In-sample simulation results are identical to those for the UCB+D model in Tables 1.4 and A.4.7. I simulate the out-of-sample results as follows. I first fit the UCB+D model for each lab with odd years of observed decisions. Using those estimated parameters, I simulate each lab's full input allocation history and output, in odd and even years. The rest of the notes of Table A.4.7 apply.

Table A.4.9: Counterfactual outcomes, no exploration, other labs

Lab	Counterfactual model	Projects attempted	Unique publications	Citations	Downloads (millions)
JCSG	Static	10,710	1,338	1,270	32.4
		(-74%)	(-17%)	(-15%)	(-16%)
	Baseline model	40,881	1,607	1,495	38.4
MCSG	Static	14,883	668	856	15.6
		(-81%)	(-67%)	(-66%)	(-66%)
	Baseline model	77,503	2,040	2,524	46.3
NYSGRC*	Static	5,203	247	1,026	5.9
		(-91%)	(-61%)	(-60%)	(-60%)
	Baseline model	59,734	626	2,579	14.7

Note: Each simulation uses $\hat{\theta}_{Xl}$ from parameter estimates of the UCB+D model for the corresponding lab. These estimates are available upon request. Results are averaged from three simulations of the model. Output from the baseline model are identical to those for the UCB+D model in Appendix Table A.4.7. Parentheses show percentage differences as compared to the baseline model. Table shows results from the three other large labs. See Table 1.5 for results from NESG. *I note data problems related to NYSGRC. For this particular lab, more than half of the trials that produced structures either miss key stage dates or have those dates in wrong orders (for example, publication is at an earlier date than previous stages). As a result, I am not able to correctly simulate the dates and outcomes of different stages of trials. Though this set of simulation results looks quite nice, please take them with a grain of salt.

Appendix B

Appendix to Chapter 2

B.1 Bilateral Bargaining Model

In this section, we formalize the common intuition that the demand for and usage of data at the application layer alters investment incentives to interconnect at the internet layer with a simple theoretical model of bilateral bargaining between network operators. The model largely draws from Besen et al. (2001). Though this model abstracts away many issues, such as interdependence of interconnection decisions, customers' choices of networks and the rich set of considerations different types of networks have in making interconnection decisions, it is parsimonious and delivers neat analytical solutions of the bargaining outcome and the amount of transfers.

First let there be two network operators O_1 and O_2 . The two networks decide whether to interconnect. Let mass M_1, M_2 account for the combined value of each network's content and users, and the value of its customers not reachable through the other network. So M_1 is the value reachable through O_1 or O_1 's customers and not reachable through O_2 or O_2 's customers. Let I_1, I_2 be the combined value of all content and users on the Internet not reachable through the other network or its customers. M_1 is a subset of I_1 and M_2 is a subset of I_2 . When network i is a large transit provider, I_i would be equal to all content and users on the Internet minus M_j . Examples of value are a content provider's video content, and the ISP's video subscribers.

For transit providers and content delivery networks, customers would be other networks depending on them to connect to other parts of the Internet. For governments, private companies and universities, their customers are just themselves. Assume that from O_1 's perspective, forming a peer-to-peer interconnection with O_2 would allow O_1 to reach mass M_2 more efficiently. O_1 can in term generate revenue from its

customers due to improved service. Assume also that O_2 would reach mass M_1 more efficiently under the peer-to-peer agreement. If O_1 is the provider, forming a provider-to-customer link with O_2 would allow O_1 to reach M_2 more efficiently. While from O_2 's perspective, forming a customer-to-provider link with O_1 would allow O_2 to access I_1 .¹

Let $f(m)$ be function of revenue collected by the network operator O_i per unit mass of its customers, where m represents the mass of customers in the internet O_i is able to reach in a reliable and efficient manner for its customers. For networks whose customers are themselves, we can think of f as the benefit of connecting their networks to the Internet in monetary values. Assume f is an increasing function and is concave. Let $C_{p2p}(M_1 + M_2)$ be the cost of a p2p interconnection and $C_{p2c}(I_1 + M_2)$ be the cost of a p2c interconnection between O_1 and O_2 , which are increasing functions in the total masses that depend on the interconnection. C is concave, evident from decreasing per Mbps interconnection fees in this industry. Moreover $\frac{dC_{p2p}(m)}{dm}|_{m=\tilde{m}} < \frac{dC_{p2c}(m)}{dm}|_{m=\tilde{m}}$ for all m , reflecting the fact that p2p agreements have more rapidly declining per Mbps cost than p2c agreements and significantly reduce cost of interconnection especially when m is large. Let τ be any additional cost associated with negotiating an agreement.

