HARVARD | BUSINESS | SCHOOL

DOCTORAL PROGRAMS

DISSERTATION ACCEPTANCE CERTIFICATE

(To be placed in First Library Copy)

The undersigned, approved by the Chair of the Doctoral Programs, have examined a dissertation entitled

The Digital Commons: Tragedy or Opportunity? The Effect of Crowdsourced Digital Goods on Innovation and Economic Growth

presented by Francis Edward Nagle

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

Signature .....

Shane Greenstein, *Co-Chair*

Signature .....

Marco Iansiti, *Chair*

Signature .....

Carliss Y. Baldwin

Signature .....

Karim R. Lakhani

Signature .....

Feng Zhu

Date April 10, 2015

WYSS | SOLDIERS FIELD | BOSTON, MA 02163 | Ph 617.495.6101 | Fx 617.496.2203 | doctoralprograms@hbs.edu | GEORGE F. BAKER FOUNDATION

The Digital Commons: Tragedy or Opportunity? The Effect of Crowdsourced Digital Goods on Innovation and Economic Growth

A dissertation presented

by

Francis Edward Nagle

In partial fulfillment of the requirements

for the degree of

Doctor of Business Administration

in Technology and Operations Management

Harvard Business School

Boston, Massachusetts

April 2015

Dissertation Advisors: Francis Edward Nagle
Professors Shane Greenstein, Marco Iansiti,
Carliss Baldwin, Karim Lakhani, Feng Zhu

The Digital Commons: Tragedy or Opportunity? The Effect of Crowdsourced Digital Goods on Innovation and Economic Growth

Abstract

The classic economic concept of the tragedy of the commons occurs when individuals overuse a public good, resulting in the complete depletion of the good. Comparatively, in the digital world public goods are non-rival and essentially infinitely abundant. However, the nearly infinite supply of a public digital good can still be tragic, albeit in a different manner. For example, the rise of the free crowdsourced digital good Wikipedia essentially destroyed billions of dollars of economic value in the encyclopedia industry. Despite this apparent destruction of value, the reduction in prices for many digital goods also represents a great opportunity. Firms are increasingly relying on the crowd to help shape future products, provide value for their customers, and build software crucial to the firm's production process. This phenomenon is leading to a weakening of firm boundaries and a change in the nature of the firm's innovative processes. My dissertation is comprised of four studies that explore this phenomenon to better understand the transformative nature of the digital commons.

The first chapter, "Innovating Without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero" (w/ Elizabeth Altman, and Michael Tushman), explores how technological progress and reductions in information costs are leading firms to increasingly engage with external digital communities. In particular, firms are increasingly engaging with networks of developers, external labor marketplaces, and users, with the latter frequently occurring through the process of crowdsourcing. This engagement leads to a weakening of firm boundaries such that the locus of innovation and value creation moves outside the boundaries of the firm. The increase in this phenomenon motivates a reevaluation of many traditional theories of how firms organize and innovate. Specifically, we

consider how shifts in information costs affect the classic organizational concepts of firm boundaries, business models, interdependence, leadership, identity, search, and intellectual property. In turn, these effects on the firm's organization alter how the firm innovates.

The second chapter, "Digital Dark Matter and the Economic Contribution of Apache" (w/ Shane Greenstein) examines the impact of crowdsourced digital goods at a macro-level. We show that due to its reliance on price to measure value, GDP calculations do not account for "digital dark matter", digital goods and services that are non-pecuniary and effectively limitless inputs into production. We scan 1% of the 1.5 billion IP addresses in the United States to measure the types of web servers businesses and individuals employ. We estimate the value of the free and open source nature of the predominant web server, Apache, by comparing it to the closest pecuniary alternative, Microsoft's Internet Information Services (IIS) server. Our analysis shows that the lack of price for the Apache server leads to an underestimation of GDP by upwards of $12 billion. Although this is the value from only one piece of digital dark matter, this miscalculation represents a large proportion of all software sales and significantly alters economic growth projections.

The third chapter, "Crowdsourced Digital Goods and Firm Productivity: Evidence from Open Source Software", empirically measures the firm-level productivity impact of managers' decisions to use non-pecuniary digital inputs from the crowd. Existing literature examining the impact of IT on productivity does not account for investments in such goods, as their use cannot properly be captured by traditional measurement methods based on price. Therefore, their contribution to the firm's production process is currently unexplored, despite mounting evidence that firms are increasingly relying on these types of inputs. Employing data from a survey of technology use at nearly 2,000 firms over 10 years, I find that a 1% increase in the amount of non-pecuniary open source software (OSS) used by a firm leads to a .073% increase in productivity. This translates to a $1.35 million increase in productivity for the average firm in my sample. This is more than double the magnitude of the coefficient on investments in traditional pecuniary IT capital. I find that this effect is greater for larger firms and for firms in the services industry. I use inverse probability weighting, instrumental variables, firm-fixed

effects and data on managerial quality from the World Management Survey to add support to a causal interpretation of these results.

The final chapter of my dissertation, "Organizational Learning Through Contributing to Public Goods: Evidence from Open Source Software," builds on the concepts developed in the other three to explore how firms that engage external communities and contribute to the development of crowdsourced digital goods enhance their ability to extract value from technology-related inputs via increased learning about how these complex goods operate. This study explores this mechanism by using data on firm contributions to Linux, an OSS operating system that is an important public digital good created via crowdsourcing. Using coarsened exact matching and inverse probability weighting to address endogeneity concerns, this study shows that firms who contribute to the development of OSS capture more productive value from the use of OSS than their non-contributing peers through a process similar to absorptive capacity. Further, this learning has a spillover effect that allows contributing firms to capture more productive value from all of their IT investments, not just OSS.

Together, the results of these four studies show that the digital commons can help create a great deal of economic value, but that this value is difficult to measure via standard economic methods that rely on price to reflect value. These results have important strategic implications for managers and policy makers to consider as organizations increasingly engage with external communities and ecosystems to innovate and create value.

Table of Contents

Acknowledgements ix

List of Tables xv

List of Figures xv

Introduction 1

Chapter 1: Innovating Without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero 7

1.1 Introduction 8

1.2 Information Constraints Reduction 11

1.2.1 Information Processing 12

1.2.2 Information Storage 15

1.2.3 Information Communication 17

1.3 Engaging Communities 20

1.3.1 Labor Marketplaces 22

1.3.2 Developer Ecosystems 23

1.3.3 User-Generated Contributions 25

1.4 Organizational and Strategic Implications 27

1.4.1 Boundaries 30

1.4.2 Strategy and New Business Models 36

1.4.3 Interdependence and Community Engagement 38

1.4.4 Leadership 41

1.4.5 Identity 44

1.4.6 Search 46

1.4.7 Intellectual Property 48

1.5 Impact on Innovation 50

1.5.1 Variation 51

1.5.2 Selection 53

1.5.3 Retention (by Communities) 55

1.6 Future Directions and Research Opportunities 56

Chapter 2: Digital Dark Matter and the Economic Contribution of Apache 66

2.1 Introduction 67

2.2 Digital Dark Matter: Framework 71

2.2.1 Institutional background 71

2.2.2 Measuring the gains: Omission 74

2.2.3 Measuring the gains: Attribution 76

2.3 The shadow value of Apache HTTP Server 80

2.3.1 The shape of the server economy 80

2.3.2 Substitution with pecuniary goods 83

2.3.3 Economic importance of Apache 85

2.3.4 The economic size of the ecosystem supported 86

2.3.5 The Rate of Return 87

2.4 Concluding thoughts and future research 93

Chapter 3: Crowdsourced Digital Goods and Firm Productivity: Evidence from Open Source Software 98

3.1 Introduction 99

3.2 Crowdsourced Digital Goods and the Returns to Information Technology 105
3.2.1 Free and Open Source Software as an Input into Productivity 105
3.2.2 Institutional Context: The Free and Open Source Software Movement 108
3.3 Theory and Hypothesis Development 110
3.3.1 Risks of Using Non-Pecuniary OSS 111
3.3.2 Benefits of Using Non-Pecuniary OSS 113
3.3.3 Moderating Effect of Firm Size 116
3.3.4 Moderating Effect of Industry 117
3.3.5 Additional Moderating Effects 117
3.4. Empirical Methodology 118
3.4.1 Estimation Models 119
3.4.2 Identification Strategy 121
3.5 Data 125
3.5.1 Variable Construction 126
3.5.2 Descriptive Statistics 131
3.6 Results and Discussion 134
3.6.1 Three-Factor Productivity Analysis 135
3.6.2 Propensity to Adopt Non-Pecuniary OSS 136
3.6.3 Baseline Regression Results 139
3.6.4 Instrumental Variable Regression Results 140
3.6.5 Moderators and Split-Sample Analysis 142
3.6.6 Robustness Checks 144
3.7 Conclusion 152

Chapter 4: Organizational Learning Through Contributing to Public Goods: Evidence from Open Source Software 162
4.1 Introduction 163
4.2 Theory and Hypotheses 165
4.2.1 Gaining Value from Information Technology Usage 166
4.2.2 Crowdsourced Technology and the Firm 167
4.2.3 Absorptive Capacity from Firm Contributions to Crowdsourced Technologies 168
4.2.4 Spillovers Beyond OSS from Learning by Contributing 170
4.3 Empirical Methodology 171
4.3.1 Estimation Models 172
4.3.2 Identification Strategy 174
4.4 Data 178
4.4.1 Variable Construction 180
4.4.2 Descriptive Statistics 186
4.5 Results 188
4.5.1 Propensity to Contribute to OSS 188
4.5.2 Benefits of Contribution to OSS 191
4.5.3 Spillovers Benefits to All IT Usage 194
4.6 Conclusion 196

Appendix A: The Shape of the Server Economy 202

Appendix B: Substitutability of Apache and IIS 209

Dedication

To Jenna,

Without your love and support,

I never could have made it this far.

Acknowledgements

Isaac Newton famously wrote, “If I have seen further it is by standing on the shoulders of giants.” This dissertation could not have happened without the help and support of many whom I consider to be giants. I am thankful for all of their contributions along the way and am forever in their debt.

First, I thank my dissertation co-chairs, Shane Greenstein and Marco Iansiti for all of their hard work and guidance throughout the last five years. Shane and I met over lunch in the first year of my doctoral program when he was visiting HBS to give a seminar. Our discussion at that lunch eventually led to Chapter 2 of this dissertation, which scratched the surface of a deeply important question in the digital economy, and in turn set the stage for Chapters 3 and 4. As Shane was a professor at Northwestern, our collaboration for the first few years was remote, but his insights into the research process and what it means to be a scholar have been incredibly formative in my progression as a doctoral student. I always say that Shane showed up at exactly the right time twice in my doctoral career – the first was that lunch, and the second was when he decided to take a visiting professorship at HBS during my fourth year. Although his remote guidance was excellent, the ability to knock on his door and have an impromptu conversation about life, research, and everything in between is what truly allowed for my research agenda to flourish and for this dissertation to come together. Shane is one of the very few economists who truly understands the transformative nature of digital technology, and I thank him for imparting much of that wisdom to me.

Marco and I first met in May the year before I started the doctoral program at HBS at the TOM department research day. He gave a presentation on his work on platforms and ecosystems in the Internet browser industry and I knew that my decision to come to HBS was the right one. Throughout my time at HBS he has offered constant encouragement to ensure that my research was not only academically rigorous, but also relevant to practitioners. His work takes a detailed knowledge of technology, applies a refined lens of academic inquiry, and presents the results in a manner that is useful for managers, academics, and policy makers. I can only hope that my work will have such a great impact on the day-to-day functioning of digital businesses (which is increasingly all businesses) as his does.

In addition to Shane and Marco, I had the privilege of having three additional amazing scholars on my dissertation committee: Carliss Baldwin, Karim Lakhani, and Feng Zhu. Carliss's work on the modularity of technology has been fundamental in shaping many of my ideas about the modularity of the firm, which, although not fully expounded upon in this dissertation, represent an important aspect of my future research agenda. Further, the quality of my work benefited greatly from her numerous readings of my job market paper and her incredibly insightful comments about both specific methodology and big-picture ideas for all of my work. Karim's deep knowledge of the open source software community was incredibly helpful throughout the dissertation process. His insights into the broader open and user innovation world helped my work speak to a broader audience that truly understands the changing nature of how firms innovate. Feng's experiences bridging the IT/IS and Strategy domains in the management literature were incredibly valuable as I too tried to walk the line between these two disciplines.

Further, his expertise at finding interesting questions that can be econometrically well-identified has helped me set a high bar for the goals of my own work.

Beyond my dissertation committee, I had the good fortune to engage with many other stellar professors during my time at HBS. Mike Tushman's sage advice and guidance led to the important conceptual piece that is Chapter 1. Further, his stellar reputation in the academic world opened many doors for me, even though he was not on my committee. His dedication to doctoral students is a shining example of what all academics should strive for. Kristina McElheran was instrumental in my early development as a doctoral student. Not only did she serve on my field exam committee and introduce me to the broader IS academic community, but her attention to empirical and econometric detail taught me a great deal about how to do high-quality research. Mike Toffel truly went above and beyond the call of duty with his constant support of the TOM doctoral program and its participants. I thank Ramon Casadesus-Masanell for serving on my field exam committee and introducing me to the mainstream strategy literature. Lisa Singh and Keith Pilbeam helped me discover my love of research long before I came to HBS. I thank Christoph Riedl for being an amazing co-author and sharing numerous insights from his experiences on the IS job market.

My work on Bitcoin with Magnus Torfason and Misiek Piskorski helped me understand the close relationship between technology and sociology. The encouragement and guidance of Scott Stern and Erik Brynjolfsson helped me better understand the connection of my work to the economics of innovation and IT literatures. Jan Hammond and Rob Huckman showed me how research can be brought into the classroom and how topics of a quantitative nature can be taught

via the case method. I thank Ananth Raman for reminding me that as hard as being an academic may be, it beats having a “real job”. Mihir Desai, Kathleen McGinn, Dennis Yao, and Rakesh Khurana were all instrumental in shaping the HBS doctoral program so that it emphasizes asking deep questions that have relevance to managers and answering them in a rigorous manner. Without Jen Mucciarone from the HBS Doctoral Programs Office, I never would have been able to make it through the first year of the program, let alone finish this dissertation and graduate. The tireless efforts of the rest of the Doctoral Programs Office, especially John Korn, Dianne Le, Marais Young, LuAnn Langan, and Karla Cohen allowed myself and all of the other doctoral students to succeed and flourish during our time at HBS.

At times, being a doctoral student can be a very demanding life. I am very grateful that I had such an excellent cohort to go through the process with. The help and support of Elizabeth Altman, Megan Lawrence, Sarah Wolfolds, Kate Barasz, Bhavya Mohan, Lingling Zhang, Sean Shin, Shelley Li, and Ryann Manning made that process a lot more enjoyable. The many, many conversations that Liz and I had about the changing nature of technology and innovation helped form my research agenda. I am very grateful that Liz asked me to join her and Mike Tushman on the project that eventually became Chapter 1 of this dissertation and I look forward to many future collaborations and conversations. In addition to my cohort, I am very grateful for the HBS doctoral students that came before me. Anil Doshi, Ryan Buell, Hila Lifhitz-Assaf, Ariel Stern, Bill Schmidt, Sen Chai, Ethan Bernstein, and Clarence Lee were all extremely helpful in offering guidance throughout the program. I thank all of those involved in the HBS Digitial Initiative, especially Colin, Matt, and Caroline, for helping to enhance the importance of the digital economy at HBS and beyond.

The papers in this dissertation benefited greatly from a wide range of individuals. Mary Tripsas offered great advice as Liz, Mike, and I developed Chapter 1. For Chapter 2, I thank Justin Ng for research assistance and am grateful for the many useful comments we received from audiences at the NBER, Kristina McElheran, Dan Sichel, Ashish Arora, and two anonymous referees. For Chapter 3, I thank Raffaella Sadun, John Van Reenen, and Nick Bloom for sharing their data from the World Management Survey. Further, I am grateful for helpful comments from Shane Greenstein, Carliss Baldwin, Yochai Benkler, Raj Choudhury, Anil Doshi, Marco Iansiti, Ohchan Kwon, Karim Lakhani, Kristina McElheran, Hart Posen, Scott Stern, Neil Thompson, Mike Toffel, Joel West, and Feng Zhu. Additional helpful comments were received from participants at ACAC 2014, AEA 2015, AOM 2014, AOM 2014 BPS Dissertation Consortium, CCC 2014, Charles River Conference 2014, DRUID 2014, HBS TOM DBA Seminar 2014, HBS TOM Alumni Conference 2014, NYU Engelberg Center Conference on Knowledge Commons 2014, OUI 2014, SMS 2014, and ZEW ICT Conference 2014. Helpful comments were also received from seminar participants at Bocconi University, Boston College, Carnegie Mellon University, Columbia Business School, Harvard Business School, IESE Business School, McGill University, Temple University, University College London, University of Maryland, University of Pennsylvania, and University of Southern California. Chapter 4 would not have been possible without the generous provision of data from The Linux Foundation, the data-wrangling efforts of Greg Kroah-Hartman, and comments from Rory McDonald.

Last, but certainly not least, I would like to thank my friends and family for their encouragement throughout my doctoral career. My friends from Washington, DC and Boston are too numerous to list here, but I am forever grateful for your support throughout my life. In particular, Mark, Sean, Pat, and Dan have kept me sane over the last five years. Thanks to Cleo, my cat, for keeping me company during the many hours that went into writing this dissertation. To my fiancée's family, thank you for constantly pushing me to do better in your unique Paone way – loudly, and with lots of gesticulating. You have loved me as your own and for that I am grateful. Thank you to my sister, Rachel: I am lucky to have such a great sister and friend and I cherish the years we spent growing up together in both Massachusetts and DC. Thank you to my parents, Arlene and Frank, for all of the sacrifices you made so that Rachel and I could chase our dreams. With a mother who loves technology and a father who loves business and entrepreneurship, it is no surprise that my dissertation is on the business of technology. Thank you both for imparting your ethic of hard work upon me.

Finally, thank you to Jenna. Without your love and support (and editing skills) I never would have gotten this far. I am forever grateful for your willingness to listen to me ramble on about my research and for offering me unending reassurance through the many times I questioned whether it was going to amount to anything. You have never doubted me, even when I doubted myself. As a musician you, more than most, understand the sometimes negative effects of the digital transformation this dissertation explores. Your strength living through what I am studying has pushed me on throughout this process and I could not have done this without you. Thank you.

List of Tables

Table 1.1 Engaging With Communities With and Without Information Constraints 27
Table 1.2 Organizational and Strategic Characteristics With and Without Information Constraints 31
Table 1.3 Innovating With and Without Information Constraints 51
Table 2.1 Contribution of Apache to GDP, simulation, Billions of Dollars 92
Table 3.1 Open Source Operating Systems 131
Table 3.2 Descriptive Statistics 132
Table 3.3 Correlation Matrix 133
Table 3.4 Industry Breakdown 134
Table 3.5 Three-Factor Productivity Results 136
Table 3.6 Predicting Adoption of Non-Pecuniary OSS 138
Table 3.7 Covariate Balance 138
Table 3.8 Baseline Regressions 140
Table 3.9 IV Regressions 142
Table 3.10 Moderator and Split-Sample Regression Results 145
Table 3.11 Robustness Checks 151
Table 4.1 Contributing Firms and their Non-Contributing Matches 180
Table 4.2 Open Source Operating Systems 184
Table 4.3 Descriptive Statistics 187
Table 4.4 Correlation Matrix 188
Table 4.5 Predicting Contribution to OSS 190
Table 4.6 Covariate Balance 190
Table 4.7 Benefits of Contribution to OSS 191
Table 4.8 Benefits of Contribution: OSS Breakdown by Type 193
Table 4.9 Spillovers from Contribution to IT Usage 195
Table 4.10 Spillovers from Contribution Intensity to IT Usage 196
Table A.1 Top 25 Counties for Server Software Use 203
Table A.2 Server use Among Top Level Domain Names 205
Table A.3 Server use Among Top 15 Second Level Domain Names Among Com 206
Table A.4 Server use Among Top 15 Second Level Domain Names Among Net 207
Table A.5 Server use Among Top 15 Second Level Domain Names Among Edu 208

List of Figures

Figure 1.1 MIPS per US Dollar Over Time (Source: Koh & Magee, 2006) 14
Figure 1.2 Megabits per US Dollar Over Time (Source: Koh & Magee, 2006) 17
Figure 1.3 Bandwidth per Cable Length per US Dollar (Source: Koh & Magee 2006) 19
Figure 1.4 Typology of Communities 21
Figure 3.1 Examples of Software on the Free/Open Spectrum 110

Introduction

The classic economic concept of the tragedy of the commons (Hardin, 1968) occurs when individuals overuse a public good, resulting in the complete depletion of that good. Comparatively, in the digital world public goods are non-rival and essentially infinitely abundant. However, the nearly infinite supply of a public digital good can still be tragic, albeit in a different manner. Take for example the case of the encyclopedia. In 1990, Encyclopedia Britannica had $650 million in revenue, but was sold six years later for $135 million. This destruction of value was due to the digitization of encyclopedias, first Microsoft's Encarta followed by Wikipedia, the crowdsourced digital encyclopedia that is free. In just over a decade, the revenues of the encyclopedia industry disappeared in an apparently massive destruction of value due to digitization in general, and the digital commons in particular. At first glance, this may appear to be simply a modern day example of creative destruction (Schumpeter, 1942), but the non-pecuniary nature of most of the crowdsourced digital goods entering the market makes the destructive impact of these innovations more apparent than the creative one. For creative destruction to lead to economic growth that can be captured by traditional means, such as gross domestic product (GDP), the newly created good that is destroying the old good must have a price through which the value of using the new good can be captured. Without an ability to capture the value of the new good, only the destructive portion of the creative destruction process is captured, essentially creating a transparent economy. This causes further problems when trying to understand the importance of investments in human capital, research and development, and technological improvements on future economic growth (Arrow, 1962; Aghion and Howitt, 1992; Griliches, 1979; Romer, 1990).

Despite this apparent destruction of value, the reduction in prices for many digital goods also represents a great opportunity. While the encyclopedia industry may no longer be adding a significant amount to GDP directly, Wikipedia and other online knowledge repositories are certainly still allowing for the accumulated wisdom of prior generations to be passed on and used in a productive manner. More broadly, the decrease in communication costs allowed by rapid technological progress is allowing firms to increasingly rely on the free efforts of the crowd to help shape future products, provide value for their customers, and build software crucial to the firm's production process. This phenomenon is leading to a weakening of firm boundaries and a change in the nature of the firm's innovative processes. My dissertation is comprised of four studies that explore this phenomenon to better understand the transformative nature of the digital commons and the effect of crowdsourced digital goods on innovation and economic growth.

Examining this phenomenon is critical at this point in human history as technological progress and digital forces increasingly alter our notions of two important concepts – what a firm is and what a product is. Since firms have existed, their primary purpose has been to create new ideas, produce them as tangible products, and distribute them to consumers. For much of modern history, the boundaries of the firm were well defined and all three steps of this process often occurred within one organization since transaction-costs for coordinating across organizations were prohibitively expensive (Williamson, 1975). However, technological progress has led to a reduction in many of these costs such that firm boundaries are weakening and firms are now able to more easily contract with other firms and even individuals, including users. This same progress has allowed individuals to more easily work together and apply collective intelligence to create goods and services governed under a commons model (Ostrom, 1990). In aggregate,

this has led to a modularization of the firm (Baldwin and Clark, 2003) where firms must increase their engagement with partners in their ecosystem (Iansiti and Levien, 2004) and co-invent with their users (Bresnahan and Greenstein, 1996), in a manner that leads to the locus of innovation being pushed outside the boundaries of the firm (Lakhani, Lifshitz-Assaf, and Tushman, 2013) such that firms become platforms for coordinating third-parties leading to hypercompetition in more concentrated industries (Brynjolfsson, McAfee, Sorrell, and Zhu, 2008).

In addition to changing our notion of what a firm is, technological progress is also changing our notion of what a product is. Prior to the digital age, products were traditionally considered to be physical goods. With the dawn of the digital age, information goods also became products that firms regularly produced. However, we are entering an era where physical goods are starting to become information goods, such that the form and function of a product are now separate (Yoo, 2013). Due to the advent of 3D printing, the process of informationization of physical goods is occurring in everything from houses to pizzas to human body parts. Therefore, the lessons learned in this dissertation about the impact of crowdsourced digital goods with a marginal cost of zero may very well apply to all information goods, and many physical goods, in the near future.

The first chapter of the dissertation, "Innovating Without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero" (w/ Elizabeth Altman, and Michael Tushman), explores how technological progress and reductions in information costs are leading firms to increasingly engage with external digital communities. In particular, firms are increasingly engaging with networks of developers, external labor

marketplaces, and users, with the latter frequently occurring through the process of crowdsourcing. This engagement leads to a weakening of firm boundaries such that the locus of innovation and value creation moves outside the boundaries of the firm. The increase in this phenomenon motivates a reevaluation of many traditional theories of how firms organize and innovate. Specifically, we consider how shifts in information costs affect the classic organizational concepts of firm boundaries, business models, interdependence, leadership, identity, search, and intellectual property. In turn, these effects on the firm's organization alter how the firm innovates.

The second chapter, "Digital Dark Matter and the Economic Contribution of Apache" (w/ Shane Greenstein) examines the economic impact of crowdsourced digital goods at a macro-level. We show that due to its reliance on price to measure value, GDP calculations do not account for "digital dark matter", digital goods and services that are non-pecuniary and effectively limitless inputs into production. We scan 1% of the 1.5 billion IP addresses in the United States to measure the types of web servers businesses and individuals employ. We estimate the value of the free and open source nature of the predominant web server, Apache, by comparing it to the closest pecuniary alternative, Microsoft's Internet Information Services (IIS) server. Our analysis shows that the lack of price for the Apache server leads to an underestimation of GDP by upwards of $12 billion. Although this is the value from only one piece of digital dark matter, this miscalculation represents a large proportion of all software sales and significantly alters economic growth projections.

The third chapter, "Crowdsourced Digital Goods and Firm Productivity: Evidence from Open Source Software", empirically measures the firm-level productivity impact of managers' decisions to use non-pecuniary digital inputs from the crowd. Existing literature examining the impact of IT on productivity does not account for investments in such goods, as their use cannot properly be captured by traditional measurement methods based on price. Therefore, their contribution to the firm's production process is currently unexplored, despite mounting evidence that firms are increasingly relying on these types of inputs. Employing data from a survey of technology use at nearly 2,000 firms over 10 years, I find that a 1% increase in the amount of non-pecuniary open source software (OSS) used by a firm leads to a .073% increase in productivity. This translates to a $1.35 million increase in productivity for the average firm in my sample. This is more than double the magnitude of the coefficient on investments in traditional pecuniary IT capital. I find that this effect is greater for larger firms and for firms in the services industry. I use inverse probability weighting, instrumental variables, firm-fixed effects and data on managerial quality from the World Management Survey to add support to a causal interpretation of these results.

The final chapter of the dissertation, "Organizational Learning Through Contributing to Public Goods: Evidence from Open Source Software," builds on the concepts developed in the other three to explore how firms that engage external communities and contribute to the development of crowdsourced digital goods enhance their ability to extract value from technology-related inputs via increased learning about how these complex goods operate. This study explores this mechanism by using data on firm contributions to Linux, an OSS operating system that is an important public digital good created via crowdsourcing. Using coarsened exact

matching and inverse probability weighting to address endogeneity concerns, this study shows that firms who contribute to the development of OSS capture more productive value from the use of OSS than their non-contributing peers through a process similar to absorptive capacity. Further, this learning has a spillover effect that allows contributing firms to capture more productive value from all of their IT investments, not just OSS.

The goal of this dissertation is to shine light on an important phenomenon and to help bring order out of the chaos of the digital world. Better understanding these aspects of the transformative effects of the digital economy may also contribute to a growing literature on organizations and the changing nature of work. In particular, as digitization leads to more firms structured as platforms whose business models result in the gamification or leisurification of work, people are increasingly doing work for free. A deeper understanding of this phenomenon may help to explain puzzles related to wage inequality and the wealth gap. Together, the results of the four studies in this dissertation show that the digital commons can help create a great deal of economic value, but that this value is difficult to measure via standard economic methods that rely on price to reflect value. These results have important strategic implications for managers and policy makers to consider as organizations increasingly engage with external communities and ecosystems to innovate and create value.

Chapter 1: Innovating Without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero

Elizabeth J. Altman, Frank Nagle, and Michael L. Tushman

ABSTRACT

Innovation traditionally takes place within an organization's boundaries and with selected partners. This Chandlerian approach is rooted in transaction costs, organizational boundaries, and information challenges. Information processing, storage, and communication costs have been an important constraint on innovation and a reason why innovation takes place inside the organization. However, exponential technological progress is dramatically decreasing information constraints, and in many contexts, information costs are approaching zero. This chapter discusses how reduced information costs enable organizations to engage communities of developers, professionals, and users for core innovative activities, frequently through platforms, ecosystems, and incorporating user innovation. When information constraints drop dramatically and the locus of innovation shifts to the larger community, there are profound challenges to the received theory of the firm and to theories of organization and innovation. Specifically, this chapter considers how shifts in information costs affect organizational boundaries, business models, interdependence, leadership, identity, search, and intellectual property.

Keywords: *managing innovation, information costs, information constraints, communities, organization boundaries, technological progress, platforms and ecosystems, user innovation*

Modern business enterprise is easily defined . . . it has two specific characteristics: it contains many distinct operating units and it is managed by a hierarchy of salaried executives.

(Chandler, 1977, p. 1)

What characterizes the networked information economy is that decentralized individual action—specifically, new and important cooperative and coordinate action . . . — plays a much greater role than it did. . . . The declining price of computation, communication, and storage have, as a practical matter, placed the material means of information and cultural production in the hands of a significant fraction of the world's population.

(Benkler, 2006, p. 3)

1.1 Introduction

Information is expensive to process, store, and communicate—at least, that has been the prevailing assumption upon which most of our organizational theories rely. Because information has been hard to gather and process, firms have emerged as hierarchical and control-based organizations (Chandler, 1962). Leaders have developed strategies to compensate for the difficulties of obtaining and processing data. Business models have been built with the underlying assumption that information costs are high (e.g., Tushman & Nadler, 1978). However, with the exponential growth in information processing, storage, and communication abilities, this is all changing. Information costs are rapidly approaching zero, and the constraints associated with information processing are disappearing. Organizations now have the ability to engage with external communities in unprecedented ways. This decrease in information processing costs is having a decentralizing impact on the locus of innovation and, in turn, on how organizations manage their innovation processes.

In this new information context, institutional logics (Friedland & Alford, 1991; Thornton & Ocasio, 1999; Thornton, Ocasio, & Lounsbury, 2012) revolving around Chandler's (1962) hierarchy and control-centric management, which have prevailed in firms such as General Electric (GE), are being challenged by new logics centered on openness, sharing, and external engagement (Benkler, 2006).[1] Recognizing that new doors are opening as information flows more freely than ever before, incumbent organizations are grappling with how and when to address these new logics. For example, in the summer of 2013, GE launched two online three-dimensional (3D) printing contests, which they referred to as quests, inviting entrepreneurs and organizations to submit new designs for aircraft engine brackets and advanced materials production capabilities (General Electric Company, 2013).

Adopting these new logics, and engaging more deeply with communities, has substantive implications for how firms organize and innovate. As we see with GE's call for inputs related to design and production capabilities, the locus of innovation for incumbent firms has begun to move from within the firm to communities beyond its full control. Evidence of this shift and the tension it is creating can also be seen as firms engage with labor/task marketplaces (e.g., oDesk, eLance, TopCoder), developer ecosystems (e.g., Apple's App Store), and user-generated contributions (e.g., open source software, user review websites). All three of these community engagements allow for reductions and blurring of firm boundaries and call into question what the firm does and what resources it owns. As we discuss throughout this chapter, this tension between a Chandlerian logic and a more open and community-centric logic challenges many of

[1]Throughout this chapter, we adopt the definition of institutional logics put forth by Thornton and Ocasio (1999, p. 804) as the "socially constructed, historical pattern of material practices, assumptions, values, beliefs, and rules by which individuals produce and reproduce their material subsistence, organize time and space, and provide meaning to their social reality." This definition embraces both the material and the symbolic and encompasses both formal and informal rules for decision makers.

the assumptions underlying the strategic and organizational research that has been treated as foundational wisdom in management scholarship.

To explore the implications of these phenomena, we start by discussing information processing, storage, and communication and note dramatic increases in capabilities coupled with substantial decreases in costs. Recognizing that cost reductions have enabled wide engagement with external communities, we present a typology of communities, emphasizing those enabled by information cost reduction: labor marketplaces, developer ecosystems, and user-generated contributions. Engagement with these communities involves parties outside the firm heavily participating in, or influencing, innovative processes and product offerings managed by the firm.

We then consider how information costs approaching zero and engagement with external communities affect firm organization and strategy. We investigate what happens with respect to organization boundaries, business models, interdependence, leadership, identity, search, and intellectual property (IP) when organizations engage with communities for capabilities core to their innovative processes. Before concluding, we explore the impact of these organizational and strategic shifts on innovative processes. Utilizing the classic evolutionary process model of variation, selection, and retention, we identify ways in which engagement with communities shapes the path of innovation at each step of the process. We suggest that when information constraints drop dramatically and the locus of innovation shifts to the larger community, there are profound challenges to the received theory of the firm and to theories of organizations and innovation. We conclude with thoughts for how these changes present opportunities for research on innovation and organizations.

1.2 Information Constraints Reduction

Just over 50 years ago, in 1961, the IBM 1301 disk drive, which could store 28 MB of information, cost $115,500 (almost $900,000 in 2013 dollars).[2] In late 2013, Hewlett-Packard's cloud service offered 500 GB (500,000 MB) of storage, almost 18,000 times the capacity, for free.[3] This massive drop in price for information storage costs is representative of the reduction in information costs in general.

Together, information processing, storage, and communication represent the three primary components of information usage. Costs for these three components represent important constraints on how information can be used to drive innovation (Maskell, 2000). As engineers, scientists, and others involved in technology development continue to push the boundaries of their craft, and thus increase technological efficiency, they generate exponential growth rates and price decreases for all three of these components. Recent assessments estimate that information processing capabilities grow at an annual rate of 58%, information storage capabilities at 23%, and capacity for information communication at 28% (Hilbert & López, 2011).

Although the costs for information usage are dropping, not everyone is able to take full advantage of this reduction. First, use of many free services is predicated on access to computing devices and infrastructure. In many parts of the United States and the world, disadvantaged populations have limited access to such devices and infrastructure due to the so-called digital divide (Greenstein & Prince, 2007; Norris, 2001; Warschauer, 2003). Second, although we

[2]IBM archives. Retrieved http://www-03.ibm.com/ibm/history/exhibits/storage/storage_1301.html on December 15, 2014.

[3]The 500 GB of free storage is valid for 90 days. Retrieved from https://www.hpcloud.com/free-trial on June 5, 2013 .

present examples in which information costs have dropped to zero, these frequently occur at scales useful only for individuals or very small firms (e.g., Google Drive's free storage is only 15 GB; larger capacities are offered for a fee to larger enterprises). Although costs for larger firms have also dropped dramatically, large-scale information operations can still be expensive.

Third, whereas the costs of the three primary components of information usage may be approaching zero, there are many complementary assets that are required to fully capture the business value of the information. For example, as firms gather more data from their customers, they require more data scientists to manage the data and extract useful insights from it. Likewise, electricity costs for running and cooling massive data warehouses have started to affect firms' bottom lines (Koomey, 2008). We keep these three caveats in mind as we explore how the capacity for information processing, storage, and communication has been increasing exponentially leading to declining prices that are rapidly approaching (and in some cases have already reached) zero.

1.2.1 Information Processing

Information processing refers to the ability of a device to take information and perform calculations using it. In the computerized world, this is frequently measured by the speed of a central processing unit (CPU), which is correlated with the number of transistors that can fit in a given space on a computer chip. Moore's Law (Moore, 1965) predicts that the number of transistors that can be placed on a chip will double every 18 to 24 months. This leads to exponential growth and an associated reduction in cost per calculation, a pattern that has continued from 1971 to the present. Although some have predicted that Moore's Law is not

sustainable in the long run due to the size of transistors, which are approaching the molecular level (Latif, 2013; Merritt, 2013), new computing methods including multicore chips, DNA computing, and quantum computing should allow for Moore's Law to hold from the perspective of how many calculations can be done per second.[4]

The impact of such sustained growth is often underestimated because it is exponential. Many estimate that information processing power is passing an inflection point in its exponential growth, described by Ray Kurzweil (1999) as entering "the second half of the chessboard."[5] We are entering a period in which the increases in processing speeds will occur in a manner never imagined before (Brynjolfsson & McAfee, 2011). The effects of this exponential growth can already be seen: A modern cell phone has more processing power than all of NASA had in 1969 when humans landed on the moon (Kaku, 2012). Likewise, the processing power of a multimillion-dollar military supercomputer in 1997 could be found, less than 10 years later, in the Sony PlayStation 3 gaming console, released in 2006 for $500 (Kaku, 2012).

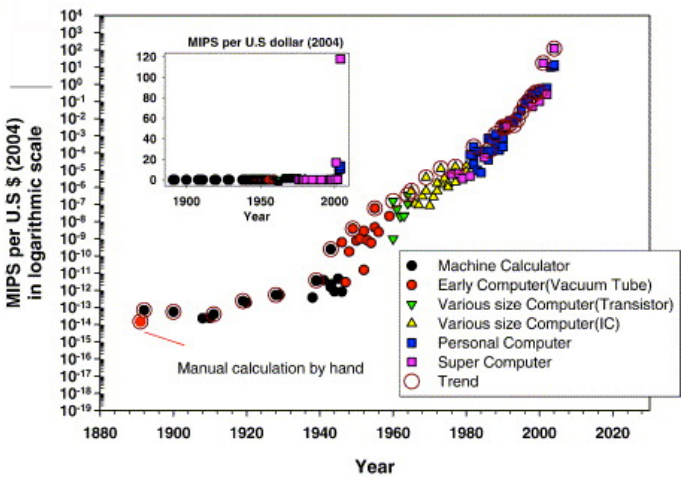
With this exponential growth in processing power has come a dramatic drop in price for a set amount of power (Figure 1.1). For example, in 1996, the best personal computers could obtain about 1 million instructions per second for each US dollar of cost (1 MIPS/$) (Koh &

[4]Multicore chips contain two or more CPUs that run in parallel. DNA computing utilizes the self-assembling nature of DNA to craft problems as half-strands of DNA, which are solved by the matching pieces of DNA. Quantum computing takes advantage of qubits, which are bits of information that can exist as both a 0 and a 1 at the same time.

[5]East Indian lore tells the story of an Indian king who loved chess so much that he offered the inventor of the game any prize he desired. The inventor asked for one grain of rice on the first square of the board, two on the second, four on the third, and so on, doubling the amount for each of the 64 squares on the board. While the amount of rice on the first half of the chessboard was large, it was within the realm of the feasible. However, the amount of rice on the second half of the board was more than all the rice in the world.

Magee, 2006), whereas today, the best personal computers can obtain about 176 MIPS/$.[6] Further, although these prices reflect the cost for cutting-edge performance, it is possible to obtain lower levels of performance for free when utilizing cloud computing services.[7] For example, Amazon Web Services EC2 provides free processing power for 1 year that runs at approximately 1,933 MIPS and HP Cloud provides free processing power for 90 days that runs at approximately 4,545 MIPS.[8] Although today's cutting-edge processing power is by no means free, the processing power that was cutting-edge for a personal computer approximately 10 years ago is now offered for free via cloud computing.

Figure 1.1 MIPS per US Dollar Over Time (Source: Koh & Magee, 2006)[9]



[6]The calculation was based on the Intel Core i7-3960X, which runs at 177,730 MIPS and could be purchased from TigerDirect.com for $1,009 in 2013.

[7]Although there are many definitions of cloud computing, we use a fairly broad definition and consider cloud computing to be the use of computer servers and services that are hosted by a third party and are accessed via the Internet. One key feature of most commonly used cloud computing platforms, including Amazon Web Services and Google Drive, is the ability for a firm to utilize more computing power, storage, and bandwidth on demand, without needing to buy and install servers within the firm.

[8]Amazon Web Services free package information, retrieved from http://aws.amazon.com/free/ on June 5, 2013. HP Cloud free package information retrieved from https://www.hpcloud.com/free-trial on June 5, 2013. MIPS calculations for both were retrieved from http://insights.wired.com/profiles/blogs/all-clouds-are-not-created-equal-2x-cpu-performance-at-nearly-the#axzz3LuAiExLF on June 5, 2013.

[9] We gratefully acknowledge permission from the authors to use Figures 1.1, 1.2, and 1.3.

1.2.2 Information Storage

The costs of information storage have also dropped dramatically. For many years, disk drives have been a common object of study for management scholars due to constant technological disruptions in this industry (e.g., Chesbrough, 2003a; Christensen, 1993, 2006). These disruptions drove an exponential growth pattern similar to that of Moore's Law for transistors. Although each generation of users frequently wonders, "How will I possibly use up all that disk space?," they always do, as technologies evolve and enable people to create increasing amounts of information that needs to be stored. Indeed, industry approximations estimate that by 2010, the amount of information created between the beginning of civilization and 2003 (5 exabytes[10]) was being created every 2 days.[11] This rapid increase of information storage allowed for the progression from text as the only practically digitizable information to pictures and eventually video becoming storable at a reasonable cost. This increased storage has led to websites such as YouTube, to which users upload 100 hours of video per minute.[12]

Not only has information storage space increased, but the portability of this storage has also grown. Magnetic tapes were followed by magnetic disks, optical disks, and flash memory. The latter now allows for up to 1 terabyte[13] of information to be carried on a device the size of a person's thumb. Flash memory was an important innovation that enabled the portable device revolution, which has led to the large-scale production and adoption of smartphones and tablets.

[10]An exabyte is 10^{18} bytes, or 1 billion gigabytes.

[11]Google CEO Eric Schmidt addressing the Techonomy 2010 conference, Lake Tahoe, California, August 6, 2010.

[12]YouTube upload statistic. Retrieved from http://www.youtube.com/yt/press/statistics.html on December 15, 2014.

[13]A terabyte is 1,000 gigabytes or 10^{12} bytes.

Such massive amounts of storage have led to a "save everything" mentality at both individual and firm levels.

Combined with increases in processing power, the ability, and thus the propensity, to save everything has led to the "big data" or data analytics phenomenon that is revolutionizing the way companies do business as they gain the ability to better understand their consumers.[14] Although basic data analytic capabilities have existed for many years, it is only through the emergence of cheap information storage that organizations can now save and analyze enough data to produce deeper and more nuanced analyses of customer behavior for use in prediction, market segmentation, and so on.

As with information processing power, the growth in information storage space has also led to a decline in the cost of storage (Figure 1.2). For example, in 2000, the cost of hard disk storage was about 140 MB/$ (Koh & Magee, 2006); today, storage on an external hard drive costs about 22,073 MB/$.[15] Further, although the largest storage devices are not free, there are a number of storage options that are free. Thumb drives holding 1 GB have become so cheap that they are regularly given out for free.

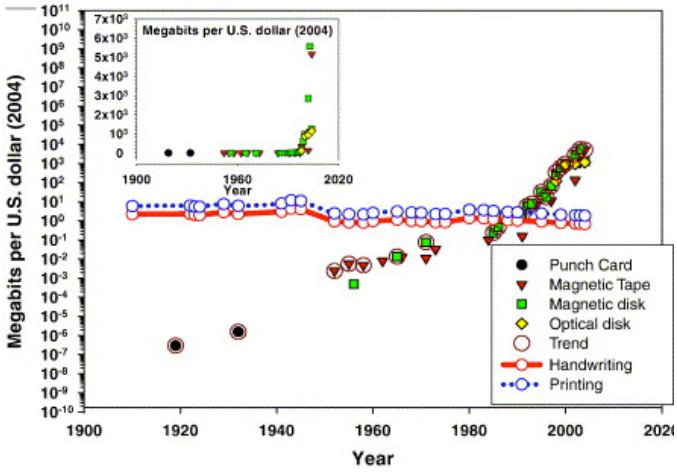
More impressively, coupling gains in storage capacity with increases in information communication power has allowed for extremely cheap, and even free, storage via the Internet.

[14]Although there are many definitions of big data and data analytics, Gartner (2013) defines big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

[15]This calculation was based on the Seagate Backup Plus 4TB External Desktop Drive, which could be purchased from TigerDirect.com for $190 in 2013.

For example, Google Drive offers 15 GB of free storage, Box offers 50 GB, and HP Cloud offers 500 GB for 90 days. A 500-GB disk drive that cost $150 five years ago is $50 today. Further, the same storage space can now be obtained through the cloud for free. These impacts on processing and storage bring down information constraints for large incumbent firms and similarly reduce information costs to essentially zero for new entrants.

Figure 1.2 Megabits per US Dollar Over Time (Source: Koh & Magee, 2006)



1.2.3 Information Communication

Information communication is the ability to move bits of data from one place to another, often from storage to processing and back. We consider this to encompass both machines communicating with each other and people communicating with each other via these machines. Although communication costs within a computer system are certainly one aspect of information communication, we focus primarily on the communication channels that move information from one device to another, namely bandwidth. The ability to move digital bits from one system to another has long relied on existing telecommunications channels, starting with phone lines and moving to cable lines and, more recently, fiber optic lines. Wireless data communication has also relied on existing channels, namely radio and cellular. In both wired and wireless domains,

bandwidth has grown exponentially since the invention of the telegraph and radio in the 1800s (Koh & Magee, 2006). This increase in communication capabilities is what allowed for the creation of the Internet and its growth into a communication channel accounting for 8% of all retail products sold in the United States (Anderson, Reitsma, Evans, & Jaddou, 2011). Ever since the invention of the precursors to the Internet in the 1960s, bandwidth has increased rapidly. For example, in 1984, the fastest modem available to a home user had a speed of 300 bits per second (bps), whereas in 2010 it was 31 Mbps, an increase of 100,000 times in just over 25 years (Nielsen, 2010).

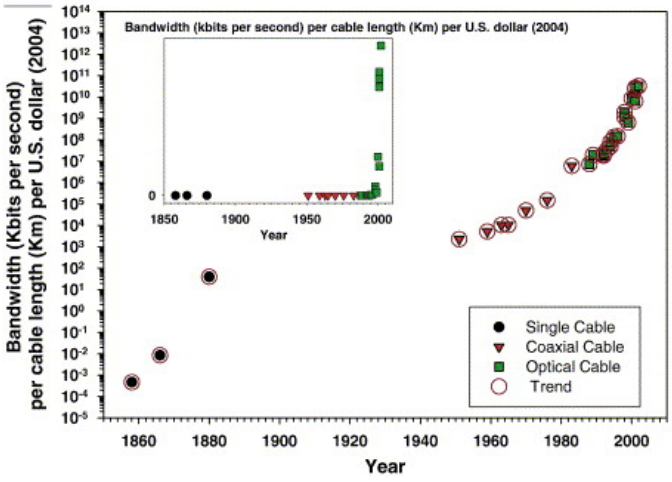
As with information processing and storage, the exponential growth of information communication has been accompanied by a rapid decline in price (Figure 1.3). Industry assessments estimate that the price per Mbps for Internet transit dropped from $1,200 in 1998 to $5 in 2010 (Norton, 2010). However, since the mid-1990s when America Online (AOL) mailed floppy discs to consumers providing free access to the Internet for a limited time, there have been avenues for free access to the Internet. Today, Google Fiber, which boasts maximum speeds of 1 Gbps, offers a free connection to the Internet with download speeds limited to 5 Mbps.[16]

It is important to note that while such cheap bandwidth is readily available in many areas of the United States, in many other areas it is very difficult to get access to high-speed Internet service, creating what many have called the "digital divide" (Greenstein & Prince, 2007; Norris, 2001; Warschauer, 2003). However, even in areas where the decreases in cost have not yet

[16]Although the monthly fee is $0, there is a one-time installation fee of $300. Information retrieved from https://support.google.com/fiber/answer/2476912 on June 6, 2013.

produced wider accessibility for broadband service, cheaper communications allow for innovations such as the delivery of agricultural market prices via text message to farmers in developing nations (Aker, 2010; Jensen, 2007). Around the world, this reduction in information communication costs has had an impact, allowing skilled workers from emerging economies to have access to developed markets via platforms such as oDesk, eLance, and TopCoder. Further, through the rise of massive open online courses (MOOCs), the reduction in information communication costs has allowed anyone with an Internet connection to gain access to high quality education in a vast array of fields. Finally, although some bandwidth may be free, 5 Mbps is not nearly enough to allow a large business to operate effectively, and therefore they must still pay for access, even if the fees are much less than only a few years ago.

Figure 1.3 Bandwidth per Cable Length per US Dollar (Source: Koh & Magee 2006)



Together, the reduction in costs of information processing, storage, and communication have led to more products that leverage modular technologies and standardized interfaces, greater engagement by consumers and other end users, and wide-scale availability of enormous computing power and comprehensive databases. This, coupled with the increased ability to

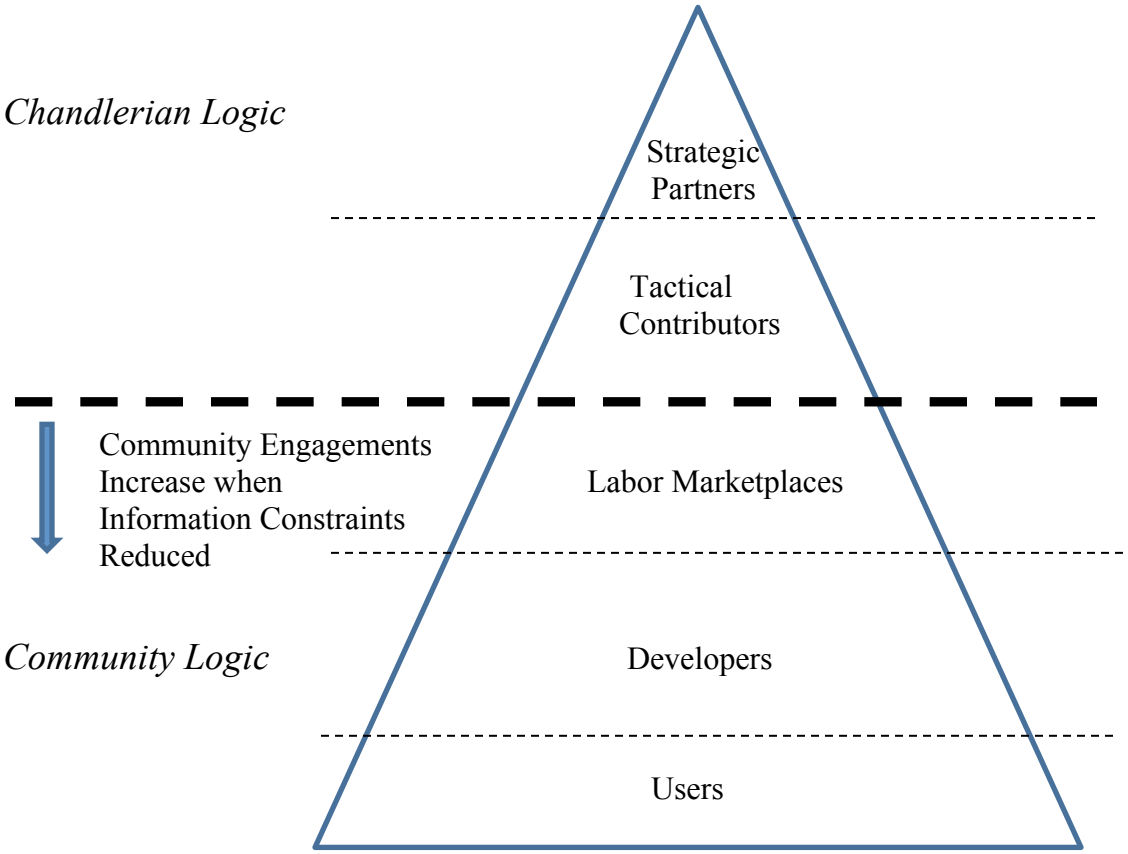
collaborate and coordinate across large distances, has produced wide-ranging effects on the way organizations create and leverage innovations as well as on fundamental organizational processes.

1.3 Engaging Communities

Organizations engage with many types of communities including customers, suppliers, partners, and complementors. One way to visualize the scale of these engagements is through the triangle shown in Figure 1.4. At the top are a small number of strategic alliances. For large technology firms, these may be multidimensional technology, service, and licensing relationships with other large firms. This type of alliance is custom-negotiated, and usually involves senior members of the executive team, possibly including the CEO. A firm will likely not have more than ten to twenty relationships of this kind that are strategic in nature. Microsoft's interaction with Intel is one example of this type of relationship (Casadesus-Masanell & Yoffie, 2007).

The next set of relationships is more tactical but still involves custom negotiations on a case-by-case basis. A relationship in this category is one in which a firm licenses a technology that it integrates into a product. A large firm might have tens of these tactical contributors but probably will not have hundreds. These relationships are usually managed by business development professionals trained to work with interfirm relationships (see the alliance literature; e.g., Gulati, 1998, 2007; Rothaermel, 2001). A mobile phone provider's relationship with a speech recognition technology provider such as Nuance is an example of this type of alliance (Nuance Communications, 2013).

Figure 1.4 Typology of Communities



Beyond these custom-negotiated relationships are community engagements enabled by reductions in information costs. In this chapter, we focus on the bottom three sections of the triangle because they include the types of engagements that are accelerating as a result of the increase in information processing capabilities and the decrease in information costs. These categories of engagement are (1) the advent of external labor and task markets, (2) the rise of developer ecosystems, and (3) the growing prevalence of user-generated contributions. Considering labor marketplaces, we examine how firms engage with parties beyond their legal control to accomplish tasks they previously would have performed internally. With developer ecosystems, we look at how complementary firms provide value to end-users. With user-generated contributions, we consider how firms engage users to contribute value. Organizations

that use labor marketplaces might have many interactions with individual external workers contributing to a project. Organizations with developer ecosystems may have hundreds, thousands, or potentially more than a million developer relationships. Organizations that interact with users could have millions of contact points. In Table 1.1, we summarize how engaging external labor, developers, and users changes with and without information constraints.

Reductions in information processing, storage, and communication costs make these relationships not only feasible but also attractive, though they need to be managed in an entirely different way from those in the top two sections of the triangle. Institutional logics that revolve around openness and sharing become essential, but they differ from the prevailing logics of hierarchy and control. Firms need to grapple with how to manage these multiple logics as they cope with an array of complex community engagements. These interactions create challenges (e.g., contrasting logics, more user input than a firm can easily process) and opportunities (e.g., introducing benefits from entities beyond those directly controlled by the firm). Studying these phenomena may prompt us to think differently about innovation, organizations, and our classic theories that explain them. Innovation is no longer occurring primarily within a firm; rather, organizations now engage with others who also innovate in ways that improve the organization's products, experiences, and value. These interactions result in new behaviors to create, capture, and select innovations while also introducing fundamentally new managerial challenges.

1.3.1 Labor Marketplaces

Labor marketplaces, also known as task marketplaces, are multi-sided platform-based businesses that allow firms and individuals that have specific tasks to find people to accomplish

those tasks. Tasks posted on the most popular of these platforms (e.g., oDesk, eLance, TopCoder) include everything from website design to language translation and marketing. Sometimes also referred to as "the human cloud" and considered the next generation of outsourcing after information technology (IT) and offshore outsourcing, these marketplaces comprise an ecosystem of platforms linking virtual workers with employers who hire them on an as-needed basis.

The recent rise of these platforms is substantial, with growth in global revenue amounting to 53% for 2010 and 74% for 2011 (Kaganer, Carmel, Hirscheim, & Olsen, 2013). Addressing some original concerns about transparency, quality control, and coordination in these labor relationships, these marketplaces now have mechanisms to allow hiring managers to monitor contractors' work as well as standardized contracts and dispute resolution services (Needleman, 2010). Task platforms allow a firm to rely on external parties for much of its labor supply in a way that was previously not possible before information technologies enabled the collaboration and communication feasible today. As we discuss in the next section, this reliance on external labor has important implications for organizational and strategic decisions.

1.3.2 Developer Ecosystems

Technology developments enable firms to deploy goods that are increasingly modular, with open interfaces allowing independent entities to contribute to end-products (Baldwin & Clark, 2000). Although many firms design and develop self-contained products that provide a complete user experience, increasingly more products require after-market applications or accessories to deliver full value (Adner, 2012). In using labor marketplaces, organizations

engage external parties directly and hire resources to further their missions. In contrast, when they build developer ecosystems, organizations enable external parties (developers) to create complementary products (apps or accessories) that customers acquire either directly from the external parties or through a marketplace.

Prevalent examples of firms with developer ecosystems are those that offer smartphones, tablets, and other devices that users customize with apps and accessories. Beyond consumer products, this same phenomenon exists in other industries, such as medical diagnostic devices. Welch Allyn traditionally provided integrated systems to doctors' offices and hospitals allowing medical practitioners to measure blood pressure, temperature, and so on. Today, it offers a platform system to which doctors and hospitals can add modules and apps provided by other firms (Welch Allyn, 2011, 2013).

The widespread availability of apps is driven by underlying reductions in information costs. Firms are able to leverage today's ease of processing and communication to open interfaces to their products, providing application programming interfaces (APIs) and software development kits (SDKs) and encouraging other firms to contribute to their products. Consumers are able to easily download apps to improve products they purchase, and market evidence indicates that they are doing so in large numbers. In May 2013, Apple announced that 50 billion apps had been downloaded from its App Store, which offers more than 850,000 apps in 155 countries for a suite of iPhone, iPad, and iPod Touch products (Apple Inc., 2013). In Facebook's second quarter 2013 earnings release, to benefit its 1.15 billion monthly active users, it announced that more than 100,000 apps had been built (Facebook, 2013). Complementary firms

(such as app developers) are able and incentivized to develop these apps because they have easy access to product information through developer websites and ease of distribution through app stores and other means. Enticed by the prospects of serving enormous markets, and equipped with enabling technologies and documentation, developers invest in creating apps and accessories for other firms' products. Firms and their complementary developers and accessory providers need to employ institutional logics consistent with operating in a world that is highly open and decentralized with significant sharing and interdependence.

1.3.3 User-Generated Contributions

As the drastic reduction in information costs has made it easier to engage an ecosystem of developers, it has also made it easier for organizations to connect with the users of their products and services (Von Hippel, 2009). In explaining this phenomenon, Benkler (2006, p. 5) highlights "the rise of effective, large-scale cooperative efforts—peer production of information, knowledge, and culture." Indeed, in some cases, such as open source software, users have become the entirety of the organization developing the product. In these cases, the creative contributors no longer reside inside an organization. Rather, they exist in a loosely affiliated community with its own set of operating procedures and norms that have developed to govern behaviors (O'Mahony & Ferraro, 2007; Shah, 2006).

Many open source software projects started within an organization and then were taken over by a group of users after the code base was opened. For example, Apache began as a federally funded research project and is now a fully open source project that runs more than 50% of websites on the Internet (Greenstein & Nagle, 2014). In a survey of large organizations, 50%

of respondents said they use open source software in their business, and another 28% said they are considering using it (Trapasso & Vujanic, 2010).

Although these types of open source software projects exist in an entirely community-based self-governing organizational form (Benkler, 2006), in more traditional firms there are increasing examples in which user-generated contributions provide firms with free inputs. For example, user-generated product and service reviews on Amazon, TripAdvisor, and Yelp help drive sales and profits of reviewed firms and products (Duan, Gu, & Whinston, 2008; Liu, 2006; Luca, 2011). Further, companies such as Threadless rely on users for idea generation and selection (designing products and determining which products are most likely to be successful in the market) (Lakhani & Kanji, 2008). All of these activities (open source software, user-generated reviews, and user idea generation and selection) are enabled by reductions in information costs.

As information costs drop sharply and all three types of community engagement increase, sharply inconsistent logics emerge within incumbent firms. Incumbents need to balance operating in their traditional internally focused mode with an approach that is more externally oriented and inclusive. They need to manage competing logics that will be more pervasive than ever before (Lounsbury, 2007). Table 1.1 summarizes how these three types of communities (labor, developers, and users) change as the environments in which they operate move from a world where information is constrained to one in which information constraints are essentially nonexistent.

Table 1.1 Engaging With Communities With and Without Information Constraints

	With Information Constraints	Without Information Constraints
Labor	• All internal to the firm, or specialized contracting through temp agencies and contractors • Long-term engagements and large-scale projects • Difficult performance quality control and monitoring	• Labor marketplaces • Micro-jobs enabled • Community rating schemes and digital monitoring
Developers	• Organization-to-developer contracting • Select few high-maintenance relationships between organizations and developers • Significant case-by-case IP considerations and negotiations • Embedded applications ("pre-loads") executed by engineering teams	• User-to-developer contracting • Many arm's length developer relationships governed by simple click-through licenses • IP licensing tailored for engagement with high volume of organizations (e.g., automated websites for contracts) • App store applications ("post-loads") by third-party developers
Users	• Users engage almost exclusively through customer service representatives • Inputs are primarily customer complaints or repair requests • External inputs are avoided	• Users provide inputs across functional organizations (e.g., to engineering and marketing) contributing to full design process • Inputs include product design suggestions, manufacturing ideas, and so on • External inputs are embraced as a valuable part of product design and delivery

1.4 Organizational and Strategic Implications

Organizations that flourished during the industrial age focused their energy on managing physical assets. The constraints they battled related to physical goods, production challenges, and employment issues. In contrast, organizations during the information age leverage sophisticated information technologies to manage their resources and pursue product development. Incumbent firms reach beyond traditional organizations and interact with individuals, firms, and communities to create offerings integrating contributions from a variety of sources. They undergo structural transitions to operate in a networked information economy characterized by decentralized action by individuals cooperating and coordinating through distributed nonproprietary, non-market strategies (Benkler, 2006).

The effects of this new economy span organizational and institutional levels. As these firms engage beyond their boundaries, they outgrow the strategies, business models, and organizational processes theorists have been studying for decades and challenge their institutional logics. Whereas previously they managed based on a Chandlerian logic that emphasized hierarchy and control (Chandler, 1977; Thornton & Ocasio, 1999), firms today balance multiple logics that incorporate peer production, information sharing, data access, and free goods. As they modify their institutional logics in response to new strategies and organizational transitions (Gawer & Phillips, 2013), they undergo institutional work, which Lawrence and Suddaby (2006, p. 215) defined as "the purposive action of individuals and organizations aimed at creating, maintaining and disrupting institutions."

Take, for example, research and development (R&D), an institutionalized category with well-understood meaning and value in society beyond the work it encompasses (Meyer & Rowan, 1977). As information constraints decrease, categories of activities change in terms of work processes, symbols, and myths that surround them, creating challenges for institutionalized rules. For example, whereas R&D used to be performed almost entirely by professionals employed within a firm, it can now be a joint activity spanning internal experts and external contributors.

In the context of increased community engagement and enhanced roles for user contributions, institutional entrepreneurs (Battilana, Leca, & Boxenbaum, 2009; Greenwood & Suddaby, 2006; Maguire, Hardy, & Lawrence, 2004) are increasingly found outside traditional

boundaries of firms. One example is social networks, which were originally a means for students to connect with each other and now have evolved to become, among other things, a primary venue for sharing photographs as well as a useful setting for firms to garner insights into consumer sentiment (Nagle, 2013). This change was largely driven by user innovators rather than members of existing firms.

Another example is the evolving role of quality assurance (QA) departments. In the days of mainframe computing, a QA department would be responsible for extensive testing of mainframe software before release. Today, users provide immediate feedback to software firms, so the role of QA professionals includes developing and managing mechanisms to collect and manage quality-related feedback from users. At the extreme, in community-centric peer production contexts such as Wikipedia, the QA role has been entirely shifted to the community (Piskorski & Gorbatai, 2013), further challenging institutionalized norms.

These community-based innovation processes affect a range of topics associated with strategy, innovation, and organization theory. These topics include organizational openness (Boudreau, 2010; Chesbrough, 2003b), community engagement (Lakhani, Lifshitz-Assaf, & Tushman, 2013; O'Mahony & Lakhani, 2011), user innovation (Lakhani & Von Hippel, 2003; von Hippel, 2009), networked economies (Benkler, 2006; Castells, 1996), and other related topics such as multi-sided markets (Hagiu & Spulber, 2013; Parker & Van Alstyne, 2005), and social media (Piskorski, 2013).[17] Regardless of where one falls on the spectrum of views related to these topics, or to which version of openness or community engagement one subscribes, they

[17] For a broad overview of the technology and innovation management literature, see Altman, Nagle, & Tushman, 2013.

all clearly have organizational implications. These include the effects on firm boundaries, strategy and new business models, interdependence and community engagement, leadership, identity, search, and IP. Table 1.2 shows how these organizational and strategic characteristics vary as information processing, storage, and communication become virtually free.

1.4.1 Boundaries

The concept of firm boundaries and what is considered inside versus outside the control of a firm (March & Simon, 1958; Pfeffer & Salancik, 1978; Thompson, 1967) is challenged as information constraints decrease and firms become more community-centric (Gulati, Puranam, & Tushman, 2012; Lakhani et al., 2013). Gulati et al. (2012, p. 573) introduced the notion of meta-organizations comprised of "networks of firms or individuals not bound by authority based on employment relationships, but characterized by a system-level goal." They developed a typology based on degrees of stratification and permeability of boundaries. These organization types, all of which bring together autonomous entities into an interconnected system, are largely enabled because information costs are so modest. Researchers have also explored the porosity of boundaries under various circumstances (Santos & Eisenhardt, 2005), and alliance researchers such as Dyer and Singh (1998) have considered the strategic value of relationships between alliance partners and networks. Yet, there remains substantial opportunity for research that considers the effects on organization boundaries as information constraints approach zero and community engagement becomes more prevalent.

A reliance on external labor leads to a weakening of firm boundaries. Task marketplaces reduce an organization's need to hire internal employees by providing a marketplace with

Table 1.2 Organizational and Strategic Characteristics With and Without Information Constraints

	With Information Constraints	Without Information Constraints
Boundaries	• More employees inside organization because it is less expensive to include them within the organization than to contract externally • Difficult to find appropriate person for job • Hold-up problems exist because individuals with specific skills have power over the organization • Organizations contract with firms providing services, rather than with individuals, thus difficult to fire underperforming individuals outside organization boundaries • Vertical and horizontal integration attractive strategic alternatives because market costs tend to be expensive • Organizations incur costs and risks associated with internal computing assets for innovation	• Fewer employees within organization because it is easy to contract with external employees when organization needs more human resources • Easy to find appropriate person in the community, so coordination costs decrease with matching efficiencies • Hold-up problems reduced because there is efficient marketplace with large supply of highly skilled people • Organization-to-individual contracts are the norm, so it is easy to fire a temporary individual • Vertical and horizontal integration less attractive strategic alternatives because market transactions are less expensive • Organizations can pool risk and costs associated with computing by using cloud computing
Strategy and New Business Models	• Organizations own or tightly contract for the assets they need • Digital goods (e.g., software) are expensive to produce, and user inputs are virtually impossible to capture • Differentiation is straightforward when resources are unique to the organization • Strength of organization resides in owned resources and skills • Difficult to conduct corporate entrepreneurship because of shared resources • Entrepreneurial organizations need to build capabilities internally to compete	• Assets are free and open; organizations leverage what they need • Free digital goods (e.g., open source software, user reviews, and ideas) are widely available for the organization to leverage • Differentiation is hard when leveraging widely available common public goods • Organization strength resides in skills and knowledge processing, not in owned resources • Corporate entrepreneurs can leverage labor markets, cloud computing, and so on, to create their own space inside the organization • Entrepreneurial organizations, including solopreneurs, can cost effectively engage external resources allowing them to highly specialize

Interdependence and Community Engagement	• Organization owns and controls computing resources for innovation • Organization internally owns resources critical to accomplishing its mission • Outputs created by the organization and/or partners with whom it is tightly contractually bound, so organization controls own destiny • Developers contract case-by-case with individual organizations and engage in strategic relationships • Accessories and applications created using resources owned by the organization	• Organization does not control, and is reliant upon, cloud computing partner to provide innovation resources • Organization contracts externally for resources critical to accomplishing its mission • Outputs created by partners with loose affiliations, so organization has high interdependence with many entities • Developers join ecosystems, must comply with ecosystem rules, and become reliant upon success of the platform • Accessories and applications created by resources residing outside the organization
Leadership	• Hierarchy and control are primary means of managing external parties (agents) through contracts • Organization must incur expenses to monitor all agents (partners) • Administrators must satisfice because they are choosing from bounded options • Leaders operate in a hierarchy • Engagement with outside communities restricted to particular staff members engaging with limited communities (e.g., disgruntled customers)	• Adopt community logic and incorporate behavioral incentives, influence, and persuasion as primary means of managing external parties (agents) • Communities via review mechanisms provide monitoring and quality control role at drastically reduced costs • Administrators satisfice less because they have more and broader options • Leaders must manage in communities • Engagement with outside communities to harness external creativity becomes central element across functions (e.g., R&D, marketing)
Identity	• Dimensions of internal organizational identity focus on internal development (e.g., R&D excellence) • External organizational identity is associated with the organization • Professional identity is associated with internal development and creativity	• Dimensions of internal organizational identity shift to emphasize engaging communities (e.g., developer evangelism) • External organizational identity (image) encompasses both the organization and related communities • Professional identity is associated with engaging external communities, sourcing, and selecting creative outputs

Search	• Local search is predominant • Search is expensive and thus there is limited rational choice in decision making • A challenge for exploitation is that gathering user feedback to incrementally improve products is hard • Exploration is hard because it is difficult to engage in distant search (hard to cast a wide net)	• Distant search, particularly leveraging communities, is predominant • Search is cheap, so decision making can be more rational • Exploitation is easier due to enhanced user feedback (e.g., localization) • Exploration is easier because distant search is cheaper
Intellectual Property (IP)	• Organizations protect IP with various legal mechanisms such as patents, trademarks, copyrights, and trade secrets • When organizations engage in interorganization collaborations, they execute traditional cross-licensing IP contracts • Without access to free digital goods, organizations need to either create or buy resources, both of which have well-defined ownership and IP implications	• IP considerations become very tricky, and organizations need to consider who owns inputs as well as outputs • Licensing involves various types of open source and public goods licenses • Availability of free digital goods provides opportunities for organizations to source resources without cost but introduces challenges related to ownership and IP

standardized contract terms and efficient matching of tasks to task performers. The matching mechanisms allow task performers to very clearly showcase their skills and portfolios of past projects, while also allowing organizations to concretely define tasks they need completed (Kaganer et al., 2013). Standardized contracts are designed to let two parties negotiate price, time for completion, and task details while covering issues such as IP and task monitoring in a consistent way. Traditionally, hierarchies are utilized to limit coordination and contracting costs (Coase, 1937; Jensen & Meckling, 1976; Thompson, 1967; Williamson, 1975). However, task platforms allow organizations to limit these costs by using markets instead of hierarchies to execute tasks.

For organizations engaging with task marketplaces, the two primary risks are projects not being completed and IP leaks (Kaganer et al., 2013). However, the scale of these marketplaces makes it possible for organizations to engage in redundant projects, which decreases failure risk. Further, task performer reputations are publicly available, incentivizing performers to complete projects that garner good feedback from their employers. To manage IP concerns, organizations employ multiple strategies such as breaking tasks into small subunits such that any individual contributor does not have enough information to make a leak valuable. Further, the high volume of individual task performers participating in labor marketplaces results in competition, which allows organizations to seek qualified individuals, test their services, and easily contract with a different person if the first is unsatisfactory. This reduces the importance of hold-up problems (Klein, Crawford, & Alchian, 1978) because organizations contract with individual contractor employees rather than hiring an outsourcing organization. Hart and Moore (1990) noted a distinct difference between firms hiring employees directly and those contracting with

outsourcing firms. When hiring employees, firms can fire individuals who underperform. In contrast, when outsourcing with third-party contractors, firms cannot address problems with individual workers. Task marketplaces eliminate this problem because individuals are contracted on a discrete basis, and thus contracts can be managed individually.

Activities enabled by reductions in information constraints and broader engagement with communities of complementors and developers also allow for a reduction in the need for vertical and horizontal integration, and thus organization size. Transaction cost economics (TCE) maintains that firms come into existence when the costs of a transaction in the market are higher than the costs of performing the same transaction within a firm (Coase 1937; Williamson 1981). However, when user-generated contributions are freely supplied, the costs of transactions are essentially zero, and therefore it is no longer logical to have these activities located within a firm. For example, because the creative agency Victors and Spoils relies on crowdsourcing to develop advertising campaigns, it does not need to employ as many creative designers as a traditional firm. Although it has long been known that firm boundaries shrink as IT (Malone, Yates, & Benjamin, 1987; Brynjolfsson, Malone, Gurbaxani, & Kambil, 1994; Hitt, 1999) and the Internet (Afuah, 2003) reduce information costs and associated transaction costs, few studies have considered what happens to organizations when information costs, and thus transaction costs, essentially vanish.

Cloud computing similarly leads to potential reductions in firm boundaries by decreasing information costs and allowing organizations to rely on external parties for critical needs (e.g., a powerful set of IT tools for innovation). Traditionally, risk reduction has been an important

reason for firms to conduct activities internally (Chandler, 1962). However, by allowing organizations to rapidly scale their computing needs, cloud computing greatly reduces the risks associated with purchasing large and expensive servers. Cloud computing allows an organization to offload the risk of overbuilding computing capacity by contracting with a third party who pools capacity demand with that of other organizations (Simchi-Levi, Kaminsky, & Simchi-Levi, 1999).

1.4.2 Strategy and New Business Models

As organizations leverage more free and open assets (e.g., open source software, user reviews and ideas), it becomes less clear what assets an organization needs to own and how it differentiates itself from competitors. When information constraints were high, these assets were expensive to produce, and user inputs were essentially impossible to capture. Now, these goods are widely available, and organizations can leverage them to accomplish their goals. However, organizations also need to re-think their basis of competitive differentiation. Perhaps the knowledge and strategies for utilizing such free and open assets will become the most important assets of an organization, and perhaps the only assets it truly owns (Teece, 2007). Consequently, an organization's most valuable assets, the knowledge and information within the organization (Arrow, 1975; Teece, 1982) and the mechanisms through which this knowledge is processed (Tushman & Nadler, 1978), will become the largest avenues for sustainable competitive advantage.

Taking advantage of these new assets and modes of competition requires the adoption of new strategies and business models and/or the modification of more traditional ones

(Chesbrough & Appleyard, 2007; Dahlander & Gann, 2010). With information costs decreasing, community engagement increasing, and new opportunities related to opening and expanding boundaries, organizations need to supplement existing business models with new approaches that capture the creativity and inventiveness of external innovators, such as those related to developer ecosystems, labor marketplaces, and user contributions. Crowdfunding, in which organizations search for funding by engaging with a wide community of potential investors, is an example of an emerging business practice in which organizations can also capture resources from external parties through taking advantage of dramatically reduced information constraints. Entrepreneurship provides a business approach that by its nature leverages scarce resources and thus thrives as information costs decrease and more resources become available with much less investment. Within large organizations, the entrepreneurial model can be mimicked through corporate entrepreneurship, in which small groups within organizations can enable mature incumbent organizations to explore new and innovative areas while continuing to exploit existing capabilities (Bresnahan, Greenstein, & Henderson, 2011).

Another business model enabled by inexpensive information capabilities is the rise of "solopreneurs," individual entrepreneurs who can build entire companies without ever hiring internal employees. Solopreneurs, such as AllergyEats and SociallyActive, no longer need to acquire large amounts of capital to buy servers and IT support, formerly an important barrier to entry; rather, they rely on cloud computing. Further, solopreneurs can utilize labor marketplaces to perform functions that previously would have required entire departments. Website design, marketing, and even sales can all be contracted out to external parties via task marketplaces. Additionally, these types of organizations can engage their users as sources of content and

direction. Although solopreneurs have existed throughout history, drastic reductions in information costs are allowing them to have a broader impact that helps them compete with larger, established organizations by focusing on their core competencies (Prahalad & Hamel, 1990) in highly specialized entrepreneurial ventures.

1.4.3 Interdependence and Community Engagement

The Internet and peer production processes function as effectively as they do because of adoption of new technical and organizational architectures combining contributions from diverse providers (Benkler, 2006). These architectures have as a defining characteristic their ability to deal with interdependencies among modular components. As Internet-based technologies become more pervasive throughout core business processes, incumbent organizations and institutions will continue to adopt new institutional logics consistent with the new processes (Thornton et al., 2012). As these organizations participate more broadly in peer-production processes, contribute to sharing communities, and generally engage in more modern forms of community interaction, they will need to develop organizational processes that embrace interdependence and community engagement.

Coordination and integration are challenges organizations face as a result of this increased interdependence and more complex logics. Okhuysen and Bechky (2009) addressed these topics and considered the creation of integrative conditions for coordination, such as accountability, predictability, and common understanding. In ecosystems incorporating community engagement, the conditions for accountability are sometimes unclear. For example, when a platform owner decides to upgrade technologies it is unclear whether the platform owner

is responsible for maintaining backward compatibility to protect all developers and for how long it would need to do so. The extent to which platform owners need to provide predictable technology roadmaps is also debatable. To leverage reduced information constraints and build and maintain a developer ecosystem, an organization needs to focus on the questions associated with these coordination mechanisms (Adner, 2012).

Interdependencies vary depending on the type of entity with which the focal organization is engaging. Organizations have interdependencies with suppliers with whom they contract directly (e.g., cloud computing, IT service providers). They also have interdependencies with complementors. Both types of interdependencies have significant implications for organizations related to how they consider and manage firm boundaries (March & Simon, 1958; Pfeffer & Salancik, 1978; Santos & Eisenhardt, 2005; Thompson, 1967). And, both increase as information constraints decrease and organizations engage with communities more broadly.

Complementor interdependencies are becoming more frequent and complex as product design, development, and deployment are evolving, particularly as more modularized products are introduced into the world with open interfaces ready for additions by other organizations (Baldwin & Clark, 2000). Formerly, product development efforts were primarily internal or occurred through a network of closely affiliated suppliers and strategic alliance partners, but when organizations build and engage with communities, the product experience is developed in conjunction with organizations operating outside the central organization's legal and economic boundaries. The central organization may exert control in terms of regulating distribution of

products through app store requirements or branding programs (such as Apple's "Made for iPhone" logo), but complementors act and innovate independently.

An example of complementors' actions influencing a central organization is privacy breaches by Facebook application developers (Steel & Fowler, 2010). Developers disclosed users' personally identifiable information (PII). Users were infuriated with Facebook. In fact, Facebook was not releasing data; app developers were releasing information after users opted in to using the apps. However, the perception was that Facebook was releasing user information. Facebook was harmed by actions of complementors they did not control.

With lower information constraints, organizations are enabled to develop and grow ecosystems and encourage communities, consisting of either organizations or individuals, to invest on their behalf. An example is a smartphone maker that encourages app developers and accessory providers to create products that work with its particular smartphones. This creates interdependencies between the phone maker and the app and accessory providers in which both become dependent on each other for business success. The smartphone provider needs apps and accessories to be available so that its product is attractive to consumers. The app and accessory providers need the smartphone provider to make available sufficient advance information so they can create compelling complementary products. Additionally, app and accessory providers must address the risk that smartphone providers might introduce new models rendering existing apps and accessories obsolete. The app or accessory organization has no control over a situation that could potentially lead to a significant negative impact such as high inventory scrap costs.

Interdependence among various members of an ecosystem also leads to risks being shared. From the perspective of the focal organization, there is a diversification of risk to developers or accessory providers. From the vantage point of an app developer or accessory provider participating in an ecosystem, there is risk associated with decisions the focal organization might make to the detriment of the accessory provider. However, these risks are usually justified by the great benefits that also exist from potential growth of the overall market.

1.4.4 Leadership

As information costs dramatically decrease and organizations engage more actively and comprehensively with communities of all types, leaders are faced with new challenges, and new leadership styles emerge. Roles transition from directing work in a traditional hierarchy (Chandler, 1977) to sourcing and organizing contributions in a more interdependent loose affiliation of communities. This is true for interactions within incumbent organizations (managing employees), outside the organization (managing suppliers and complementors), and in the newer community-based organizational forms. As Benkler (2006, p. 67) explained regarding the large-scale Linux operating system development process, "a certain kind of meritocratic hierarchy is clearly present. However, it is a hierarchy that is very different in style, practical implementation, and organizational role than that of the manager in the firm."

Because of increased access to information, leaders no longer can use asymmetries of information as a significant source of control. Herbert Simon (1945/1997) outlined considerations related to the creation of an administrative organization and highlighted the notion of influencing staff members (beyond just directing them). This is even more relevant when staff

members have the same or better access to information and information processing than managers. Similarly, in a context where user-generated contributions play a significant role in product development and brand management, leaders need to influence not only staff members but also those in the community who contribute work, reviews, and other resources to projects.

Leaders also need to manage and orchestrate interactions with ecosystem members, and the form of management cannot be one of traditional hierarchy and control because the members are independent entities outside the organization. Instead, leaders need to use incentives and persuasion, frequently referred to as "developer evangelism" by practitioners in this arena, to convince developers to invest in their products. Developer conferences, websites, tools, and cross-promotions are all means that leaders can use to influence developers to invest valuable resources on behalf of their organization as they expand their search for innovative solutions beyond their boundaries (Rosenkopf & Nerkar, 2001).

Illustrating the importance of engaging individuals, Samsung has long had a developer program through which developers can obtain product information online and attend local conferences. Expanding this activity, Samsung hosted a worldwide developer conference in October 2013. The conference website invited participants to "Engage with industry leaders; Collaborate with fellow developers; Learn about new Samsung tools and SDKs; Create what's next" (http://samsungdevcon.com/sdc13/). This highlights the importance that Samsung's leadership is placing both on building relationships with ecosystem members worldwide and also on the role they need to play in fostering community interactions among members.

Beyond considering influence and persuasion, Simon's (1945/1997, p. 199) notion of an administrator as one who satisfices, choosing actions that are satisfactory or "good enough," is worth reconsidering when inputs are from large external communities. To what extent do administrators need to satisfice when the solutions from which they are choosing come from external communities widely diverse in functional expertise, geography, motivations, and experiences? No longer are managers bound by inputs from their employees and close partners; rather, they may be able to get closer to the economic model of maximizing decision making when search extends beyond the boundaries of their organization to large-scale communities.

Furthermore, top management team operations and roles (Finkelstein & Hambrick, 1996) may be affected by changes as a result of decreasing information constraints. Just as individuals might be affected by shifts in the relative importance of roles when firm boundaries shift and interdependence increases, so too might dynamics within top management teams change. For example, as developer communities become increasingly important, the roles of team members who create and nurture these communities might also increase in importance. However, in a management team where product development professionals have traditionally held sway, shifting power to business development staff might be a difficult transition for a leadership team. Additionally, the openness associated with more community engagement may introduce top management team challenges related to managing paradoxes and contradictions as leaders aim to protect traditional proprietary advantages while embracing creative innovative inputs from external parties (Smith & Tushman, 2005).

Moreover, across the organization, shifts to broader external community engagement, sharing, and openness may introduce challenges related to roles and functional responsibilities. In the past, primary engagement with external communities was largely restricted to particular staff members, such as customer service personnel. Now, in cases where sharing with external parties becomes important and more pervasive, other functional areas (such as product development) might need to interact directly with external parties and process their inputs (e.g., suggestions from users).

Monitoring costs, a central topic in the TCE discourse (Williamson, 1981), vary in the context of interdependent communities. One might initially think that monitoring costs would increase as the number of developers in an app store increases. In fact, through network effects, the more popular an app store becomes, with an increasing number of apps, the larger the community of users it develops, and that community then contributes reviews to the marketplace, which serve as a form of monitoring. In practice, a conglomeration of developers monitors all the individual developers. Therefore, not only does lack of information constraints allow for production of complementary goods by parties outside the organization, it also allows for monitoring and quality control of these goods for free by users. Leaders may no longer need to manage organizations of individuals monitoring outputs but rather organizations of individuals nurturing and managing the community that monitors outputs.

1.4.5 Identity

Organizational identity research encompasses both an internal perception of organizational identity (Albert & Whetten, 1985) and an external conception, which is

sometimes referred to as an organization's image (Dutton & Dukerich, 1991). As information constraints decrease and the locus of innovation moves outside the organization, both internal and external conceptions of organizational identity may be challenged. With respect to internal organizational identity, as an organization transitions from creating innovations entirely internally to sourcing and selecting innovations externally, it may change from considering itself as primarily a research-based organization to being one that delivers innovative product experiences regardless of where they are sourced. This may lead to changes in which functions have the most power in an organization, potentially shifting the power base from engineers to business development professionals or vice versa, depending on the nature of the organization.

Relative to external identity, an organization may change from presenting itself as primarily a technology-led product organization to a services-based one. It may move from having an organizational identity centered on the organization alone to one that encompasses both the organization and its related communities (e.g., its developer ecosystem). In both cases, the organization's identity may be threatened and undergo a transition as a result of transitions prompted by technological changes (Tripsas, 2009).