Assume any disruption to data exchange between O_1 and O_2 is only sustained during bargaining² and customers do not change their networks during bargaining or in response to the bargaining outcome. We also hold fixed the interconnection agreements between either of O_1 , O_2 and all other networks. Assume these other agreements allow each network to access mass G_1 and G_2 . O_1 and O_2 can either form an agreement with one of the three agreement types: (a) a p2c agreement where O_1 is a provider to O_2 , (b) a p2p agreement, (c) a c2p agreement where O_1 is a customer to O_2 , or take the outside option (d) no agreement. In practice, as the relative masses and bargaining power of the two networks strongly influence the type of agreement formed³, we first assume networks compare one of (a), (b), (c) with the outside option (d), rather than comparing all of the four options simultaneously, and derive comparative statics. We then discuss potential substitutions between agreement types.

¹Note M_1 , M_2 , I_1 and I_2 are specific to the negotiation between O_1 and O_2 . If O_1 and O_2 form a p2c agreement where O_1 is the provider, under a negotiation between O_1 and another network O_3 , M_2 becomes part of O_1 's combined value of content and users M'_1 .

²In the event of no agreement between O_1 and O_2 , customers in M_1 and M_2 experience less efficient service in reaching I_2 and I_1 . In practice, data usually takes a longer and inefficient path through a series other networks between O_1 and O_2 .

³For reference, if we measure M_1 , M_2 and I_1 purely in terms of the number of IP addresses and let O_2 be the smaller network in an agreement, the average ratio of M_2 to M_1 is 0.81 for p2p agreements, while the average ratio of M_2 to the full routed IP address space (which is close to I_1 , given the relatively small size of M_2) is 0.00016 for p2c agreements.

B.1.1 Peer-to-Peer Agreements (p2p)

When the two networks have relatively similar masses and bargaining power, they consider either a p2p agreement or no agreement. The bargaining outcome according to the noncooperative bargaining theory with short times between offers is approximately the same as that of the Nash bargaining model, provided the payoff each earns during the period of disruption is treated as the Nash threat point (Binmore, Rubinstein, & Wolinsky (1986), Besen et al. (2001)). The total surplus to be divided when O_1 and O_2 reach an agreement is $M_1f(G_1 + M_2) + M_2f(G_2 + M_1) - C_{p2p}(M_1 + M_2) - \tau$, while the threat point is $(M_1f(G_1), M_2f(G_2))$. We further assume when the mass M_i of a network O_i increases, this change has a higher impact on O_j 's threat point value than on the cost of interconnection, that is $\frac{d(M_jf(G_j+M_i))}{dM_i}|_{M_i=\tilde{M}} > \frac{dC_{p2p}(M_i+M_j)}{dM_i}|_{M_i=\tilde{M}}$ for all M_i, M_j .

O_1 and O_2 would decide to interconnect if the gains from agreement

$$\begin{aligned} g &= M_1f(G_1 + M_2) + M_2f(G_2 + M_1) - C_{p2p}(M_1 + M_2) - \tau \\ &\quad - M_1f(G_1) - M_2f(G_2) \geq 0. \end{aligned} \tag{B.1}$$

At a noncooperative bargaining outcome, the two networks divide equally any gains relative to the threat point, so the resulting bargaining payoff for network O_1 is

$$\begin{aligned} \pi_1 &= \frac{1}{2} [M_1f(G_1 + M_2) + M_2f(G_2 + M_1) - C_{p2p}(M_1 + M_2) - \tau \\ &\quad + M_1f(G_1) - M_2f(G_2)]. \end{aligned} \tag{B.2}$$

and for network O_2 is

$$\begin{aligned} \pi_2 &= \frac{1}{2} [M_1f(G_1 + M_2) + M_2f(G_2 + M_1) - C_{p2p}(M_1 + M_2) - \tau \\ &\quad - M_1f(G_1) + M_2f(G_2)]. \end{aligned} \tag{B.3}$$

With interconnection, O_1 would be able to earn a revenue of

$M_1f(G_1 + M_2)$ from M_1 and needs to share half the cost of the interconnection $\frac{1}{2}[C_{p2p}(M_1 + M_2) + \tau]$. Let O_1 's profit be

$\rho_1 = M_1f(G_1 + M_2) - \frac{1}{2}[C_{p2p}(M_1 + M_2) + \tau]$, then the excess

$$\pi_1 - \rho_1 = \frac{1}{2}M_2[f(G_2 + M_1) - f(G_2)] - \frac{1}{2}M_1[f(G_1 + M_2) - f(G_1)]. \tag{B.4}$$

is the negotiated net payment from O_2 to O_1 . Define $h(M) = \frac{[f(G+M)-f(G)]}{M}$, then O_1 receives a positive

payment from O_2 if and only if $h(M_1) - h(M_2) > 0$. In such a case, O_1 and O_2 are in a paid peering agreement. When $h(M_1) - h(M_2) = 0$, the two networks are in a settlement-free peering agreement.