Identity spans levels of analysis considering both individuals and organizations (Gioia, 1998). Both of these identity types may shift as organizations transform, and the two may influence each other (Fiol, 2002). How employees identify with their organization and with their professions is likely to be challenged as the locus of innovation moves outside the organization. When much of the innovation included in an organization's product offering is being sourced externally, do employees have the same level of pride in their organization? As engineers

transition from considering themselves creators of innovations to evaluators of others' innovations, is there also a potential threat to their professional identities (Ibarra, 1999; Lifshitz-Assaf, 2013)? Must organizations hire people with different profiles when the roles of people within R&D include much greater levels of interaction with external communities? Professional identities are increasingly associated with engaging external communities, sourcing, and selecting creative outputs rather than with internal development and creativity when an organization is more focused on external engagement. Both individual and organizational identities provide powerful lenses through which we can study these changes. Further, organizational identity research could likely benefit from examples that link changes associated with information constraints reduction, such as product-to-platform transitions, with identity transitions (Altman & Tripsas, 2015).

1.4.6 Search

Search and decision making (Cyert & March, 1963) are relevant topics to reconsider with respect to organizations and communities in the context of minimal information constraints. A fundamental underpinning of rational choice theory is that there is a cost associated with gathering better information. In his behavioral model of limited rational choice, Simon (1955, p. 112) tied these costs to aspiration levels of individuals and then built his argument on the idea that a "behaving organism does not in general know these costs" and thus cannot be fully rational in its decision making. In the world of social media, users employ tags, "like" buttons, and hashtags to signify their approval (or disapproval) of content.[18] Through these mechanisms, they

[18]Tags are keywords included in the metadata of text that make it easier to search. Like buttons are a small button that allows a user to indicate that they approve or agree with an action or statement by another user. Hashtags are the # symbol followed by a keyword or phrase within a block of text to allow for easier searching and grouping.

self-organize into communities supporting particular ideas. These freely created groups exist and are searchable by entities looking for trends and insights into popular culture. When we have free contributions (e.g., user reviews), costs associated with searching for better information are greatly reduced.[19] This reduction in constraints enables individuals to meaningfully operate in less boundedly rational ways and thereby adopt a classic welfare-maximizing approach to decision making.[20]

At an organizational level, absorptive capacity is understood to characterize an organization's ability to exploit external knowledge as a function of its prior related knowledge and is dependent on the structure of communication between the organization and its environment (Cohen & Levinthal, 1990). In a world of free contributions from individuals and self-organized groups, it is not clear whether the gatekeeper and boundary-spanning roles in traditional R&D organizations (Allen, 1977; Tushman, 1977), which are important for absorptive capacity, maintain the same functions or possibly morph into more of a curatorial or distributor role, managing inputs from the community at large. Although community contributions increase alternatives available to managers and introduce new complexity into the search process, on balance these changes present an enormous opportunity for leaders to make better decisions from better alternatives.

[19]We recognize that these reviews can potentially be manipulated by the organization or individual of focus and thus must be monitored. Nevertheless, these reviews are having sizable impacts across business models and industries and thus are relevant to this discussion.

[20]We acknowledge also that we are assuming individuals can easily process information without bias, but we believe this is a reasonable enough assumption to make this point.

At an organizational level related to search, innovative organizations continually strive to balance the challenges and trade-offs of exploiting existing knowledge while also exploring new opportunities (March, 1991). Within product development particularly, search behavior varies in terms of both how organizations re-use existing knowledge and how widely they look for new knowledge (Katila & Ahuja, 2002). User-generated contributions can apply in modes of both exploitation and exploration. In the exploitation mode, user-generated contributions can extend the reach of an existing product through localization efforts. A specific example is when organizations enable users to localize products for particular markets and then capture these localizations for the benefits of other users, as Facebook does when it relies on users to translate its site into non-English languages. User-generated contributions and developer interactions offer even greater opportunities in an exploration mode because they dramatically increase the available search area. When an organization casts a wide net for user contributions and developer applications, it dramatically increases its ability to explore new alternatives. If managed properly, these contributions allow the organization to gain important insights into how products are used. Further, engaging with users and developers leads to products that better satisfy the needs of users and are therefore more widely adopted.

1.4.7 Intellectual Property

Decreased information constraints, greater engagement with communities, and a shifting locus of innovation lead to strategic considerations regarding how organizations manage IP. When innovation and the accompanying invention were conducted entirely within the boundaries of an organization, the situation was relatively straightforward. Organizations protected IP through legal mechanisms such as patents, trademarks, copyrights, and trade secrets. When they

engaged in interorganization collaboration, they executed appropriate licensing contracts to document ownership and usage rights of the IP created during that relationship.

Organizations, individuals, and groups of users all need to understand IP considerations in a world where organizations regularly solicit inputs and then incorporate these contributions into product offerings (Harhoff, Henkel, & von Hippel, 2003). Beyond determining who owns outputs (which is a challenge in itself), organizations need to be concerned about verifying ownership of inputs. When a user leaves a suggestion on a feedback forum and the organization integrates that suggestion into the next version of a product, does the user have any ownership rights? And, how can the organization be certain that the user did not steal that idea and its implementation from someone else and thus whether the user has the rights to contribute it in the first place? Similarly, when open source software is used to develop proprietary software (e.g., Mac OS X is based on the open source BSD Unix kernel), one must carefully consider how that particular open source license is framed (O'Mahony, 2003). Further, when cloud computing resources are used to develop important innovations, clear ownership agreements with the cloud provider must be in place. The full scope of strategic implications and considerations related to IP in a world of external resources, app developers, and user-generated contributions are well beyond the purview of this chapter. However, it is clear that increases in processing capabilities and reduction in information constraints create novel and complex challenges for IP attorneys and the leaders and individuals with whom they work. They may even call into question the utility of IP laws for spurring innovation (Benkler, 2006; Jaffe & Lerner, 2004).

In summary, while many of the traditional organizational and strategic theories do not necessarily fail as information costs approach zero, several of the assumptions that underlie these theories may no longer apply. Therefore, in all of the areas discussed (boundaries, strategy and new business models, interdependence and community engagement, leadership, identity, search, and IP), research is required to understand how organizations shift strategic visions to account for the reduction in information constraints. However these shifts occur, it is clear that the process of innovation will be significantly altered.

1.5 Impact on Innovation

Scholars often use evolutionary process models, incorporating variation, selection, and retention as lenses through which to view innovation (Campbell, 1960; O'Reilly & Tushman, 2008; Staw, 1990; Tushman & O'Reilly, 1996). We employ this framework to help better understand how the reduction of information constraints affects innovation. Variation is the process through which individuals, organizations, communities, and institutions take existing problems and explore potential solutions through a process of experimentation. In a world without information constraints, the locus of this innovative process shifts from being centered within an organization to more broadly encompassing organizations, individuals, and communities. Selection is the process through which competing alternatives are evaluated and the dominant solution is chosen and brought to market. Finally, although the classic evolutionary view of retention is that of a hereditary process of distributing the selected attributes to the next generation, we instead use the term to mean retention and adoption by the community of users (or potential users). In all three of these stages, dramatic reductions in information processing, storage, and communication costs allow individuals and communities to be more engaged in the

innovation process than previously was possible. In Table 1.3, we compare these three innovation stages in contexts with and without information constraints.

Table 1.3 Innovating With and Without Information Constraints

	With Information Constraints	Without Information Constraints
Variation	• R&D conducted internally and with select partners • Long prototype and pilot cycles • Inputs from internal domain-specific experts • Reseller models do not encourage complementary innovation • Computing tools are expensive and inaccessible	• Organization defines the problem, uses community to help generate possible solutions • Faster experimentation (lean) • Inputs from diverse disciplines (e.g., biologists answering physics problems) • Multi-sided platforms (marketplaces) create opportunities for a large variety of offerings from a community of sources • High-performance tools are available for innovators
Selection	• Management hierarchy decision making • Homogenous perspectives during evaluation • Traditional market research techniques (e.g., focus groups)	• Community-based decision making (or at least input) • Heterogeneous perspectives during evaluation • Online and field-based rapid experimentation
Retention (by Communities)	• Limited and costly communication to potential customers (e.g., traditional advertising) • Complexity in segmenting and targeting customers • Organization/customer relationship ends with product purchase (e.g., brick and mortar checkout) • Slower diffusion and difficult distribution of product offerings	• Easy and inexpensive communication to potential customers (e.g., social media) • Big data enables specific customer targeting • Organization/customer relationship starts with product purchase (e.g., account signup) • Leverage platforms and ecosystems for wide diffusion of new products (e.g., apps)

1.5.1 Variation

In settings both with and without information constraints, the process of variation is a key driver of innovation. Whereas the first movers create the variation via new innovations, all other

organizations must react to the variation. Both must manage the variation as it inevitably affects the status quo. During the variation stage, organizations conduct research and development by searching the existing solution space for a problem, use innovation tools to experiment with possible new solutions, and are open to complementary innovations that add value to the original innovation. However, as we move toward a world without information constraints, all of these activities require more engagement with communities and in some cases may be conducted by communities. Individuals are capable of performing many of these activities on their own when they are armed with the tools enabled by reductions in information constraints.

Previously, most R&D was conducted within an organization that perhaps engaged a few select partners in their innovative efforts. Now, platforms such as TopCoder and InnoCentive allow organizations, and even complex government agencies such as NASA, to focus their efforts on defining problems that are then opened to the community to help generate possible solutions (Lifshitz-Assaf, 2013). This allows organizations to seek inputs from individuals based in diverse disciplines who can engage in out-of-the-box thinking (e.g., a biologist may have the solution to a physics problem).

Powerful new tools, such as cloud computing, allow individual innovators to create solutions that previously could have been developed only within an organization with vast resources. These same tools allow all innovators (organizations, individuals, and communities) to conduct faster experimentation whenever fully detailed prototypes are not necessary to gain accurate measurements of how a product will function or be adopted. Web-based communication tools, including email, mobile phones, and sharing sites (all sometimes gathered under the term

"social media"), are also making it much easier for groups to quickly form and grow and for new types of groups to gather. As Shirky (2008, p. 20) explained in his popular book on self-organization, "We are living in the middle of a remarkable increase in our ability to share, to cooperate with one another, and to take collective action, all outside the framework of traditional institutions and organizations," all of which leads to production of knowledge that organizations can employ in their innovation efforts.

With information constraints dramatically reduced, organizations are changing how they leverage creativity of entities outside their organizations and engender ever greater levels of variation. Open and distributed innovation research provides insights into how organizations manage some of these engagements (Baldwin & von Hippel, 2011; von Hippel, 2009). In related work, the burgeoning literature on multi-sided platform-based businesses and ecosystems provides guidance for how organizations leverage complementors to increase the value of their offerings (Adner & Kapoor, 2010; Eisenmann, Parker, & Van Alstyne, 2011; Zhu & Iansiti, 2012). Although there are numerous types of multi-sided platform business models, they all enable interactions between two or more types of customers (e.g., buyers and sellers) interacting in a market (Hagiu & Wright, 2013). Transitioning to this business model may enable increased variation and better innovative outcomes, yet may also create new challenges for organizations.

1.5.2 Selection

After going through the variation process, in which firms either create or react to a new innovation, an innovating entity must select which version of an innovative solution it wants to bring to market (Lakhani et al., 2013). However, without information constraints, the

organization can engage with external communities to gain important feedback regarding what is most likely to be successful. For example, when a traditional clothing retailer, such as The Gap, must decide which designs to mass manufacture and release to the public, the decision is frequently made by the management hierarchy, with input from consumers, if any exists, filtered via a marketing or market research organization using tools such as focus groups. However, when a firm such as Threadless desires to launch a new product, it has the user community vote directly on competing designs. In this manner, Threadless already has a good sense of a product's potential consumer acceptance and demand before it manufactures the product. Organizations no longer need to rely primarily on traditional market research techniques like focus groups; they can directly engage a large subset of the user community to experiment with reactions to products before making final selections.

Similar to the variation process, engaging communities outside an organization during the selection process allows for heterogeneous perspectives to be sampled before a decision is made. This gives experts in fields outside an organization's core competencies the ability to identify potential challenges the organization might not have considered. These contributors can be professional experts, as when a biologist answers a physics-based problem on a competition website, but they can also be amateurs who have become "experts" with particular products. This often occurs with user-generated reviews: End-user customers contribute to e-commerce websites by posting product reviews, and then other customers vote on the level of helpfulness of the comment. In one recent instance, one of us received a catalog from a mail-order firm highlighting the top ten rated products on the firm's website and offering discounts on those goods. The firm was engaging users to select products on which the firm then offered a

promotional discount through its catalog, which blended the traditionally unidirectional world of mail order catalog merchandising with the digital world of customer ranking and ratings.

1.5.3 Retention (by Communities)

For an innovation to survive, the innovator must ensure that it is retained, diffused, and adopted by the community. The reduction of information constraints has important implications for the diffusion of innovations, which has been an important topic of economic inquiry for many years (see Griliches, 1957, and Rogers, 1962, for early examples and Geroski, 2000, for an overview). The reduction of information constraints speeds communication about new innovations, but this means that organizations have less room for error in early versions of products. Big data and data analytics, enabled by major information cost reductions, allow organizations to mine their existing customers' behaviors to better identify potential early adopters of new products; this can greatly improve the speed with which an innovation diffuses. However, it also causes an organization to increase its engagement with customers after they purchase the product. In many instances with today's online products, the first thing users do when they start to engage with a product is create an account with the organization selling the product. This establishes a link between the organization and the user that represents an ongoing relationship, enabling the user to provide feedback to the organization that can be integrated into the innovation process.

Further, application marketplaces (e.g., Apple's App Store, the Facebook App Center) have large captive audiences that developers want to reach. By using cloud hosting services (e.g., Heroku, Amazon Web Services), which integrate seamlessly with the marketplaces, developers

are able to quickly and widely distribute applications to an audience well beyond what they could reach without such services. Additionally, utilizing cloud computing to host innovative applications allows organizations to experiment and update software-based products without requiring users to download a new version to their desktop after every update.

Importantly, the world without information constraints not only allows for more rapid diffusion of information and physical goods but also allows for some physical goods to diffuse as rapidly as information goods via the invention of 3D printing. 3D printing enables individuals to send digital files of goods rather than sending actual physical goods. Receivers can then print their own versions of a physical good from files they have received. Sending digital information that represents a physical good is much easier (and less expensive) than sending actual goods.

1.6 Future Directions and Research Opportunities

During the time we were writing this chapter, we frequently encountered situations in which we found ourselves thinking, "This is it! This is what we are writing about! This is innovating in a world without information constraints. This is an organization acting differently because information is essentially free." An example occurred while we were researching incumbent organizations engaging with communities. One of us found GE's open innovation call for participation, thought it was well executed, and tweeted the link with reference to the source.[21] Within 15 minutes, much to our surprise, GE tweeted back. That interchange represents exactly the type of organizational change examined herein. A decade ago, this type of interchange could never have happened. In addition to the technological constraints, there were

[21]"Tweeted" in this context refers to posting an update on the twitter.com website to a community of followers.

organizational ones, particularly for large, hierarchical control–centric organizations. Before GE, or any large organization, distributed text publicly, it would need to go through an onerous approval loop. Today, embracing new tools and approaches enabled by reduced information constraints, GE has changed how it engages with the world and is publicly posting multiple tweets per hour, chatting with consumers and potential innovators.

In this chapter, we explored the implications of information processing, storage, and communication costs approaching zero. We showed that the reduction of these costs allows organizations to engage with communities of laborers, developers, and users, and that this engagement leads to shifts in fundamental assumptions of traditional organizational theory. In turn, these organizational shifts lead to new innovation methods. What we see with the simple social media interchange just described, and the phenomena from which it derived, is the instantiation of these shifts.

The changes described herein lead to opportunities for theoretical and empirical research. From a theoretical standpoint, the existing assumptions that many fundamental organizational theories are built upon may no longer be accurate portrayals of a world without information constraints. Although the theories may still be valid, there are open questions as to which of them remain relevant in the modern world. From an empirical standpoint, it is logical to focus on changes to existing business models and development of entirely new ones. Mature organizations are struggling with new levels of interdependency and complexity as they share and engage more broadly and attempt to manage multiple logics simultaneously. Entrepreneurial organizations are emerging with entirely new approaches to managing innovation. These organizations and

institutions are undergoing significant transitions, at multiple levels of analysis, which neither practitioners nor scholars fully understand.

Quantitative and qualitative research methods should be employed to improve our knowledge of these phenomena and their theoretical implications. We see a wealth of research questions related to these studies. In particular, the value of free contributions by users also deserves further research. Is this value accounted for in productivity and growth measurements? Do organizations that utilize such free inputs have higher rates of return than their competitors? What drives users to contribute such free labor? Further, when traditionally product-centric organizations transition to platform-based marketplaces leveraging today's environment with de minimis information constraints, what are the organizational and strategic ramifications? To what extent is organizational identity involved in these types of transitions? Can it help with the transition, or is it always a hindrance? How do organizations that participate in another organization's ecosystem balance their need to differentiate with the requirements of compliance when they are part of a community? These questions stem from the observation that we are living in a world where information is no longer expensive to process, store, or communicate, and this opens a world of innovation opportunities for individuals, organizations, and institutions.

References

Adner, R., & Kapoor, R. (2010). Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strategic Management Journal, 31*(3), 306–333.

Adner, R. (2012). *The wide lens.* New York, NY: Portfolio/Penguin.

Afuah, A. (2003). Redefining firm boundaries in the face of the Internet: Are firms really shrinking? *Academy of Management Review, 28*(1), 34–53.

Aker, J. C. (2010). Information from markets near and far: Mobile phones and agricultural markets in Niger. *American Economic Journal: Applied Economics, 2*(3), 46–59.

Albert, S., & Whetten, D. A. (1985). Organizational identity. *Research in Organizational Behavior, 7,* 263–295.

Allen, T. J. (1977). *Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization.* Cambridge, MA: MIT Press.

Altman, E.J., Nagle, F., & Tushman, M. (2013). Technology and Innovation Management. In R.W. Griffin (Ed.), *Oxford Bibliographies in Management*. New York: Oxford University Press.

Altman, E. J., & Tripsas, M. (2015). Product to platform transitions: Implications of organizational identity. In C. Shalley, M. Hitt, & J. Zhou (Eds.), *Oxford handbook of creativity, innovation, and entrepreneurship: Multilevel linkages.* Oxford, UK: Oxford University Press.

Anderson, J., Reitsma, R., Evans, P. F., & Jaddou, S. (2011). *Understanding online shopper behaviors.* Cambridge, MA: Forrester Research.

Apple Inc. (2013). *Apple's App Store marks historic 50 billionth download.* Retrieved from http://www.apple.com/pr/library/2013/05/16Apples-App-Store-Marks-Historic-50-Billionth-Download.html.Arrow, K. J. (1975). Vertical integration and communication. *The Bell Journal of Economics, 6*(1), 173–183.

Baldwin, C. Y., & Clark, K. B. (2000). *Design rules.* Cambridge, MA: MIT Press.

Baldwin, C. Y., & von Hippel, E. (2011). Modeling a paradigm shift: From producer innovation to user and open collaborative innovation. *Organization Science, 22,* 1399–1417.

Battilana, J., Leca, B., & Boxenbaum, E. (2009). How actors change institutions: Towards a theory of institutional entrepreneurship. *The Academy of Management Annals, 3*(1), 65–107.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom.* New Haven, CT: Yale University Press.

Boudreau, K. (2010). Open platform strategies and innovation: Granting access vs. devolving control. *Management Science, 56*(10), 1849–1872.

Bresnahan, T. F., Greenstein, S. M., & Henderson, R. M. (2011). *Schumpeterian competition and diseconomies of scope: Illustrations from the histories of Microsoft and IBM* (Working paper no. 11-077). Boston, MA: Harvard Business School Strategy Unit.

Brynjolfsson, E., Malone, T. W., Gurbaxani, V., & Kambil, A. (1994). Does information technology lead to smaller firms? *Management Science, 40*(12), 1628–1644.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy.* Lexington, MA: Digital Frontier Press.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review, 67*(6), 380–400.
Casadesus-Masanell, R., & Yoffie, D. B. (2007). Wintel: Cooperation and conflict. *Management Science, 53*(4), 584–598.
Castells, M. (1996). *The rise of the network society.* Cambridge, MA: Blackwell Publishers.
Chandler, A. D. (1962). *Strategy and structure: Chapters in the history of the industrial enterprise.* Cambridge, MA: MIT Press.
Chandler, A. D. (1977). *The visible hand: The managerial revolution in American business.* Cambridge, MA: Belknap.
Chesbrough, H. W. (2003a). Environmental influences upon firm entry into new sub-markets: Evidence from the worldwide hard disk drive industry conditionally. *Research Policy, 32*(4), 659–678.
Chesbrough, H. W. (2003b). *Open innovation: The new imperative for creating and profiting from technology.* Boston, MA: Harvard Business School Press.
Chesbrough, H. W., & Appleyard, M. M. (2007). Open innovation and strategy. *California Management Review, 50*(1), 57–76.
Christensen, C. M. (1993). The rigid disk drive industry: A history of commercial and technological turbulence. *The Business History Review, 67*(4)c 531–588.
Christensen, C. M. (2006). The ongoing process of building a theory of disruption. *Journal of Product Innovation Management, 23*(1), 39–55.
Coase, R. H. (1937). The nature of the firm. *Economica, 4*(16), 386–405.
Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.
Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm.* Englewood Cliffs, NJ: Prentice-Hall.
Dahlander, L., & Gann, D. M. (2010). How open is innovation? *Research Policy, 39*(6), 699–709.
Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *Journal of Retailing, 84*(2), 233–242.
Dutton, J. E., & Dukerich, J. M. (1991). Keeping an eye on the mirror: Image and identity in organizational adaptation. *Academy of Management Journal, 34*(3), 517–554.
Dyer, J. H., & Singh, H. (1998). The relational view: Cooperative strategy and sources of interorganizational competitive advantage. *Academy of Management Review, 23*(4), 660–679.
Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal, 32*(12), 1270–1285.
Facebook. (2013, July 24). *Facebook reports second quarter 2013 results.* Retrieved from http://investor.fb.com/releasedetail.cfm?ReleaseID=780093
Fiol, C. M. (2002). Capitalizing on paradox: The role of language in transforming organizational identities. *Organization Science, 13*(6), 653–666.
Finkelstein, S., & Hambrick, D. C. (1996). *Strategic leadership: Top executives and their effects on organizations.* Minneapolis/St. Paul, MN: West Publishing.
Friedland, R,. & Alford, R. R. (1991). Bringing society back in: Symbols, practices, and institutional contradictions. In W. W. Powell & P. J. DiMaggio (Eds.), *The new*

institutionalism in organizational analysis (pp. 232–263). Chicago, IL: University of Chicago Press.
Gartner Inc. (2013). "Big data." *Gartner IT Glossary.* Retrieved from http://www.gartner.com/it-glossary/big-data/.
Gawer, A., & Phillips, N. (2013). Institutional work as logics shift: The case of Intel's transformation to platform leader. *Organization Studies, 34*(8), 1035–1071.
General Electric Company. (2013). *The Finest Print: GE Challenges Innovators to Design Jet Engine Parts, Print in 3D Complex Healthcare Components.* Retrieved from http://www.gereports.com/post/77131814030/the-finest-print-ge-challenges-innovators-to
Geroski, P. A. (2000). Models of technology diffusion. *Research Policy, 29,* 603–625.
Gioia, D. A. (1998). From individual to organizational identity. In D. A. Whetten & P. C. Godfrey (Eds.), *Identity in organizations: Building theory through conversations* (pp. 17–31). Thousand Oaks, CA: Sage Publications.
Greenstein, S., & Nagle, F. (2014). Digital dark matter and the economic contribution of Apache. *Research Policy*, 43(4), 623-631.
Greenstein, S., & Prince, J. (2007). The diffusion of the Internet and the geography of the digital divide in the United States. In R. Mansell, C. Avgerou, D. Quah, & R. Silverston (Eds*.), Oxford handbook on information and communication technologies.* New York, NY: Oxford University Press.
Greenwood, R., & Suddaby, R. (2006). Institutional entrepreneurship in mature fields: The Big Five accounting firms. *The Academy of Management Journal, 49*(1), 27–48.
Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica, 25*(4), 501–522.
Gulati, R. (1998). Alliances and networks. *Strategic Management Journal, 19*(4), 293–317.
Gulati, R. (2007). *Managing network resources: Alliances, affiliations and other relational assets.* New York, NY: Oxford University Press.
Gulati, R., Puranam, P., & Tushman, M. (2012). Meta-organization design: Rethinking design in interorganizational and community contexts. *Strategic Management Journal, 33*(6), 571–586.
Hagiu, A., & Spulber, D. (2013). First-party content and coordination in two-sided markets. *Management Science, 59*(4), 933–949.
Hagiu, A., & Wright, J. (2013). *Marketplace or reseller?* **(**Working paper no. 13-092). Boston, MA: Harvard Business School Strategy Unit.
Harhoff, D., Henkel, J., & von Hippel, E. (2003). Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. *Research Policy, 32*(10), 1753–1769.
Hart, O., & Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy, 98*(6), 1119–1158.
Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332*(6025), 60–65.
Hitt, L. M. (1999). Information technology and firm boundaries: Evidence from panel data. *Information Systems Research, 10*(2), 134–149.
Ibarra, H. (1999). Provisional selves: Experimenting with image and identity in professional adaptation. *Administrative Science Quarterly, 44*(4), 764–791.

Jaffe, A. B., & Lerner, J. (2004). *Innovation and its discontents: How our broken patent system is endangering innovation and progress, and what to do about it.* Princeton, NJ: Princeton University Press.
Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the South Indian Fisheries Sector. *The Quarterly Journal of Economics, 122*(3), 879–924.
Jensen, M., & Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305–360.
Kaganer, E., Carmel, E., Hirscheim, R., & Olsen, T. (2013). Managing the human cloud. *MIT Sloan Management Review 54*(2), 23–32.
Kaku, M. (2012). *Physics of the future.* New York, NY: DoubleDay.
Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. *The Academy of Management Journal, 45*(6), 1183–1194.
Klein, B., Crawford, R., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics, 21*(2), 297–326.
Koh, H., and Magee, C. L. (2006). A functional approach for studying technological progress: Application to information technology. *Technological Forecasting and Social Change 73,* 1061–1083.
Koomey, J. G. (2008). Worldwide electricity used in data centers. *Environmental Research Letters 3*(3). doi:10.1088/1748-9326-3-3-034008
Kurzweil, R. (1999). *The age of spiritual machines.* New York, NY: Viking Penguin Books.
Lakhani, K. R., & Kanji, Z. (2008). *Threadless: The business of community* (Multimedia/video case 608–707). Boston, MA: Harvard Business School.
Lakhani, K., Lifshitz-Assaf, H., & Tushman, M. L. (2013). Open innovation and organizational boundaries: Task decomposition, knowledge distribution, and the locus of innovation. In A. Grandori (Ed.), *Handbook of economic organization: Integrating economic and organization theory* (pp. 355–382). Cheltenham, UK: Edward Elgar.
Lakhani, K. R., & Von Hippel, E. (2003). How open source software works: "Free" user-to-user assistance. *Research Policy 32*(6), 923–943.
Latif, L. (2013, April 2). AMD claims 20-nm transition signals the end of Moore's Law. *The Inquirer*
Lawrence, T. B., & Suddaby, R. (2006). Institutions and institutional work. In S. R. Clegg, C. Hardy, T. Lawrence, & W. R. Nord (Eds.), *Handbook of organization studies* (2nd ed., pp. 215–254). London, England: Sage Publications.
Lifshitz-Assaf, H. (2013). *From problem solvers to solution seekers: Dismantling Knowledge Boundaries at NASA* (Working paper). Boston, MA: Harvard Business School.
Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office. *Journal of Marketing,* 70(3), 74–89.
Lounsbury, M. 2007. A tale of two cities: Competing logics and practice variation in the professionalizing of mutual funds. *Academy of Management Journal, 50*(2), 289–307.
Luca, M. (2011). *Reviews, reputation, and revenue: The case of Yelp.com* (Working paper no. 12-016). Boston, MA: Harvard Business School.
Maguire, S., Hardy, C., & Lawrence, T. B. (2004). Institutional entrepreneurship in emerging fields: HIV/AIDS treatment advocacy in Canada. *The Academy of Management Journal, 47*(5), 657–679.

Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM*, 30(6): 484-497.
March, J. G., & Simon, H. A. (1958). *Organizations.* New York, NY: Wiley.
March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87.
Maskell, P. (2000). Social capital, innovation, and competitiveness. In S. Baron, J. Field, & T. Schuller (Eds.), *Social capital: Critical perspectives.* Oxford, UK: Oxford University Press.
Merritt, R. (2013, May 23). Broadcom: Time to prepare for the end of Moore's Law. *EE Times*. Retrieved from http://www.eetimes.com/document.asp?doc_id=1263256.
Meyer, J. W. & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology, 83*(2), 340–363.
Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117.
Nagle, F. (2013). *Predicting firm value based on social media sentiment about competitors* (Working paper). Boston, MA: Harvard Business School.
Needleman, S. E. (2010, June 21). Managing at a distance: New websites help managers at small companies keep closer track of their freelancers' work. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/SB10001424052748703862704575100221456493514.
Nielsen, J. (2010). *Nielsen's law of Internet bandwidth.* Retrieved from http://www.nngroup.com/articles/law-of-bandwidth/.
Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide.* Cambridge, UK: Cambridge University Press,
Norton, W. B. (2010). *Internet transit prices—Historical and projected* [Peering White Papers Series]. Retrieved from http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php.
Nuance Communications. (2013). *Second quarter fiscal 2013 earnings announcement* [Prepared conference call remarks], (p. 21).
Okhuysen, G. A., & Bechky, B. A. (2009). Coordination in organizations: An integrative perspective. *The Academy of Management Annals, 3*(1), 463–502.
O'Mahony, S. (2003). Guarding the commons: How community-managed software projects protect their work. *Research Policy, 32,* 1179–1198.
O'Mahony, S., & Ferraro, F. (2007). The emergence of governance in an open source community. *Academy of Management Journal 50*(5), 1079–1106.
O'Mahony, S., & Lakhani, K. R. (2011). Organizations in the shadow of communities. In C. Marquis, M. Lounsbury, & R. Greenwood (Eds.), *Research in the Sociology of Organizations, Vol. 33: Communities and organizations*. Bingley, UK: Emerald Group Publishing.
O'Reilly, C. A., III, & Tushman, M. L. (2008). Ambidexterity as a dynamic capability: Resolving the innovator's dilemma. *Research in Organizational Behavior 28,* 185–206.
Parker, G. G., & Van Alstyne, M. W. (2005). Two-sided network effects: A theory of information product design. *Management Science 51*(10), 1494–1504.
Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: A resource dependence perspective.* New York, NY: Harper & Row.

Piskorski, M. J. (2013). *Social strategy: Why social media platforms work and how to leverage them for competitive advantage.* Princeton, NJ: Princeton University Press.
Piskorski, M. J., & Gorbatai, A. (2013). *Testing Coleman's social-norm enforcement mechanism: Evidence from Wikipedia* (Working paper no. 11-055). Boston, MA: Harvard Business School Strategy Unit.
Prahalad, C. K., & Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review, 68*(3), 79–91.
Rogers, E. M. (1962). *Diffusion of innovations.* New York, NY: The Free Press.
Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287–306.
Rothaermel, F. T. (2001). Incumbent's advantage through exploiting complementary assets via interfirm cooperation. *Strategic Management Journal, 22*(6–7), 687–699.
Santos, F. M., & Eisenhardt, K. M. (2005). Organizational boundaries and theories of organization. *Organization Science, 16*(5), 491–508.
Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science 52*(7), 1000–1014.
Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations.* New York, NY: Penguin Press.
Simchi-Levi, D., Kaminsky, S., & Simchi-Levi, E. (1999). *Designing and managing the supply chain: Concepts, strategies, and cases.* Lexington, MA: McGraw-Hill.
Simon, H. A. (1945/1997). *Administrative behavior: A study of decision-making processes in administrative organization* (4th ed.). New York, NY: The Free Press.
Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.
Smith, W. K., & Tushman, M. L. (2005). Managing strategic contradictions: A top management model for managing innovation streams. *Organization Science, 16*(5), 522–536.
Staw, B. M. (1990). An evolutionary approach to creativity and innovation. In M. A. West & J. L. Farr (Eds.), *Innovation and creativity at work: Psychological and organizational strategies* (pp. 287–308). Oxford, England: John Wiley & Sons.
Steel, E., & Fowler, G. A. (2010, October 18). Facebook in privacy breach: Top-ranked applications transmit personal IDs, a *Journal* investigation finds. *Wall Street Journal*. Retrieved from http://www.wsj.com/articles/SB10001424052702304772804575558484075236968.
Teece, D. J. 1982. Towards an economic theory of the multiproduct firm. *Journal of Economic Behavior and Organization 3*(1), 39–63.
Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal, 28*(13), 1319–1350.
Thompson, J. D. (1967). *Organizations in action: Social science bases of administrative theory.* New York, NY: McGraw-Hill.
Thornton, P. H., & Ocasio, W. (1999). Institutional logics and the historical contingency of power in organizations: Executive succession in the higher education publishing industry, 1958–1990. *American Journal of Sociology, 105*(3), 801–843.
Thornton, P. H., Ocasio, W., & Lounsbury, M. (2012). *The institutional logics perspective: A new approach to culture, structure and process.* Oxford, UK: Oxford University Press.

Trapasso, E., & Vujanic, A. (2010). *Investment in open source software set to rise, Accenture survey finds* [Press release]. Retrieved from http://newsroom.accenture.com/article_display.cfm?article_id=5045.
Tripsas, M. (2009). Technology, identity, and inertia through the lens of "The Digital Photography Company." *Organization Science, 20*(2), 441–460.
Tushman, M. L. (1977). Special boundary roles in the innovation process. *Administrative Science Quarterly, 22*(4), 587–605.
Tushman, M. L., & Nadler, D. A. (1978). Information processing as an integrating concept in organizational design. *Academy of Management Review, 3*(3), 613–624.
Tushman, M. L., & O'Reilly, C. A., III. (1996). Ambidextrous organizations: Managing evolutionary and revolutionary change. *California Management Review, 38*(4), 8–30.
Von Hippel, E. (2009). Democratizing innovation: The evolving phenomenon of user innovation. *International Journal of Innovation Science 1*(1), 29–40.
Warschauer, M. (2003). *Technology and social inclusion: Rethinking the digital divide.* Cambridge, MA: MIT Press.
Welch Allyn. (2011). Welch Allyn introduces High-Performance Healthcare technology at Medica 2011: Medical device manufacturer will showcase its newest line of platform-based connected and configurable solutions, demonstrate telehealth solution with Cisco [Press release]. Retrieved from http://www.businesswire.com/news/home/20111115005064/en/Welch-Allyn-Introduces-High-Performance-Healthcare%E2%84%A2-Technology-Medica#.VI4rWv10w5s.
Welch Allyn. (2013). *Developer information for Device Connectivity Software Developer Kit.* Dusseldorf, Germany: Welch Allyn.
Williamson, O. (1975). *Markets and hierarchies: Analysis and antitrust implications.* New York, NY: Free Press.
Williamson, O. (1981). The economics of organization: The transaction cost approach. *The American Journal of Sociology, 87*(3), 548–577.
Zhu, F., & Iansiti, M. (2012). Entry into platform-based markets. *Strategic Management Journal, 33*(1), 88–106.

Chapter 2: Digital Dark Matter and the Economic Contribution of Apache

Shane Greenstein and Frank Nagle

ABSTRACT

Researchers have long hypothesized that research outputs from government, university, and private company R&D contribute to economic growth, but these contributions may be difficult to measure when they take a non-pecuniary form. The growth of networking devices and the Internet in the 1990s and 2000s magnified these challenges, as illustrated by the deployment of the descendent of the NCSA HTTPd server, otherwise known as Apache. This study asks whether this experience could produce measurement issues in standard productivity analysis, specifically, omission and attribution issues, and, if so, whether the magnitude is large enough to matter. The study develops and analyzes a novel data set consisting of a 1% sample of all outward-facing web servers used in the United States. We find that use of Apache potentially accounts for a mismeasurement of somewhere between $2 billion and $12 billion, which equates to between 1.3 percent and 8.7 percent of the stock of prepackaged software in private fixed investment in the United States and a very high rate of return to the original federal investment in the Internet. We argue that these findings point to a large potential undercounting of the rate of return from IT spillovers from the invention of the Internet. The findings also suggest a large potential undercounting of "digital dark matter" in general.

Keywords: *Open source, Apache, economic measurement, digital economics*

2.1 Introduction

Astrophysicists draw on the term "dark matter" to describe the unseen parts of the universe. Many artifacts, such as the rotational speed of galaxies and gravitational effects, indicate the presence of dark matter, although measuring its existence directly can be difficult. Economists need a similar label for some innovative building blocks of the digital economy that standard tools cannot measure. *Digital dark matter* can serve as the phrase for these digital goods and services that are non-pecuniary and effectively limitless, and serve as inputs into production. They are hybrids of public goods and private investments. This study develops an example that illustrates the potential for the growth and importance of these inputs and their impact. By understanding the value of one specific example of digital dark matter, we aim to better understand the size of the mismeasurement that occurs due to the presence of digital dark matter.

The growth of networking devices and the Internet in the 1990s and 2000s magnified the challenges affiliated with measuring digital dark matter. After decades of development under the auspices of the Department of Defense and the National Science Foundation (NSF), the NSF privatized the Internet backbone in the first half of the 1990s. Software and standards affiliated with operating TCP/IP networks migrated into widespread commercial use. Additionally, in 1991 Tim Berners-Lee made available the basic building blocks of the World Wide Web, supporting its use and development by founding the World Wide Web Consortium in 1994. Its use became common, and formed the basic software infrastructure for a wide range of new forms of electronic commerce and new media.

This study examines one part of these larger events, the deployment of the descendants of the National Center for Supercomputing Applications (NCSA)[22] HTTPd server, today known as Apache. It was one of two notable pieces of NCSA software, the Mosaic browser[23] being the other one. Both inventions moved into widespread use in the middle of the 1990s, continued to evolve thereafter, and subsequently became essential for online commercial activities. Apache's experience deserves academic scrutiny because, in part, it is convenient to examine. Though no publically available data provides a definitive estimate of the size of the Apache economy, it is believed to be the second largest open source project after Linux. It is so large that it has left more observable traces than many other examples of digital dark matter, albeit, such traces are not easy to find.

This study contains two sections. It initially reviews the practices surrounding Apache's deployment, and extends existing measurement theory to this setting, showing how Apache's experience could produce omission and attribution issues. The paper next develops a quantitative approach to address the open question raised by the first section, namely, whether the attribution and measurement issues are large. This study develops a novel dataset, based on a one-percent sample of all "outward facing" web servers used in the United States (we give a more precise definition below). Our quantitative approach using non-proprietary information is an important innovation in this study. The "best" information is collected for private purposes, is closely guarded (Netcraft, 2012), and, in any event, is not publically available for statistical scrutiny by researchers.

[22] The NCSA is one of the four original supercomputing centers funded jointly by the NSF and state governments. It was founded in 1984 to help address the scientific research needs of the future.

[23] Together, the HTTPd server and the Mosaic browser propelled the World Wide Web forward with the HTTPd server acting as a content publisher and the Mosaic browser acting as a content reader.

Using principles of GDP measurement (Nordhaus, 2006), the study estimates the monetary value of the stock of servers. The value is compared to different benchmarks, and we conclude that the estimated value is large. We find that Apache potentially accounts for a mismeasurement of somewhere between $2 billion and $12 billion, which equates to between 1.3 percent and 8.7 percent of the stock of prepackaged software in private fixed investment in the United States. We also provide some arguments for why the estimates should tend towards the higher end of this range. After estimating the value of Apache, we calculate the rate of return for federal investments in the technologies that led to the creation of the Internet. By using our value of Apache as the only output from these investments, we are necessarily underestimating the true rate of return. However, even with this significant underestimation, we still find a rate of return between 10.5% and 19%. We argue that these findings point to a large potential undercounting of the rate or return from research output affiliated with university and federal funding for the Internet.

The study contributes to two literatures. First, it contributes to the underdeveloped literature on measuring the spillovers from the invention of the Internet. Supporters of federal funding for research often cite the Internet as an example of the best-case scenario, presuming that federal funded research led to public goods with large societal benefit (Greenstein, 2011). Despite much broad interest in measuring the economic gains from the invention and deployment of publically funded inventions (See e.g., David, Hall, and Toole, 2000), no estimate exists for the benefits the Internet conferred to the economy. Digital dark matter is principally to blame for this gap in knowledge, as there is little appropriate data for distinguishing the contribution of the Internet from contributions from general advances in ICTs (Greenstein, 2012). This is an unfortunate gap in knowledge considering the research on the origins and

creation of the Internet (Mowery and Simcoe, 2002) and the contribution of all information technology to productivity gains over the last several decades (Brynjolfsson, 1993; Barua, Kriebel, and Mukhopadhyay, 1995; Barua and Byungtae, 1997; and Brynjolfsson and Hitt, 2003). This is also unfortunate in light of the large body of literature that has examined the important contribution of information technology to productivity growth (Jorgenson, Ho, and Stiroh, 2005; Brynjolfsson and Saunders, 2009; and Tambe and Hitt 2012). The gap is also somewhat inconsistent with other evidence indicating the Internet appears responsible for altering the economic landscape in the late 1990s,[24] and contributed to creating new processes in the economy that had long lasting consequences.[25]

We also contribute to an extensive literature on mismeasurement of economic activity and productivity growth (Nordhaus, 2006; Corrado, 2011; Syverson, 2011). Our study contributes to this literature by showing that mismeasurement of Apache has reduced the estimated contribution of IT to productivity growth. For instance, were it measured like other software Apache should be regarded as an important contributor to economic growth, large enough to have merited investing in the research to create it.

These two contributions together focus attention on a larger unaddressed topic. The micro-mechanisms that create measurement issues for economic accounting of open source software are not unique to Apache. They are common to several Internet inventions that diffused

[24] Forman, Goldfarb and Greenstein (2003) estimate that by the year 2000 approximately 88% of US business establishments with over 100 employees had equipment for basic Internet functions, such as email and browsing, while 12% had evidence of upgrades to enhancing their business processes with Internet functionality. In many industries the former was well over 90%, and the latter was well over 20%. Forman, Goldfarb and Greenstein (2012) find evidence that this upgrade in enterprise use of the Internet was affiliated with major changes in the wage structure across the United States.

[25] For example, recent industry assessments estimate that approximately 8% of all retail products sold in the United States are now sold via the Internet (Anderson, Reitsma, Evans, Jaddou, 2011).

into commercial use without formal market transactions and licenses, and where open source institutions supported deployment and use. Other prominent examples from this time period are Linux, software built around TCP/IP, and the World Wide Web (Greenstein, 2010). Further, while Linux and Apache are two of the most recognized open source software projects, there are many others that play an important role in the digital economy but are not accounted for in any productivity measures, such as Perl, PHP, or Firefox, as well as a creative common license in a not-for-profit setting, such as in Wikipedia. While the study offers only a specific estimate of digital dark matter in Apache's case, we think it also illustrates a much broader issue with wide applicability. The study shows why the problem is large in one specific instance, and offers one approach for framing vexing measurement issues in general.

Section 2.2 provides a general framework for thinking about Apache's experience and the affiliated measurement issues. Section 2.3 describes the novel data and calculations that hint at the scale of the mismeasurement. Section 2.4 concludes.

2.2 Digital Dark Matter: Framework

This section discusses the institutional setting that created Apache. It then discusses the omission and attribution issues created for productivity measurement by Apache's widespread diffusion.

2.2.1 Institutional background

Apache descended from software invented at the NCSA at the University of Illinois, which also was the home of the Mosaic browser. Apache arose from server software that worked with Mosaic. It was called the NCSA HTTPd server. This was the most widely used HTTP (Hypertext Transfer Protocol) server software in the research-oriented "early-days" of the

Internet. The server was a collection of technologies that supported browsing and use of Web technologies.

While the University of Illinois successfully licensed the Mosaic browser for millions of dollars,[26] its licensing of the HTTPd server software did not enjoy a similar experience. In part this was because the server software first became available for use as shareware, with the underlying code available to anyone, without restriction. Many Webmasters took advantage of the shareware by adding improvements as needed or by communicating with the lead programmer, Robert McCool. McCool, however, left the University (along with others) to work at Netscape in the middle of 1994, and thereafter webmasters and web participants lost their coordinator.

By early 1995 there were eight distinct versions of the server in widespread use, each with some improvements that the others did not include. These eight teams sought to coordinate further improvements. They combined their efforts, making it easier to share resources, share improvements, and build further improvements on top of the (unified) software. The combination

[26] Notably, the University of Illinois did license the Mosaic browser to a third party, who licensed it to over one hundred other firms, including Microsoft. Netscape never licensed it. Many of the programmers involved in the project left the university in April 1994 and founded Netscape, then got into a dispute with the University over some ownership rights (initially over the ownership of the name "Mosaic"), and they reprogrammed their commercial browser from scratch. They never paid any licensing fees. In its third year Netscape sold over $500 million dollars of software. It is widely agreed that Netscape's entry was a catalyst for Microsoft's accelerated development of a browser. Those events, in conjunction with Apache's diffusion, catalyzed the entry of thousands of new startups in complementary applications. Though there is no doubt that the licensing revenue collected by Mosaic was a tiny fraction of the value created, which is consistent with this study's theme, fully developing that observation would involve a wider array of historical detail and analysis beyond this study's limited scope.

of eight versions was called *Apache* (ostensibly because it was "a patchy web server"[27]), and, informally at first and more formally over time, the group adopted the practices of open source.

As has been documented elsewhere, Apache grew into a very large open source project, widely used in private firms to support electronic commerce.[28] Apache became an essential component in the customer-facing commercial transactions of many firms, as well as in the procurement activities supported by electronic commerce. Further, Apache is used as the base for many other commercial products, such as the IBM HTTP Server, which comes bundled with the IBM WebSphere Application Server. Today it is widely used across the globe, and is regarded as the second most popular open source project used by businesses, after Linux.[29] Additionally, Apache is disproportionately used to host web sites that receive large amounts of traffic. 57% of the million busiest web sites are hosted on Apache. The next closest server is nginx at 15%. [30]

The lack of prices became essential to the operation and success of the project, and, as we show below, this creates potential measurement issues.[31] The absence of pecuniary transactions first arose at the beginning of Apache's existence, when the HTTPd server moved from

[27] In a later interview Brian Behlendorf, one of the founders of Apache, acknowledges the pun, but claims it did not motivate his initial thoughts about naming the project Apache. He states " It just sort of connoted: 'Take no prisoners. Be kind of aggressive and kick some ass.'" McMillan (2000).

[28] The Apache Software Foundation, which was founded to support the Apache HTTPd project, has since created a wide array of other open source projects that add additional unquantified value to the Internet ecosystem. However, the HTTPd project remains the largest project and therefore is the primary focus of our inquiry.

[29] See http://httpd.apache.org/ABOUT_APACHE.html, accessed March 2011, or the similar account in Mockus, Fielding, and Herbsleb (2002).

[30] See the "Market share of the top million busiest sites" section of http://news.netcraft.com/archives/2013/09/05/september-2013-web-server-survey.html.

[31] The Apache Software Foundation argues that the lack of price encourages the commitment of the community, and this community would likely fall apart if its products were not free. "Why Apache Software Is Free," http://httpd.apache.org/ABOUT_APACHE.html (accessed July 11, 2011).

universities to commercial use without formal commercial licenses. It continued as Apache emerged as an open source project based on the HTTPd server, and relied upon donations and a community of users who provided new features for free. As with other open source software, Apache eschews standard marketing/sales activities, instead relying on word-of-mouth and other non-priced communication online. Like other open source organizations, Apache also does not develop large support and maintenance arms for their software, although users do offer free assistance to each other via mailing lists and discussion boards (Lakhani and von Hippel, 2003; West and Lakhani, 2008; Lerner and Schankerman, 2010).

2.2.2 Measuring the gains: Omission

What potential economic measurement issues could result from this invention's deployment? If any major issues arise, they arise from the measurement of the software's contribution to production. Two categories of issues need attention, a problem affiliated with *omission* and another affiliated with *attribution*.

Normal procedures of economic accounting omit Apache as input into production or into stocks of capital. Normal economic measurement focuses on measuring transactions taking place in markets, and presumes that transactions involve a positive price (Nordhaus, 2006). Without explicit attention, normal procedures presume that unpriced activities are nonmarket activities. In sum, like other open source software, the prices and revenue for Apache are zero.

Though open source is *not* singled out as an example by Nordhaus (2006), this setting fits one of the settings he outlines as problematic, namely what Nordhaus labels a "near-market good." He discusses omission errors that arise when standard procedures presume that a zero price is affiliated with non-market activity, but real economic activity creates goods that have a

value, but no price. This setting fits Nordhaus' description in many respects. Creating Apache code relied on the equivalent of donations for support. These may come in the form of explicit donations from firms who provide personnel time and firm capital, or it may come from programmers devoting leisure time to open source activity. It also may come in the form of in-kind or unacknowledged donations of capital or services, such as computer time and hosting facilities. Further, the software also contributes to producing more or better output that may appear unaccounted for.

There are also important differences with the examples discussed in Nordhaus. In this case, some of the activities affiliated with Apache can be measured. Like other widely used open source software, third party firms perform many complementary support functions. This activity typically involves consultants, independent programmers, and providers of bridging software between open source software and commonly used proprietary software.[32] This activity of complementary actors is a key part of the open source ecosystem (West, 2003). Most of that activity will involve market transactions and positive prices. In addition, to obtain service from Apache a firm might have to make considerable investments, using paid personnel, including training personnel to install Apache and conduct ongoing operations, and customizing and adapting Apache to the unique needs of the enterprise. Finally, firms also might purchase

[32] We also note that similar issues pertain to licensed software, though a considerable variety applies there as well. Licensing can be on a per-CPU, per-employee, or per-copy basis. In most other respects, investment activities with personnel and customization and a complementary ecosystem remain the same. A key difference may be the size and operations of the network that has grown up around the standardized commercial software, especially when proprietary firms subsidize those operations with tools and technical support. See Lerner and Schankerman (2010).

hardware for deployment, and potentially additional hardware to accommodate large-scale use.[33] Such expenditure would appear as an operating expense.

We will argue that the presence of open source software, specifically, and digital dark matter, more broadly, raises the potential for attribution and omission biases in productivity analysis. The problem with omission bias is readily transparent. For example, studies that measure the importance of IT to economic growth (e.g. Jorgenson, Ho, and Samuels 2013) could be underestimating the existing stock of IT due to the non-pecuniary nature of digital dark matter. Further, productivity studies that seek to understand the impact of investments in IT on a firm's output (e.g. Brynjolfsson, 1993; Byrne, Oliner, and Sichel, 2013) could be undercounting investments in IT that are unpriced. Our analysis below (Section 2.3.5) shows that Apache alone produces a large omission bias, on the order of billions of dollars. The issues with attribution bias are subtler, however, and merit a deeper discussion.

2.2.3 Measuring the gains: Attribution

To understand the mechanisms behind omission and misattribution, consider the standard productivity model.

Begin with this representation:

$$Y_{it} = A_{it} * f(L_{it}, K_{it}, IT_{it}),$$

[33] While at any point in time there must be a strong association between the number of Apache web servers in use and the number of hardware machines acting as servers, that association does not imply a fixed or constant Leontiff production function over time between the number of Apache servers and the amount of hardware in a firm or industry. There need not be as strong an association between the number of web pages and number of Web servers deployed, for example. One Apache web server can support many web pages, and that has grown over time. In addition, the software improves through software upgrades after new version releases, yielding improvement with no hardware expenditure. Improvement also may arise from better practices at complementary processes within the network, such as mirror servers. Hence, many users have enjoyed functional upgrades without any change in their own hardware.

where Y is output for firm i at time t[34], which results from a production function with arguments for (L) labor, (K) capital stock, and (IT) information technology capital stock, and A is an unmeasured contributor to firm efficiency. In the standard Cobb-Douglas production model this becomes

$$\ln(Y_{it}) = A_{it} + a*\ln(L_{it}) + b*\ln(K_{it}) + g*\ln(IT_{it}).$$

where, typically, the natural log of each side is taken. This results in an equation that can be used for regression estimates. In typical analyses, growth is measured by improvement over time, namely, $Y_{it} - Y_{i,t-1}$, and productivity is measured as multifactor productivity (Corrado, 2011, Syverson, 2011, Byrne et al, 2013). Because usage of open source software by a firm does not have a specific pecuniary measure, there is no mechanism for such usage to enter the equation as an input variable on the right hand side. This results in several possible scenarios of misattribution:

- *Growth without cause.* One scenario for misattribution arises if firms experience growth without hiring more labor, and seemingly without paying for more IT capital or L or K or, for that matter, any visible service. This can happen when Apache code improves and users receive updates at no expense. In this case some firms grow without appearing to change their inputs. Growth will be attributed to A, because of the appearance of more productivity that cannot be attributed to growth in inputs.[35] This scenario resembles a scenario discussed in Syverson (2011), misattribution due

[34] This type of analysis can be implemented at the industry level (Stiroh, 2002), but for simplicity, we carry it through at the firm level.

[35] A similar scenario arises when donations by firms lead to an increase in output prices at many firms. If the price increase eventually leads to an increase in revenue, this would lead to a growth in Y improperly attributed to A.

to externalities from the local environment, which is analogous to firms relying on the quasi-public goods created by the open source community.[36] Syverson argues that the gains could appear to be disembodied technical change, not attributable to any specific input.[37]

- *Growth attributed to the wrong input.* Another scenario for misattribution arises if a large fraction of firms employ Apache software and another fraction makes no investment in Apache, and those investing in Apache invest in labor to support a new release or upgrade.[38] In that case, the firms using open source software will experience an increase in output, Y, and an increase in L. They will show no measured change in IT capital. Non-Apache users do not show any change in Y, L, or IT. Normal productivity analysis will then attribute output growth to the growth of L, even though it is due to increases in unmeasured IT capital.[39]

[36] The mismeasurement is analogous to mismeasuring an improving public good. In her analysis of the various types of protections used in OSS, for example, O'Mahony (2003) highlights this analogy and finds it is an important driver of legal efforts of OSS projects to protect their work.

[37] Or, as in Tambe and Hitt (2012), problems could arise from mismeasurement of labor, which lacks adjustments for human capital affiliated with supporting the software, or for the extent to which labor relies on the community to enhance their productivity. Tambe and Hitt (2012) also points out that measurement error may occur due to the differences between labor-based and capital-based estimates of IT productivity.

[38] Higher labor expenditure could arise either from the need to hire more workers or compensate workers more for their efforts. Though the prevailing view in industry is that open source labor receives higher compensation, there is only limited evidence for this belief. There is some evidence that contributions to open source projects yield increases in pecuniary compensation (see e.g., Hann, Roberts, Slaughter, and Fielding, 2002; Hann, Roberts and Slaughter, 2013). However, the evidence is limited to whether contributors gain monetary rewards, not whether an otherwise equivalent worker gains premiums on their wages for Apache-specific skills in comparison to others. The monetary gains from contributions are consistent with the existence of the premium, but cannot serve as an estimate of its size.

[39] This can cause particular problems in cross-sectional analysis since growth may be measured accurately for some firms and inaccurately for other firms. An interesting variant in this scenario arises from deploying a new web server, which generates purchase of hardware upon which to run Apache. That generates an increase in Y, L and IT among Apache users, but the real measure of IT will be lower than the actual level. The growth will be attributed to

- *Competition between open source and commercial software leading to misattribution:* The third scenario is related to the two scenarios described above. Consider a situation – observed in the data below – where a large fraction of firms invest in Apache software while another large fraction use functionally equivalent software from a commercial firm. Both firms will also invest in more labor, with the firms using Apache software making similar or larger increases in expenditure for labor than those investing in commercial software.[40] All firms experience an increase in Y. Both users experience a growth in L, while the commercial software users experience a larger increase in IT because they paid for the software. Normal productivity analysis will then attribute some part of the growth to L and IT and some growth to A for the firm using Apache.[41]

That explanation also illustrates the omission and attribution problems in tracing the gains to the economy from federally funded research if the gains diffuse into the economy as unpriced inventions, as Apache did. Many of the costs to developing Apache were incurred as part of the research to support the development of the Internet at NCSA. Those were monetary costs and real economic costs. Most of the gains, however, were not recorded – either omitted or

---

both the L and IT. In such an instance, IT expenditure will appear especially productive due to the unmeasured complementary software input.

[40] If the labor for open source software cost the same or less, in addition to open source software costing nothing, and yielded outcomes equivalent to the commercial software, then the commercial software would fade from being used at all. This is not what we observe in the data. Though Apache is the largest service software for Web commerce, functionally-equivalent software from commercial firms has achieved substantial market share, especially from Microsoft. For this situation to be sustainable as market equilibrium, labor expenditure for open source software has to be higher than that for commercial software. A related possibility is general resistance to using open source software or some other distaste for it, or, equivalently, a taste for some attribute affiliated with pecuniary products, which would lead some potential users to pecuniary products for reasons other than labor costs.

[41] An interesting variant arises when Apache labor gets a premium. Then Apache users experience a larger growth in L than commercial software users, but a smaller growth in IT. If most firms are Apache users then standard estimates will attribute much of the gains to L and not enough to IT.

misattributed – because the software took the form of open source, and the code improved without any explicit costs or transactions.

Further, the scenarios above only consider the spillovers from direct usage of Apache as an input into production. They do not account for the spillovers that occur when a competing product, such as Microsoft's Internet Information Services (IIS), add a feature by imitating a similar feature developed for Apache. Nor does this include further gains from enabling the entry of complementary applications.

While the omission and attribution issues discussed above are possible and likely, that does not settle whether they are large and important. The next section addresses the question: Is the evidence about unmeasured value of Apache software large enough to suggest the attribution and measurement issues are important economic issues?

2.3 The shadow value of Apache HTTP Server

To demonstrate the potential impact of digital dark matter, we will calculate the shadow value of the Apache HTTP Server market by considering the price of substituting the non-pecuniary Apache HTTP Server with the pecuniary Microsoft IIS. Although we could have also considered the impact of substituting Microsoft IIS for nginx, the second most popular open source web server, as well, we chose to limit our analysis to only one product, as this adequately illustrates the core point.

2.3.1 The shape of the server economy

Although data on the number of websites hosted via Apache HTTP Server is readily available in a public manner (Netcraft, 2012), data on the number of actual Apache HTTP Servers used is not. Additionally, existing public data does not clearly identify the

location/country for these servers. However, because web servers are primarily used to host public web pages, and are therefore directly reachable via the Internet, we were able to collect information on the number of Apache HTTP Servers used to serve public web pages in the US. Because Apache HTTP Servers can be used internally by organizations, our calculation of the number of Apache HTTP Servers that serve public web pages can be considered a lower bound on the number of actual Apache HTTP Servers in use. Furthermore, a number of different network architectures –load balancing, elastic/cloud computing, and so on – allow for multiple web servers to run on one IP address, which would also lead to our collection method yielding an underestimate of the true number of Apache HTTP Servers.

We first identified the full list of IPv4[42] addresses registered to U.S. organizations. To do this, we utilized information published by the American Registry for Internet Numbers, the organization responsible for managing the distribution of IPv4 addresses in the United States. As of October 15, 2011, there were 1537.37 million IPv4 addresses allocated in the United States. It was too costly to scan every one of these IPv4 addresses, so we took a random sampling of 15,865,522 addresses, which is just over 1 percent of the entire U.S. IPv4 space. For each IPv4 address in our sample, we checked to see if the system was running a web server. If it was, we determined whether the server ran Apache, Microsoft IIS, or anything else including unidentified servers.[43]

[42] IPv4 is version 4 of the Internet Protocol and is currently the most widely used protocol for routing Internet traffic. It is in the process of being replaced by IPv6, but at the time the data was collected all IPv6 addresses also used a backwards-compatible IPv4 address.

[43] The details are straightforward for someone technically skilled in web programming and administration, albeit tedious to report in this context. This method will identify "outward" facing servers, but will systematically undercount any server used entirely for internal purposes. Hence, it is necessarily an underestimate of all Apache HTTP Server software in use. Further details about the process are available from the authors, upon request.

This method will generate a sample of server use and its characteristics, which otherwise is not available. It has one principal drawback. One server may support a large or small number of pages. This method will be proportional to Apache's actual importance in the economy when the size of use is uncorrelated with our measurement strategy (i.e., no selection bias), and our sample size is large. We look for selection issues in the sample, and do not find any symptoms of such issues (Appendix A). This feature of our method also makes us cautious about inference from small sample sizes, as it will be when analysis focuses on narrow geographies or industries.

Of the 15,865,522 addresses in our sample, we found that 195,885 (1.23 percent) were running a web server.[44] Of these 195,885 web servers, 44,211 (22.57 percent) were running Apache and 24,222 (12.37 percent) were running Microsoft IIS.[45] If we extrapolate these numbers to the full U.S. IPv4 space, we estimate that there are 18,981,268 outward-facing web servers in the United States, 4,284,049 of which are running Apache Web Server.[46] Appendix A gives an analysis of the servers in our sample set, including geographic location and top-level domain distribution.

[44] The other 98.77% of the IPs scanned were either inactive or were devices that were not web servers on standard TCP ports.

[45] Apache and IIS account for 34.94% of all web servers in our sample. The remaining web servers were either unable to be correctly identified or were running a different web server such as nginx or a proprietary web server. For example, Google has developed its own internal web server that it uses in place of a publicly available web server.

[46] Continuing this extrapolation to the entire range of IP addresses in the world, of which there are 3.706 billion that are not reserved, there would be 10,288,264 Apache servers in the world. Based on Netcraft's publicly released data on websites, (see news.netcraft.com/archives/2011/12/09/december-2011-web-server-survey.html) that translates into 33 websites per Apache server. This is plausible because the number of web pages per web server must be very skewed. While some Apache servers serve only a single website, many are used by hosting facilities and host hundreds of websites.

2.3.2 Substitution with pecuniary goods

We seek to put a monetary value on the Apache HTTP Server by comparing it with the most widely used proprietary and pecuniary choice. We follow Nordhaus (2006), who states that (p. 146) "…the price of market and nonmarket goods and services should be imputed on the basis of the comparable market goods and services," and (p. 151) valuation "…should rely on available market and behavioral data wherever and whenever possible." At the time of this study a number of proprietary source web servers exist, the most prevalent of which is Microsoft's IIS. IIS's most obvious cost as a substitute for Apache HTTP Server is pecuniary. IIS is shipped for free with Microsoft's Windows Server 2008 operating system, the price of which varies greatly.[47] Appendix B discusses the substitutability of Apache and IIS.

At the time of this study the price for Windows Server 2008 R2 Standard is $1,029 for five licenses, Windows Server 2008 R2 Enterprise is $3,999 for twenty-five licenses, and Windows Server 2008 R2 Datacenter Edition is $2,999 for one license. The most bare-bones version of Windows Server 2008, called the Windows Web Server 2008, is priced at $469. This version of Server 2008 is intended purely for "the development and deployment of Internet-facing Web sites and services."[48] Finally, IIS also comes installed with Windows 7, which can be purchased for as low as $119.99. However, Windows 7 is not designed to be used as a production scale web server and it is unlikely that any company hosting a public website would use this version of Windows.

[47] http://www.microsoft.com/windowsserver2008/en/us/pricing.aspx (accessed July 11, 2011).

[48] http://www.microsoft.com/windowsserver2008/en/us/pricing.aspx (accessed July 11, 2011).

What is a representative price for IIS? We utilize three of the above price points to understand the range of possible prices. On the cheap end, we consider Windows 7, which can cost as low as $119.99, albeit, it also possesses too little functionality to be of practical use. On the high end, we consider Windows Server 2008 R2 Datacenter Edition, which costs $2,999 for one license. Finally, we can consider the bare-bones version Windows Web Server 2008 in the middle,[49] and is currently priced at $469. These three price points allow us to construct a range of possible values for the shadow value of Apache HTTP Server. [50]

With our estimate of the number of Apache Web Servers publically reachable in the United States, we can compute a pecuniary cost of replacing all of these Apache Web Servers with Microsoft IIS. Based on the valuations of Microsoft license fees as mentioned above, the cost of replacing all publically reachable Apache Web Servers in the United States would be between $514 million and $12.8 billion, with a middle estimate of $2 billion.

As previously mentioned, this middle number should be considered a lower bound, because it is based solely on web servers that are attached to the public Internet and does not account for web servers on corporate Intranets or in private use, servers that are behind load balancers or other configurations where multiple servers may exist on one IP address. In addition, the valuation we employ – namely, the present price of IIS – reflects the presence of

[49]We consider Windows Web Server 2008 a comparable match with Apache HTTP Server because it exhibits the closest functionality set and the other versions have additional functions that Apache HTTP Server does not provide. However, it should be noted that most, if not all, of these additional functions can be replicated by free open source software. For example, the operating system functionality is equivalent to the Linux operating system, which is open source and free.

[50] This procedure follows standard GDP measurement principles. It is not a valuation of user gains from employing Apache. Standard revealed preference suggests, for example, that the valuation of IIS by the infra-marginal IIS users would be higher than the market price, and similarly, valuation by infra-marginal Apache users should be higher as well. Conventional GDP measurement does not use consumer surplus. Rather, it uses the marginal valuation.

this differentiated competition. In the absence of any other price, we have to presume that this price reflects the marginal value of the software.[51] Further, although we consider the Windows Web Server 2008 to be the most similar to Apache, Apache is disproportionately used to host the busiest websites (Netcraft, 2012). Hence, there are reasons to think the functionality of Apache tends towards the functionality of the higher end Datacenter version of IIS.

2.3.3 Economic importance of Apache

Is the estimate of the value of Apache a large or a small number? It depends on whether it is compared to sales or investment. First, consider sales. Of the $357 billion (2010 dollars) in software sales by U.S. firms in 2010, $257 billion went to private fixed investment.[52] By this yardstick, the stock of Apache software in the United States is as much as 5 percent (12.8/257) of software sales. However, this compares a stock to a flow, so some readers might consider it like comparing apples with oranges.

Consider a benchmark against investment. Of the $295 billion of software invested in by U.S. firms in 2010, $81 billion (or just over 27 percent) was prepackaged.[53] If that ratio holds for investment stocks, then the stock of prepackaged software in the United States was $146 billion

[51] These prices reflect the current state of the market, where the market leading good (Apache) is unpriced. We do recognize that the presence of differentiated competition lends doubt to the assumption that prices reflect marginal value. Microsoft may not have pricing power in setting the price for IIS. It seems possible and plausible that the price of IIS would be higher if Apache was a priced good. Nonetheless, we follow Nordhaus' dictum to use observed prices, and not counterfactual prices. This is another reason why our calculations could be considered an underestimate of the value of a single Apache server.

[52] "GDP and Final Sales of Software," Bureau of Economic Analysis, http://www.bea.gov/national/info_comm_tech.htm (accessed October 2011).

[53] "Software Investment and Prices, by Type," Bureau of Economic Analysis, http://www.bea.gov/national/info_comm_tech.htm (accessed October 2011). The vast majority of software investment is "custom software" or "own-account," namely, software built by a third party, such as a consultant, or software built by in-house employees.

dollars in 2010.[54] By that yardstick, Apache software is bounded by as much as 8.7 percent (12.8/146) of the measured capital stock of packaged software, or as little as 1.3 percent (2/146).

This illustration suggests the scale of the issue is more than merely a rounding error, particularly when one considers the ubiquity of other widely used, free open source software. We conclude it is likely that the sum of a few of these cases reaches a significant fraction of the total value of the packaged software capital stock and in turn results in a significant impact on overall U.S. GDP.

2.3.4 The economic size of the ecosystem supported

Apache can be viewed through another lens, as a part of the large ecosystem that supports Internet activity. Are our estimates large or small in relation to the value of Internet activities?

Apache is one of several complementary components that together provide Internet services. How important a component is Apache? Consider these comparisons. The size of Internet access revenue in the United States in 2009 (the last year of reliable data) is $59.6 billion,[55] and the size of US online advertising revenue in 2009 is approximately $21 billion.[56]

---

[54] This is 27 percent of the total stock of software in the United States (under nonresidential equipment and software), which was $533 billion in 2010. See "Fixed Assets and Consumer Durable Goods for 1997-2010," http://www.bea.gov/scb/pdf/2011/09%20September/0911_fixed-assets.pdf (accessed October 2011).

[55] 2009 Service Annual Survey Data, Information Sector Services-NAICS 51, does not provide a direct estimate of online access revenue, but it lists four categories of access revenue in four tables: Table 3.3.6. Wired telecommunications carriers (NAICS 5171); Table 3.3.9. Wireless and other telecommunications carriers (NAICS 517212); Table 3.3.12. Cable and other program distribution (NAICS 5175); and 3.4.1. Internet service providers (NAICS 518111), http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

[56] 2009 Service Annual Survey Data, Information Sector Services - NAICS 51, does not provide a direct estimate of online advertising revenue, but it lists three categories in three tables: Table 3.3.5. Internet publishing and broadcasting (NAICS 516); Table 3.4.1. Internet service providers (NAICS 518111); and Table 3.4.2. Web search portals (NAICS 518112). The latter table does not provide an estimate for 2009, but it does provide sufficient data for 2008 and other categories in 2009 to make an educated guess at its level. For the estimate above, that guess was $14 billion. http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

That number combines access revenue from both households and businesses, and it includes $10.1 billion of wireless Internet access revenue. Compared to the revenue it helps produce, the $2 billion to $12 billion of Apache software appears significant.

Now consider another benchmark: the value of Apache in comparison to the size of the software market. The size of system software revenue in 2009 was $48 billion, though personal computer software comprised the largest category, and it was not comparable to Apache. The enterprise and mainframe software revenue together amounted to $26 billion.[57] Against that, the $2 or $12 billion of Apache software appears quite large, albeit, a reader could worry about comparing apples with oranges, once again. This comparison mixes different time scales, as we are comparing sales of one year to replacing the entire stock of Apache.

Of course, neither of these comparisons is precise. With estimates of the replacement cycle for Apache it would be possible to translate the stock into service flows, or their equivalent.

2.3.5 The Rate of Return

Another way to calculate the value of Apache is to place it in a cost benefit framework. We provide two different calculations on the rate of return. In each case we use the same estimate for the cost of creating the Internet, but different methods for estimating the benefits. The second calculation uses productivity estimates, much like those discussed above. Due to data limitations, at best we can make an estimate in the "ballpark." The exercise offers evidence of

[57] 2009 Service Annual Survey Data, Information Sector Services - NAICS 51, Table 3.1.6. Software Publishers (NAICS 5112), http://www.census.gov/services/sas_data.html#NAICS%2048/49 (accessed November 2011).

the large value of Apache specifically, and digital dark matter more broadly. It also provides evidence that is useful for thinking about misattribution.

During NSF's management (approximately 1985–1995) the agency invested $200 million dollars in creating Internet related technology (Leiner, Cerf, Clark, Kahn, Kleinrock, Lynch, Postel, Roberts, Wolff, 2003). That budget covered the costs of the Internet backbone, operational expenses, and the supercomputer centers, of which NCSA was one. Apache originated out of NCSA around 1995, so the rate of return calculations will start counting benefits after 1995.

Before the NSF programs, DARPA funded most of the early invention related to the Internet in the 1970s and early 1980s. While DARPA's financial commitment was considerable, no historian has made a precise estimate of the size of the commitment to networking, which was one of several programs it funded. Government secrecy about DARPA's specific projects and budgets prevents any historian from uncovering further details. Nonetheless, we do know something useful. The entire expenditure for the IPTO, the agency within DARPA that funded most of the Internet, did not exceed approximately $500 million over its entire existence (1963–1986), and the funding for what became the Internet was but one of many IPTO projects (Norberg, O'Neill, Freedman, 1996). To be conservative we add another $200 million to the costs of creating the Internet, which is much more than likely.

This calculation includes the most direct costs for creating the Internet, and excludes numerous other costs for research. For example, this estimate of costs does not include a range of other experiments in computer science that NSF paid for (out of different budgets) and from

which the general community of researchers learned. We are comfortable with this because it provides a number that is likely more than the actual amount of the direct costs.

That sets up the first cost-benefit calculation. Above we estimated that Apache is worth between 2 and 12.2 billion dollars in 2012, seventeen years later. As noted above we consider 2 and 12.2 to be implausibly low and high. So for this cost/benefit calculation we consider two more plausible levels, $7.1B and $10B, where the former is the midpoint between the two numbers and the latter is as high as we consider plausible. What constant rate of growth would generate such levels of benefit after $400M investment seventeen years earlier? Using the current dollar numbers, the former generates a rate of growth of 17% and the latter 19%. While inflation (which averaged 2-3% a year in this period) would diminish some of that gain, it is quite high for only one output from NCSA, and there were many more benefits than just this one.

A second cost-benefit calculation examines the scale of economic growth attributable to Apache, using productivity gains from investment in software, much as discussed above. For such a calculation we use the model of Byrne, Oliner and Sichel (2013) (hereafter BOS), because it distinguishes between the gains to labor productivity from distinct IT inputs in the US economy – IT hardware, software and communications. Using this model we ask the question: if additional software were added to the capital stock of software, how much economic growth would it generate?[58] In each year it generates more income, and that continues to accumulate after 1995. We consider the total through 2012. This estimate must make assumptions about the rate of growth of Apache, which we do not observe directly.

[58] Specifically, this uses Equation 1 in BOS, and the estimates in Table 1. The simulation asks what would happen if Apache had pecuniary value.

To illustrate how this calculation works we begin with one year, the estimate for 2012. In 2012 BOS estimate that software accounts for 0.16 of labor productivity growth (out of 1.56 total labor productivity growth per year). Software accounts for only 3.75% of the income share in that year, but software makes a big contribution to productivity growth in comparison to the size of software in use.[59] Our low and high estimates for the value of Apache would place it at 1.3% and 1.8% of the stock of software in 2012.[60] While the increase in labor productivity due to Apache has to be comparatively small, the US economy is so large that even a small improvement can yield a substantial economic gain. In this case, we assume that labor does not increase, but only software does, and simulate how much GDP would grow with more software. This model estimates that additional software stock would generate a low/high estimate of $1.0B and $1.8B in additional income in 2012.[61]

Apache also generated income in all the years between 1995 and 2012. For those estimates we must make an assumption about the rate of growth in Apache servers over this period, which we did not observe directly. We make an estimate from public data on the growth of web pages supported by Apache over this time period.[62] It shows web pages grow at one rate over the 1990s, and then grow at a slower rate in the last decade, with Apache generally supporting 60% of all web pages. This history suggests that a constant rate of growth over all

[59] As shown in Equation 1 in BOS, each component's contribution to labor productivity arises from multiplication of the income share and the growth in that input. Our simulation is equivalent to asking how the contribution would change if the income share for software increased without any increase in labor. Thanks to Dan Sichel for patiently walking us through the steps of the calculation so it remained consistent with the estimates in Table 1 from BOS.

[60] This uses the same source of data, as noted above.

[61] This estimate follows BOS and estimates the labor productivity gain on non-farm private income.

[62] See Netcraft.com

seventeen years is probably slower than the true rate of growth, so we make the assumption of linear growth in order to be conservative.

Next we estimate the incremental contribution of Apache for each year, just as we did for 2012, but now we ask a slightly different question. A rate of return calculation addresses the question: "How much economic activity did the investment in the Internet generate by 2012?" Such a calculation requires aggregating all future benefits into the same dollar units. We assume a 10% discount rate on the future from 1995, so we can add up the contemporary dollars.[63] This assumption will weight short-term gains against those that come many years later, as Table 2.1 shows (comparing line A to line C, or comparing line B to line D). Then we add up those gains, and calculate that Apache generated economic activity equivalent to between $2.6B and $4.5B by the end of 2012. On an investment of $400M, that is a rate of return between 10.5% and 14%. Table 2.1 shows the components that went into that calculation. Once again, that is quite high for only one output from NCSA.

There are important qualifications to this second set of estimates. First, the estimates in BOS also examine the effect of productivity gains in the production of IT, as distinct from capital deepening, but we did not use those in our calculation. If Apache generated productivity gains to production, the above estimates would not capture those gains.[64]

Second, no reader should quote the rate return on Apache as precisely 10-14%. The estimate depends on assumptions, such as of linear growth and 10% discounting. A small

[63] Note that this calculation takes all values in contemporary values and discounts from that value by 10% per year, blending all price level corrections and forecasting into one value in 1995 terms.

[64] As noted by the literature, to the extent that open source investments were capitalized into the private value of firms, some of this would have been accounted for in other capital. The BOS estimates do not use intangible capital as another form of capital, however, so this is not a concern.

Table 2.1 Contribution of Apache to GDP, simulation, Billions of Dollars

Year	1996	1997	1998	1999	2000	2001	2002	2003
A	0.04	0.08	0.13	0.19	0.25	0.30	0.36	0.43
B	0.07	0.14	0.23	0.32	0.43	0.53	0.63	0.75
C	0.03	0.10	0.20	0.32	0.46	0.63	0.80	0.99
D	0.06	0.17	0.34	0.55	0.80	1.08	1.38	1.71
2004	2005	2006	2007	2008	2009	2010	2011	2012
0.52	0.50	0.58	0.65	0.71	0.74	0.82	0.93	1.04
0.90	0.86	1.00	1.13	1.23	1.28	1.42	1.61	1.80
1.19	1.36	1.54	1.73	1.91	2.08	2.25	2.42	2.59
2.06	2.36	2.67	2.99	3.30	3.59	3.89	4.18	4.48

A: Additional GDP in that year, low estimate, contemporary dollars
B: Additional GDP in that year, high estimate, contemporary dollars
C: Accumulation of GDP up until that year, low estimate, discounted 10% per year
D: Accumulation of GDP up until that year, high estimate, discounted 10% per year
Source: Author's calculation, based on Byrne, Oliner, Sichel (2013). See text.

increase in the rate of discount would mildly lower the rate of return, as would later growth in Apache. Similarly, a lower rate of discount would mildly increase the rate of return, as would sooner growth in Apache. So why make such an estimate? These estimates provide a good sense of the scale of the gains, and small changes in those assumptions do not alter the quality of the answer. Anything in this range has to yield a large rate of return, which is our point.

Third, these are estimates on a counter-factual in order to illustrate the scale of importance of additional software, had it been accounted for like any other asset. Because it is a small change in the value of assets it is possible to approximate its effect with this simulation. Were Apache measured properly the other estimates of the contribution of other IT capital would change as well. In that light, this exercise yields one other insight. Because Apache is such a small percentage of total capital, the attribution biases associated with mismeasuring this asset, therefore, appear to be small. This implies that Apache alone, as one piece of open source software, produces large omission bias but does not produce attribution bias. That also suggests

that proper measurement of open source software would produce omission bias, but it leaves open the question of whether it would produce a major attribution bias.

We would stress that our estimates are necessarily below the true value. Despite this underestimate, these numbers, as well as the estimates from the first cost-benefit analysis, indicate that the return on investment for Apache was quite high. Since this is only one program, this leads us to conclude that the returns from federal R&D invested in the Internet must be underestimated.

2.4 Concluding thoughts and future research

In this study we argued that digital dark matter is an important issue to consider in the online economy. Like other private assets, digital dark matter acts at times like an input into the production of a pecuniary good, and regular investment extends functionality or delays obsolescence. Like a public good, more than one user can employ digital dark matter nonexclusively. In contrast to many private assets or public goods, something other than market prices shapes the extent of investment and use. Finally, even when visible, digital dark matter is measured indirectly at best. Omission and attribution errors are possible, even likely.

We illustrated these observations by focusing on one prominent case, Apache, which is a key piece of software in the operation of the Internet. We argued that Apache contributes value to the online economy, and that this value could be quite large, and that it is not currently captured through standard GDP measurement. We find evidence that the omission biases are significant, but the attribution biases are not. Our estimates also imply that, were we to add additional open source software, we could reach a significant fraction of the total value of

packaged software sales. Once again, we conclude that this evidence suggests that it is likely that open source software significantly contributes to omission bias.

These findings point to a large potential undercounting of "digital dark matter" and related IT spillovers from university and federal funding. Apache's experience focuses attention on a broader set of open source software projects, such as Linux, the software built around IETF standards, the World Wide Web, PERL, or a creative common license in a not-for-profit setting, such as Wikipedia. Every project took a distinct institutional form, but shares similar potential for omission and attribution errors.

While open source software is certainly an important piece of digital dark matter, we speculate that similar concerns about measurement may arise in other activities where digital goods and services are non-pecuniary, effectively limitless, and serve as inputs into production. For example, user contributed content powers websites as diverse as Twitter, Yelp, and YouTube, but these free "inputs" from users go unmeasured by standard productivity measurement. As another example, digitized blueprints, many of which are non-pecuniary, have become widely available for 3D printing, and as that activity grows, these prints will contribute to production, despite their lack of price.

We speculate that the effect of omission biases are likely to increase as information costs approach zero and firms rely more on non-pecuniary digital inputs from communities of users and developers (Altman, Nagle, and Tushman, 2013). Although such quantification may be difficult to attain directly, we have shown that indirect methods of estimating this value are possible. More precise and broad-based estimates may be used to create GDP calculations that more accurately reflect the true production of the U.S. economy, resulting in policies that are

more suited to the reality of the online economy. We foresee such studies shedding light on the measurement of the gains from research and development in universities that diffused into commercial use as part of open source software and in many other ways. Such quantification may also lead to a better understanding of the impact of free and open source software on the economy as a whole.

These concerns lead to a number of open questions. If the undercounting of digital dark matter leads to mismeasurement of productivity, does it also lead to underinvestment – both public and private - in projects that create digital dark matter? Would demand for digital dark matter products decrease significantly if they were pecuniary? We also wonder how digital dark matter shapes a variety of online activities where these and related products are common, such as online news, entertainment, scientific inquiry, educational and reference activities, and business operations.

References

Altman, Elizabeth, Frank Nagle, and Michael Tushman. 2013. "Innovation without Information Constraints: Organization, Communities, and Innovation When Information Costs Approach Zero," in Oxford *Handbook of Creativity, Innovation, and Entrepreneurship*, edited by Michael Hitt, Christina Shalley, and Jing Zhou. Oxford University Press.

Anderson, Jacqueline, Reineke Reitsma, Patti Freeman Evans, and Samantha Jaddou. 2011. Understanding Online Shopper Behaviors. *Forrester Research*.

Barua, Anitesh, Charles H. Kriebel, and Tridas Mukhopadhyay. 1995. Information technologies and business value: An analytic and empirical investigation. *Information Systems Research* 6(1): 3-23.

Barua, Anitesh, and Byungtae Lee. 1997. The information technology productivity paradox revisited: A theoretical and empirical investigation in the manufacturing sector. *International Journal of Flexible Manufacturing Systems* 9(2): 145-166.

Brynjolfsson, Erik. 1993. The productivity paradox of information technology. *Communications of the ACM* 36 (12): 66–77.

Brynjolfsson, Erik, and Lorin M. Hitt. 2003. Computing productivity: Firm-level evidence. *Review of Economics and Statistics* 85(4): 793-808.

Brynjolfsson, Erik, and Adam Saunders. 2009. *Wired for innovation: how information technology is reshaping the economy.* MIT Press.

Byrne, David, Stephen Oliner, and Daniel Sichel. 2013. Is the Information Technology Revolution Over? *International Productivity Monitor, Centre for the Study of Living Standards,* Vol 25, pp 20-36. Spring.

Corrado, Carol, 2011. Communications Capital, Metcalfe's Law, and U.S. Productivity Growth, The Conference Board, EPWP #11-01. March.

David, P.A., Bronwyn H. Hall, and Andrew A. Toole. 2000. Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy* 29 (4-5): 497-529.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2003. Which Industries use the Internet? in (ed) Michael Baye, *Organizing the New Industrial Economy,* Elsevier. Pages 47-72.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2012. The Internet and Local Wages: A Puzzle, *American Economic Review.* February. 102(1), pages 556-575.

Greenstein, Shane. 2010. Innovative Conduct in U.S. Commercial Computing and Internet Markets. in *Handbook on the Economics of Innovation*, edited by Bronwyn Hall and Nathan Rosenberg. Burlington: Academic Press. Pp. 477-538.

Greenstein, Shane. 2011. Nurturing the Accumulation of Innovations: Lessons from the Internet. in *Accelerating Innovations in Energy. Insights from Multiple Sectors,* edited by Rebecca Henderson and Richard Newell. Chicago: University of Chicago Press. Pp. 189-224.

Greenstein, Shane, 2012. The absence of data for measuring the economic impact of IT in the US. in *Regulation and Performance of Communications and Information Networks,* Edited by Gary Madden, Gerry Faulhaber, and Jeffery Petchey, Edward Elgar Press; Cheltenham, UK. Pp. 328-344.

Hann, I., Roberts, J., Slaughter, S. 2013. All Are Not Equal: An Examination of the Economic Returns to Different Forms of Participation in Open Source Software Communities. *Information Systems Research,* April 2013.