Now suppose O_1 is a network serving customers in the EU while O_2 is some other network outside the EU that connected with O_1 before the GDPR was implemented. We can work out the comparative statics for changes in bargaining outcomes following changes in model parameters due to the GDPR. We consider two different changes in model parameters: (a) a decrease in M_1 , and (b) an increase in τ . Goldberg et al. (2019) shows large and significant 10% decline in recorded page views, visits, orders and revenue of EU customers after the implementation of the GDPR. Jia et al. (2021) show decline in venture capital investment in technology start-ups, particularly in the total amounts raised across funding deals, the number of deals, and the amount raised per individual deal. The effects are especially pronounced for newer and data-related ventures. Both papers provide some evidence of decline in the mass of EU customers, both in terms of the number of users and the amount of content supplied to the rest of the Internet. This change is represented by a decrease in M_1 in our model. As the new legislation rolled out, it creates uncertainty in the business environment and additional burden in making sure both interconnecting parties and their customers are GDPR-compliant, increasing bargaining frictions. We represent this change by an increase in τ in our model.

Taking the derivative of the gains from agreement with respect to M_1 , we have

$$\frac{dg}{dM_1} = f(G_1 + M_2) - \frac{dC_{p2p}(M_1 + M_2)}{dM_1} - f(G_1) + M_2 \frac{df(G_2 + M_1)}{dM_1} > 0. \quad (\text{B.5})$$

It is also easy to show $\frac{d[h(M_1) - h(M_2)]}{dM_1} > 0$. Together, these derivatives imply two changes when M_1 decreases: (1) Gains from agreement fall. If the gains fall below zero, the agreement between O_1 and O_2 breaks. (2) O_1 receives a reduced amount of transfer from O_2 , though we do not observe transfers in our data. Using similar derivations, an increase in τ would also imply higher chance of termination of the interconnection agreement, though it does not have an effect on the transfers.

B.1.2 Provider-to-Customer Agreements (p2c)

When O_1 has substantially more mass and bargaining power than O_2 , the networks consider either a p2c agreement where O_1 is the provider or no agreement. Using the same set of assumptions as above for the p2p

agreements, O_1 and O_2 would decide to interconnect if the gains from agreement

$$g = M_1 f(G_1 + M_2) + M_2 f(G_2 + I_1) - C_{p2c}(I_1 + M_2) - \tau - M_1 f(G_1) - M_2 f(G_2) \geq 0. \quad (\text{B.6})$$

The negotiated net payment from O_2 to O_1 is

$$\pi_1 - \rho_1 = \frac{1}{2} M_2 [f(G_2 + I_1) - f(G_2)] - \frac{1}{2} M_1 [f(G_1 + M_2) - f(G_1)] > 0. \quad (\text{B.7})$$

Suppose O_1 is a transit provider in EU and a significant portion of I_1 are EU users and content. The GDPR might result in a decrease in I_1 . Taking the derivatives of Equations B.6 and B.7 with respect to I_1 , we derive two changes when I_1 decreases: (1) Gains from agreement fall. If the gains fall below zero, the agreement between O_1 and O_2 breaks. (2) O_1 receives a reduced amount of transfer from O_2 , though we do not observe transfers in our data.

If we instead suppose O_2 is a EU network seeking access to I_1 and the GDPR decreases M_2 , we take the derivatives of Equations B.6 and B.7 with respect to M_2 and derive two changes: (1) Gains from agreement fall. If the gains fall below zero, the agreement between O_1 and O_2 breaks. (2) O_1 receives an increased amount of transfer from O_2 , though we do not observe transfers in our data.

B.1.3 Substitution between Agreement Types

When O_1 has a larger mass than O_2 , it is possible the two networks decide between a p2c agreement where O_1 is the provider and a paid p2p agreement. The two networks would enter a paid p2p agreement if Equation B.1 holds and the gains from a p2p agreement are greater than the gains from a p2c agreement

$$g_\Delta = M_2 [f(G_2 + M_1) - f(G_2 + I_1)] + C_{p2c}(I_1 + M_2) - C_{p2p}(M_1 + M_2) \geq 0 \quad (\text{B.8})$$

Suppose O_1 is a EU network and the GDPR negatively impacts both M_1 and I_1 . $\frac{dg_\Delta}{dM_1} > 0$, implying that a decrease in M_1 , holding all else fixed, would make it more likely for the two networks to enter a p2c agreement. However, $\frac{dg_\Delta}{dI_1} < 0$, implying that a decrease in I_1 , holding all else fixed, would make it more likely for the two networks to enter a p2p agreement. The overall effect is unclear and depends on the relative

changes to the masses M_1 and I_1 offered, their prices and O_2 's revenue function.⁴ Suppose instead O_2 is a EU network and the GDPR negatively impacts M_2 . Taking the derivative, $\frac{dg_{\Delta}}{dM_2}$ can either be positive or negative, depending on the cost functions, O_2 's revenue function, and the relative masses M_1 and I_1 .

In summary, this simple model formalizes the intuition that negative impacts of the GDPR on the application layer negatively impact European networks' bargaining positions. European networks would have fewer agreements of all three types and receive a reduced amount of transfers. The effect of the GDPR on the potential substitutions between agreement types is unclear.

B.2 Data Appendix

In this section, we provide additional information about our data sources and data collection techniques. Our data comes from various data sources collected and compiled by the Center of Applied Internet Data Analysis (CAIDA) at the University of California, San Diego. Since 1998, CAIDA has been studying interconnectivity of the Internet by actively probing the Internet using its monitors placed at various vantage points around the world. Its current flagship active measurement infrastructure, Archipelago, collects interconnectivity data on the IP-address-level from more than 200 monitors located on 6 continents in over 60 countries. A list of current Archipelago monitor locations can be found at <https://www.caida.org/projects/ark/locations/>.

CAIDA also collaborates with many organizations and compiles data collected from their monitors. Most notably, it collaborates with the Route Views Project at the University of Oregon and the Réseaux IP Européens Network Coordination Centre (RIPE NCC) in Europe to collect routing tables for network-level paths. A list of Route Views monitors can be found at

<http://www.routeviews.org/routeviews/index.php/collectors/>. A list of RIPENCC monitors can be found at <https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris/ris-raw-data>.

Moreover, CAIDA gathers records of network registration information from the world's five regional Internet registries (RIRs), allowing us to identify countries (or territories) of organizations that own individual networks. The dataset is available through the link:

<https://www.caida.org/data/as-organizations/>. The five RIRs are:

⁴An intuitive way to understand this situation is to use the second-degree price discrimination framework. One can view the p2c agreement as the product with a larger quantity and a higher price and the p2p agreement as the product with a smaller quantity and a lower price. The choice between the two products depends on the consumer's preferences as well as the structure of non-linear pricing.

- The African Network Information Center (AFRINIC)
- The American Registry for Internet Numbers (ARIN)
- The Asia-Pacific Network Information Center (APNIC)
- The Latin America and Caribbean Network Information Center (LACNIC)
- The Réseaux IP Européens Network Coordination Centre (RIPE NCC)

Our main data on the network-level interconnection agreements comes from the routing tables, while our IP-address-level interconnection points for each agreement come from the active probes. The data extraction process is explained in the Data section in the main text.

A number of key variables in this study come from a dataset called *AS Relationships*, as in the computer science field, an independently operated network connected to the Internet is referred to as an Autonomous System (AS). This dataset is available through the link <http://www.caida.org/data/as-relationships/>.

To construct the AS Relationships dataset, CAIDA collects BGP tables from its partner monitors placed at various vantage points across the Internet and peered directly with networks' BGP routers, typically major ones with large numbers of routes stored, at Internet exchange points. Network-to-network connection agreements are then extracted from routing paths announced in these BGP tables. Then the agreements are annotated with inferred agreement types. The inference algorithm draws from Gao (2001); Subramanian et al. (2002); Erlebach et al. (2002); Battista et al. (2003); Xia & Gao (2004); Dimitropoulos et al. (2005, 2007).

Our IP-address-level interconnection points within each agreement come from the dataset *IPv4 Prefix-Probing*. This dataset is available through the link https://www.caida.org/data/active/ipv4_prefix_probing_dataset.xml.

To keep visibility consistent throughout our sample periods, we extract agreements only from a set of monitors that operated throughout our sample periods, January 2015–June 2019 for AS Relationships and December 2015–June 2019 for IPv4 Prefix-Probing. Moreover, we dropped all of the affected interconnections due to configuration changes in three RIPENCC monitors in October 2018. To make these sample restrictions, we use nonpublic versions of the datasets which include monitor identifiers for each observation of interconnection.