Hann, I., Roberts, J., Slaughter, S., and R. Fielding. 2002. Delayed returns to open source participation: An empirical analysis of the Apache HTTP Server Project. Working Paper.
Jorgenson, Dale, Mun Ho, and Jon Samuels. 2013. Economic Growth in the Information Age: A Prototype Industry-Level Production Account for the United States, 1947-2010. Working Paper.
Jorgenson, Dale, Mun Ho, and Kevin Stiroh. 2005. *Information Technology and the American Growth Resurgence*. MIT Press; Cambridge, MA.
Lakhani, K., and E. von Hippel. 2003. How open source software works: "Free" user-to-user assistance. *Research Policy* 32 (6): 923-943.
Leiner, Barry, Vint Cerf, David Clark, Robert Kahn, Leonard Kleinrock, Daniel Lynch, Jon Postel, Larry Roberts, Stephen Wolff, 2003, A Brief History of the Internet, Version 3.32. Reston Virginia, The Internet Society, http://www.isoc.org/internet/history/brief.shtml.
Lerner, J., and M. Schankerman. 2010. *The comingled code, open source and economic development*. Cambridge, MA: MIT Press.
McMillan, Robert. 2000. Apache Power. *Linux Magazine.* April 15. http://www.linux-mag.com/id/472/, accessed September, 2013.
Mockus, A., Fielding, R. T., and J. D. Herbsleb. 2002. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology* 11 (3): 309-346.
Mowery, D. and T. Simcoe. 2002. Is the Internet a US Invention? An Economic and Technological History of Computer Networking. *Research Policy*, 31(8-9): 1369-87.
Netcraft. 2012. June 2012 Web Server Survey. *Netcraft*. Retrieved from http://news.netcraft.com/archives/2012/07/03/june-2012-web-server-survey.html, accessed September, 2013
Norberg, Arthur, Judy O'Neill, and Kerry Freedman. 1996. *Transforming Computer Technology, Information Processing for the Pentagon, 1962-1986.* Baltimore, MD., Johns Hopkins University Press.
Nordhaus, William D., 2006, "Principles of National Accounting for Nonmarket Accounts," in editors, Dale W. Jorgenson, J. Steven Landefeld, and William D. Nordhaus, *A new Architecture for the US National Accounts,* University of Chicago Press.
O'Mahony, S. 2003. Guarding the commons: how community managed software projects protect their work. *Research Policy* 32 (7) 1179-1198.
Stiroh, Kevin J., 2002. "Information Technology and the U.S. Productivity Revival: What do the Industry Data Say?" *American Economic Review,* 92 (5), pp 1559-1576. December.
Syverson, C. 2011. What Determines Productivity? *Journal of Economic Literature* 49(2): 326-365.
Tambe, P., and L. M. Hitt. 2012. The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research* 23(3-1): 599-617.
West, Joel. 2003. How open is open enough? Melding proprietary and open source platform strategies. *Research Policy* 32 (7): 1259-1285.
West, J., and K. R. Lakhani. 2008. Getting clear about communities in open innovation. *Industry & Innovation* 15 (2): 223-231.

Chapter 3: Crowdsourced Digital Goods and Firm Productivity: Evidence from Open Source Software

Frank Nagle

ABSTRACT

As crowdsourced digital goods become more widely available and more frequently used as key inputs by firms, understanding the impact they have on productivity becomes of critical importance. This study measures the firm-level productivity impact of one such good, non-pecuniary (free) open source software (OSS). The results show a positive and significant return to the usage of non-pecuniary OSS that has gone unmeasured in prior studies of the economics of IT and is not solely due to cost savings. The study addresses the endogeneity issues inherent in productivity studies by using inverse probability weighting, an instrumental variable approach, firm fixed effects, and data on management quality from the World Management Survey to add support for a causal interpretation. Across firms, a 1% increase in the amount of non-pecuniary OSS used by a firm leads to a .073% increase in productivity. This translates to a $1.35 million increase in value-added production for the average firm in the sample. This is more than double the magnitude of the coefficient on investments in traditional pecuniary IT capital. This effect is greater for larger firms and for firms in the services industry. These findings suggest that firms willing to take on the risks associated with non-pecuniary OSS reap benefits from collective intelligence and labor spillovers. Further, the results indicate that existing studies underestimate the amount of IT used at the firm.

Keywords: *productivity of information technology, open source software, user innovation, crowdsourced digital goods*

3.1 Introduction

As the digital age progresses, information goods are easier and easier to reproduce at costs that are rapidly approaching zero. Coupled with decreases in communication costs, this has made it easier for groups of individuals, frequently referred to as the crowd, to produce digital goods that are freely distributed to users who do not pay a monetary price. Wikipedia, the online crowdsourced encyclopedia, is a frequently cited example of this phenomenon, although there are many other examples including open source software (OSS), crowdsourced innovation tournaments, and the digitization of consumers' opinions via online review sites and social media. The same information cost decreases that enable the production of these goods also enable firms to use these crowdsourced goods as inputs into production. Recent research has shown that firms are increasingly relying on these types of goods to drive innovation and production (Baldwin and von Hippel, 2011; Lakhani, Lifshitz-Assaf, and Tushman, 2012; Corrado and Hulten, 2013; Altman, Nagle, and Tushman, 2014).

This trend is also widely discussed in the popular press as technology giants like Apple, Google, and Facebook increase their reliance on crowdsourced digital goods to grow their innovative and productive efforts (Sorkin and Peters, 2006; Asay, 2013; Finley, 2013). However, it is not only technology focused companies that are relying on the crowd - Ford, Pepsi, Walmart, and a host of other well-known non-IT brands use free inputs from the crowd to help drive their bottom line (Horovitz, 2013; McCue, 2013; Phipps, 2014). Additionally, these same crowd-based technologies are allowing small start-ups to have a large impact, even when they are capital constrained, due to a reliance on free crowdsourced digital goods as inputs. OSS, the empirical focus of this study, is a particularly important example of a crowdsourced digital good

as more than 50% of firms now use or contribute to OSS (Black Duck, 2014) and billions of venture capital dollars are pouring into the OSS ecosystem (Black Duck, 2014; Forrest, 2014; Hamilton, 2014; Lunden, 2014). Further, due to the rise of mobile operating systems such as Android and iOS, more than 50% of all computing devices are now based on OSS (Yarow, 2013).

Despite the growing importance of crowdsourced digital goods as inputs into production, measuring the value they help create can be difficult. In a classic Schumpeterian creative destruction process (Schumpeter, 1942), these new goods destroy old business models while creating new opportunities for growth. For example, the introduction of Wikipedia destroyed much of the market for pecuniary encyclopedias (both paper and digital). At the same time, Wikipedia has provided great societal value. However, as with all crowdsourced digital goods, this value is difficult to measure for two primary reasons. First, because these goods are frequently free, standard productivity measures, which rely on price to reflect value, do not properly capture these increasingly critical inputs. Second, because such goods are often distributed under licenses that allow for unlimited copying, it is unknown exactly how widespread they are. Despite the increasing prominence of crowdsourcing, these measurement challenges have prevented researchers from analyzing how its impact varies across different firms and market environments. Further, it has been suggested that integrating such resources into the firms production process can be more costly than comparable non-crowdsourced inputs (Giera and Brown, 2004), and consequently their use could have a negative impact on productivity. Therefore, the goal of this paper is to answer the following question: what is the impact of non-pecuniary crowdsourced digital goods on firm productivity? After answering this

broad question, the paper seeks to answer the related question: What are the firm-level determinants of the productivity impact of such goods?

As the production, and productive use, of such goods increases, the answer to these questions becomes more interesting and more important. Recent research has shown that the increased use of unpriced goods of both a digital (Greenstein and Nagle, 2014) and non-digital (Bridgman, 2013) nature may be an important factor in understanding recent trends in Gross Domestic Product (GDP). Non-pecuniary digital goods can cause standard GDP measures to greatly underestimate the true productivity of a nation and its firms. These same mismeasurement issues can lead firms and managers to underestimate the importance of including crowdsourced digital goods as key inputs into their productive and innovative processes. While some leading firms, like Google and Facebook, have embraced the crowd and the free labor and content it provides, others have shied away from relying on such inputs due to concerns about reliability, sharing with competitors, and the costs of restructuring business models to add the user directly into the production and innovation process.

In addition to productivity-related implications, the reliance on, and contribution to, crowdsourced goods also has implications for firm competitive strategy. In a world where a firm must rely on actors outside of its boundaries for valuable inputs, and at the same time must consider contributing internally developed code to the world, co-opetition (Brandenburger and Nalebuff, 1996; Afuah, 2000) becomes an increasingly important concept. As firms' competitors increase their reliance on crowdsourced digital goods, understanding how these goods contribute to productivity and for what types of firms they are the most useful becomes increasingly

important to allow managers to make the right decisions regarding the crowd. Finally, understanding the productive implications of free digital goods scratches the surface of the broader issue of all digital goods, which essentially have a marginal cost of zero, and are therefore likely priced below their actual value.

To understand how usage of such non-pecuniary digital inputs affects firm productivity, this paper first discusses why such goods could have a positive or negative impact on productivity and then considers what firm characteristics are likely to determine the degree of this impact. To test the resultant competing hypotheses, it utilizes a dataset that measures the usage of one particularly important non-pecuniary crowdsourced digital good, open source software (OSS) operating systems. OSS is an important digital good that is produced by a community of tens of thousands of users and is frequently distributed free of charge. Thus it is exactly the type of non-pecuniary digital input that is uncounted in GDP and other productivity measures. This data is combined with firm financial data and productivity measures to allow for the application of a classic Cobb-Douglas production function analysis to understand the role of non-pecuniary IT inputs in firm-level productivity. This is a standard methodology for estimating the value of IT (Brynjolfsson and Hitt, 1996; Dewan and Min, 1997; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013), although non-pecuniary OSS is normally not accounted for in such frameworks. Due to sample selection and endogeneity concerns, inverse probability weighting, a method similar to propensity score matching, is used to construct a setting more like that of an experiment. Panel fixed effects and instrumental variables are also utilized to allow for a more causal interpretation of the results. Further, for a sub-sample of the firms, data from the World Management Survey (Bloom, Sadun, and Van Reenen, 2012) is used

to show that there is no correlation between firm management quality and the use of non-pecuniary OSS, indicating that the full sample results are not proxying for management quality.

The results show that firms that use non-pecuniary OSS have higher levels of productivity than those that do not. They also show that increased usage of non-pecuniary OSS has a positive and significant impact on firm productivity. This makes intuitive sense since firms that use non-pecuniary IT are able to tap into the collective intelligence of the crowd through spillovers from free labor. The primary effect is robust to various endogeneity concerns, allowing for a causal interpretation of the results. The estimates indicate that a 1% increase in the amount of non-pecuniary OSS used by a firm leads to a .073% increase in productivity when comparing firms against other firms. The average value added for the firms in the sample is $1.846 billion; this indicates that a 1% increase in the number of non-pecuniary OSS operating systems leads to a $1.35 million increase in value-added production (or profits) for the average firm. This effect size is more than double the size of the coefficient on traditional pecuniary IT capital. This effect is greater for larger firms and for firms in the services sector (versus those in the manufacturing sector). The main effect is of a similar order of magnitude as other IT-related inputs. Because the study measures only non-pecuniary OSS operating systems, it does not capture other firm investments in non-pecuniary OSS, thus the main effect is likely a lower bound for the true effect of all non-pecuniary OSS on productivity. Further, the results indicate that it is not only the lack of cost of such software that provides a benefit to the firm. Indeed, if the non-pecuniary OSS were assigned a cost similar to that of other pecuniary operating systems, it would still have a significant positive effect. Finally, the results indicate that current studies underestimate the amount of IT at the firm.

This paper seeks to add insights to two important bodies of literature: the user innovation literature and the returns to IT literature. The user innovation literature (e.g., von Hippel, 1986, Chatterji and Fabrizio, 2014), in particular that which is centered on OSS (e.g., Kogut and Metiu, 2001; Lerner and Tirole, 2002; Lakhani and von Hippel, 2003; West and Lakhani, 2008), focuses primarily on supply side questions, e.g. why do individuals and firms contribute time and resources to the development of OSS, with almost no literature focusing on the demand and usage side of the OSS market. At the same time, the literature on the returns to IT investment (e.g., Brynjolfsson and Hitt, 1996; Tambe and Hitt, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013) focuses almost exclusively on IT investments of a pecuniary nature, completely missing investments in non-pecuniary IT, such as OSS. This paper contributes to both of these bodies of work by filling these important gaps in the literature and shedding light on the underestimation of IT used by the firm, and therefore the underestimation of the productivity impact of non-pecuniary IT. Understanding the impact of such goods on firm productivity not only helps to contribute to the broad literature on the determinants of productivity[65], but also shows that user innovation is no longer a rare phenomenon and is becoming a key input into firm productivity and innovation. Additionally, the paper offers insights for practitioners that can be utilized to increase the profitability of the firm's operations and gain competitive advantage by using crowdsourced goods as inputs. Finally, for policy makers, the results encourage policies that incentivize production of public digital goods as a method for increasing firm and, in turn, national productivity.

This paper is laid out as follows. Section 3.2 discusses the existing gap in the user innovation

[65] See Syverson, 2011 for an over view of this literature.

and productivity of IT literatures and then presents a brief history of OSS operating systems. Section 3.3 develops competing hypotheses about whether the use of non-pecuniary OSS has a positive or negative impact on firm productivity, and considers firm characteristics that are determinants of this effect. Section 3.4 constructs the models used in the estimation and discusses strategies for dealing with the sample selection and endogeneity issues in the study. Section 3.5 details the dataset on OSS usage and firm production and the construction of the main variables. Section 3.6 presents the results and discusses their implications, and Section 3.7 concludes.

3.2 Crowdsourced Digital Goods and the Returns to Information Technology

One of the oldest and most successful crowdsourced digital goods is open source software and this will be the empirical setting of this analysis. Therefore, this section first reviews prior research on crowdsourced digital goods and user innovation as well as research on the returns to IT investments. In doing so, an important gap is identified at the intersection of these two literatures, motivating the primary research question. Then, this section gives a brief history of the development of the two most widely used OSS operating system, GNU/Linux and BSD, both of which play an integral part in today's modern IT ecosystem.

3.2.1 Free and Open Source Software as an Input into Productivity

As early as the 1980's, production by users has been a topic of interest in the management field (von Hippel, 1986). While such production is by no means limited to the digital world, it is here that user innovation is frequently studied, primarily in the realm of OSS. However, most of the academic work on OSS has been focused on exploring supply side mechanisms – why do

users contribute to OSS (Benkler, 2002; Lerner and Tirole, 2002; West and Lakhani, 2008, Athey and Ellison, 2014), how do users join OSS projects (von Krogh, Spaeth, and Lakhani, 2003), how do users help each other contribute to OSS (Lakhani and von Hippel, 2003), and how do OSS communities organize to protect their intellectual property (O'Mahony, 2003) and to guard against free-riding (Baldwin and Clark, 2006). Research on the supply side has also been extended to better understand why firms release some of their proprietary code as OSS (Harhoff, Henkel, and von Hippel, 2003; von Hippel and von Krogh, 2003; Lerner, Pathak, and Tirole, 2006; Henkel, 2006; Fosfuri, Giarratana, and Luzzi, 2008; Lerner and Schankerman, 2010; Casadesus-Masanell and Llanes, 2011). Despite the abundance of literature on the supply side of OSS, there is almost no literature on the demand side of OSS[66] – who uses it, why do they use it, and are there productivity benefits to using it remain unanswered questions. This is despite the fact that OSS, and – more broadly – non-pecuniary, community-based user-production, has been identified as an increasingly important input into the business models of firms in both academic literature (Krishnamurthy, 2005; Baldwin and von Hippel, 2011; Lakhani, Lifshitz-Assaf, and Tushman, 2012; Altman, Nagle, and Tushman, 2014; Greenstein and Nagle, 2014) and popular literature (Howe, 2008; Shirky, 2008).

Although the productivity related value of OSS usage has not been directly investigated, there is a significant body of literature examining the impact of IT usage on productivity at both the firm and country levels. This literature has shown that the rate of return for investments in IT is positive and significant (Brynjolfsson and Hitt, 1996; Athey and Stern, 2002) and productivity boosts from investments in IT are frequently mistaken for intangible firm-specific benefits

[66] The one notable exception is Lerner and Schankerman (2010), which explores the cross-country differences in demand for OSS usage. However, their analysis does not examine the returns to OSS usage and does not include the US.

(Brynjolfsson, Hitt, and Yang, 2002; Syverson, 2011; Tambe, Hitt, and Brynjolfsson, 2011). Studies have also shown that IT-producing and using industries contributed a disproportionately large amount to the economic growth experienced in the US, particularly from 1995-2004 (Jorgenson, 2001; Jorgenson, Ho, and Stiroh, 2005). In addition to spending on IT capital, spending on IT labor has also been found to boost firm productivity (Tambe and Hitt, 2012). Further, participation in networks of practice adds IT related knowledge spillovers that increase productivity (Huang, Ceccagnoli, Forman, and Wu, 2013). However, it has been found that not all firms receive the same return on IT investment (Aral and Weill, 2007) and that the returns to IT investment are not as strong as they once were (Byrne, Oliner, Sichel, 2013). An important aspect of all such studies is that they measure IT investment via dollars spent on software, hardware, labor, or a combination of the three. Since most OSS does not have a price directly associated with it,[67] it is not properly factored into such calculations. This mismeasurement of "digital dark matter" has been shown to be on the order of billions of dollars for one piece of OSS in the US alone (Greenstein and Nagle, 2014) and the inclusion of intangibles[68] and non-pecuniary production have been shown to significantly alter GDP calculations (Corrado, Hulten, and Sichel, 2009; Bridgman, 2013). Because of this measurement issue, OSS is not properly included in current productivity calculations, and therefore the productive value of OSS is currently unknown.

Despite the vast literatures that exist in these two areas, there is a noticeable dearth of

[67] Although some literature exists analyzing the total-cost of ownership (TCO) when comparing open and closed source software (e.g., MacCormack, 2003; Varian and Shapiro, 2003; Russo et al, 2005; Wheeler, 2005; Fitzgerald, 2006), a consensus has not been reached and this literature does not explore the productivity implications of the two types of software, just the costs of employing it. The analysis in this study will control for the costs of employing either type of software by including labor and capital costs in the analysis. This allows for the measurement of the impact of the software itself even though the TCO question is not directly addressed.

[68] Intangible assets include intellectual property, user-generated content, organizational capital, and human capital.

literature that addresses the intersection, leaving an open question this paper attempts to answer: What is the impact of OSS on firm productivity? After establishing a baseline answer to this question, the paper further considers the firm-level differences in extracting productivity value from OSS, allowing for a better understanding of the productivity implications of non-pecuniary crowdsourced digital goods.

3.2.2 Institutional Context: The Free and Open Source Software Movement

Although the concept of free and open source software developed as part of the early computer culture, it was not formalized until 1983 when Richard Stallman founded the GNU Project[69] to create a computer operating system that gave users the freedom to share and modify the software, unlike the predominant operating system at the time, UNIX, which was proprietary and closed-source software. Two years later, Stallman founded the Free Software Foundation (FSF), a non-profit organization designed to encourage the creation and dissemination of software with unrestrictive licenses, including the GNU General Public License (GPL), which continues to be the most widely used software license for free software. The FSF emphasizes that it uses the word "free" to mean "liberty, not price", encapsulated in the pithy slogan "free as in free speech, not as in free beer."[70] However, the software released under this license is frequently also offered at a price of zero. This ambiguity later led to Eric Raymond's call for the use of the term "open source" instead of "free" (Raymond, 1998).

As the GNU Project progressed, it was successful in creating most of the middle and upper layers (user interface) of the operating system. However, very little work had been finished for

[69] GNU is a recursive acronym for "GNU's Not UNIX".

[70] http://www.gnu.org/philosophy/free-sw.html, retrieved on February 23, 2014.

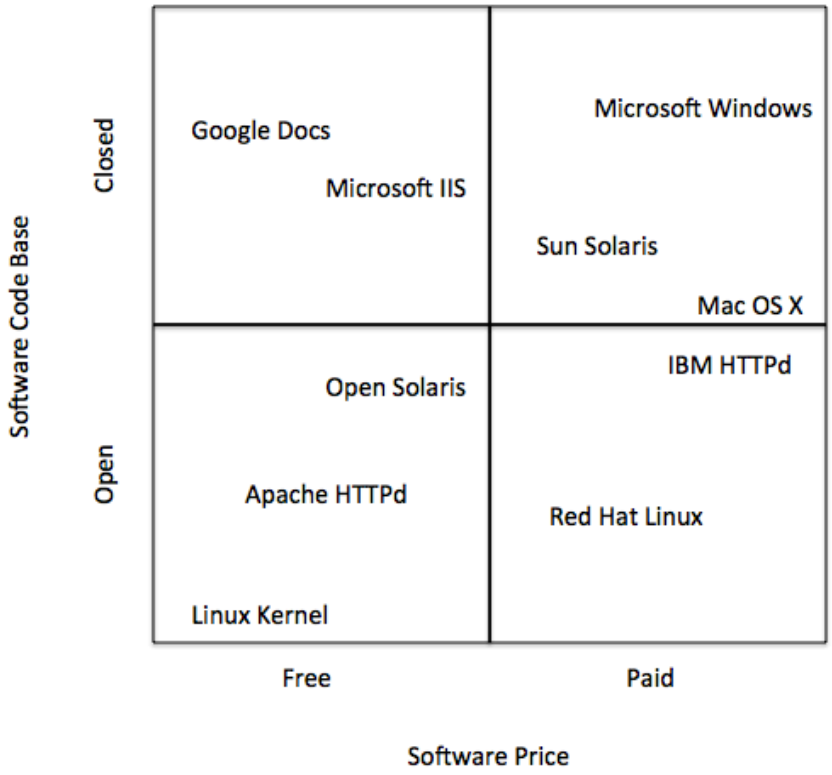
the lowest layers, known as the kernel, of the operating system. In 1991, Linus Torvalds released the Linux kernel to take the place of the incomplete GNU kernel. GNU developers rapidly latched on to the Linux kernel and the combination of the Linux kernel and GNU software on top of it became the basis for most free and open source operating systems in use today. The other main free and open source operating system is the Berkeley Software Distribution (BSD) operating system, which was initially proprietary until a variant of version 4.3 was released as open source in 1989 under the terms of the BSD License, which allowed for redistribution provided the BSD License was included. Both GNU/Linux and BSD rely on a community of mostly unpaid contributors to maintain and upgrade the code base.[71] From 2005 to 2013, nearly 10,000 developers contributed to the Linux Kernel (Corbet, Kroah-Hartman, McPherson, 2013). From 1993-2014, FreeBSD, one of the largest BSD distributions, had nearly 1,000 core developers and nearly 3,000 contributors (FreeBSD, 2014).

Since these early operating systems were released, there has been a flood of free and open source software projects that are either a variant of these operating systems or are applications that run on top of them, such as the vast array of projects maintained by the Apache Software Foundation. Although unrestricted non-pecuniary software is at the core of the free and open source software movement, many companies have structured profitable business models on top of this software. Common examples include Red Hat, which offers its own Linux distribution and charges for customer support, the IBM HTTP Server, which is built on the open source Apache HTTP Server and is included with the IBM WebSphere Application Server, and Apple's

[71] Although historically such OSS projects relied primarily on unpaid contributors, larger projects are increasingly receiving contributions from coders who are paid by their company to contribute to the code base. However, from the perspective of the OSS project, these contributions are unpaid since the project does not pay the coders directly. Further, during the timeframe of the empirical setting in this paper, widespread corporate contributions to OSS were limited.

Mac OS X, which is built on the FreeBSD operating system. Figure 3.1 gives various examples of operating systems and software that fall on different dimensions of price and the openness of the code base.

Figure 3.1 Examples of Software on the Free/Open Spectrum



3.3 Theory and Hypothesis Development

As shown in Figure 3.1, when a firm considers a software investment, it must make decisions along two important dimensions: price and whether the code base is open or closed. Compared to closed and pecuniary software, using free and open source software can be risky, but it can also provide a number of additional benefits. This section discusses these risks and benefits and develops competing hypotheses about the baseline productivity impact of using non-

pecuniary OSS as well as further hypotheses about the characteristics of the firm that moderate the main effect.

3.3.1 Risks of Using Non-Pecuniary OSS

Compared to pecuniary and closed source software, non-pecuniary OSS can be a risky investment. This section discusses the largest of these risks, including the fact that free software is not costless, there is no guaranteed technical support or technical path, OSS has security concerns not present in closed source software, and there is no contractual relationship allowing for recourse if something goes wrong.

When considering implementing new software, the allure of "free" software can be great for any capital constrained firm. However, firms run the risk of assuming that implementing such software will be costless. The price of the software itself does not truly represent the total cost of ownership (TCO) of the investment. Indeed, although there is a diversity of opinions, the consensus in the literature on the TCO of software is that the actual cost for software is negligible when compared to the hardware and labor costs of implementing, using, and maintaining it (e.g., Varian and Shapiro, 2003; Russo et al, 2005; Wheeler, 2005; Fitzgerald, 2006). In a review of the literature on TCO, MacCormack (2003) finds that the one fact most TCO studies can agree on is that the purchase price of a piece of software represents less than 10% of all of the costs that go into using that software. Therefore, one of the most salient benefits of non-pecuniary OSS, may actually be misleading and may lead to long-term costs that are 5% to 20% higher than those of proprietary closed-source software (Giera and Brown, 2004).

In addition to the direct monetary costs of supporting it, non-pecuniary OSS[72] is often seen as riskier than pecuniary software for a number of reasons. First, because a collective of users, rather than a central producer, creates non-pecuniary OSS, there is rarely official technical support for the products. While some users do offer help by creating manuals or answering user questions (Lakhani and von Hippel, 2003), there is no guarantee that a user's question will ever be answered because they do not have a service agreement with any vendor (Woods and Guliani, 2005). Relatedly, although larger OSS foundations, like the Linux Foundation and the Apache Foundation, employ commons-based governance structures (Ostrom, 1990; O'Mahony and Ferraro, 2007), there is no guarantee that the OSS project will be continuously developed and supported. Likewise, even if the project is continuously maintained, there is no guarantee about the features and technical path of future versions (Kogut and Metiu, 2001).

From a security standpoint, the openness of the underlying code in OSS allows anyone to examine it for security vulnerabilities. Although Linus's Law[73] would predict that the open nature of the code would be a benefit from a security perspective, recent widespread vulnerabilities in OSS integral to the operation of the Internet and Linux have shown that these bugs are not always caught early in the development process.[74] Perhaps the most concerning risk of all is the lack of a contractual relationship between a firm using non-pecuniary OSS and any

[72] The focus of this research is primarily on non-pecuniary OSS. The availability of pecuniary products, like Red Hat Linux, which build on non-pecuniary OSS is important, but the risks associated with these products is lower due to the contractual relationship a customer has with the vendor, which greatly mitigates these risks.

[73] Linus's Law is attributed to Eric Raymond (1999), but named after the founder of Linux, Linus Torvalds. Linus's Law states "Given enough eyeballs, all bugs are shallow," which implies that the more people who look at the code, the more likely bugs are to be found and fixed.

[74] The Heartbleed security bug was introduced into the OpenSSL cryptography library in December 2011, and was not noticed and fixed until April 2014. As of May 8, 2014, more than 300,000 public web servers were still vulnerable to the issue (Graham, 2014). The Shellshock security bug was introduced into the Bash Shell in 1992, and was not noticed and fixed until September 2014. The Bash Shell is used in nearly all Unix-style operating systems, including Linux and BSD, the latter of which is the basis of the Mac OS X operating system.

one entity responsible for the development of such software, which leaves the firm with no one to sue when something goes wrong. There are no service level agreements (SLAs) for non-pecuniary OSS, which means the use of such software is riskier than pecuniary software where such agreements exist.

The view of non-pecuniary OSS as a risky decision led to the commonly used phrase "No one ever got fired for buying Microsoft."[75] This phrase became popular in the technology industry as customers were increasingly willing to pay a premium for software from big name firms they could trust. In aggregate, the various risks laid out above could have a negative impact on the productivity of the firm. Formally,

H1a: The usage of non-pecuniary OSS at a firm has a negative impact on firm productivity.

3.3.2 Benefits of Using Non-Pecuniary OSS

Despite all of the risks discussed above, non-pecuniary OSS can also provide a number of benefits to the firms willing to take on these risks. These benefits include reduced upfront costs, collective intelligence of the crowd, and greater flexibility to alter and enhance the code base.

The most salient benefit of using non-pecuniary OSS is the free nature of the software. Although, as discussed above, the actual cost of software is minimal compared to the costs of implementing, the fact remains that firms using non-pecuniary OSS are paying less for their

[75] This phrase actually started about IBM in the 1970's, long before OSS. However, it was ported to Microsoft in the 1990's as OSS started to gain traction in the marketplace. Interestingly, IBM later invested heavily in OSS and built some of its products on top of OSS. However, IBM but offered large support contracts and SLAs, removing many of the risks associated with the use of non-pecuniary OSS.

software than their competitors using pecuniary software. However, since this cost reduction is rather small, if there is a measurable positive effect of non-pecuniary OSS on firm productivity, it is likely that the free nature of the software is not the only mechanism driving this effect.

Beyond being free, the crowdsourced nature of non-pecuniary OSS can have an important effect on the quality of software development. A pithy quote from the technology industry helps to illuminate this potential benefit of non-pecuniary OSS – "No matter who you are, most of the smartest people work for someone else." This quote, known as Joy's Law, highlights the fact that regardless of how big and powerful a company is, it can never hire all of the best and brightest people.[76] This is the modern-day interpretation of earlier arguments by von Hayek (1945), who pointed out that knowledge is distributed throughout society and cannot be fully aggregated in one central body. In the software development world, this means that code developed within a closed firm cannot benefit from the intelligence of anyone outside of the firm (Kogut and Metiu, 2001; von Hippel and von Krogh, 2003). Non-pecuniary OSS projects address this problem by allowing anyone to contribute to the development of the underlying code base. Indeed, as mentioned above, nearly 10,000 individuals contribute to the Linux kernel, while less than 1,000 individuals contributed to all of Windows 7 (Schofield, 2008), and only one team of less than 40 people created the Windows 8 kernel (Sinofsky, 2011). Therefore, the use of OSS allows a firm to harness the labor efforts of a wide collective of individuals. Further, as individuals' motives for contributing are primarily intrinsic (Lerner and Tirole, 2002), any benefits by firms using the software can be seen as positive externalities via spillovers from the labor contributions of the crowd.

[76] This statement is from a speech Bill Joy, the co-founder of Sun Microsystems, gave in 1990, and was first mentioned in print by Gilder (1995).

Although collective intelligence and the wisdom of crowds is often associated with completing simple problems, recent research has shown that the crowd can also be successful in solving more complex problems (Woolley et al, 2010; Woolley and Fuchs, 2011; Yi et al, 2012), including software development (von Hippel and von Krogh, 2003). Further, collective intelligence represents an important mechanism for enhancing the knowledge inputs of the firm, which have been shown to contribute to productivity (Hulten, 2010).

The open nature of non-pecuniary OSS has the added benefit of allowing firms to avoid hold-up problems. If a firm relies on closed or pecuniary software built on OSS, it cannot control the path of development and is therefore subject to hold-up by the developer. However, if a firm relies on non-pecuniary OSS and they need a specific function, they can contribute the code themselves (Schwarz and Takhteyev, 2011). This flexibility allows for the firm to more efficiently use its software once it is deployed within the enterprise (Woods and Guliani, 2005).

Like many investment opportunities a firm must make, the decision to invest in non-pecuniary OSS allows firms that are willing to take on higher levels of risk to obtain higher levels of reward. For many firms, the risks of relying on non-pecuniary OSS are too high and they therefore rely on pecuniary software. However, the firms that are willing to take on the risks associated with non-pecuniary OSS allows them to obtain the benefits of tapping into the collective intelligence of the crowd, leading to productivity spillovers from the free external

labor and knowledge[77] that support the non-pecuniary OSS ecosystem as well as the more flexible nature of OSS. Therefore, firms that use non-pecuniary OSS should obtain a net positive effect on productivity:

H1b: An increase in the amount of non-pecuniary OSS used at a firm has a positive impact on firm productivity.

3.3.3 Moderating Effect of Firm Size

Due to differences in capital constraints, it is likely that firm size will play a role in determining the productive impact of non-pecuniary OSS. For very small firms, non-pecuniary OSS can play a critical role in allowing the IT capability of the firm to ramp up quickly, without expensive outlays for pecuniary software. However, as firms grow, it is likely they will not be able to fully support a non-pecuniary OSS infrastructure themselves, and will therefore rely on external consulting firms to take the place of the support that comes with pecuniary software. On the other hand, larger firms have the capacity for greater economies of scale[78] and can therefore obtain greater returns from their IT investments as well as any consulting activities to help implement an OSS infrastructure. Together, this implies a U-shaped relationship between firm size and productivity returns to non-pecuniary OSS that is high for very small firms, drops for medium sized firms, and increases for larger firms. Due to data restrictions and the sample only consisting of public firms, it is only possible to test the latter portion of this relationship and the former is therefore left for future research. This leads to the following formal hypothesis:

[77] While it is true that some firms who use non-pecuniary OSS also contribute back to the creation of these products, even these firms benefit from the external labor contributed by other firms and individuals, which they do not pay for. A deeper analysis of this relationship is left for future research.

[78] There may be a concern that if larger firms disproportionately use non-pecuniary OSS, then the use of OSS could simply be proxying for economies of scale. However, it is possible to control for firm size when estimating the effect of OSS on productivity. Controlling for this effect should allow for it to be ruled out as an alternative explanation to the main effect of non-pecuniary OSS.

H2: For public firms, the productivity impact of non-pecuniary OSS is more positive (less negative) for larger firms than for smaller firms.

3.3.4 Moderating Effect of Industry

IT related inputs frequently require higher levels of human capital for value extraction. This is especially the case for software that is not supported by a vendor, as is the case with non-pecuniary OSS. Accordingly, prior research (Dewan and Min, 1997; Huang, Ceccagnoli, Forman, and Wu, 2013) has shown that the output elasticity of IT is lower in firms that are in the less human capital intensive manufacturing sector compared to those that in the services sector. Since non-pecuniary OSS is an important piece of the IT ecosystem, this relationship should hold for it as well.

H3: Compared to firms in the manufacturing sector, firms in the services sector will obtain higher (less negative) returns from the use of non-pecuniary OSS.

3.3.5 Additional Moderating Effects

Although some research has speculated a labor-premium for IT workers who understand OSS, this has not yet been shown to be true in all cases.[79] However, since OSS is less frequently used than pecuniary software, the skills to operate and maintain OSS are more niche. Therefore, it is possible that IT workers who are capable of operating and maintaining OSS are of a higher quality than those who are not. Were this true, then the presence of OSS would indicate higher quality labor, which would result in additional productivity as an indirect consequence of the use of OSS. However, estimating this effect is difficult due to the misattribution issues associated

[79] Hann et al. (2002) and Hann, Roberts, and Slaughter (2013) show that not all participants in OSS receive higher wages in their jobs, but they do find that OSS contributors with managerial responsibilities in the OSS community receive up to an 18% increase in wages.

with non-pecuniary IT investments (Greenstein and Nagle, 2014). For example, comparing the elasticity of labor to productivity for firms who use OSS to those who do not may result in a higher return to IT labor for firms using OSS. However, these results would be observationally equivalent to the results if misattribution was the cause because the misattribution discussed above could result in the same shift in elasticity, but for a different reason (namely that the OSS is unaccounted for). To properly disentangle these effects, detailed data on IT labor inputs would be necessary. Such data is not currently available. Therefore, it is not possible test for this effect in the current setting.

Likewise, if non-pecuniary OSS were of a higher quality than its pecuniary counterpart, then firms using OSS would gain an increase in productivity due to the difference in quality of inputs. However, this too is difficult to disentangle from the misattribution effect. If this effect were driving the increase in productivity, comparing the elasticity of IT-software capital between firms who do and do not use OSS would again be observationally equivalent to the case where the misattributed value of OSS increases the coefficient for IT capital. Therefore, testing this relationship is left for future research.

3.4. Empirical Methodology

This section describes the empirical methodology employed to test the hypotheses developed above. First, it describes the estimation model, which is consistent with other models of the productivity of IT, but accounts for non-pecuniary digital inputs. Then, it discusses identification concerns due to sample selection and endogeneity as well as the methodologies employed to address these concerns. These methods include inverse probability weighting, instrumental

variables, and firm fixed effects.

3.4.1 Estimation Models

The dataset will measure capital, labor, and various IT inputs. Before describing this data in detail, it is useful to review the model and estimation approach of the paper. In the economics of IT literature, the standard method of estimation is the classic Cobb-Douglas Production function modified to include IT (Brynjolfsson and Hitt, 1996; Dewan and Min, 1997; Tambe and Hitt 2012; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013):

$$Y_{it} = {K_{it}}^{\alpha} {L_{it}}^{\beta} IT_{it}^{\gamma} A_{it} \quad (1)$$

where Y_{it} is the production of firm i in time t, ${K_{it}}^{\alpha}$ is the amount of non-IT capital stock, and ${L_{it}}^{\beta}$ is the amount of non-IT labor. IT_{it}^{γ} is the amount of IT capital stock and A_{it} is a firm-specific efficiency multiplier that captures intangible assets such as management skill or institutional knowledge and learning. In earlier literature, IT capital and IT labor have been combined into a single variable; however, more recent literature has shown a differing effect of these two inputs (Tambe and Hitt 2012). Therefore, the primary specification separates the two, but a robustness check is performed with them combined.

$$Y_{it} = {K_{it}}^{\alpha} {L_{it}}^{\beta} ITK_{it}^{\gamma_1} ITL_{it}^{\gamma_2} A_{it} \quad (2)$$

Value-added productivity (VA_{it}) is substituted for sales as a measure of output to remove concerns about trends in the economy or demand shocks (Brynjolfsson and Hitt 2003) and then the log of each side is taken to obtain:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln ITK_{it} + \gamma_2 \ln ITL_{it} + \varepsilon_{it} \quad (3)$$

Taking the natural log of each side results in coefficients that are equivalent to a firm's output elasticity to a given input. This allows for an interpretation of the coefficients as the percentage change in VA_{it} for a one percent change in the value of the given input. Unobserved differences in firm-level efficiency are captured in the error term. This baseline model is consistent with the most current total-factor productivity models of productivity measurement that account for IT usage (e.g., Tambe and Hitt 2012; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013). However, all of these models rely on the assumption that the price of the inputs reveals their importance into production. For example, one-hour of labor that costs $15 will have less of an effect on output than one-hour of labor that costs $20. What such models cannot account for is when the value of an input is priced at $0 (such as non-pecuniary OSS). Such an input is essentially uncounted in such models and can lead to misattribution of production at the macro-level in a variety of ways (Greenstein and Nagle, 2014). To account for this properly, a measure of a firm's utilization of non-pecuniary open source software, $non\_pecuniary\_OSS_{it}$, in a given period is added to the specification. Non-pecuniary OSS must be separated from pecuniary OSS because the latter is already measured by current productivity methods since it has a price.[80] The measurement of non-pecuniary OSS is described in the data section below. To allow for consistent interpretation, the natural log of this measure is used. This results in the following equation:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln ITK_{it} + \gamma_2 \ln ITL_{it} + \gamma_3 \ln non\_pecuniary\_OSS_{it} + \varepsilon_{it} \quad (4)$$

[80] As mentioned above, an important aspect of the OSS movement is the ability to build pecuniary software on top of non-pecuniary OSS. For example, Red Hat Enterprise Linux is built on the open source Linux kernel, but is not free due to the additional functionality and support Red Hat provides. Conversely, a product like Mandrake Linux is both open source and non-pecuniary. Therefore, pecuniary OSS is considered differently than non-pecuniary OSS.

Using equation 4 as the preferred estimation equation, an estimate of the impact of non-pecuniary OSS usage can be obtained.

3.4.2 Identification Strategy

In an ideal experiment, one would randomly assign firms from the full population of US firms to use or not use non-pecuniary OSS at varying levels of intensity. However, such an experiment is infeasible and therefore observational data, discussed in the next section, is used. Like all studies of the impact of IT on productivity using observational data, this analysis is subject to both sample selection bias and endogeneity. Sample selection is a potential threat to identification due to the fact that the dataset (discussed below) undersamples firms that use non-pecuniary OSS. This could result in incorrect estimation of coefficients for the population. A second threat to identification is the fact that firms endogenously decide whether or not to use non-pecuniary OSS. If firms that are, for example, better managed are both more likely to use non-pecuniary OSS and have higher levels of productivity, then the relationship between non-pecuniary OSS and productivity could not be interpreted as causal due to simultaneity bias. Further, this could lead to an incorrect estimation of the size of the effect. Both of these concerns prevent a complete answer to the primary question that can be used to make recommendations to managers. Additionally, to understand the determinants of how OSS impacts productivity, a believable baseline must be established. Therefore, the paper employs a number of methods that help to address both of these concerns.

Inverse-Probability Weighting

First, inverse-probability weighting (IPW) (Horvitz and Thompson, 1952) is utilized to

address the issue of sample selection bias. This increases the consistency of the estimator (Wooldridge, 2007) in a manner similar to Heckman correction (Heckman, 1976, 1979), but with fewer assumptions (Wooldridge, 2002; Young and Johnson, 2009). This is necessary because the dataset (discussed below) undersamples firms that use OSS, which can adversely affect the estimation procedure. IPW also helps address endogeneity concerns and allows for the results to be interpreted as causal, in a manner similar to matching, by balancing the dataset between treatment and control groups to identify the direct effect of the independent variable (Hirano, Imbens, and Ridder, 2003; Hogan and Lancaster, 2004; Cole and Hernan, 2008; Huber, 2013).

IPW is similar to propensity score matching, but allows for full use of all existing observations. This makes IPW more efficient than matching, which drops observations that do not have a close match. The first step is to predict the propensity of a firm to adopt non-pecuniary OSS based on observables. To do this, a Probit function is used to predict the likelihood of treatment (adoption of non-pecuniary OSS) based on observables. In addition to the four primary input variables ($ITK_{it}, ITL_{it}, K_{it}, L_{it}$), the model also uses two constructed variables estimating the number of pecuniary OSS operating systems and closed source operating systems at the firm ($pecuniary\_OSS_{it}$ and $closed_{it}$). These additional variables help to account for the amount of other operating systems used by the firm, which could be an important predictor of non-pecuniary OSS adoption. The propensity function looks as follows:

$$\Pr(T = 1) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln ITK_{it} + \gamma_2 \ln ITL_{it} + \gamma_3 \ln pecuniary\_OSS_{it} + \gamma_4 \ln closed_{it} + \varepsilon_{it} \quad (9)$$

The coefficients from the propensity function are then used to predict the likelihood of a

given firm to adopt non-pecuniary OSS, $\hat{T}$. This allows for the construction of a weighting such that firms who have adopted (are treated, $T = 1$), are assigned a weight of the inverse of their propensity to adopt, $1/\hat{T}$, and firms who have not adopted ($T = 0$), are assigned a weight of the inverse of 1 minus their propensity to adopt, $\frac{1}{1-\hat{T}}$. These weights are then used to adjust the regression results to account for the sample selection bias such that firms who adopt and do not adopt are equally weighted in the regression results. This is similar to a propensity score matching procedure where each adopting firm is matched with a non-adopting firm that has a similar likelihood of adopting, based on observables, but does not require dropping observations that do not have a good match. Therefore, the resulting estimation can be interpreted as a causal effect similar to that of a randomized experiment, but without actually randomizing adoption (Hirano, Imbens, and Ridder, 2003; Hogan and Lancaster, 2004; Cole and Hernan, 2008; Huber, 2013).