We drop networks owned by a number of small island countries, Andorra, Central African Republic, Eritrea, North Korea and Vatican City from our sample due to these countries' very small overall number of connections with the rest of the Internet. Our EEA subsample includes networks owned by organizations headquartered in the 31 EEA member countries as well as networks owned by EU-wide organizations. For networks owned by EU-wide organizations, their countries of origin are shown as "EU" in network registration records. We include these networks in the EEA subsample for the purpose of our empirical analysis. The resulting total number of countries and territories in our sample is 200. A complete list of countries and territories in our sample is presented in Table B.2.1.

Table B.2.1: List of Countries and Territories

EEA	Non-EEA OECD	Non-EEA Non-OECD	
Austria	Australia	Afghanistan	Somalia
Belgium	Canada	Albania	South Africa
Bulgaria	Chile	Algeria	Sri Lanka
Croatia	Israel	Angola	Sudan
Cyprus	Japan	Argentina	Suriname
Czech Rep.	Korea (Rep.)	Armenia	Swaziland
Denmark	Mexico	Aruba	Syria
Estonia	New Zealand	Azerbaijan	Taiwan
Finland	Switzerland	Bahamas	Tajikistan
France	Turkey	Bahrain	Tanzania
Germany	USA	Bangladesh	Thailand
Greece		Barbados	Timor-Leste
Hungary		Belarus	Togo
Iceland		Belize	Tonga
Ireland		Benin	Trinidad and Tobago
Italy		Bermuda	Tunisia
Latvia		Bhutan	Turkmenistan
Liechtenstein		Bolivia	Uganda
Lithuania		BiH	Ukraine
Luxembourg		Botswana	UAE
Malta		Brazil	Uruguay
Netherlands		Brunei	Uzbekistan
Norway		Burkina Faso	Vanuatu
Poland		Burundi	Venezuela
Portugal		Côte d'Ivoire	Viet Nam
Romania		Cabo Verde	Virgin Is.
Slovakia		Cambodia	Yemen
Slovenia		Cameroon	Zambia
Spain		Cayman Is.	Zimbabwe
Sweden		Chad	
UK		China	
EU		Colombia	
		Congo (DR)	Mozambique
		Costa Rica	Myanmar
		Cuba	Namibia
		Djibouti	Nepal
		Dominica	New Caledonia
		Dominican Rep.	Nicaragua
		Ecuador	Niger
		Egypt	Nigeria
		El Salvador	Oman
		Eq. Guinea	Pakistan
		Ethiopia	Palestine, State of
		Faroe Is.	Panama
		Fiji	Papua New Guinea
		Fr. Guiana	Paraguay
		Fr. Polynesia	Peru
		Gabon	Philippines
		Gambia	Puerto Rico
		Georgia	Qatar
		Ghana	Reunion
		Gibraltar	Russia
		Greenland	Rwanda
		Grenada	St. Kitts and Nevis
		Guadeloupe	St. Lucia
		Guam	St. Vincent
		Guatemala	San Marino
		Guernsey	Saudi Arabia
		Guinea	Senegal
		Guyana	Serbia
		Haiti	Seychelles
		Honduras	Sierra Leone
		Hong Kong	Singapore
		Morocco	Solomon Is.

Note: The list of EEA countries and territories includes the 31 EEA member countries and the EU, as some networks are owned by EU-wide organizations. The list of non-EEA OECD countries and territories includes 11 countries. The list of non-EEA non-OECD countries and territories includes 157 countries and territories. There were no changes to EEA or OECD membership status during our study period.

B.3 Robustness Checks

B.3.1 Results on Levels

Table B.3.1: The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e \times EEA$	-24.00 (15.12)	-0.48 (0.30)	-1.22 (1.84)
$POST_a \times EEA$	-33.79 (27.52)	-0.79 (0.01)	0.93 (0.02)
Group dummies	country pairs	country pairs	country pairs
Time dummies	months	months	months
Clusters	418	6,751	880
R^2	0.999	0.903	0.990
Observations	22,572	364,554	47,520

Notes: The dependent variable is $numAg_{ijt}$. The variable $numAg_{ijt}$ is rectangularized as described in Table 2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Column (1) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by an EEA country. All regressions include month dummies and country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Table B.3.2: The GDPR's impact on the number of agreements by EEA and non-EEA OECD countries, by counterparty and agreement type

	Counterparty is non-EEA OECD			Counterparty is non-EEA non-OECD			Counterparty is EEA		
	Provider (1)	Peer (2)	Customer (3)	Provider (4)	Peer (5)	Customer (6)	Provider (7)	Peer (8)	Customer (9)
$POST_e \times EEA$	-12.06 (7.84)	-2.76 (2.19)	-11.99 (7.83)	-0.03 (0.10)	-0.32 (0.20)	-0.10*** (0.04)	0.77 (0.59)	-0.53 (0.97)	0.84 (0.64)
$POST_a \times EEA$	-19.2 (14.51)	-1.68 (2.59)	-18.94 (14.51)	-0.19** (0.08)	-0.57 (0.49)	-0.114*** (0.04)	0.29 (0.75)	-0.31 (1.50)	0.587 (0.67)
Group dummies	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months	ctry pairs months
Time dummies									
Clusters	473	418	473	6,751	6,751	6,751	1,376	880	1,376
R^2	0.999	0.987	0.999	0.878	0.840	0.954	0.993	0.971	0.993
Observations	25,542	22,572	25,542	364,554	364,554	364,554	74,304	47,520	74,304

Notes: The dependent variable is $numProvAg_{ijt}$ for columns (1), (4), (7), $numPeerAg_{ijt}$ for columns (2), (5), (8), and $numCustAg_{ijt}$ for columns (3), (6), (9). The dependent variables are rectangularized as described in Table 2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Columns (1), (2), (3) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by a non-EEA OECD country. Columns (4), (5), (6) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by a non-EEA OECD country. Columns (7), (8), (9) include observations when the treatment/control party is a network owned by an EEA/non-EEA OECD country and is the provider, peer, or customer to the counterparty network owned by an EEA country. All regressions include month dummies and country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the * 10% level, ** 5% level, *** 1% level.

Table B.3.3: The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA and non-EEA OECD countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e \times EEA$	2.281 (2.326)	9.393 (8.956)	-1.734 (2.650)
Group dummies	network pairs	network pairs	network pairs
Time dummies	weeks	weeks	weeks
Clusters	128	522	307
R^2	0.640	0.701	0.742
Observations	2,593,805	494,374	1,886,031

Notes: The dependent variable is $numAgIP_{ijt}$. $POST_e$ is an indicator variable equal to 1 if the observation is made after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if the observation is made after the GDPR was approved. Column (1) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party of the agreement is a network owned by an EEA or non-EEA OECD country and the counterparty is a network owned by an EEA country. Only agreements present for at least 150 weeks are used. The GDPR approval date Apr 2016 is close to the sample starting date Dec 2015, so $POST_a \times EEA$ is not included in the regressions. All regressions include week dummies and network pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Table B.3.4: The GDPR's impact on additional outcomes

	$numAgNtwrk_{kt}$	$numNtwrk_{it}$	$NtwrkCustCone_{kt}$
	(1)	(2)	(3)
$POST_e \times EEA$	0.039 (0.183)	-94.84 (91.45)	0.017 (0.754)
$POST_a \times EEA$	0.167 (0.264)	-111.70 (105.1)	1.479 (0.995)
Group dummies	networks	countries	networks
Time dummies	months	quarters	months
Clusters	43	43	43
R^2	0.963	0.998	0.978
Observations	1,275,236	1,275,236	1,275,236

Notes: The dependent variables are noted in column headers. Only observations from networks or countries in the EEA or OECD and present throughout Jan 2015 – June 2019 are used for regressions. All regressions include time dummies and group dummies. All regressions cluster standard error by country of ownership of network. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

B.3.2 First Differences

Table B.3.5: The GDPR's impact on the number of agreements by EEA countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e$	−0.001 (0.002)	−0.000 (0.000)	−0.002 (0.001)
$POST_a$	−0.002* (0.001)	−0.000* (0.000)	−0.003** (0.001)
Group dummies	country pairs	country pairs	country pairs
Clusters	418	6,751	880
R^2	0.007	0.008	0.006
Observations	18,656	266,272	27,984

Notes: The dependent variable is $\log(numAg_{ijt} + 1) - \log(numAg_{ijt-1} + 1)$. The variable $numAg_{ijt}$ is rectangularized as described in Table 2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if time t is after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if time t is after the GDPR was approved. Column (1) includes observations when one party is a network owned by an EEA and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party is a network owned by an EEA and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party is a network owned by an EEA and the counterparty is a network owned by an EEA country. All regressions include country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Table B.3.6: The GDPR's impact on the number of agreements by EEA countries, by counterparty and agreement type