Instrumental Variables

Two instrumental variables that exogenously shift a firm's likelihood of using non-pecuniary OSS are used to further address endogeneity concerns. Both instruments are constructed based on the non-pecuniary OSS adoption habits of firms that are similar (in industry or geography) to the focal firm, but whose adoption decision is exogenous to the firm itself. Such firms face supply conditions similar to the focal firm and are therefore likely to be affected by similar shocks to supply. This is similar to instruments that have been used for other studies of the digital economy (e.g., Forman, Goldfarb, and Greenstein, 2005). Importantly, most firms in the sample were founded before OSS diffused widely. Therefore, the firm's decisions to operate in a specific industry and locate in a specific geography are independent of OSS adoption patterns.

The first instrument is a measure of the mean non-pecuniary OSS usage of other firms within a given firm's 2-digit Standard Industrial Classification (SIC) industry within the same year. The amount of non-pecuniary OSS usage by the firms in a firm's same industry exogenously affects that firm's propensity for using non-pecuniary OSS primarily through labor. Employees of firms in a given industry are likely to interact with other firms in their industry through conferences and job movement. Therefore, in industries where there is widespread use of non-pecuniary OSS, a given firm is more likely to use non-pecuniary OSS.

The second instrument is a measure of the mean non-pecuniary OSS usage by other establishments within a given firm's county within the same year. Similarly to industry, geographically close firms also face supply conditions similar to the focal firm. Specifically, the availability of IT labor familiar with OSS in a local area is likely to affect the firm's decision to adopt OSS. The availability of this labor is greater in areas where other firms are already using OSS. Therefore, the amount of non-pecuniary OSS usage by the firms in a firm's local geography may exogenously shift that firm's propensity for using non-pecuniary OSS, but does not directly affect the firm's productivity level.

Panel Data Methods

Finally, since the data is panel data, firm fixed effect models can be used to estimate the effect at individual firms. However, because an individual firm is likely to only change from not using non-pecuniary OSS to using it once, fixed effects are only used when looking at continuous adoption of non-pecuniary OSS. This helps identify the effect as it relies on within-

firm variation in usage of non-pecuniary OSS rather than across firm variation. Further, to control for unobserved time and industry trends, the models uses year fixed effect and industry fixed effect at the 1-digit SIC level. The latter is only used when the 2-digit SIC instrument is not in use to avoid perverse instrumentation. The combination of these approaches helps eliminate unobserved firm, time, or industry effects that may bias the results. In aggregate, the identification strategy adds significant weight to a causal interpretation rather than just a correlational one.

3.5 Data

The data breaks into two primary areas: OSS usage and financial statements, both of which are at the firm level. Data on which firms are using OSS comes from the Harte Hanks IT Survey – a survey of IT usage by multiple sites at over 10,000 firms from 2000-2009. This database is used frequently in studies of the impact of IT on firm-level productivity (Brynjolfsson and Hitt, 2003; Forman, 2005; Forman, Goldfarb, and Greenstein, 2005; Forman, Goldfarb, and Greenstein, 2008; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013; McElheran, 2014). The Harte Hanks survey asks site-level IT managers questions about the types of IT (both hardware and software) used at the site as well as the number of IT employees at the site. In cases where Harte Hanks does not interview all sites within a firm, the average values for sites that are interviewed is assigned to sites that are not interviewed. This allows for the construction of firm level values that account for all sites within the firm.

The Harte Hanks data is augmented with detailed firm financial data. In particular, firm expenditures on labor (IT and non-IT) and capital (IT and non-IT) as well as firm revenues and

costs of materials. For public firms, this information is available via Standard and Poor's Compustat database. The firm's stock ticker symbol is used to match the Harte Hanks data to the Compustat data. In this manner, sites within the Harte Hanks database that are owned by different firms in different years (e.g., through mergers or acquisitions) will be associated with the correct parent firm and therefore the correct financial data. Although the Harte Hanks database contains information on over 10,000 firms, the final sample uses only public firms as the model requires additional financial information filed in the firm's 10-K. This reduces the sample size to 1,850 firms, and indicates that the results can best be applied to public firms. The sections below detail how these two datasets are used to construct the variables discussed in the previous section. All monetary values are converted to 2009 dollars using an appropriate deflation index and are reported in millions of dollars.

3.5.1 Variable Construction

Value-Added (VA_{it})

The dependent variable is constructed using a method consistent with prior literature (e.g., Dewan and Min, 1997; Brynjolfsson and Hitt, 2003; Huang, Ceccagnoli, Forman, and Wu, 2013). First, yearly operating costs (XOPR in Compustat) are deflated by the BLS Producer Price Index by stage of processing for intermediate materials, supplies, and components. Then deflated IT labor and non-IT labor (defined below) are both subtracted from the operating costs. The result is then subtracted from yearly sales (SALE in Compustat) deflated by the BEA Gross Domestic Product Price Index for gross output for private industries.

IT Capital (ITK_{it})

Most prior literature in the field constructs a combined measure of IT Capital that includes both the value of IT hardware at the firm and three times the value of IT labor at the firm due to the importance of IT labor being used for internal software development efforts, the result of which is a capital good (Brynjofsson and Hitt, 1996; Hitt and Brynjofsson, 1996; Dewan and Min, 1997; Huang, Ceccagnoli, Forman, and Wu 2013).[81] However, recent literature has shown that IT labor can have a separate effect from IT capital (Tambe and Hitt, 2012). Therefore, the primary analysis uses separate IT capital and IT labor variables. Later, the combined variable is tested for robustness purposes and the results are shown to be consistent.

To calculate IT Capital, the market value of the IT stock is estimated by multiplying the number of PCs and Servers at the firm (from Harte Hanks[82]) by the average value of a PC or Server that year from The Economist Intelligence Unit Telecommunications Database. The BEA Price Index for computers and peripherals is then used to deflate this value. This method is consistent with prior work in this area (e.g., Brynjofsson and Hitt, 1996; Huang, Ceccagnoli, Forman, and Wu 2013). Because the costs of the IT Capital are being imputed, a robustness check using the raw number of PCs and servers will be run and shows that the results are consistent.

[81] Ideally, the portion of the IT budget that is spent on software in addition to hardware would be included. However, software expenditures are combined with other capital expenditures in firm 10-K reporting. Therefore, while purchased software cannot be separated from other firm purchases, the cost of such software is captured in the non-IT Capital variable. Further, internal software development efforts will be captured in the IT Labor variable. This methodology is consistent with prior literature (e.g., Brynjofsson and Hitt, 1996; Huang, Ceccagnoli, Forman, and Wu 2013). Additionally, the high correlation between purchased software and hardware expenditures helps to mitigate concerns about not having software expenditure data.

[82] For most firms, Harte Hanks only surveys a sample of the sites within the firm. In such cases, the average number of PCs and Servers at the sites that are in the survey is multiplied by the total number of sites in the firm to obtain the total number of PCs and Servers in the firm. The same procedure is used for calculating the number of IT employees and the number of each type of operating system at the firm.

IT Labor (ITL_{it})

The value of IT labor is calculated by taking the number of IT workers at each firm (from Harte Hanks[83]) and multiplying by the mean annual wage for all Computer and Mathematical Science Occupations[84]. The BLS Employment Cost Index for wages and salaries for private industry workers is then used to deflate this value. Because the cost of the IT labor is being imputed, a robustness check using the raw number of IT employees will be run and shows that the results are consistent.

Non-IT Capital (K_{it})

The K_{it} variable is constructed by taking the yearly Gross Total Property, Plant and Equipment (PPEGT in Compustat), deflating it by the BLS price index for Detailed Capital Measures for All Assets for the Private Non-Farm Business Sector, and then subtracting the deflated value of IT Capital (defined above).

Non-IT Labor (L_{it})

Non-IT Labor is constructed using the total number of employees at the firm (EMP in Compustat) and subtracting the number of IT employees (from Harte Hanks) to obtain the total number of non-IT employees. This is then multiplied by the mean annual wage of all occupations[85] that year. The BLS Employment Cost Index for wages and salaries for private

[83] Harte Hanks reports the number of IT employees at each site as a range so the average value of the range is used. The ranges are 1-4, 5-9, 10-24, 25-49, 50-99, 100-249, 250-499, and 500 or More.

[84] Obtained from the Bureau of Labor and Statistics: http://www.bls.gov/oes/2009/may/oes_nat.htm#15-0000.

[85] Obtained from the Bureau of Labor and Statistics, for example the data for 2009 can be found here: http://www.bls.gov/oes/2009/may/oes_nat.htm#00-0000.

industry workers is then used to deflate this result. This method of calculation is consistent with prior studies on IT productivity (Bloom and Van Reenen, 2007; Bresnahan, Brynjolfsson, and Hitt, 2002; Brynjolfsson and Hitt 2003). However, because the cost of labor is being imputed, a robustness check with the raw number of non-IT employees is run and shows that the results are consistent.

Non-Pecuniary Open Source Software Usage

To measure the intensity of non-pecuniary OSS usage at the firm, the number and type of operating systems used at the firm is measured. Although operating systems are certainly not the only non-pecuniary OSS used at the firm, they are important and frequently indicate the wider use of non-pecuniary OSS. Further, the Harte Hanks survey asks firms what type of operating systems they use, but does not always capture other types of non-pecuniary OSS. Because this only captures non-pecuniary OSS operating systems, the dataset necessarily underestimates the amount of non-pecuniary OSS used at the firm. Therefore, the estimates should be considered a lower bound on the impact of non-pecuniary OSS to the firm.

In addition to constructing a measure of non-pecuniary OSS operating systems, measures of pecuniary OSS and closed-source operating systems are also constructed for use in predicting the propensity of a firm to adopt non-pecuniary OSS. These three measures ($non\_pecuniary\_OSS_{it}$, $pecuniary\_OSS_{it}$, and $closed_{it}$) are constructed by calculating the total number of each type of operating system at the firm (from Harte Hanks). The Harte Hanks data does not report the precise number of operating systems in use at a given firm. It does, however, report the different types of operating systems used at each site at the firm. These operating

systems are classified into three categories: non-pecuniary OSS, pecuniary OSS, or closed source. Table 3.1 shows the OSS operating systems in the dataset.[86] All other operating systems are labeled as "closed". Harte Hanks also reports whether each operating system is for a PC or a server as well as the total number of PCs and servers at each site. Therefore, for each site, the number of PC operating systems is evenly split over the total number of PCs at the site. The same is done for servers. This yields an estimate of how many instances of a given type of operating system exist at the site. This is then aggregated to the firm level and divided by the number of sites at the firm in the Harte Hanks database to obtain an average per site. Finally, this average is multiplied by the total number of sites in the firm to obtain a firm-wide imputation of the number of each type of operating system. As the resulting numbers are estimates, the analysis begins by only using a binary indicator of the presence of non-pecuniary OSS at the firm. The estimated number of operating systems will then allow for a more granular interpretation of the primary effect.

Because the number of operating systems in any of the three categories can potentially be zero (e.g., that category of operating system is not in use at the firm), one is added to the number of operating systems in each category before taking the natural log as the natural log of zero is undefined. Although there are many firms that have zero non-pecuniary and pecuniary OSS operating systems, there is a high degree of skewness in these numbers (as shown in the descriptive statistics below). Therefore, adding a one before taking the natural log should not significantly bias the results.

[86] Although some non-pecuniary OSS operating systems, such as Debian, are offered at a nominal pecuniary price by third-party vendors for the convenience of the distribution being pre-loaded on a CD or DVD, they are included in the non-pecuniary column as the full distribution is downloadable for free via the distribution's website. Additionally, although Apple's Mac OS X is built on BSD, it behaves more like a closed operating system than one that is pecuniary, but built on OSS, like Red Hat. Robustness checks were run against this assumption with no change to the primary results.

Table 3.1 Open Source Operating Systems

Pecuniary OSS Operating Systems	Non-Pecuniary OSS Operating Systems
Red Hat Linux SUSE Linux SCO Linux TurboLinux	Berkeley Software Distribution (BSD) Debian Conectiva Fedora FreeBSD Gentoo Linux Linux Kernel Mandrake Linux NetBSD OpenBSD Ubuntu

3.5.2 Descriptive Statistics

Table 3.2 shows the descriptive statistics of the firms in the dataset. There are 12,244 firm/year observations from 1,850 firms in the dataset.[87] The ranges vary greatly for all variables and demonstrate the breadth of the firms in the sample. This breadth allows for results that are more generalizable than many other studies of this kind, which only focus on Fortune 1000 companies. However, due to the Harte Hanks sampling methodology, larger firms are overrepresented in the sample and very small firms (e.g., startups) are not in the sample. Additionally, because of the reliance on 10-k data for financial information, all firms in the sample are public firms, which tend to be medium or large. For example, as shown in Table 3.2, the smallest company in the sample (Matec Corp.) had sales of $2.7 million in its lowest selling year. Comparatively, the largest firm (Exxon Mobil Corp.) had sales of $425 billion. Therefore, results should be interpreted as applying to medium and large firms. The firms in the dataset also

[87] This results in an average of 6.6 observations per firm. The panel is unbalanced because Harte Hanks does not survey every firm in every year. However, this is still a large enough number of observations per firm to conduct a fixed effect analysis and does not adversely affect the pooled analysis.

have a wide range of the type and intensity of IT use. The mean number of closed source operating systems at a firm is 5,026.755 while the mean number of non-pecuniary OSS and pecuniary OSS operating systems are much lower at 182.253 and 181.172, respectively. Looking deeper into the data, there are 3,527 observations where firms use at least one non-pecuniary OSS operating system. For these 3,527 observations, the average number of non-pecuniary OSS operating systems is 632.635. 7,341 observations use no OSS (pecuniary or non-pecuniary) at all. Only 10 observations use exclusively OSS (pecuniary or non-pecuniary).

Table 3.2 Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
$sales_{it}$	12244	5951.913	18793.42	2.694	425071
VA_{it}	12244	1845.747	5471.536	.006	154608
ITK_{it}	12244	8.279	48.687	.001	3165.154
ITL_{it}	12244	19.783	46.141	.184	835.876
K_{it}	12244	4243.141	14840.1	.113	305797.1
L_{it}	12244	851.044	2838.818	.028	91149.09
$non\_pecuniary\_OSS_{it}$	12244	182.253	1264.606	0	65690
$pecuniary\_OSS_{it}$	12244	181.172	2983.13	0	207646
$closed_{it}$	12244	5026.755	18304.15	0	1176977

Values for monetary variables are in millions of deflated US dollars. Values for operating systems are in number of computers at the firm running operating systems in that category.

Table 3.3 shows the correlation matrix. As to be expected, K_{it} and L_{it} have a fairly high correlation with value-added productivity since they are the primary inputs into the production function. Additionally, it is notable that the correlations between non-pecuniary OSS and the other two types of operating systems, pecuniary OSS and closed, are fairy low, while the correlation between pecuniary OSS and closed is comparatively high.

Table 3.3 Correlation Matrix

	VA_{it}	ITK_{it}	ITL_{it}	K_{it}	L_{it}	$nonpecuniary\ OSS_{it}$	$pecuniary\ OSS_{it}$	$closed_{it}$
VA_{it}	1.0000							
ITK_{it}	0.2989	1.0000						
ITL_{it}	0.4659	0.4444	1.0000					
K_{it}	0.7461	0.1910	0.3921	1.0000				
L_{it}	0.7846	0.1948	0.3561	0.4378	1.0000			
$nonpecuniary\ OSS_{it}$	0.1846	0.0986	0.3264	0.1448	0.1541	1.0000		
$pecuniary\ OSS_{it}$	0.1389	0.6848	0.2205	0.0594	0.0956	0.0384	1.0000	
$closed_{it}$	0.4089	0.9339	0.6007	0.3024	0.2758	0.1744	0.5472	1.0000

Table 3.4 shows the breakdown of observations by industry. While 48% of the observations are from the manufacturing industry, there is also good representation from other key industries, such as finance (14%), services (14%), and trade (11%). Further, Table 3.4 shows the percentage of firms within the industry that use non-pecuniary OSS or any type of OSS operating system. The percentage of firms in an industry using non-pecuniary OSS varies between 17.82% and 34.78%, with an average of 28.81% and has no major outliers. The percentage of firms in an industry using any OSS varies between 26.49% and 71.43%. However, this maximum should be considered an outlier because SIC 0 has a low number of observations. Therefore, the more realistic range is between 26.49% and 47.20%, with an average of 40.04%.

Table 3.4 Industry Breakdown

1-Digit SIC	Description	Frequency	Percent of all firms	Percent of firms using non-pecuniary OSS	Percent of firms using any OSS
0	Agriculture, Forestry, and Fishing	21	0.17	33.33	71.43
1	Mining and Construction	650	5.31	25.38	31.23
2-3	Manufacturing	5,879	48.02	31.25	43.35
4	Transportation, Communications, Electric, Gas, and Sanitary Services	927	7.57	27.83	37.32
5	Wholesale and Retail Trade	1,397	11.41	17.82	26.49
6	Finance, Insurance, and Real Estate	1,694	13.84	25.27	37.07
7-8	Services	1,676	13.69	34.78	47.20
9	Public Administration	0	0	0	0
All		12,244	100	28.81	40.04

3.6 Results and Discussion

This section presents the results of the empirical analysis and discusses the interpretation of these results in light of the hypotheses. First, basic three-factor productivity results are compared to those of other studies to confirm the consistency of the data and methods with prior research. Then, the results from the propensity score analysis, the first stage of the inverse-probability weighting method, are presented. These weightings are then used to obtain baseline regression results for the impact of non-pecuniary OSS on firm productivity. An instrumental variable approach is then employed to enhance the causal interpretation of these results. A number of moderator and split-sample analyses are then conducted to better understand the firm characteristics that are important determinants of the primary results. Finally, several robustness checks are considered to confirm that various assumptions are not driving the results.

3.6.1 Three-Factor Productivity Analysis

Before delving into the results on open source usage, the results of the baseline regression are presented to compare the elasticities of the three main productivity inputs with other existing studies. To properly achieve this comparison, the combined measure of IT Capital that is consistent with prior studies is used, rather than the separated measures used in the primary analysis. Table 3.5 shows the results of the basic three-factor productivity analysis. Models 1-3 use Ordinary Least Squares (OLS) regression with increasingly restrictive fixed effects, while Model 4 uses panel regression with firm fixed effects and Model 5 uses panel regression with random effects. For all models, the standard errors are robust and clustered by firm to account for any serial correlation in the error terms since the dataset contains multiple observations of the same firm over different time periods (Angrist and Pischke, 2009; Imbens and Kolesar, 2012). The high R^2 values are characteristic of such productivity studies. The confidence intervals of the coefficients in models 4 and 5 overlap with those of Huang, Ceccagnoli, Forman, and Wu (2013), whose methodology this study most closely resembles. However, the coefficients on non-IT capital are slightly higher than theirs, likely because their sample size is only companies in the Fortune 1000, while this study casts a wider net. Further, the column 4 coefficient on IT capital is very similar to that of Brynjolfsson and Hitt (2003) in their 1-year difference model with year and industry controls. The coefficients in column 4 are also very similar to the fixed effect estimate of Tambe and Hitt (2012), although the IT capital coefficient is slightly lower, likely because they are calculating their coefficient based solely on IT labor. These similarities help to add support to the validity of the dataset used in this study. The similarities also imply that if support is found for the hypotheses above, then the estimates in the prior literature are likely suffering from either attribution or omission bias.

Table 3.5 Three-Factor Productivity Results

DV: Value-Added (VA_{it})	1	2	3	4	5
Model	OLS	OLS	OLS	FE	RE
IT Capital (IT_{it})	.098*** (.008)	.066*** (.008)	.055*** (.008)	.030*** (.007)	.035*** (.006)
Non-IT Capital (K_{it})	.317*** (.012)	.314*** (.012)	.299*** (.012)	.082** (.034)	.270*** (.014)
Non-IT Labor (L_{it})	.631*** (.014)	.649*** (.014)	.671*** (.015)	.745*** (.035)	.699*** (.017)
Constant	.308*** (.040)	.234** (.045)	.298* (.163)	1.313*** (.169)	.379*** (.010)
Year fixed effect?	N	Y	Y	Y	Y
Industry fixed effect (SIC2)	N	N	Y	Y	Y
Number of firm/year observations	12244	12244	12244	12244	12244
Number of firms (groups)	1850	1850	1850	1850	1850
R^2 (between for panel)	0.898	0.913	0.917	0.903	0.930

***p<.01, **p<.05, *p<.1. All standard errors are clustered at the firm level. All variables are the natural log of the underlying variable.

3.6.2 Propensity to Adopt Non-Pecuniary OSS

As discussed previously, propensity scores are used to estimate the likelihood a firm adopts non-pecuniary OSS based on observables. The presence of non-pecuniary OSS in a firm-year observation is predicted based on the four primary input variables ($ITK_{it}, ITL_{it}, K_{it}, L_{it}$) as well as the two constructed variables estimating the number of pecuniary OSS operating systems and closed source operating systems at the firm ($pecuniary\_OSS_{it}$ and $closed_{it}$). These additional variables help to account for the technology usage of the firm. This method relies on firm observables to predict the propensity to adopt non-pecuniary OSS. Traits of the firm that are unobservable through a firm's financial reports, such as management quality, may also have an impact on the firm's propensity to adopt. However, as will be shown in a robustness check in Section 3.6.6, for a subset of the firms in this study that are also in the World Management Survey dataset (Bloom, Sadun, and Van Reenen, 2012), management quality does not predict use of non-pecuniary OSS.

The results of the propensity estimation are shown in Table 3.6. These results show there is a significant negative coefficient on ITK_{it} indicating that firms who spend more on IT Capital are less likely to adopt non-pecuniary OSS. This supports the theory that non-pecuniary OSS is a substitute for other IT, rather than a complement. However, there is a positive and significant coefficient on ITL_{it}, indicating that firms with larger IT staffs are more likely to adopt non-pecuniary OSS. Although interesting, it is difficult to interpret these results as causal due to the inherent endogeneity and potential omitted variable bias. However, they allow for the construction of the inverse-probability weighting discussed above, such that the remaining results are adjusted for sample bias and can be interpreted in a more causal manner.

Table 3.7 shows the resulting improvement of the balance in the sample after applying the IPW. Panel A shows the covariate balance without weighting. The t-statistics indicate that the adopting firms in the sample are significantly different from those that are non-adopters when comparing the four primary production inputs. Panel B shows the covariate balance after weighting. Here, the balance is much better and for all inputs except IT Capital, the balance drastically improves. While the IT Capital balance is still concerning, the use of weighting is primarily to deal with sample selection. This motivates the additional use of an instrumental variable approach. Although IPW improves the ability to interpret the resulting coefficients as causal, the instrumental variable approach helps to diminish any concerns of the covariate balance in the weighted sample presenting a threat to causal identification.

Table 3.6 Predicting Adoption of Non-Pecuniary OSS

DV: Binary adoption of OSS	1
Model	Probit
IT Capital (ITK_{it})	-.426*** (.065)
IT Labor (ITL_{it})	.200*** (.020)
Non-IT Capital (K_{it})	.019 (.015)
Non-IT Labor (L_{it})	.029 (.021)
$pecuniary\_OSS_{it}$	.022** (.011)
$closed_{it}$	.431*** (.073)
Constant	-4.092*** (.073)
Number of firm/year observations	12244
Number of firms (groups)	1850
Pseudo - R^2	0.085
Wald chi^2	373.06

***p<.01, **p<.05, *p<.1. All standard errors are clustered at the firm level. All variables are the natural log of the underlying variable.

Table 3.7 Covariate Balance

	Panel A			Panel B		
	Unweighted Sample			Weighted Sample		
	Adopters	Non-Adopters	t-stat	Adopters	Non-Adopters	t-stat
IT Capital (ITK_{it})	10.567	7.354	3.31	3.935	10.191	4.27
IT Labor (ITL_{it})	37.044	12.810	27.08	12.523	17.086	1.35
Non-IT Capital (K_{it})	7231.559	3032.545	14.30	3129.700	3728.530	1.42
Non-IT Labor (L_{it})	1300.912	668.785	11.22	673.788	826.972	2.56
Number of firm/year observations	3,527	8,717		3,527	8,717	

Values reported are the means of the adopting or non-adopting firms. Panel A presents the unweighted OLS regression of the given variable on non-pecuniary OSS adoption. Panel B presents the weighted OLS regression of the given variable on non-pecuniary OSS adoption.

3.6.3 Baseline Regression Results

Table 3.8 presents the estimation results using pooled OLS regressions without instrumental variables but with inverse-probability weighting. Columns 1 and 2 show the results when considering non-pecuniary OSS as a binary variable - do firms use non-pecuniary OSS or not. Column 1 shows a positive and significant coefficient of 0.059 on the use of non-pecuniary OSS. However, this effect becomes not significant when adding in the industry fixed effect in Column 2. These results are encouraging, although not conclusive due to the lack of granularity over how much non-pecuniary OSS a firm uses. Columns 3 and 4 show results for a similar analysis, but use a continuous measure of how many non-pecuniary OSS operating systems a firm uses. Here, the coefficient is slightly smaller than the binary coefficient, which makes intuitive sense, but it remains stable and significant when adding in the industry fixed effect. Columns 5 and 6 show a similar, although slightly larger, effect when considering only firms who have adopted at least one non-pecuniary OSS operating system. By only using firms that have adopted non-pecuniary OSS, the results in these two columns can be interpreted in a slightly more causal manner than the prior results as they compare firms who have all made the decision to adopt non-pecuniary OSS and therefore estimate the impact of the amount of non-pecuniary OSS adopted on productivity. However, caution must be applied in interpreting any of the results in Table 8 as causal as they only rely on IPW for dealing with endogeneity. The results in the following section use IPW as well as instrumental variables to additional add support for a causal interpretation.

Table 3.8 Baseline Regressions

DV: Value-Added (VA_{it})	1	2	3	4	5	6
Model	Pooled OLS	Pooled OLS	Pooled OLS	Pooled OLS	Pooled OLS	Pooled OLS
Adoption Measure	Binary	Binary	Continuous	Continuous	Continuous	Continuous
IT Capital (ITK_{it})	0.017 (0.022)	-0.001 (0.015)	0.012 (0.022)	-0.005 (0.014)	0.003 (0.035)	-0.043 (0.024)
IT Labor (ITL_{it})	0.024 (0.018)	0.026** (0.013)	0.028 (0.018)	0.028** (0.013)	0.032 (0.025)	0.030 (0.019)
Non-IT Capital (K_{it})	0.303*** (0.023)	0.288*** (0.022)	0.302*** (0.023)	0.286*** (0.023)	0.297*** (0.036)	0.283*** (0.033)
Non-IT Labor (L_{it})	0.663*** (0.018)	0.694*** (0.019)	0.660*** (0.018)	0.695*** (0.019)	0.651*** (0.030)	0.708*** (0.031)
$non\_pecuniary\_OSS_{it}$	0.058* (0.031)	0.067** (0.034)	0.016*** (0.006)	0.016** (0.008)	0.021*** (0.008)	0.026** (0.011)
Constant	0.329** (0.147)	0.191 (0.124)	0.338** (0.152)	0.199 (0.127)	0.401 (0.241)	0.149 (0.225)
Year fixed effect?	Y	Y	Y	Y	Y	Y
Industry fixed effect (SIC2)?	N	Y	N	Y	N	Y
Number of firm/year observations	12244	12244	12244	12244	3530	3530
Number of firms (groups)	1850	1850	1850	1850	946	946
R^2	0.925	0.934	0.928	0.925	0.936	0.945

***p<.01, **p<.05, *p<.1. Standard errors are clustered at the firm level. All variables are the natural log of the underlying variable. All regressions are weighted with inverse-probability weightings based on the propensity of the firm to adopt non-pecuniary OSS. Columns 5 and 6 only use firms that have adopted non-pecuniary OSS as the sample.

3.6.4 Instrumental Variable Regression Results

Having found a positive and significant result in the baseline regressions, the instrumental variables discussed above are now used in a two-stage least-squares framework to help further address endogeneity concerns. The results of this analysis are shown in Table 3.9. The first-stage F-statistics are above 10 for all models, adding support to the choice of instruments. Columns 1 and 2 show the results when pooling observations and considering adoption of non-pecuniary OSS in a binary manner. These columns show a larger coefficient on the binary usage of non-pecuniary OSS that is highly significant both when using only the industry instrument (column

1) and when using both instruments (column 2). Likewise, when considering adoption in a continuous manner, columns 3 and 4 show strong positive coefficients on the amount of non-pecuniary OSS used by the firm. Since the dependent variable is a natural log, the coefficient on $non\_pecuniary\_OSS_{it}$ in column 4 indicates that a 1% increase in the use of non-pecuniary OSS results in a .073% increase in productivity (as measured by value-added). The average value added for the firms in the sample is $1.846 billion; this indicates that a 1% increase in the number of non-pecuniary OSS operating systems leads to a $1.35 million increase in production output for the average firm. This effect is more than double the size of the coefficient on all IT capital found in columns 4 and 5 of Table 3.5. The negative coefficient on IT Capital (ITK_{it}) is characteristic of such analyses (Huang, Ceccagnoli, Forman, and Wu, 2013) due to the high level of correlation between IT related variables.[88] Column 5 reports the results when using a firm fixed-effect specification such that it is measuring the within firm variation of non-pecuniary OSS usage. The coefficient is again positive and statistically significant. Together, these results add significant support for H1b rather than H1a, indicating that the adoption of non-pecuniary OSS has a positive impact on firm productivity. Notably, the coefficients on non-pecuniary OSS are larger when using the IV methodology, indicating that overlooking the endogeneity concerns discussed above biases the baseline regression results towards zero. This is not surprising because of the geographic and industry differences that can effect the technology decisions of the firm.

[88] This is especially the case when using the continuous measure of non-pecuniary OSS operating systems as the number of operating systems and the number of computers is highly correlated. Since the IT Capital variable is not instrumented in this estimation, it acts as a control and therefore the negative coefficient should not be interpreted as causal.

Table 3.9 IV Regressions

DV: Value-Added (VA_{it})	1	2	3	4	5
Model	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	FE 2SLS
Adoption Measure	Binary	Binary	Continuous	Continuous	Continuous
IT Capital (ITK_{it})	-0.039 (0.030)	-0.010 (0.027)	-0.039 (0.032)	-0.020 (0.029)	-0.175** (0.078)
IT Labor (ITL_{it})	0.088*** (0.031)	0.055** (0.024)	0.080*** (0.029)	0.061** (0.025)	0.045*** (0.015)
Non-IT Capital (K_{it})	0.302*** (0.021)	0.302*** (0.022)	0.298*** (0.022)	0.299*** (0.022)	-0.128 (0.114)
Non-IT Labor (L_{it})	0.649*** (0.019)	0.656*** (0.018)	0.639*** (0.019)	0.647*** (0.018)	0.834*** (0.058)
$non\_pecuniary\_OSS_{it}$	0.813*** (0.276)	0.428** (0.173)	0.107*** (0.035)	0.073*** (0.025)	0.407** (0.200)
Constant	0.388*** (0.142)	0.434*** (0.144)	0.548*** (0.133)	0.529*** (0.137)	-
Year fixed effect?	Y	Y	Y	Y	Y
SIC2 Instrument (for $non\_pecuniary\_OSS_{it}$)	Y	Y	Y	Y	Y
County Instrument (for $non\_pecuniary\_OSS_{it}$)	N	Y	N	Y	N
Number of firm/year observations	12244	12244	12244	12244	12244
Number of firms (groups)	1850	1850	1850	1850	1850
First Stage F-test	26.74	22.73	28.64	19.15	9.80
R^2	0.898	0.918	0.906	0.913	0.478

***p<.01, **p<.05, *p<.1. Standard errors are clustered at the firm level for models 1-4 and are conventional GLS for model 5. All variables are the natural log of the underlying variable. All regressions are weighted with inverse-probability weightings based on the propensity of the firm to adopt non-pecuniary OSS.

3.6.5 Moderators and Split-Sample Analysis

After establishing the primary effect, the preferred specification (the pooled instrumental variable analysis with the continuous measure of non-pecuniary OSS) is used to calculate various moderator and split-sample results to better understand the determinants of the main effect. For specifications that include an interaction term, the interaction of the two instruments with the moderator is also used to ensure a causal interpretation is still plausible. Table 3.10 shows the results of this analysis. Column 1 shows the effect of using open source interacted with the size

of the firm, measured by the natural log of yearly employees. A positive coefficient on the interaction term indicates a positive relationship between firm size and the effect of OSS usage on firm productivity.[89] This finding adds support for H2.

Columns 2 and 3 break down the analysis by industry showing the manufacturing sector (column 2) and the services sector (column 3). Consistent with H3, these results show that services firms have a much greater output elasticity for non-pecuniary OSS than manufacturing firms. Interestingly however, when lagging the use of non-pecuniary OSS by one year, the coefficient for manufacturing firms becomes positive, but not significant. When lagging usage by two years, the coefficient for manufacturing firms becomes positive and significant at the 10% level, indicating that non-pecuniary OSS can also have a positive impact on firms in the manufacturing sector, it just takes longer for these benefits to accrue.[90] Column 4 shows the analysis when removing firms in the finance industry (SIC code 6) as their financial reporting methods often differ from other types of companies. However, removing these firms does not significantly alter the main results, indicating that the main effect is not being driven by financial reporting methods. Column 5 shows the analysis when removing firms in the agriculture and mining industries, as their use of IT differs from most other industries. However, removing these firms does not significantly alter the main results.

Finally, columns 6-8 consider the importance of IT at the industry level. Jorgenson, Ho, and Stiroh (2005), show that the importance of IT to productivity is higher in industries that are

[89] As mentioned above, the dataset focuses on medium to large public firms, so small firms in this sample are still larger than many private firms or startups.

[90] The results of this lagged analysis are not included to save space. However, they are available from the author upon request.

either IT-producing or IT-using when compared to industries that are neither. Columns 6-8 separate the industries into these three categories based on the same industry classification as Jorgenson, Ho, and Stiroh (2005). The baseline analysis for this breakdown was inaccurately measured due to large standard errors, and therefore a one-year lag of the use of non-pecuniary OSS, as well as the instrumental variables, is used. The full impact of IT often takes longer than one year to materialize (Brynjolfsson and Hitt, 2003). This phenomenon is explored further in the next section. As seen by the coefficients in columns 6-8, non-pecuniary OSS has a strong effect on the productivity of IT-using and IT-producing industries, while it appears to have no effect on firms in neither of those groups. This is consistent with the findings in Jorgenson, Ho, and Stiroh (2005). Interestingly, the point estimate for the impact of non-pecuniary OSS is higher for firms in IT-using industries than it is for IT-producing industries. However, the confidence intervals overlap so it is difficult to interpret this in any meaningful way.

3.6.6 Robustness Checks

As with any empirical estimation, the estimation strategy is founded on a number of assumptions that may affect the outcome of the analysis. Therefore, this section considers a number of robustness checks against some of these assumptions to ensure they are not directly leading to the results discussed above. Due to space constraints, only the results of the preferred specification (the pooled instrumental variable analysis with the continuous measure of non-pecuniary OSS) are shown for each robustness check in Table 3.11.

Production Input Assumptions

As mentioned in Section 3.5, IT Labor and IT Capital are separated, rather than including them in a combined variable, as is standard in the economics of IT literature (Brynjofsson and

Table 3.10 Moderator and Split-Sample Regression Results

DV: Value-Added (VA_{it})	1	2	3	4	5	6	7	8
Model	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS
IT Capital (ITK_{it})	0.003 (0.034)	0.042 (0.053)	-0.173*** (0.054)	-0.039 (0.032)	-0.028 (0.029)	-0.027 (0.050)	-0.126 (0.077)	0.055 (0.040)
IT Labor (ITL_{it})	0.011 (0.027)	-0.008 (0.035)	0.137*** (0.046)	0.053** (0.027)	0.070*** (0.025)	0.056 (0.038)	0.068* (0.036)	-0.002 (0.018)
Non-IT Capital (K_{it})	0.287*** (0.030)	0.187*** (0.058)	0.179*** (0.049)	0.295*** (0.027)	0.269*** (0.022)	0.241*** (0.053)	0.292*** (0.025)	0.335*** (0.021)
Non-IT Labor (L_{it})	0.101 (0.168)	0.794*** (0.040)	0.798*** (0.043)	0.667*** (0.021)	0.676*** (0.017)	0.717*** (0.065)	0.674*** (0.037)	0.616*** (0.024)
$non\_pecuniary\_OSS_{it}$	-0.251* (0.141)	-0.009 (0.046)	0.194** (0.086)	0.091*** (0.028)	0.083*** (0.026)	0.107*** (0.042)	0.177** (0.071)	-0.004 (0.073)
ln (emp)	0.501*** (0.193)							
$non\_pecuniary\_OSS_{it}$ * ln (emp)	0.035** (0.015)							
Constant	-0.663 (0.913)	0.465** (0.211)	0.450*** (0.156)	0.453*** (0.151)	0.550*** (0.136)	0.448*** (0.143)	0.631*** (0.177)	0.453*** (0.086)
Year fixed effect?	Y	Y	Y	Y	Y	Y	Y	Y
Sample Restriction	-	Manuf.	Services	Excluding finance industries	Excluding agriculture and mining	IT-Producing Industries	IT-Using Industries	Non-IT Using or Producing Industries
SIC2 Instrument (for $non\_pecuniary\_OSS_{it}$)	Y	Y	Y	Y	Y	Y	Y	Y
County Instrument (for $non\_pecuniary\_OSS_{it}$)	Y	Y	Y	Y	Y	Y	Y	Y
Number of firm/year observations	12244	5880	1677	10555	11574	1168	4515	4714
Number of firms (groups)	1850	863	316	1644	1764	238	798	832
First Stage F-test	-	17.05	7.77	19.13	18.44	19.141	12.011	5.705
R^2	0.929	0.926	0.928	0.928	0.924	0.954	0.875	0.921

***p<.01, **p<.05, *p<.1. Standard errors are clustered at the firm level. All variables are the natural log of the underlying variable. All regressions are weighted with inverse-probability weightings based on the propensity of the firm to adopt non-pecuniary OSS. Columns 6, 7, and 8 use a one-year lag of OSS usage and instruments.

Hitt, 1996; Hitt and Brynjofsson, 1996; Dewan and Min, 1997; Huang, Ceccagnoli, Forman, and Wu 2013). Therefore, to confirm the separation of these variables does not have an impact on the results, a combined IT variable consistent with the prior literature is considered. This variable

consists of the deflated value of IT Capital plus three times the deflated value of IT Labor. Using this combined variable instead of the separate IT Capital and IT Labor variables, both the baseline and the IV regressions are re-estimated. In all cases, the results for the coefficient on non-pecuniary OSS were substantively similar. In all cases, the coefficient is consistently positive and significant, and in almost all cases the confidence interval of the coefficients overlaps when comparing the results for the combined IT variable and the separated variables. The results of this robustness check with the preferred specification are shown in column 1 of Table 3.11. This adds support to the robustness of the primary results against concerns that using the more granular separation of the two variables drove the results.

Average prices and wages for a given input in a given year are used to impute the costs of many of the primary input variables. As discussed in Section 3.5, the IT Labor, non-IT Labor, and IT Capital variables are all imputed based on the raw number of IT employees, non-IT employees, and computers and the yearly average for IT worker wages, non-IT worker wages, and prices for PCs and servers, respectively. To confirm that the results are robust against the assumption that these averages apply to all firms in a similar manner, all regressions are re-run using only the raw numbers for the inputs, rather than the imputed cost of each input. Again, in all cases, the coefficient on non-pecuniary OSS is consistently positive and significant, and in most cases the confidence interval of the coefficient overlaps when comparing the results for the imputed cost variables with those of the raw input variables. The results of this robustness check with the preferred specification are shown in column 2 of Table 3.11. This adds support to the robustness of the primary results against concerns that imputing the cost drove the results.

The inclusion of non-pecuniary OSS operating systems as a raw number in the regressions makes comparing the size of the effect to other inputs un-intuitive, as the other inputs are all measured in dollars. Therefore, the price of a pecuniary operating system, Microsoft Windows, for that year is used to estimate the value of each non-pecuniary OSS operating system.[91] The BEA computer price index is then used to deflate this value. The cost of replacing the non-pecuniary OSS operating systems at each firm with this pecuniary alternative is then estimated in a method similar to that of Greenstein and Nagle (2014), who perform the same estimation for the non-pecuniary OSS web server Apache. Although there is wide variance in the functionality and quality of operating systems, this rough estimate allows for a comparison of dollars to dollars, rather than dollars to number of operating systems. The result is shown in column 3 of Table 3.11. The resulting coefficient is significant and positive and is greater than the coefficient for IT Capital found in the more restrictive models in columns 4 and 5 of Table 3.5. This is encouraging as it indicates that the value of non-pecuniary OSS is on a similar order to that of other IT-related inputs. However, its effect is greater than these less risky inputs, adding further support to the primary hypotheses.

Timing of OSS Factors

Prior studies have shown that the full effect of IT on productivity can take 5-7 years to be realized due to the organizational changes that must occur for the full effect of IT to be realized (Brynjolfsson and Hitt, 2003). Therefore, the analysis in column 4 shows the preferred specification with a 6-year lag of the amount of non-pecuniary OSS used. To account for this lag, the instruments are lagged by 6 years as well. The coefficient on lagged non-pecuniary OSS is

[91] Prices for Microsoft Windows are based on the latest version of Windows in a given year and are gathered from various industry publications at the time of release.

larger than in the preferred specification, although the confidence intervals overlap. Similar results occur for lags up to 6 years, but are not show due to space constraints. These results indicate that investments in non-pecuniary OSS in year's past have an effect that spills over to the productivity of the current year.

Relatedly, the implementation of the instrumental variables is such that the instruments are constructed for the same year as the observation being estimated. It is quite possible that it is the adoption of non-pecuniary OSS in prior years by other firms in the same county or industry that influences the likelihood of a given firm to adopt. Therefore, a robustness check is run using a 1-year lag of both instruments, rather than the same year. The results for the preferred specification are shown in column 5 of Table 3.11. The resulting coefficient on non-pecuniary OSS is positive and significant and the confidence interval overlaps with that of the coefficient from the primary specifications. Therefore, the primary results are robust to this concern.

Estimation Methodology

There may be a concern that all results shown from the IV regressions have inverse probability weighting applied. To confirm that the results from the IV regressions are not only the result of the weighting, column 6 in Table 3.11 shows the results of the primary specification with no weighting, but with both instruments. The results show that the coefficient on non-pecuniary OSS is still positive and significant. Further, the confidence intervals of this coefficient overlap with those of the primary specification, indicating that the use of IPW is not interfering with the application of the instruments.

There may also be a possible concern that the results are driven by local industry agglomeration or knowledge spillovers, which have been shown to have an important effect on innovation (Jaffe, Trajtenberg, and Henderson, 1993; Furman, Porter, and Stern, 2002). This is of a particular concern as the second IV is based on county. Therefore, column 7 in Table 3.11 shows the results of the primary specification with a county-fixed effect and without the county IV. The coefficient on non-pecuniary OSS continues to be positive and significant, adding support to the robustness of the primary results against such concerns.

There may also be concerns with the use of IPW rather than a more standard matching methodology. Therefore, as a robustness check, I also use the nearest-neighbor matching methodology of Abadie and Imbens (2006). Using a nearest-neighbor match based on all observables used in the prior regressions, I construct a matched sample based on the binary use of non-pecuniary OSS. I then use this matched sample to estimate the sample average treatment effect (SATE) at 0.165 with a standard error of 0.025. This positive and statistically significant coefficient again offers support for the validity of my primary results.

Identification of OSS Effect

Concerns may arise that the effect found in the primary analysis is just that of an accounting nature, that the results are simply because non-pecuniary OSS is free and therefore it is not accounted for. While this may be true to some degree, the TCO literature discussed above has argued that the actual cost of software is so small compared to the implementation costs (hardware and labor), that it is almost negligible. Therefore, any residual effect found in this analysis should not be primarily due to an accounting issue, but instead to the firm benefiting

from spillovers due to crowd intelligence. However, to further rule this alternative explanation out, an analysis is run that includes both non-pecuniary OSS and pecuniary OSS. One would expect that the coefficient on such a variable may be slightly smaller than non-pecuniary OSS alone as many of the risks, and likewise the benefits, associated with pecuniary OSS are lower. This is indeed what is found in column 8. The coefficient on the combined OSS is slightly lower than that on non-pecuniary OSS alone, although the confidence intervals overlap.

An additional concern may be that the use of non-pecuniary OSS is correlated with unobservable managerial practices that are likely to increase productivity. Although the primary data set does not allow ruling out such simultaneity bias, additional data from the World Management Survey (Bloom, Sadun, and Van Reenen, 2012) is used to confirm this is not driving the results.[92] The World Management Survey (WMS) asks a wide array of firms about their management practices every few years starting in 2004. 183 of the 1,850 firms from the main dataset for this paper appear at least once in the WMS dataset. Although this is far from a complete overlap, it does represent nearly 10% of the firms in the dataset. There are 247 firm/year observations that overlap from two datasets. To increase the amount of overlap, results from the WMS data are carried one year forward and one year backwards, except where the firm is actually surveyed in consecutive years. For example, the results from a firm surveyed in 2004 are carried to both 2003 and 2005. This allows for the expansion of the number of firm/year observations to 650. Although this method assumes firm management practices do not change significantly within a one-year time window, this assumption is consistent with results from firms that were surveyed multiple times. The firms that appear in both datasets are used to test

[92] The author is grateful to Nick Bloom, Raffaella Sadun, and John Van Reenen for allowing access to the WMS dataset.

Table 3.11 Robustness Checks

DV: Value-Added (VA_{it})	1	2	3	4	5	6	7	8
Model	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS	Pooled 2SLS
IT Capital (ITK_{it})			-0.016 (0.028)	-0.029 (0.020)	0.011 (0.030)	-0.036* (0.021)	-0.057* (0.034)	-0.025 (0.015)
IT Labor (ITL_{it})			0.059** (0.025)	0.058*** (0.019)	0.045* (0.027)	0.031** (0.013)	0.050*** (0.016)	0.052** (0.022)
Non-IT Capital (K_{it})	0.298*** (0.021)	0.297*** (0.022)	0.301*** (0.022)	0.316*** (0.014)	0.297*** (0.024)	0.313*** (0.012)	0.306*** (0.015)	0.301*** (0.022)
Non-IT Labor (L_{it})	0.643*** (0.018)		0.651*** (0.018)	0.607*** (0.017)	0.639*** (0.019)	0.653*** (0.015)	0.661*** (0.017)	0.653*** (0.018)
$non\_pecuniary\_OSS_{it}$	0.073*** (0.024)	0.068*** (0.025)		0.091*** (0.008)	0.061*** (0.017)	0.110*** (0.036)	0.093** (0.048)	0.062*** (0.015)
IT Capital and Labor combined	0.051*** (0.013)							
# of PCs and Servers		-0.013 (0.029)						
# of IT employees		0.040 (0.025)						
# of non-IT employees		0.668*** (0.019)						
Imputed cost for non-pecuniary OSS			0.040*** (0.015)					
Constant	0.485*** (0.122)	-1.666*** (0.136)	0.484*** (0.139)	0.431*** (0.070)	0.578*** (0.153)	0.515*** (0.051)	1.336*** (0.104)	0.458*** (0.146)
Year fixed effect?	Y	Y	Y	Y	Y	Y	Y	Y
County fixed effect?	N	N	N	N	N	N	Y	N
Robustness Check	Combined IT capital and 3x labor	Raw # for ITL, non-ITL, and ITK	Imputed price for OSS	6-year lag of OSS use	1-year lag of instruments	No IPW	County fixed-effect	All OSS variable
SIC2 Instrument (for $non\_pecuniary\_OSS_{it}$)	Y	Y	Y	Y	Y	Y	Y	Y
COUNTY Instrument (for $non\_pecuniary\_OSS_{it}$)	Y	Y	Y	Y	Y	Y	N	Y
Number of firm/year observations	12244	12244	12244	3670	10397	12244	12244	12244
Number of firms (groups)	1850	1850	1850	1182	1718	1850	1850	1850
First Stage F-test	22.38	19.52	20.72	165.68	25.08	36.74	40.88	19.27
R^2 (between)	0.920	0.922	0.920	0.934	0.928	0.900	0.931	0.920

***p<.01, **p<.05, *p<.1. Standard errors are clustered at the firm level. All variables are the natural log of the underlying variable. Regressions in columns 1-5 and 7-8 are weighted with inverse-probability weightings based on the propensity of the firm to adopt non-pecuniary OSS.

the correlation between management practices and the use of OSS (both pecuniary and non-pecuniary). The results indicate that an increase in the quality of a firm's management practices is uncorrelated with the decision to use non-pecuniary or pecuniary OSS.[93] This result is consistent when using the 247 firm/year direct observations or the 650 imputed observations. Further, it is consistent when examining the binary or continuous use of OSS, and when controlling for the production inputs of the firm ($ITK_{it}, ITL_{it}, K_{it}, L_{it}$). Indeed, when running a regression of the binary or continuous usage of OSS on production inputs and the WMS measure of management quality, the coefficient on the latter is negative, but not significant. This indicates that the quality of a firm's management is uncorrelated with the firm's decision to use OSS. Therefore, concerns of simultaneity bias due to management quality can be alleviated.

3.7 Conclusion

The results of this study show that the use of non-pecuniary OSS does indeed have an impact on the productivity of the firm, and that this impact is positive. The effect is consistently positive in all specifications that account for sample selection and endogeneity via inverse probability weighting, instrumental variable analysis, and firm fixed effects. This effect exists when considering the use of non-pecuniary OSS at both a binary and continuous level such that both the usage and the amount of non-pecuniary OSS used positively affect productivity. The effect is still positive and significant when considering within firm variation through a firm-fixed effect model. Because the use of non-pecuniary OSS is only measured via operating systems, other firm investments in non-pecuniary OSS are not captured. Therefore, the true effect of all non-pecuniary OSS is likely greater than the effect found in this study.

[93] The full tables of results are not shown to save space, but are available from the author upon request.

Digging further into the main effect by exploring various split sample analyses reveals that larger firms (based on employees) gain a larger benefit from increased usage of non-pecuniary OSS. However, due to the sample construction, even the smallest firms are still rather large. It is quite possible, even likely, that the use of non-pecuniary OSS has an even larger effect for firms that are very small and therefore capital-constrained. However, due to data constraints, the effect of non-pecuniary OSS on small companies, technology related start-ups in particular, is left for future research. Finally, consistent with other literature on the productivity of IT, this study finds that services firms have a higher output elasticity of non-pecuniary OSS than manufacturing firms. These findings, as well as the risks associated with adopting non-pecuniary OSS discussed above, help explain why not all firms are using what, at first glance, appears to be a free input.

Although endogeneity is always a concern in productivity studies, this study takes many steps to help rule out this bias to allow for the results to be interpreted in a causal manner. All of the regression results use fixed effects for year. This helps to rule out alternative explanations due to trends over time. In all specifications inverse probability weighting is used to generate an analysis similar to that of a matched sample strategy. With this statistically rigorous matching method, the primary finding of a positive causal effect of non-pecuniary OSS usage on productivity holds. Additionally, in some specifications firm fixed effects are used so that a firm is compared with itself over time. Finally, the use of instrumental variables allows for a proper identification of the effect within this panel framework. As mentioned above, the complete identification strategy adds a significant amount of weight to a causal interpretation of the findings, rather than just a correlational interpretation.

The findings have important implications for researchers, practitioners, and policy makers. For researchers, the results draw additional attention to the mismeasurement that occurs when firms use non-pecuniary OSS (and, more generally, non-pecuniary crowdsourced digital goods) as inputs into production. The results indicate that current studies underestimate the amount of IT at the firm. Future studies of productivity, especially the productivity of IT, should account for these non-pecuniary inputs, rather than misattributing them to firm intangible effects. This is especially important as information costs are increasingly approaching zero and the amount of non-pecuniary crowdsourced digital inputs firms use is likely to rise in the coming years. For practitioners, the results indicate that firms of all sizes may enhance their productivity by increasing the amount of OSS they employ in their production process, although larger firms may benefit more than medium sized firms due to economies of scale. Similarly, firms in the services sector may benefit more than those in the manufacturing sector. For policy makers, the results indicate that federal funding of OSS and other publicly available digital goods could enhance the productivity of firms. While other studies have shown that federal investments in such goods can have a high rate of return based on the value of the goods themselves (Greenstein and Nagle, 2014), the results of this study indicate that such goods can also boost the productivity of the firms that use them. However, as shown in the moderator and split sample results, not all firms benefit to the same degree.

Despite the in-depth analysis of the determinants of the productivity effect of non-pecuniary OSS, a handful of open questions remain. First, some firms contribute back to the production of OSS and other public digital goods. It is unclear whether these contributions help the firm gain more out of using these inputs, or if the firm is needlessly giving away proprietary information.

Second, limitations of the dataset do not allow for the measurement of the importance of non-pecuniary crowdsourced digital goods for the productivity of very small firms and start-ups. It is likely that the credit constraints of such firms lead to an even higher reliance upon, and productive impact from, such goods. Finally, the analysis has been constrained to focus on digital goods that are free. It is quite likely that the price of all digital goods, which have a marginal cost of zero, does not properly reflect their value to production. Therefore, the broader implications for the productivity impact of all digital goods remain an interesting area for future research.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235-267.

Afuah, A. (2000). Do your co-opetitors' capabilities matter in the face of technological change?. *Strategic Management Journal*, 21, pp. 387-404.

Altman, E., Nagle, F., & Tushman, M. (2014). Innovating without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero. In Oxford Handbook of Creativity, Innovation, and Entrepreneurship, edited by Michael A. Hitt, Christina Shalley, and Jing Zhou. Oxford University Press.

Angrist, J.D., & Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Aral, S., & Weill, P. (2007). IT assets, organizational capabilities, and firm performance: How resource allocations and organizational differences explain performance variation. *Organization Science*, *18*(5), 763-780.

Asay, M. (2013). Is Facebook The World's Largest Open Source Company? *ReadWrite.* Retrieved from http://readwrite.com/2013/10/17/is-facebook-the-worlds-largest-open-source-company on October 31, 2014.

Athey, S., & Ellison, G. (2014). Dynamics of Open Source Movements. *Journal of Economics & Management Strategy*, *23*(2), 294-316.

Athey, S. & Stern, S. (2002) The impact of information technology on emergency health care outcomes. *RAND Journal of Economics*, 33(3), 399-432.

Baldwin, C. Y., & Clark, K. B. (2006). The architecture of participation: Does code architecture mitigate free riding in the open source development model?. *Management Science*, *52*(7), 1116-1127.

Baldwin, C., & Von Hippel, E. (2011). Modeling a Paradigm Shift: From Producer Innovation to User and Open Collaborative Innovation. *Organization Science*, *22*(6), 1399–1417.

Benkler, Y. (2002). Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale Law Journal*, 369-446.

Black Duck Software. (2014). The Eighth Annual Future of Open Source Survey. Retrieved from https://www.blackducksoftware.com/future-of-open-source on Oct. 31, 2014.

Bloom, N. & J. Van Reenen. (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics*. **122**(4) 1351-1408.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). Americans Do IT Better: US Multinationals and the Productivity Miracle. *American Economic Review*, 102(1), 167-201.

Brandenburger, A. M., & Nalebuff, B. J. (2011). *Co-opetition*. Random House LLC.

Bresnahan, T.F., E. Brynjolfsson, & L.M. Hitt. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*. **117**(1) 339-376.

Bridgman, B. (2013). *Home Productivity*. Bureau of Economic Analysis Working Paper 2013-03.

Brynjolfsson, E., & Hitt, L. (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management Science*, *42*(4), 541-558.

Brynjolfsson, E., & Hitt, L. M. (2003). Computing productivity: Firm-level evidence. *Review of economics and statistics*, *85*(4), 793-808.

Brynjolfsson, E., Hitt, L. M., & Yang, S. (2002). Intangible assets: Computers and organizational capital. *Brookings papers on economic activity*, *2002*(1), 137-198.
Byrne, D., Oliner, S., & Sichel, D. (2013). Is the information technology revolution over? *Available at SSRN 2240961*.
Casadesus-Masanell, R., & Llanes, G. (2011). Mixed Source. *Management Science*, *57*(7), 1212–1230.
Chatterji, A. K., & Fabrizio, K. R. (2013). Using users: When does external knowledge enhance corporate product innovation?. *Strategic Management Journal*.
Cole, S.R., & Hernan, M.A. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168 (6), pp. 656-664.
Corbet, J., Kroah-Hartman, G. & McPherson, A. (2013). Linux Kernel Development: How Fast it is Going, Who is Doing It, What They are Doing, and Who is Sponsoring It (2013 Edition). Linux Foundation Whitepaper.
Corrado, C., & Hulten, R. (2013). Innovation Accounting. In Measuring Economic Sustainability and Progress, edited by Dale W. Jorgenson, J. Steven Landefeld, and Paul Schreyer. University of Chicago Press.
Corrado, C., Hulten, C., & Sichel, D. (2009). Intangible capital and US economic growth. *Review of Income and Wealth*, *55*(3), 661-685.
Dewan, S., & Min, C. K. (1997). The substitution of information technology for other factors of production: A firm level analysis. *Management Science*, *43*(12), 1660-1675.
Finley, K. (2013). Apple's Operating System Guru Goes Back to His Roots. *Wired*. Retrieved from www.wired.com/2013/08/jordan-hubbard/ on October 31, 2014.
Fitzgerald, B. (2006). The transformation of open source software. *MIS Quarterly*, 587-598.
Forman, C. (2005). The corporate digital divide: Determinants of Internet adoption. *Management Science*, *51*(4), 641-654.
Forman, C., Goldfarb, A., & Greenstein, S. (2005). How did location affect adoption of the commercial Internet? Global village vs. urban leadership. *Journal of Urban Economics*, *58*(3), 389-420.
Forman, C., Goldfarb, A., & Greenstein, S. (2008). Understanding the inputs into innovation: Do cities substitute for internal firm resources?. *Journal of Economics & Management Strategy*, *17*(2), 295-316.
Forrest, C. (2014). Salil Deshpande: Software Engineer, Venture Capitalist, Open Source Investor. *TechRepulic*. Retrieved from http://www.techrepublic.com/article/salil-deshpande-software-engineer-venture-capitalist-open-source-investor/ on October 31, 2014.
Fosfuri, A., Giarratana, M. S., & Luzzi, A. (2008). The penguin has entered the building: The commercialization of open source software products. *Organization Science*, *19*(2), 292-305.
FreeBSD Contributor List. https://www.freebsd.org/doc/en/articles/contributors/article.html, retrieved on October 27, 2014.
Furman, J. L., Porter, M. E., & Stern, S. (2002). The determinants of national innovative capacity. *Research policy*, *31*(6), 899-933.
Gilder, G. (1995). The Coming Software Shift. *Forbes*, August 28, 1995.
Giera, J., & Brown, A. (2004). The Costs and Risks of Open Source – Debunking the Myths. Forrester Research Whitepaper.
Graham, R. (2014). 300k vulnerable to Heartbleed two months later. *Errata Security*. Retrieved from http://blog.erratasec.com/2014/06/300k-vulnerable-to-heartbleed-two.html on October 1, 2014.

Greenstein, S., & Nagle, F. (2014). Digital Dark Matter and the Economic Contribution of Apache. *Research Policy* 43, pp.623-631. (Prior version released as National Bureau of Economic Research Working Paper 19507).
Hamilton, D. (2014). Mirantis Gains $100M in the Largest Series-B Investment Round in Open-Source Software History. Retrieved from http://www.thewhir.com/web-hosting-news/mirantis-gains-100m-largest-series-b-investment-round-open-source-software-history on October 31, 2014.
Hann, I., Roberts, J., and Slaughter, S. (2013). All Are Not Equal: An Examination of the Economic Returns to Different Forms of Participation in Open Source Software Communities. *Information Systems Research* 24(3), pp. 520-538.
Hann, I., Roberts, J., Slaughter, S. and Fielding, R. (2002). Economic Incentives for Open Source Projects: Can Participation be Explained by Career Concerns? *Proceedings of the 22nd International Conference on Information Systems (ICIS)*, Barcelona, Spain, December 2002.
Harhoff, D., Henkel, J., & Von Hippel, E. (2003). Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. *Research Policy*, *32*(10), 1753-1769.
Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4* (pp. 475-492). NBER.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.
Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, *35*(7), 953-969.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.
Hitt, L. M., & Brynjolfsson, E. (1996). Productivity, business profitability, and consumer surplus: three different measures of information technology value. *MIS Quarterly*, 121-142.
Hogan, J.W., & Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13, pp 17-48.
Horvitz, D.G., & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, pp. 663-685.
Horovitz, B. (2013). Crowdsourcing rules for Super Bowl ads. *USA Today*. Retrieved from http://www.usatoday.com/story/money/business/2013/01/19/crowdsourcing-super-bowl-commercials-doritos-lincoln-pepsi/1842937/ on October 31, 2014.
Howe, J. 2008. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business.* Crown Business, New York.
Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2013). IT Knowledge Spillovers and Productivity: Evidence from Enterprise Software. *Available at SSRN 2243886*.
Huber, M. (2013). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 2013.
Hulten, C. (2010). Decoding Microsoft: Intangible Capital as a Source of Company Growth. National Bureau of Economic Research (NBER) Working Paper 15799.
Imbens, G. W., & Kolesar, M. (2012). Robust Standard Errors in Small Samples: Some Practical Advice. *NBER Working Paper w18478*.

Jaffe, A. B., & Trajtenberg, M. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, *108*(3).
Jorgenson, D. W. (2001). Information technology and the US economy. *The American Economic Review*, *91*(1), 1-32.
Jorgenson, D. W., Ho, M. S., & Stiroh, K. J. (2005). Productivity, Volume 3: Information Technology and the American Growth Resurgence. *MIT Press Books*, *3*.
Kogut, B., & Metiu, A. (2001). Open-source software development and distributed innovation. *Oxford Review of Economic Policy*, *17*(2), 248-264.
Krishnamurthy, S. (2005). "An Analysis of Open Source Business Models," in Perspectives on Free and Open Source Software, J. Feller, B. Fitzgerald, S. Hissam, and K. Lakhani (eds.), MIT Press, Cambridge, MA, 2005, pp. 279-296.
Lakhani, K., & Von Hippel, E. (2003). How open source software works: "free" user-to-user assistance. *Research Policy*, *32*(6), 923–943.
Lakhani, K., Lifshitz-Assaf, H., & Tushman, M. (2012). Open innovation and organizational boundaries: the impact of task decomposition and knowledge distribution on the locus of innovation in *Handbook of Economic Organization: Integrating Economic and Organization Theory*, A. Grandori (ed.), Edward Elgar Publishing, Northampton, MA, pp. 355-382.
Lerner, J., Pathak, P. A., & Tirole, J. (2006). The dynamics of open-source contributors. *The American Economic Review*, 114-118.
Lerner, J., & Schaknerman, M. (2010). The comingled code: Open source and economic development. *MIT Press Books*.
Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal of Industrial Economics*, *50*(2), 197–234.
Lunden, I. (2014). Alfresco Raises A Fresh $45M to Fuel Open-Source Enterprise Content Management. *TechCrunch.* Retrieved from http://techcrunch.com/2014/08/21/alfresco-raises-a-fresh-45m-to-fuel-open-source-enterprise-content-management/ on October 31, 2014.
MacCormack, A. (2003). Evaluating Total Cost of Ownership for Software Platforms: Comparing Apples, Oranges, and Cucumbers. AEI-Brookings Joint Center for Regulatory Studies Related Publication, April 2003.
McCue, T.J. (2013). For Motor Company Sees Open Source. *Forbes.* Retrieved from http://www.forbes.com/sites/tjmccue/2013/01/10/ford-motor-company-sees-open-source/ on October 31, 2014.
McElheran, K. S. (2014). Delegation in Multi-Establishment Firms: Adaptation vs. Coordination in I.T. Purchasing Authority. *Journal of Economics & Management Strategy*, 23 (2), 225-258.
O'Mahony, S. (2003). Guarding the commons: how community managed software projects protect their work. *Research Policy*, *32*(7), 1179–1198.
O'Mahony, S., & Ferraro, F. (2007). The emergence of governance in an open source community. *Academy of Management Journal*, *50*(5), 1079-1106.
Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge university press.
Phipps, S. (2014). Walmart's investment in open source isn't cheap. *InfoWorld*. Retrieved from http://www.infoworld.com/article/2608897/open-source-software/walmart-s-investment-in-open-source-isn-t-cheap.html on October 31, 2014.
Raymond, Eric. (1998). Goodbye, "free software"; hello, "open source". Retrieved from http://www.catb.org/~esr/open-source.html on February 23, 2014.

Raymond, E. (1999). The cathedral and the bazaar. *Knowledge, Technology & Policy*, *12*(3), 23-49.
Russo, B., Braghin, B., Gasperi, P., Sillitti, A., and Succi, G. (2005). Defining TCO for the Transition to Open Source Systems. *Proceedings of the First International Conference on Open Source (OSS2005)*, pp. 108-112.
Schofield, J. (2008). How many people make Windows 7? *The Guardian*. Retrieved from http://www.theguardian.com/technology/blog/2008/aug/19/howmanypeoplemakewindows7 on October 29, 2014.
Schumpeter, J. 1942. The Process of Creative Destruction. Chapter VII, pp. 81-86 in *Capitalism, Socialism, and Democracy.* Harper & Row, New York, NY.
Schwarz, M., & Takhteyev, Y. (2011). Half a Century of Public Software Institutions". *Journal of Public Economic Theory*, *12*(4), 609-639.
Shirky, C. (2008). Here Comes Everybody: The Power of Organizing Without Organizations. Penguin Press, New York.
Sinofsky, S. (2011). Introducing the team. *Microsoft Developer Network Blog*. Retrieved from http://blogs.msdn.com/b/b8/archive/2011/08/17/introducing-the-team.aspx on October 29, 2014.
Sorkin, A. & Peters, J. (2006). Google to Acquire YouTube for $1.65 Billion. *The New York Times.* Retrieved from http://www.nytimes.com/2006/10/09/business/09cnd-deal.html on October 31, 2014.
Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature*, 49(2), pp. 326-365.
Tambe, P., & Hitt, L. M. (2012). The Productivity of Information Technology Investments : New Evidence from IT Labor Data. *Information Systems Research*, *23*(3), 599–617.
Tambe, P., Hitt, L., & Brynjolfsson, E. (2011). The Price and Quantity of IT-Related Intangible Capital. Working paper.
Tambe, P., Hitt, L. M., & Brynjolfsson, E. (2012). The Extroverted Firm: How External Information Practices Affect Innovation and Productivity. *Management Science*, *58*(5), 843–859.
Varian, H. R., & Shapiro, C. (2003). Linux adoption in the public sector: An economic analysis. *Manuscript. University of California, Berkeley*.
Von Hayek, F. A. (1945). The use of knowledge in society. *The American economic review*, 519-530.
Von Hippel, E. (1986). Lead Users: A Source of Novel Product Concepts. *Management Science*, *32*(7), 791–805.
Von Hippel, E., & Von Krogh, G. (2003). Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science*, *14*(2), 209-223.
Von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, *32*(7), 1217–1241.
West, J., & Lakhani, K. R. (2008). Getting clear about communities in open innovation. *Industry and Innovation*, *15*(2), 223-231.
Wheeler, D. (2005). Why Open Source Software/Free Software (OSS/FS, FLOSS, or FOSS)? Look at the Numbers! available online at http://www.dwheeler.com/oss_fs_why.html.
Woods, D., & Guliani, G. (2005). Open Source for the Enterprise: Managing Risks, Reaping Rewards. O'Reilly Media.

Wooldridge, J. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1, pp. 117-139.
Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281-1301.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686-688.
Woolley, A. W., & Fuchs, E. (2011). PERSPECTIVE-Collective Intelligence in the Organization of Science. *Organization Science*, *22*(5), 1359-1367.
Yarow, J. (2013). Microsoft's Biggest Problem In One Chart. *Business Insider*. Retrieved from http://www.businessinsider.com/microsofts-biggest-problem-in-one-chart-2013-9 on October 31, 2014.
Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive science*, *36*(3), 452-470.
Young, R. & Johnson, D. R., (2009). A Comparison of Four Methods for Handling Missing Secondary Respondent Data. *Paper presented at the American Sociological Association Annual Meeting.*

Chapter 4: Organizational Learning Through Contributing to Public Goods: Evidence from Open Source Software

Frank Nagle

ABSTRACT

As technological progress lowers barriers to communication and coordination, organizations are increasingly relying on external resources as key inputs into productivity and innovation. Frequently these resources are public goods that are available for their competitors to use as well. Counter-intuitively, some firms even pay their employees to contribute to the creation of these goods. One possible explanation for this behavior is that contributing to the creation of these public goods allows a firm to learn how to better capture value from them and therefore increase its competitive advantage over its competitors. This study explores this mechanism by using data on firm contributions to open source software (OSS), an important public digital good that is created through crowdsourcing. Using coarsened exact matching and inverse probability weighting to address endogeneity concerns, this study shows that firms who contribute to the development of OSS capture more productive value from the use of OSS than their non-contributing peers. Further, this learning has a spillover effect that allows contributing firms to capture more productive value from all of their IT investments. This effect increases as firms increase the intensity of their contributions. These results have important strategic implications for managers to consider as they increasingly engage with external communities and ecosystems.

Keywords: *organizational learning, public goods, user innovation, open source software, crowdsourced digital goods*

4.1 Introduction

As technological improvements lower barriers to communication and coordination across organizations, the boundaries between organizations are progressively weakening. This weakening leads organizations to increasingly engage with externally sourced goods, especially digital goods that are produced through crowdsourcing efforts. Since these goods are frequently public and can therefore be used freely by anyone, organizations need to be able to capture more value out of using these goods than their competitors to increase their competitive advantage and likelihood of succeeding in the marketplace. One potential method for increasing value capture in these scenarios is by learning how to more efficiently use such goods. Since these goods are crowdsourced and public, it is possible that contributing to their creation may be one mechanism for firms to enhance their learning. Therefore, the goal of this paper is to answer the question of whether organizations learn from their contribution to the creation of crowdsourced digital goods in a manner that allows them to better capture productive value from the use of such goods. Further, the paper examines whether the benefits of this learning spillover also enhance the productive value of all IT investments by the organization.

To examine this phenomenon, this study uses data on firm contributions to open source software (OSS). Pairing data on contributions to Linux, the most widely used open source project in the world, with information on firm usage of IT allows for an analysis of the learning benefits that occur when a firm contributes to OSS. First, data from public firm financial statements is used to estimate a firm's productivity return to investments in IT, including OSS. Then, using coarsened exact matching (CEM), firms that contribute to Linux are matched to firms that do not contribute. Inverse probability weighting is then used to add further balance to the sample. A

matched sample analysis is then run to estimate the increased returns to productivity from the use of OSS that contributors gain and non-contributors do not. Finally, the question of whether the learning that occurs through contribution to OSS has a spillover effect to the productivity of all IT at the firm is then addressed.

The results of the empirical analysis show that after a firm contributes to Linux, it gains a productivity benefit from its use of OSS that is 11% higher than that of a matched firm that does not contribute. A more detailed analysis of the OSS usage at the firm reveals that the majority of these gains come from the firm's use of non-pecuniary (free) OSS rather than OSS the firm pays for. Limited evidence is found to support the argument that the learning from contributing to OSS has a spillover effect to all of the firm's investments in IT, not just OSS. Contributing firms have a slightly higher return to all investments in IT, and investments in IT labor in particular. Measuring contribution at a more granular level – numbers of contributors or contributions – rather than just a binary level reveals that firms who contribute more to OSS gain more from their use of IT.

This study contributes to a growing body of literature that examines the impact of firms increasingly engaging with their external ecosystem for core innovative processes (Altman, Nagle, and Tushman, 2015; Baldwin and von Hippel, 2011; Iansiti and Clark, 1994; Lakhani, Lifshitz-Assaf, and Tushman, 2012). It also contributes to the large literature on a firm's ability to integrate external knowledge through the absorptive capacity process (Cohen and Levinthal, 1989; Cohen and Levinthal, 1990). However, unlike much of the absorptive capacity literature, which focuses on *internal* investments to be better able to integrate *external* resources, this study

shows that firm investments in *external* development of OSS allows them to better extract value from both *external* resources (other OSS) and *internal* resources (IT capital and labor they already own). Further, the results of this study help to explain why some firms engage in the seemingly irrational behavior of paying their employees to contribute to the development of public goods that their competitors can use.

The following section lays the theoretical groundwork for the primary argument and establishes five testable hypotheses. Section 4.3 discusses the empirical methodology used to test these hypotheses. Section 4.4 presents the data and the construction of the various measures used in the empirical analysis. Section 4.5 discusses the results of the analysis and Section 4.6 concludes.

4.2 Theory and Hypotheses

This section first considers some of the existing literature on why firms gain value from investments in information and digital technologies. Then it discusses how crowdsourced digital goods, including open source software, allow a firm to contribute to the value creating process of developing such goods. An argument is then made that such contributions also allow the firm to better capture value from using those crowdsourced goods in a manner similar to absorptive capacity from investments in research and development (R&D). Finally, the argument is extended to examine how the learning from this process can allow firms to gain more from all of their technology investments, not only those of a crowdsourced nature.

4.2.1 Gaining Value from Information Technology Usage

For decades, arguments have been made on both sides of the debate about the value of information technology investments (Barua, Kriebel, and Mukhopadhyay, 1995; Brynjolfsson and Hitt, 1996; Brynjolfsson, McAfee, Sorell, and Zhu, 2008; Byrne, Oliner, and Sichel 2013; Gordon, 2002; Jorgenson, 2001; Jorgenson, Ho, and Stiroh, 2005). This debate is very similar to that around the value of investing in R&D.[94] To date, the general consensus has shown that IT can be valuable for firm productivity, but to capture the full value IT has to offer, firms must have the right people and organizational structure (Aral and Weill, 2007; Tambe and Hitt, 2012). Further, IT systems have become so complex that few individuals fully understand how they work (Attewell, 1992; Fichman and Kremerer, 1997). Additionally, investments in IT can be quite expensive and often take as many as seven years of use to reap the full benefits of those investments (Brynjolfsson and Hitt, 2003). Together these various contingencies on capturing value from IT are due to the fact that IT is expensive and is often a black box that is not fully accessible to the firm. While the recent increase in reliance on cloud computing has decreased the cost of IT in many ways, it has also dramatically increased the black box nature of the technology behind "the cloud". Therefore, even though IT is a resource that the firm purchases and uses internally, in many ways it remains an external resource since the inner workings are not fully understood by the firm.

In the R&D literature, the difficulties in integrating such external resources have been considered for a long time. The concept of "absorptive capacity" was developed to capture the ability of a firm to "recognize the value of new information, assimilate it, and apply it to

[94] See Griliches (1979) for an early overview of this debate and Doraszelski and Jaumandreu (2013) for a more recent look at the debate.

commercial ends" (Cohen and Levinthal, 1989; Cohen and Levinthal 1990). Although this concept has been applied to the ability of firms to integrate information technology in a productive manner (Garcia-Morales, Ruiz-Moreno, and Llorens-Montes, 2007) the black box nature of closed source technology can make this process difficult. Firms cannot truly engage with the black box in a meaningful way, so gains from absorptive capacity can be limited. This fact, combined with the high costs of closed source IT, have led many firms to turn to crowdsourced technologies which are more open, and often cheaper, than their closed source counterparts.

4.2.2 Crowdsourced Technology and the Firm

Crowdsourcing as a method for creating technology in general, and software in particular, has been a crucial part of the development process since the early days of IT. In the software world, the concept was formalized in 1983 when Richard Stallman founded the GNU Project[95] to create the first computer operating system that gave users the freedom to share and modify the software. With the addition of the Linux Kernel to the GNU operating system in 1991, OSS took off rapidly and diffused widely into the production practices of firms. An important characteristic of the opens source ecosystem, and a partial reason for its acceptance in production environments, is the ability to take free and open software as a base and add functionality and support to create a product that is priced and is not fully open. For example, the Linux Kernel (which is free and open source) is the base for Red Hat Linux (which is not free and not fully open). The lower cost and more open nature of OSS has led many firms to integrate it into their

[95] GNU is a recursive acronym for "GNU's Not UNIX". UNIX was the predominant operating system at the time and its license structure did not allow for users to examine or alter the underlying code.

core infrastructure that has led to productivity benefits (Nagle, 2015). However, why firms would contribute to the creation of such goods is less obvious.

There are many incentives for individuals to contribute to crowdsourced technology including reputation w/ peers, job market, and pure enjoyment (Benkler, 2002; Lerner and Tirole, 2002; West and Lakhani, 2008, Athey and Ellison, 2014). There are also many incentives for firms to open their technology to competitors (Aksoy-Yurdagul, 2015; Aksoy, Fosfuri, and Giarratana, 2011; Casadesus-Masanell and Llanes, 2011; Fosfuri, Giarratana, and Luzzi, 2008; Harhoff, Henkel, and von Hippel, 2003; von Hippel and von Krogh, 2003; Henkel, 2006; Lerner, Pathak, and Tirole, 2006; West and Gallagher, 2006). The case of the electric car company Tesla opening its patents to competitors is an often-cited example of this behavior. Doing so allowed Tesla's architecture to become the industry standard leading to a dramatic increase in the value creation in the industry, but at the same time letting Tesla capture a great deal of this value. However, the incentives for firms to engage in co-opetition (Brandengurger and Nalebuff, 2011) by contributing to existing projects that are already open have gone underexplored. At first glance, such an activity may seem irrational – why should a firm pay its own employees to write code that will be used by its competitors? One potential rationale for this behavior is that by doing so, firms are able to enhance their absorptive capacity for extracting value from using the technologies they are helping to create.

4.2.3 Absorptive Capacity from Firm Contributions to Crowdsourced Technologies

In the classic absorptive capacity literature, firms increase their ability to gain productive value from external R&D efforts by increasing their own R&D efforts (Cohen and Levinthal

1989; Cohen and Levinthal, 1990). Similarly, by contributing to an existing crowdsourced IT project (rather than just opening their existing IT to others), firms position themselves to better take advantage of their use of such projects. Prior work has shown that absorptive capacity can contribute to the performance of OSS groups themselves (Daniel, Agarwal, and Stewart, 2006) When a firm pays their employees to contribute to these projects, they are essentially paying the employee to better understand the complex nature of the technology black box, since they are now helping to build it. The learning is likely to be even greater when the IT project is particularly complicated, like in the case of a computer operating system, which is known to be a very complex piece of software that has many interactions between different components (MacCormack, Rusnak, and Baldwin, 2006). This learning allows the firm to better integrate this external innovation into their own process, which increases the value they can capture from using these crowdsourced technologies. This leads to the following hypothesis:

> *Hypothesis 1: Firms that contribute to the development of open source software are able to gain greater productivity returns from the use of open source software than firms that do not contribute.*

As discussed above, not all OSS is completely open. There are many firms that build on top of OSS, add enhanced features and services, and then charge a price for the use of the resulting software. The underlying workings of this *pecuniary* OSS are often more inaccessible to the firms that use it than truly open *non-pecuniary* OSS. For example, in the operating system realm, a pecuniary OSS like Red Hat Linux can often be more opaque than a non-pecuniary OSS operating system like Gentoo Linux. Therefore, the learning obtained through contributing to the

development of OSS cannot be applied as easily to pecuniary OSS when compared to non-pecuniary OSS. Formally,

> *Hypothesis 2: Firms that contribute to the development of open source software are able to gain greater productivity returns from the use of non-pecuniary open source software than from pecuniary open source software.*

4.2.4 Spillovers Beyond OSS from Learning by Contributing

Software is one small piece of the IT ecosystem at the firm. Indeed, software costs often only represent 10% of the total cost of implementing that piece of software (MacCormack, 2003). The other 90% of the costs are predominately due to the hardware that the software is installed on and the labor used to install and maintain the software. Therefore, while the absorptive capacity benefits on the use of OSS are the most direct effect, there may be a similar effect on the rest of the IT ecosystem. It is quite likely that the learning gained by contributing to OSS has a spillover effect and allows a firm to better utilize its other IT investments. Interestingly, this is almost the reverse of absorptive capacity. In the absorptive capacity relationship, firms invest in *internal* R&D to gain more productivity from *external* technology. In the case of spillovers from learning by contributing, the firm is investing in *external* code development to gain more productivity from *internal* IT resources that it already owns. By investing in public value creation, the firm is better able to capture value from its private investments.

Hypothesis 3: Contributing to OSS development allows firms to gain more productive value from all of their IT investments.

To this point, contribution to OSS has been only considered at the binary level – firms either contribute or they do not. However, it is certainly the case that not all firms contribute to the same degree. Some firms may contribute a great deal more to the development of OSS than others. For such firms, the learning they obtain would be greater, and therefore the spillover to their ability to capture productive value from their IT investments would also be greater.

Hypothesis 4: An increase in the degree to which a firm contributes to OSS development leads to an increase in the firm's ability to gain productive value from their IT investments.

Finally, since the learning obtained by contributing to OSS development resides in the people who do the contributing, rather than in the hardware at the firm, there is likely to be a greater effect on the productivity gains from investments in IT-related labor than other IT investments.

Hypothesis 5: Contributing to OSS development allows firms to gain more productive value from their investments in IT-related labor.

4.3 Empirical Methodology

This section describes the empirical methodology employed to test the hypotheses developed

above. First, it describes the estimation model, which is consistent with other models of the productivity of IT, but accounts for contributions to crowdsourced digital goods. Then, it discusses identification concerns due to sample selection and endogeneity as well as the methodologies employed to address these concerns. These methods include coarsened exact matching and inverse probability weighting. Some of the wording in this section is identical to that in Chapter 3, but is reprinted here for continuity purposes.

4.3.1 Estimation Models

The dataset will measure capital, labor, and various IT inputs, including contributions to open source software, a widely used crowdsourced digital good. Before describing this data in detail, it is useful to review the model and estimation approach of the paper. In both the economics of IT literature, the standard method of estimation is the classic Cobb-Douglas Production function modified to include IT (e.g., Brynjolfsson and Hitt, 1996; Dewan and Min, 1997; Tambe and Hitt 2012; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013):

$$Y_{it} = {K_{it}}^{\alpha} {L_{it}}^{\beta} IT_{it}^{\gamma} A_{it} \quad (1)$$

where Y_{it} is the production of firm i in time t, ${K_{it}}^{\alpha}$ is the amount of non-IT capital stock, and ${L_{it}}^{\beta}$ is the amount of non-IT labor. IT_{it}^{γ} is the amount of IT expenditure (including IT-related labor) and A_{it} is a firm-specific efficiency multiplier that captures intangible assets such as management skill or institutional knowledge and learning. This methodology is consistent with the methodology frequently used in studies of absorptive capacity (e.g., Griffith, Redding, and Van Reenen, 2003; Knott, 2008).