	Counterparty is non-EEA OECD			Counterparty is non-EEA non-OECD			Counterparty is EEA		
	Provider (1)	Peer (2)	Customer (3)	Provider (4)	Peer (5)	Customer (6)	Provider (7)	Peer (8)	Customer (9)
$POST_e$	-0.002 (0.001)	0.001 (0.002)	-0.001 (0.001)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.001)	-0.003** (0.001)	-0.000 (0.001)
$POST_a$	-0.001 (0.001)	-0.004*** (0.001)	0.004*** (0.001)	-0.000* (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.001* (0.001)	-0.006*** (0.001)	-0.001 (0.001)
Group dummies	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs	ctry pairs
Clusters	473	418	473	6,751	6,751	6,751	1,376	880	1,376
R^2	0.007	0.007	0.009	0.006	0.007	0.005	0.007	0.004	0.007
Observations	18,656	18,656	18,656	266,272	266,272	266,272	54,272	27,984	54,272

Notes: The dependent variable is $\log(\text{numProvAg}_{it} + 1) - \log(\text{numProvAg}_{it-1} + 1)$ for columns (1), (4), (7), $\log(\text{numPeerAg}_{it} + 1) - \log(\text{numPeerAg}_{it-1} + 1)$ for columns (2), (5), (8), and $\log(\text{numCustAg}_{it} + 1) - \log(\text{numCustAg}_{it-1} + 1)$ for columns (3), (6), (9). The dependent variables are rectangularized as described in Table 2 and we add one when we take the log to account for zero values. $POST_e$ is an indicator variable equal to 1 if time t is after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if time t after the GDPR was approved. Columns (1), (2), (3) include observations when the treatment party is a network owned by an EEA and is the provider, peer, or customer to the counterparty network owned by a non-EEA OECD country. Columns (4), (5), (6) include observations when the treatment party is a network owned by an EEA and is the provider, peer, or customer to the counterparty network owned by a non-EEA non-OECD country. Columns (7), (8), (9) include observations when the treatment party is a network owned by an EEA and is the provider, peer, or customer to the counterparty network owned by an EEA country. All regressions include country pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the * 10% level, ** 5% level, *** 1% level.

Table B.3.7: The GDPR's impact on the number of IP-address-level interconnection points per agreement by EEA countries, by counterparty

	Non-EEA OECD (1)	Non-EEA Non-OECD (2)	EEA (3)
$POST_e$	-0.001 (0.001)	-0.004** (0.002)	-0.002*** (0.000)
Group dummies	network pairs	network pairs	network pairs
Clusters	128	522	307
R^2	0.012	0.017	0.010
Observations	2,205,760	1,344,324	3,487,571

Notes: The dependent variable is $\log(\text{numAgIP}_{ijt} + 1) - \log(\text{numAgIP}_{ijt-1} + 1)$. $POST_e$ is an indicator variable equal to 1 if time t is after the GDPR became effective. $POST_a$ is an indicator variable equal to 1 if time t after the GDPR was approved. Column (1) includes observations when one party of the agreement is a network owned by an EEA and the counterparty is a network owned by a non-EEA OECD country. Column (2) includes observations when one party of the agreement is a network owned by an EEA and the counterparty is a network owned by a non-EEA non-OECD country. Column (3) includes observations when one party of the agreement is a network owned by an EEA and the counterparty is a network owned by an EEA country. Only agreements present for at least 150 weeks are used. The GDPR approval date Apr 2016 is close to the sample starting date Dec 2015, so $POST_a$ is not included in the regressions. All regressions include network pair dummies. All regressions cluster standard error by country pair. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Table B.3.8: The GDPR's impact on additional outcomes

	<i>numAgNtwrk_{kt}</i>	<i>numNtwrk_{it}</i>	<i>NtwrkCustCone_{kt}</i>
	(1)	(2)	(3)
<i>POST_e</i>	−0.001*** (0.000)	−0.001 (0.002)	−0.001*** (0.000)
<i>POST_a</i>	−0.003*** (0.000)	−0.005** (0.002)	−0.001*** (0.000)
Group dummies	networks	countries	networks
Clusters	43	43	43
<i>R</i> ²	0.007	0.218	0.008
Observations	451,316	510	451,316

Notes: The dependent variables are the first differences in the logged outcomes noted in column headers. Only observations from networks or countries in the EEA and present throughout Jan 2015 – June 2019 are used for regressions. All regressions include group dummies. All regressions cluster standard error by country of ownership of network. Standard errors are in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Appendix C

Appendix to Chapter 3

Table C.1: Variable definitions

Variables	Definition
$timeToUpdateMonthEnd_{it}$	The number of months since organization i 's previous update of Apache server software
$updated_{it}$	Binary = 1 if organization i makes an update to the Apache server software in month t ; an update happens in month t either when the organization uses a higher major version in the next observed month (e.g. the organization uses Apache 2.2 then changes to Apache 2.4) or when the organization uses a higher minor version in the next observed month (e.g. the organization uses Apache 2.2.3 then changes to 2.2.4)
$hasSevereBugs_{it}$	Binary = 1 if the minor version of Apache software (e.g., Apache 2.2.3) organization i used in month t had publicly known severe security vulnerabilities
$hasAlexaRank2010_i$	Binary = 1 if the organization i 's website was ranked by Alexa in the top 1 million websites in 2010
aws_{it}	Binary = 1 if organization i 's website in month t used an IP address that belongs to AWS in 2022; data on historical AWS IP addresses is not available

Table C.1 (Continued): Variable definitions

<i>techCategories_i</i>	The number of technology categories embedded in organization <i>i</i> 's website between 2016 and 2018; examples of technology categories include analytics, e-commerce, and web frameworks; measures technological complexity
<i>tech_i</i>	The number of technologies embedded in organization <i>i</i> 's website between 2016 and 2018; examples of technologies include jQuery, Google Tag Manager, and WordPress; measures technological complexity
<i>monetizationTech_i</i>	The number of monetization technologies embedded in organization <i>i</i> 's website between 2016 and 2018; technologies in "analytics," "tag managers," "advertising networks," "marketing automation," "e-commerce," and "payment processors" categories are included; measures monetization intent
<i>newVersionAvailable_{it}</i>	Binary = 1 if a newer Apache version than organization <i>i</i> 's adopted version in month <i>t</i> was available in that month; for example, the variable is equal to 1 if organization <i>i</i> used in month <i>t</i> version 2.0.x released in month <i>t'</i> < <i>t</i> , and 2.4.y was released after <i>t'</i> and was available at <i>t</i>
<i>newMinorVersionAvailable_{it}</i>	Binary = 1 if a newer Apache minor version than organization <i>i</i> 's adopted version in month <i>t</i> was available in that month; for example, the variable is equal to 1 if organization <i>i</i> used in month <i>t</i> version 2.0.64 released in month <i>t'</i> < <i>t</i> , and 2.0.65 was released after <i>t'</i> and was available at <i>t</i>
<i>infrequentUpdater_i</i>	Binary = 1 if organization <i>i</i> did not update server software even once in the first nine observations of Apache usage (Apache on average releases a new minor version once every 4.5 months for a major version); the variable is undefined for the first nine observations of Apache usage of each organization because those observations are used to construct the variable
<i>frequentUpdater_i</i>	Binary = 1 if organization <i>i</i> updated server software two or more times in the first nine observations of Apache usage; the variable is undefined for the first nine observations of Apache usage of each organization because those observations are used to construct the variable

Table C.1 (Continued): Variable definitions

$naics_i$	Organization i 's two-digit primary NAICS code
$state_i$	State where organization i 's headquarter is located
$logEmployment_i$ (or $logEmployment_{iy}$)	Logged value of ($employment + 1$) of organization i ; for the cross-sectional Mint Global sample of organizations, a single value is associated with an organization; for the panel Compustat sample of organizations, the variable is also subscripted by year y
$logRevenue_i$	Logged value of ($revenue + 1$) of organization i ; if revenue is negative, this variable is equal to $-1 \times \log(abs(revenue) + 1)$; from Mint Global
$logCapx_{iy}$	Logged value of ($capital\ expenditure + 1$) of organization i in year y ; from Compustat
$logTotalAssets_{iy}$	Logged value of ($total\ assets + 1$) of organization i in year y ; from Compustat
$logDepreciation_{iy}$	Logged value of ($depreciation\ and\ amortization + 1$) of organization i in year y ; from Compustat
$logIncome_{iy}$	Logged value of ($income\ before\ extraordinary\ items + 1$) of organization i in year y ; if income is negative, this variable is equal to $-1 \times \log(abs(income) + 1)$; from Compustat
$logCashflow_{iy}$	Logged value of ($cashflow + 1$) of organization i in year y ; if cashflow is negative, this variable is equal to $-1 \times \log(abs(cashflow) + 1)$; from Compustat
$severeBugsReported_t$	Binary = 1 if severe securities vulnerabilities were reported to the Apache Foundation in month t

Table C.1 (Continued): Variable definitions

<i>logOrganizationsImpacted_t</i>	<i>organizationsImpacted_t</i> represents the number of organizations in our broad sample that were impacted by severe securities vulnerabilities reported to the Apache Foundation in month <i>t</i> ; <i>logOrganizationsImpacted_t</i> takes the log of (<i>organizationsImpacted_t</i> + 1)
---	--