Value-added productivity (VA_{it}) is substituted for sales as a measure of output to remove concerns about trends in the economy or demand shocks (Brynjolfsson and Hitt 2003) and then the log of each side is taken to obtain:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma \ln IT_{it} + \varepsilon_{it} \quad (2)$$

Taking the natural log of each side results in coefficients that are equivalent to a firm's output elasticity to a given input. This allows for an interpretation of the coefficients as the percentage change in VA_{it} for a one percent change in the value of the given input. Unobserved differences in firm-level efficiency are captured in the error term. This baseline model is consistent with the most current total-factor productivity models of productivity measurement that account for IT usage (e.g., Tambe and Hitt 2012; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013). As discussed in prior literature (Greenstein and Nagle, 2014; Nagle, 2015), open source software is often not captured is such analyses due to its non-pecuniary nature. Therefore, to account for this properly, a measure of a firm's utilization of open source software, OSS_{it}, in a given period is added to the specification. The measurement of OSS is described in the data section below. To allow for consistent interpretation, the natural log of this measure is used. This results in the following equation:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln IT_{it} + \gamma_2 \ln OSS_{it} + \varepsilon_{it} \quad (3)$$

Equation 3 is used as the baseline estimation equation. To test the hypotheses defined above, whether or not the firm contributes to the creation of OSS, $Contrib_i$, is introduced into the equation as a standalone dummy variable (1 if the firm contributes, 0 otherwise) and is then

interacted with OSS_{it}. Because firms start to contribute at various times during the observation period, the binary $Post_t$ term is used to capture the period after the firm's first contribution to OSS. The value of $Post_{it}$ is 1 if the firm's first contribution was this year or any prior year, and 0 otherwise. This is interacted with the other terms, resulting in a triple interaction $Contrib_i * OSS_{it} * Post_{it}$. The coefficient on this triple interaction can be interpreted as the additional benefit to using OSS obtained by firms who have contributed to the creation to OSS due to an absorptive capacity-like mechanism. The full estimation equation is as follows:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln IT_{it} + \gamma_2 \ln OSS_{it} + \gamma_3 Contrib_i + \gamma_4 Post_{it} + \gamma_5 Contrib_i * \ln OSS_{it} + \gamma_6 Contrib_i * Post_{it} + \gamma_7 \ln OSS_{it} * Post_{it} + \gamma_8 Contrib_i * \ln OSS_{it} * Post_{it} + \varepsilon_{it} \quad (4)$$

To test the second set of hypotheses, $Contrib_i$ is interacted with IT_{it} in a similar manner. The coefficient on the resulting triple interaction, $Contrib_i * IT_{it} * Post_{it}$, is interpreted as the productivity benefit that firm *i* receives at time *t* due to the learning spillover from the firm's contribution to OSS that can be applied to other IT. The full estimation equation is as follows:

$$\ln(VA_{it}) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln IT_{it} + \gamma_2 \ln OSS_{it} + \gamma_3 Contrib_i + \gamma_4 Post_{it} + \gamma_5 Contrib_i * \ln IT_{it} + \gamma_6 Contrib_i * Post_{it} + \gamma_7 \ln IT_{it} * Post_{it} + \gamma_8 Contrib_i * \ln IT_{it} * Post_{it} + \varepsilon_{it} \quad (5)$$

4.3.2 Identification Strategy

In an ideal experiment, one would randomly assign firms from the full population of US firms to not contribute to the creation of OSS or to contribute at varying levels of intensity. However, such an experiment is infeasible and therefore observational data, discussed in the next section, is used. Like all studies that use observational data, this analysis is subject to both

sample selection bias and endogeneity. Sample selection is a potential threat to identification due to the fact that the dataset (discussed below) undersamples firms that contribute to OSS. This could result in incorrect estimation of coefficients for the population. A second threat to identification is the fact that firms endogenously decide whether or not to contribute to the creation of OSS. If firms that are, for example, better managed are both more likely to contribute to OSS and have higher levels of productivity, then the relationship between contribution and productivity could not be interpreted as causal due to simultaneity bias. Further, this could lead to an incorrect estimation of the size of the effect. Both of these concerns prevent a complete answer to the primary question that can be used to make recommendations to managers. Therefore, multiple methods that help to address both of these concerns are applied.

Coarsened Exact Matching

Coarsened Exact Matching (CEM) is a matching methodology that allows for causal inference without balance checking (Iacus, King, and Porro, 2011a; Iacus, King, and Porro, 2011b). This methodology is subset of the Monotonic Imbalance Bounding (MIB) matching methods that allow for observational data to be used in a manner that mimics an experiment, allowing for a causal interpretation of the results. To apply CEM to the data in this study the decision to contribute is considered the "treatment" variable. Therefore, for all firms that have contributed to OSS, a firm that is similar based on observables, but which does not contribute, is identified. The "control" firm identified by CEM has productivity characteristics as similar as possible to the "treated" firm, including the firm's industry, value added productivity, IT expenditure, non-IT capital, non-IT labor, and usage of OSS. In cases where a good match in the same 3- or 4- digit SIC industry could not be found, a match in the 1- or 2- digit industry was

used instead. The matched firms are discussed in the data section below. After performing this one-to-one CEM match, the full data set consists of 50% firms that have contributed to OSS and 50% firms that have not contributed. Each non-contributor is assigned the same "first contribution" date as its matched contributor. Therefore, any trends or industry shocks over time are properly accounted for and the interpretation of contribution can occur in a causal manner. This methodology is similar to a diff-in-diff, although the treatment is not strictly exogenous.

Inverse-Probability Weighting

Although CEM does a good job of matching contributing firms to non-contributing firms, the matches are not always perfect and therefore it is possible the contributing firms may look slightly different than the non-contributing firms. To address this concern, as well as any issues from overall sample selection bias, inverse-probability weighting (IPW) (Horvitz and Thompson, 1952) is used. IPW increases the consistency of the estimator (Wooldridge, 2007) in a manner similar to Heckman correction (Heckman, 1976, 1979), but with fewer assumptions (Wooldridge, 2002; Young and Johnson, 2009). This is necessary because the dataset (discussed below) undersamples firms that contribute to OSS, which can adversely affect the estimation procedure. IPW also helps address endogeneity concerns and allows for the results to be interpreted as causal, in a manner similar to matching, by balancing the dataset between treatment and control groups to identify the direct effect of the independent variable (Hirano, Imbens, and Ridder, 2003; Hogan and Lancaster, 2004; Cole and Hernan, 2008; Huber, 2013).

To construct the IPW weights, the first step is to predict the propensity of a firm to contribute to OSS based on observables. To do this, a Probit function is used to predict the likelihood of

treatment (contribution to OSS) based on observables. In addition to the three primary input variables (IT_{it}, K_{it}, L_{it}), the model also uses three constructed variables estimating the number of non-pecuniary OSS operating systems, pecuniary OSS operating systems and closed source operating systems at the firm ($non\_pecuniary\_OSS_{it}$, $pecuniary\_OSS_{it}$ and $closed_{it}$). These additional variables help to account for the amount of operating systems used by the firm, which could be an important predictor of contribution to OSS. The construction of these variables is discussed in the next section. The propensity function looks as follows:

$$\Pr(T = 1) = \alpha \ln K_{it} + \beta \ln L_{it} + \gamma_1 \ln IT_{it} + \gamma_2 \ln \_non\_pecuniary\_OSS_{it} + \gamma_3 \ln pecuniary\_OSS_{it} + \gamma_4 \ln closed_{it} + \varepsilon_{it} \quad (6)$$

The coefficients from the propensity function are then used to predict the likelihood of a given firm to contribute to OSS, $\hat{T}$. This allows for the construction of a weighting such that firms who have contributed (are treated, $T = 1$), are assigned a weight of the inverse of their propensity to contribute, $1/\hat{T}$, and firms who have not contributed ($T = 0$), are assigned a weight of the inverse of 1 minus their propensity to contribute, $\frac{1}{1-\hat{T}}$. These weights are then used to adjust the regression results to account for the sample selection bias such that firms who contribute and do not contribute are equally weighted in the regression results. Therefore, the resulting estimation can be interpreted as a causal effect similar to that of a randomized experiment, but without actually randomizing adoption (Hirano, Imbens, and Ridder, 2003; Hogan and Lancaster, 2004; Cole and Hernan, 2008; Huber, 2013).

Finally, to control for unobserved time trends, the models use a year fixed effect. Further, heteroskedastic robust standard errors are used in all models, and the more restrictive clustered

standard errors are used to confirm the interpretation of the coefficients as significant. The combination of these approaches with the CEM and IPW helps eliminate unobserved firm or time effects that may bias the results. In aggregate, the identification strategy adds significant weight to a causal interpretation rather than just a correlational one.

4.4 Data

The data consists of three primary datasets: OSS usage, contributions to OSS, and financial statements, all of which are at the firm level. Data on which firms are using OSS comes from the Harte Hanks IT Survey – a survey of IT usage by multiple sites at over 10,000 firms from 2000-2013. This database is used frequently in studies of the impact of IT on firm-level productivity (Brynjolfsson and Hitt, 2003; Forman, 2005; Forman, Goldfarb, and Greenstein, 2005; Forman, Goldfarb, and Greenstein, 2008; Tambe, Hitt, and Brynjolfsson, 2012; Huang, Ceccagnoli, Forman, and Wu, 2013; McElheran, 2014). The Harte Hanks survey asks site-level IT managers questions about the types of IT (both hardware and software) used at the site as well as the number of IT employees at the site. In cases where Harte Hanks does not interview all sites within a firm, the average values for sites that are interviewed is assigned to sites that are not interviewed. This allows for the construction of firm level values that account for all sites within the firm.

Data on contributions to OSS comes from the Linux Foundation. The Linux Foundation is the non-profit organization that manages the OSS project Linux. Linux[96] is a computer operating system and is the most widely used piece of OSS in the world. Like many OSS projects, Linux tracks each individual contribution made to the code base of the software. However, Linux also

[96] For an overview of the creation of Linux and other OSS operating systems, see Nagle (2015).

tracks what firms these contributors work for.[97] Therefore, it is possible to map contributions to Linux to users of OSS operating systems from the Harte Hanks data based on the firm's name.

The Harte Hanks and Linux Foundation data are augmented with detailed firm financial data. In particular, firm expenditures on labor (IT and non-IT) and capital (IT and non-IT) as well as firm revenues and costs of materials. For public firms, this information is available via Standard and Poor's Compustat database. The firm's stock ticker symbol is used to match the Harte Hanks data to the Compustat data. In this manner, sites within the Harte Hanks database that are owned by different firms in different years (e.g., through mergers or acquisitions) will be associated with the correct parent firm and therefore the correct financial data. Although the Harte Hanks database contains information on over 10,000 firms, the final sample uses only public firms as the model requires additional financial information filed in the firm's 10-K. Further, since most firms do not contribute to Linux, they are not included in this analysis unless they are a CEM match to a contributing firm. In total, there are 34 public firms that contribute to Linux and are in the Harte Hanks dataset. Each of these firms is matched with a non-contributing firm that is also in the Harte Hanks dataset, resulting in a total of 68 firms. The contributing firms and their non-contributing matches are shown in Table 4.1. The sections below detail how the three datasets are used to construct the variables discussed in the previous section. All monetary values are converted to 2013 dollars using an appropriate deflation index and are reported in millions of dollars. Some of the wording in this section is identical to that in Chapter 3, but is reprinted here for continuity purposes.

[97] The author is grateful to Greg Kroah-Hartman of the Linux Foundation for his assistance in collecting and aggregating this data.

Table 4.1 Contributing Firms and their Non-Contributing Matches

Contributor	Match based on CEM
AMD	Micron Technology
Analog Devices	Microchip Technology Inc
Atmel Corporation	Fairchild Semiconductor International
Concurrent Computer	Astro Med Inc
Conexant Systems	Veeco Instruments Inc
Cypress Semiconductor	International Rectifier Corporation
Digi International Inc.	SS&C Technologies Inc
EMC Corp.	Cummins Inc
Exar Corp	Parlex Corp
General Electric	Textron
Harris Corporation	Stryker Corp
Hewlett Packard	Tyco International Ltd
IBM	Apple Computer Inc
ITT Industries	Parker Hannifin Corp
Intel Corporation	Kyocera International Inc
LSI Logic	Plexus Corp
Maxim Integrated Products	Thomas & Betts Corp
Mentor Graphics	Jack Henry & Associates Inc
Microsoft Corp	Pfizer Inc
NEC	Alliance Data Systems Corp
Nokia	Whirlpool Corp
Novell	Harte Hanks Inc
Oracle Corp	Gannett Company Inc
Polycom Inc.	Tekelec
Rockwell Collins	Teleflex Inc
Silicon Graphics Inc. (SGI)	Evans & Sutherland Computer Corp
ST Microelectronics	Vishay Intertechnology
Symantec Corp	Leucadia National Corp
Synopsys Inc.	Merix Corp
Texas Instruments	Amkor Technology
Unisys Corp	Fiserv Inc
Xerox Corp	Raytheon
Xilinx Inc.	TriQuint Semiconductor Inc
Yahoo Inc.	CoStar Group

4.4.1 Variable Construction

Value-Added (VA_{it})

The dependent variable is constructed using a method consistent with prior literature (e.g.,

Dewan and Min, 1997; Brynjolfsson and Hitt, 2003; Huang, Ceccagnoli, Forman, and Wu, 2013). First, yearly operating costs (XOPR in Compustat) are deflated by the BLS Producer Price Index by stage of processing for intermediate materials, supplies, and components. Then deflated IT labor and non-IT labor (defined below) are both subtracted from the operating costs. The result is then subtracted from yearly sales (SALE in Compustat) deflated by the BEA Gross Domestic Product Price Index for gross output for private industries.

IT Expenditure (IT_{it})

Prior literature in the field constructs a combined measure of IT Expenditure that includes both the value of IT hardware at the firm and three times the value of IT labor at the firm due to the importance of IT labor being used for internal software development efforts, the result of which is a capital good (Brynjofsson and Hitt, 1996; Hitt and Brynjofsson, 1996; Dewan and Min, 1997; Huang, Ceccagnoli, Forman, and Wu 2013).[98] The primary analysis uses this combined measure of IT expense. However, to properly test Hypothesis 5, hardware and labor will be split into separate variables.

To calculate the value of IT Hardware, the market value of the IT stock is estimated by multiplying the number of PCs and Servers at the firm (from Harte Hanks[99]) by the average

[98] Ideally, the portion of the IT budget that is spent on software in addition to hardware would be included. However, software expenditures are combined with other capital expenditures in firm 10-K reporting. Therefore, while purchased software cannot be separated from other firm purchases, the cost of such software is captured in the non-IT Capital variable. Further, internal software development efforts will be captured in the IT Labor variable. This methodology is consistent with prior literature (e.g., Brynjofsson and Hitt, 1996; Huang, Ceccagnoli, Forman, and Wu 2013). Additionally, the high correlation between purchased software and hardware expenditures helps to mitigate concerns about not having software expenditure data.

[99] For most firms, Harte Hanks only surveys a sample of the sites within the firm. In such cases, the average number of PCs and Servers at the sites that are in the survey is multiplied by the total number of sites in the firm to obtain the total number of PCs and Servers in the firm. The same procedure is used for calculating the number of IT employees and the number of each type of operating system at the firm. Further, Harte Hanks reports an estimated

value of a PC or Server that year from The Economist Intelligence Unit Telecommunications Database. The BEA Price Index for computers and peripherals is then used to deflate this value. This method is consistent with prior work in this area (e.g., Brynjofsson and Hitt, 1996; Huang, Ceccagnoli, Forman, and Wu 2013).

The value of IT labor is calculated by taking the number of IT workers at each firm (from Harte Hanks[100]) and multiplying by the mean annual wage for all Computer and Mathematical Science Occupations[101]. The BLS Employment Cost Index for wages and salaries for private industry workers is then used to deflate this value.

Non-IT Capital (K_{it})

The K_{it} variable is constructed by taking the yearly Gross Total Property, Plant and Equipment (PPEGT in Compustat), deflating it by the BLS price index for Detailed Capital Measures for All Assets for the Private Non-Farm Business Sector, and then subtracting the deflated value of IT Hardware (defined above).

Non-IT Labor (L_{it})

Non-IT Labor is constructed using the total number of employees at the firm (EMP in Compustat) and subtracting the number of IT employees (from Harte Hanks) to obtain the total number of non-IT employees. This is then multiplied by the mean annual wage of all

---

range of the number of PCs and Servers at the firm. These ranges are used to confirm the imputations and adjustments are made as necessary.

[100] Harte Hanks reports the number of IT employees at each site as a range so the average value of the range is used. The ranges are 1-4, 5-9, 10-24, 25-49, 50-99, 100-249, 250-499, and 500 or More.

[101] Obtained from the Bureau of Labor and Statistics: http://www.bls.gov/oes/2009/may/oes_nat.htm#15-0000.

occupations[102] that year. The BLS Employment Cost Index for wages and salaries for private industry workers is then used to deflate this result. This method of calculation is consistent with prior studies on IT productivity (Bloom and Van Reenen, 2007; Bresnahan, Brynjolfsson, and Hitt, 2002; Brynjolfsson and Hitt 2003).

Open Source Software Usage

To measure the intensity of OSS usage at the firm, the number and type of operating systems used at the firm is measured. Although operating systems are certainly not the only OSS used at the firm, they are important and frequently indicate the wider use of OSS. Additionally, since the data on contributions to OSS (discussed below) is focused on Linux, an operating system, the spillover effect will likely be more prevalent. Further, the Harte Hanks survey asks firms what type of operating systems they use, but does not always capture other types of OSS. In addition to constructing a measure of OSS operating systems, I will construct more granular measures of non-pecuniary OSS, pecuniary OSS, and closed-source operating systems for use in predicting the propensity of a firm to contribute to OSS and are used for testing Hypothesis 2. These three measures ($non\_pecuniary\_OSS_{it}$, $pecuniary\_OSS_{it}$, and $closed_{it}$) are constructed by calculating the total number of each type of operating system at the firm (from Harte Hanks). The Harte Hanks data does not report the precise number of operating systems in use at a given firm. It does, however, report the different types of operating systems used at each site at the firm. These operating systems are classified into three categories: non-pecuniary OSS, pecuniary OSS, or closed source. Table 4.2 shows the OSS operating systems in the dataset.[103] All other

[102] Obtained from the Bureau of Labor and Statistics, for example the data for 2009 can be found here: http://www.bls.gov/oes/2009/may/oes_nat.htm#00-0000.

[103] Although some non-pecuniary OSS operating systems, such as Debian, are offered at a nominal pecuniary price by third-party vendors for the convenience of the distribution being pre-loaded on a CD or DVD, they are included

operating systems are labeled as "closed". Harte Hanks also reports whether each operating system is for a PC or a server as well as the total number of PCs and servers at each site. Therefore, for each site, the number of PC operating systems is evenly split over the total number of PCs at the site. The same is done for servers. This yields an estimate of how many instances of a given type of operating system exist at the site.[104] This is then aggregated to the firm level and divided by the number of sites at the firm in the Harte Hanks database to obtain an average per site. Finally, this average is multiplied by the total number of sites in the firm to obtain a firm-wide imputation of the number of each type of operating system in a manner that accounts for sites in a firm that are not captured in the Harte Hanks survey.

Table 4.2 Open Source Operating Systems

Pecuniary OSS Operating Systems	Non-Pecuniary OSS Operating Systems
Red Hat Linux	Berkeley Software Distribution (BSD)
SUSE Linux	Debian
SCO Linux	Conectiva
TurboLinux	Fedora
	FreeBSD
	Gentoo Linux
	Linux Kernel
	Mandrake Linux
	NetBSD
	OpenBSD
	Ubuntu

Because the number of operating systems in any of the three categories can potentially be zero (e.g., that category of operating system is not in use at the firm), one is added to the number

in the non-pecuniary column as the full distribution is downloadable for free via the distribution's website. Additionally, although Apple's Mac OS X is built on BSD, it behaves more like a closed operating system than one that is pecuniary, but built on OSS, like Red Hat. Robustness checks were run against this assumption with no change to the primary results.

[104] For firms that have no reported operating systems in a given year, the proportion of the three types of operating systems for the year prior and after are used to impute values for that year.

of operating systems in each category before taking the natural log as the natural log of zero is undefined. Although there are many firms that have zero non-pecuniary and pecuniary OSS operating systems, there is a high degree of skewness in these numbers (as shown in the descriptive statistics below). Therefore, adding a one before taking the natural log should not significantly bias the results.

Open Source Software Contribution ($Contrib_{it}$)

As discussed above, data on firm contributions to OSS comes from the Linux Foundation. Every contribution to the Linux Kernel is associated with a time stamp and an email address for the person who contributed it. Since the early 2000's, the Linux Foundation has asked its contributors what firm they work for, if the firm sponsors their contributions. Therefore, each individual contribution email can be linked to a firm if that contributor was being paid by their firm to contribute to Linux. This allows for the construction of the binary variable $Contrib_{it}$, which is 1 if the firm has ever contributed and 0 otherwise. The timestamp allows for the construction of the $Post_{it}$ variable, which is 1 if the firm contributed in year *t* or earlier. Further, the $Post_{it}$ variable for non-contributing firms is set to 1 if the matched contributing firm contributed in year *t* or earlier. For example, Digi International first contributed to Linux in 2006. The best matched firm for Digi International is SS&C Technologies, who did not contribute to Linux. Therefore, for both firms, the $Post_{it}$ variable is 1 when *t* is 2006 or later and is 0 when *t* is 2005 or earlier.

In addition to the binary contribution variable, more granular continuous variables are used as well. The unique email address attribution of code contribution allows for a measure of the

total number of contributors for a firm in a given year to be constructed. Relatedly, the total number of changes to the Linux code a firm made in a given year can be constructed as well. Finally, every code change that is made requires at least one experienced contributor to "signoff" on the code, ensuring that the code has been tested and will not cause problems with the existing code base. These signoffs are also marked with an email address allowing them to be counted and attributed to firms as well. Although all three of these measures are likely to be highly correlated, their scales are quite different and allow for a more granular understanding of the effects of contribution. Because of the high skew in these measures, the natural log will be taken before adding them into the regression analysis. Therefore, the value 0.00001 is added to each observation so that no values are zero, which has an undefined natural log.

4.4.2 Descriptive Statistics

Table 4.3 shows the descriptive statistics of the firms in the dataset. There are 779 firm/year observations from 68 firms in the dataset.[105] The ranges vary greatly for all variables and demonstrate the breadth of the firms in the sample. This breadth allows for results that are generalizable despite the smaller sample size. However, due to the Harte Hanks sampling methodology, larger firms are overrepresented in the sample and very small firms (e.g., startups) are not in the sample. Additionally, because of the reliance on 10-k data for financial information, all firms in the sample are public firms, which tend to be medium or large. For example, as shown in Table 4.3, the smallest company in the sample (Astro-Med Inc.) had sales of $36.6 million in its lowest selling year. Comparatively, the largest firm (General Electric) had sales of $180 billion. Therefore, results should be interpreted as applying to medium and large

[105] This results in an average of 11.5 observations per firm. The panel is unbalanced because Harte Hanks does not survey every firm in every year. However, this is still a large enough number of observations per firm to conduct the full analysis.

firms. The firms in the dataset also have a wide range of the type and intensity of IT use. The investment in IT ranges from less than $1 million to $4.4 billion and the range for open source operating systems is from zero to 207,646. The intensity of contributing to Linux also has a large variance, with non-contributing firms having zero for all values and contributing firms having as many as 239 employees contributing up to 6,285 changes and signing off on up to 13,395 changes in one year.

Table 4.3 Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
$sales_{it}$	779	14792.52	28813.68	36.609	180929
VA_{it}	779	5256.864	10757.89	.526	79453.94
IT_{it}	779	215.441	391.059	.741	4483.846
K_{it}	779	7947.196	17557.91	3.129	130017.8
L_{it}	779	1970.068	3359.154	.659	20081.17
OSS_{it}	779	2802.169	13236.44	0	207646
Num. Contributors	738	6.069	23.311	0	239
Num. Changes	738	132.049	542.944	0	6285
Num. Signoffs	738	239.323	1108.282	0	13395

Table 4.4 shows the correlation matrix between the primary variables. As to be expected, K_{it} and L_{it} have a fairly high correlation with value-added productivity since they are the primary inputs into the production function. Additionally, it is notable that the correlation between OSS the contribution variables are fairly low, indicating that it is not necessarily the most intense users of OSS who are contributing back to it. Further, as expected, there is a high degree of correlation between the contribution intensity measures of number of contributors, changes, and signoffs.

Table 4.4 Correlation Matrix

	VA_{it}	IT_{it}	K_{it}	L_{it}	OSS_{it}	Num. Contribs	Num. Changes	Num. Signoffs
VA_{it}	1.0000							
IT_{it}	0.6214	1.0000						
K_{it}	0.8475	0.5572	1.0000					
L_{it}	0.8249	0.6947	0.7517	1.0000				
OSS_{it}	0.2374	0.5266	0.2150	0.3453	1.0000			
Num. Contributors	0.3586	0.2764	0.3351	0.4412	0.3543	1.0000		
Num. Changes	0.2930	0.2142	0.2885	0.3105	0.3597	0.9269	1.0000	
Num. Signoffs	0.2589	0.1741	0.2693	0.2557	0.3268	0.8729	0.9717	1.0000

4.5 Results

As mentioned above, the results of the CEM match of contributing firms with non-contributing firms are presented in Table 4.1. This section first presents the results of the propensity score analysis and the resultant inverse-probability weighting of the dataset. Then the analysis for the effect of contributing to OSS on the productivity impact of using OSS is presented. Finally, the results for examining the spillover effect from contributing to OSS on the productivity of all IT investments are presented.

4.5.1 Propensity to Contribute to OSS

As discussed previously, propensity scores are used to estimate the likelihood a firm contributes to OSS based on observable characteristics of the firm. A firm's contribution to OSS in a given year is predicted based on the three primary input variables (IT_{it}, K_{it}, L_{it}) as well as the three constructed variables estimating the number of non-pecuniary OSS, pecuniary OSS, and closed source operating systems at the firm ($non\_pecuniary\_OSS_{it}$, $pecuniary\_OSS_{it}$ and $closed_{it}$). These additional variables help to account for the technology usage decisions of the firm. Traits of the firm that are unobservable through a firm's financial reports, such as

management quality, may also have an impact on the firm's propensity to adopt. However, the CEM matching process discussed above helps to account for this potential confounder.

The results of the propensity estimation are shown in Table 4.5. These results show there is a positive relationship between non-IT capital (K_{it}) and contribution to OSS and a negative relationship between non-IT labor (L_{it}) and contribution to OSS. However, both of these relationships are weak and only significant at the 10% level. However, there are strong and positive relationships between the usage of both non-pecuniary and pecuniary OSS and the likelihood of contributing. This is not surprising given the fact that a firm's incentives to contribute to OSS are low if they are not using OSS. Further, although the use of the various types of operating systems was included in the CEM process, these results help show the importance of using inverse-probability weighting in addition to the CEM. Although interesting, it is difficult to interpret these results as causal due to the inherent endogeneity and potential omitted variable bias. However, they allow for the construction of the inverse-probability weighting discussed above, such that the remaining results are adjusted for sample bias and can be interpreted in a more causal manner.

Table 4.6 shows the resulting improvement of the balance in the sample after applying the IPW. Panel A shows the covariate balance without weighting. The t-statistics indicate that the adopting firms in the sample are significantly different from those that are non-adopters when comparing the three primary production inputs and their use of operating systems. Panel B shows the covariate balance after weighting. Here, the balance is much better and for all inputs the balance drastically improves. While the weighting does not perfectly address the differences

between the contributing and non-contributing firms, it does help add balance to the sample, allowing for a more causal interpretation of the regression coefficients.

Table 4.5 Predicting Contribution to OSS

DV: Binary Contribution to OSS	1
Model	Probit
IT Expense (IT_{it})	0.018 (0.071)
Non-IT Capital (K_{it})	0.365* (0.202)
Non-IT Labor (L_{it})	-0.367* (0.222)
$non\_pecuniary\_OSS_{it}$	0.127*** (0.036)
$pecuniary\_OSS_{it}$	0.141*** (0.036)
$closed_{it}$	-0.127 (0.106)
Constant	-0.066 (0.714)
Number of firm/year observations	779
Pseudo - R^2	0.132
Wald chi^2	23.28

***p<.01, **p<.05, *p<.1. Standard errors are clustered at the firm level. All variables are the natural log of the underlying variable.

Table 4.6 Covariate Balance

	Panel A			Panel B		
	Unweighted Sample			Weighted Sample		
	Contributors	Non-Contributors	t-stat	Contributors	Non-Contributors	t-stat
IT Expense (IT_{it})	265.184	162.943	3.68	232.530	196.374	1.37
Non-IT Capital (K_{it})	11000.0	4488.049	5.45	8940.516	5857.567	3.12
Non-IT Labor (L_{it})	2574.187	1332.475	5.24	2327.878	1548.152	3.63
$non\_pecuniary\_OSS_{it}$	1446.959	386.808	3.67	1006.695	603.015	1.86
$pecuniary\_OSS_{it}$	3363.791	295.486	3.42	2502.282	501.521	2.93
$closed_{it}$	21000.0	12000.0	2.45	21000.0	15000.0	1.64
Number of firm/year observations	400	379		400	379	

Values reported are the means of the contributing or non-contributing firms. Panel A presents the unweighted OLS regression of the given variable on contribution to OSS by the firm. Panel B presents the weighted OLS regression of the given variable on contribution to OSS by the firm.

Table 4.7 Benefits of Contribution to OSS

DV: Value-Added (VA_{it})	1	2	3	4	5
IT Expense (IT_{it})	0.069** (0.030)	0.069** (0.034)	0.063* (0.032)	0.063* (0.032)	0.058* (0.032)
Non-IT Capital (K_{it})	0.174** (0.080)	0.174** (0.080)	0.146* (0.081)	0.150* (0.083)	0.155* (0.080)
Non-IT Labor (L_{it})	0.843*** (0.089)	0.843*** (0.088)	0.884*** (0.092)	0.877*** (0.088)	0.869*** (0.086)
Open Source Usage (OSS_{it})		-0.000 (0.023)	-0.008 (0.027)	-0.001 (0.030)	0.034 (0.033)
Contributor ($Contrib_i$)			0.159 (0.186)	0.317 (0.302)	0.498 (0.329)
Post Contribution ($Post_{it}$)			-0.264 (0.170)	-0.276 (0.188)	-0.154 (0.185)
($Contrib_i * OSS_{it}$)				-0.038 (0.042)	-0.098* (0.056)
($Contrib_i * Post_{it}$)				-0.077 (0.175)	-0.501* (0.282)
($OSS_{it} * Post_{it}$)				0.023 (0.027)	-0.030 (0.033)
($Contrib_i * OSS_{it} * Post_{it}$)					0.113** (0.050)
Constant	0.204 (0.213)	0.204 (0.210)	0.140 (0.219)	0.120 (0.225)	0.051 (0.347)
Standard Error Type	Clustered	Clustered	Clustered	Clustered	Clustered
Year Control?	Yes	Yes	Yes	Yes	Yes
N	779	779	779	779	779
r2	0.888	0.888	0.892	0.893	0.894

***p<.01, **p<.05, *p<.1. All variables are the natural log of the underlying variable. All regressions are OLS models and are weighted with inverse-probability weightings based on the propensity of the firm to contribute to OSS.

4.5.2 Benefits of Contribution to OSS

Table 4.7 presents the estimation results of the primary specifications (equations 2, 3, and 4 from above). Column 1 reports the coefficients for the baseline three-factor productivity model including IT, non-IT capital, and non-IT labor. These coefficients are all consistent with prior research on the productivity of IT (Brynjolfsson and Hitt, 2003; Huang, Ceccagnoli, Forman, and Wu, 2013; Tambe and Hitt, 2012). Column 2 adds the measure of OSS usage at the firm. The

effect is not distinguishable from zero, although this likely stems from the measurement and endogeneity concerns that can make the direct measurement of this effect difficult (Nagle, 2015). Column 3 adds the additional individual variables $Contrib_i$ and $Post_{it}$, and column 4 adds in the interactions between these two variables and the usage of OSS. Column 5 adds the triple interaction between $Contrib_i$, OSS_{it}, and $Post_{it}$, which is the primary coefficient of interest. The positive and significant coefficient of 0.113 indicates that within the same period, firms who contribute to OSS gain 11% more productive value from the use of OSS than firms who do not contribute. This adds substantial support to Hypothesis 1, and shows that firms who contribute to OSS do indeed gain more productive value out of using OSS.

Table 4.8 shows the main results when the open source operating systems are split into pecuniary and non-pecuniary. As discussed above, not all OSS is free and the pecuniary OSS tends to be more opaque than the non-pecuniary. As shown in column 4, the coefficient on the triple interaction between contribution, non-pecuniary OSS, and the time period after the firm starts contributing is positive and significant, while the similar triple interaction for pecuniary OSS is not. This adds support to Hypothesis 2, that contribution to OSS allows firms to gain more from non-pecuniary OSS than from pecuniary OSS.

Table 4.8 Benefits of Contribution: OSS Breakdown by Type

DV: Value-Added (VA_{it})	1	2	3	4
IT Expense (IT_{it})	0.074* (0.048)	0.063 (0.051)	0.081 (0.050)	0.079 (0.051)
Non-IT Capital (K_{it})	0.174** (0.082)	0.143* (0.084)	0.176** (0.077)	0.181** (0.077)
Non-IT Labor (L_{it})	0.846*** (0.087)	0.888*** (0.090)	0.841*** (0.080)	0.831*** (0.079)
Non-Pecuniary Open Source Usage (NP_OSS_{it})	-0.007 (0.024)	-0.014 (0.024)	0.005 (0.029)	0.053* (0.029)
Pecuniary Open Source Usage (P_OSS_{it})	-0.004 (0.060)	0.001 (0.058)	-0.041 (0.072)	-0.094 (0.070)
Contributor ($Contrib_i$)		0.153 (0.154)	-0.732 (0.443)	-1.228* (0.644)
Post Contribution ($Post_{it}$)		-0.272 (0.186)	0.248 (0.365)	-0.187 (0.557)
$(Contrib_i * Post_{it})$			-0.271 (0.212)	0.454 (0.724)
$(Contrib_i * NP\_OSS_{it})$			-0.056 (0.035)	-0.130** (0.050)
$(NP\_OSS_{it} * Post_{it})$			0.030 (0.034)	-0.066* (0.035)
$(Contrib_i * P\_OSS_{it})$			0.137*** (0.063)	0.227** (0.093)
$(P\_OSS_{it} * Post_{it})$			-0.053 (0.049)	0.021 (0.076)
$(Contrib_i * NP\_OSS_{it} * Post_{it})$				0.149*** (0.053)
$(Contrib_i * P\_OSS_{it} * Post_{it})$				-0.127 (0.096)
Constant	0.230 (0.243)	0.153 (0.243)	0.395 (0.346)	0.711 (0.578)
Standard Error Type	Clustered	Clustered	Clustered	Clustered
Year Control?	Yes	Yes	Yes	Yes
N	779	779	779	779
r2	0.889	0.892	0.897	0.899

***p<.01, **p<.05, *p<.1. All variables are the natural log of the underlying variable. All regressions are OLS models and are weighted with inverse-probability weightings based on the propensity of the firm to contribute to OSS.

4.5.3 Spillovers Benefits to All IT Usage

Tables 4.9 and 4.10 show the results of estimation equation 5, which explores the relationship between contributing to OSS and the productivity returns of all IT. As shown in columns 1 and 2 of Table 4.9, there is a small, but positive coefficient for the triple interaction between contributing firms and their total IT expenditure in years after their first contribution. Similarly, columns 4 and 5 show the spillover effect on IT labor, rather than IT labor and hardware combined. Here the coefficient is again positive, indicating that firms who contribute to OSS may receive higher productivity from their IT employees. However, in all of these cases, the coefficients are not significant at the 10% level. Therefore, it is difficult to say with certainty that this relationship exists and the evidence to support Hypothesis 3 and Hypothesis 5 is limited. However, this could be partially due to the fact that the measure of contribution in this analysis is rather limited. Table 4.10 uses a more granular measure of contribution intensity, rather than a simple binary variable. In all three measures of contribution intensity – number of contributors, number of code changes, and number of code signoffs – there is a positive coefficient on the interaction with the IT expense at the firm. These coefficients are all significant at the 10% level, which offers support for Hypothesis 4, that the more firms contribute, the greater the spillover to the productivity of all IT will be.

Table 4.9 Spillovers from Contribution to IT Usage

DV: Value-Added (VA_{it})	1	2	3	4	5
IT Expense (IT_{it})	0.054 (0.041)	0.054 (0.058)			
IT Capital (ITK_{it})			0.053 (0.050)	0.066 (0.046)	0.066 (0.069)
IT Labor (ITL_{it})			0.022 (0.039)	0.010 (0.048)	0.010 (0.075)
Non-IT Capital (K_{it})	0.154*** (0.043)	0.154* (0.078)	0.165*** (0.046)	0.137*** (0.048)	0.137 (0.087)
Non-IT Labor (L_{it})	0.869*** (0.057)	0.869*** (0.089)	0.844*** (0.056)	0.875*** (0.059)	0.875*** (0.092)
Open Source Usage (OSS_{it})	-0.008 (0.011)	-0.008 (0.025)		-0.009 (0.012)	-0.009 (0.027)
Contributor ($Contrib_i$)	-0.029 (0.220)	-0.029 (0.293)		0.098 (0.161)	0.098 (0.235)
Post Contribution ($Post_{it}$)	-0.097 (0.192)	-0.097 (0.242)		-0.131 (0.143)	-0.131 (0.143)
($Contrib_i * Post_{it}$)	-0.180 (0.296)	-0.180 (0.390)		-0.144 (0.215)	-0.144 (0.286)
($Contrib_i * IT_{it}$)	0.062 (0.057)	0.062 (0.084)			
($ITC_{it} * Post_{it}$)	-0.030 (0.045)	-0.030 (0.065)			
($Contrib_i * IT_{it} * Post_{it}$)	0.018 (0.071)	0.018 (0.093)			
($Contrib_i * ITL_{it}$)				0.042 (0.057)	0.042 (0.079)
($ITL_{it} * Post_{it}$)				-0.039 (0.044)	-0.039 (0.060)
($Contrib_i * ITL_{it} * Post_{it}$)				0.020 (0.069)	0.020 (0.090)
Constant	0.186 (0.274)	0.186 (0.399)	0.410 (0.266)	0.380 (0.289)	0.380 (0.424)
Standard Error Type	Robust	Clustered	Robust	Robust	Clustered
Year Control?	Yes	Yes	Yes	Yes	Yes
N	779	779	779	779	779
r2	0.893	0.893	0.888	0.892	0.892

***p<.01, **p<.05, *p<.1. All variables are the natural log of the underlying variable. All regressions are OLS models and are weighted with inverse-probability weightings based on the propensity of the firm to contribute to OSS.

Table 4.10 Spillovers from Contribution Intensity to IT Usage

DV: Value-Added (VA_{it})	1	2	3	4	5	6
IT Expense (IT_{it})	0.086*** (0.024)	0.137*** (0.037)	0.087*** (0.024)	0.127*** (0.033)	0.087*** (0.024)	0.126*** (0.033)
Non-IT Capital (K_{it})	0.182*** (0.043)	0.183*** (0.042)	0.182*** (0.043)	0.183*** (0.043)	0.182*** (0.043)	0.183*** (0.043)
Non-IT Labor (L_{it})	0.821*** (0.055)	0.817*** (0.054)	0.821*** (0.055)	0.817*** (0.054)	0.821*** (0.055)	0.817*** (0.055)
Open Source Usage (OSS_{it})	-0.002 (0.012)	-0.004 (0.011)	-0.003 (0.011)	-0.005 (0.011)	-0.003 (0.011)	-0.005 (0.011)
Contribution Intensity ($Contrib\_Int_{it}$)	0.001 (0.006)	-0.022 (0.016)	0.002 (0.005)	-0.016 (0.013)	0.002 (0.005)	-0.016 (0.013)
($Contrib\_Int_{it} * IT_{it}$)		0.005* (0.003)		0.004* (0.003)		0.004* (0.003)
Constant	0.224 (0.250)	0.021 (0.274)	0.241 (0.246)	0.081 (0.259)	0.243 (0.246)	0.085 (0.259)
Measure of Contribution Intensity	Number of Contributors	Number of Contributors	Number of Changes	Number of Changes	Number of Signoffs	Number of Signoffs
Year Control?	Yes	Yes	Yes	Yes	Yes	Yes
N	738	738	738	738	738	738
r2	0.894	0.894	0.894	0.894	0.894	0.894

***p<.01, **p<.05, *p<.1. All variables are the natural log of the underlying variable. All regressions are OLS models and are weighted with inverse-probability weightings based on the propensity of the firm to contribute to OSS. All models use heteroskedastic robust standard errors.

4.6 Conclusion

The results of this study show that contributing to the production of a crowdsourced digital good allows a firm to learn how to better capture value from the use of that good. This learning has a spillover effect that also allows the firm to better capture value from all of its IT investments. These findings help to explain the seemingly irrational behavior of firms who pay their employees to contribute to the development of public goods that their competitors can use. Digging further into these main effects shows that the absorptive capacity increase due to contributions to OSS is stronger for investments in non-pecuniary OSS than in pecuniary OSS. This makes sense since the non-pecuniary OSS is generally more open, and therefore allows for

learning to be more easily applied. Further, the spillover effect on all IT investments is stronger for firms who contribute more than those who contribute less. Identification concerns due to the endogenous firm decision to contribute to OSS are addressed through the use of coarsened exact matching and inverse-probability weighting. This allows for a more causal interpretation of the coefficients.

This study adds important insights to the literature in both the IT and organizational learning fields. In doing so, it also extends the growing literature on user innovation and crowdsourcing by shining light on the benefits for firms when they engage with external parties to produce public goods. Unlike the classic absorptive capacity studies that show how firms may increase their productivity from integrating *external* resources by investing in *internal* development, this study shows that firms can enhance their ability to gain productive value from *internal* resources by investing in *external* development.

Despite the in-depth analysis of the effect of contributing to OSS development, open questions remain. Limitations of the dataset do not allow for the measurement of the importance of contributing to crowdsourced digital goods for the productivity of very small firms and start-ups since the firms in this study are all public. Further, data on contributions to OSS only comes from one open source project, Linux. Although Linux is the largest open source project, there are thousands of other projects that firms contribute to. Any learning that occurs from contribution to smaller open source projects goes un-captured in this study and an examination of this effect is left for future research.

References

Altman, E., Nagle, F., & Tushman, M. (2015). Innovating without Information Constraints: Organizations, Communities, and Innovation When Information Costs Approach Zero. In Oxford Handbook of Creativity, Innovation, and Entrepreneurship, edited by Michael A. Hitt, Christina Shalley, and Jing Zhou. Oxford University Press.

Aksoy-Yurdagul, D. (2015). The Impact of Open Source Software Commercialization on Firm Value. *Industry & Innovation*, 22(1), 1-17.

Aksoy, D., Fosfuri, A., & Giarratana, M. (2011). The Impact of Open Source Software on Firm Value. *Proceedings of the DRUID Society Conference, 2011.*

Aral, S., & Weill, P. (2007). IT assets, organizational capabilities, and firm performance: How resource allocations and organizational differences explain performance variation. *Organization Science*, *18*(5), 763-780.

Athey, S., & Ellison, G. (2014). Dynamics of Open Source Movements. *Journal of Economics & Management Strategy*, *23*(2), 294-316.

Attewell, P. (1992). Technology diffusion and organizational learning: The case of business computing. *Organization Science*, *3*(1), 1-19.

Barua, A., Kriebel, C. H., & Mukhopadhyay, T. (1995). Information technologies and business value: An analytic and empirical investigation. *Information systems research*, *6*(1), 3-23.

Baldwin, C., & Von Hippel, E. (2011). Modeling a Paradigm Shift: From Producer Innovation to User and Open Collaborative Innovation. *Organization Science*, *22*(6), 1399–1417.

Benkler, Y. (2002). Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale Law Journal*, 369-446.

Bloom, N. & J. Van Reenen. (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics*. **122**(4) 1351-1408.

Brandenburger, A. M., & Nalebuff, B. J. (2011). *Co-opetition*. Random House LLC.

Bresnahan, T.F., E. Brynjolfsson, & L.M. Hitt. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*. **117**(1) 339-376.

Brynjolfsson, E., & Hitt, L. (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management Science*, *42*(4), 541-558.

Brynjolfsson, E., & Hitt, L. M. (2003). Computing productivity: Firm-level evidence. *Review of economics and statistics*, *85*(4), 793-808.

Brynjolfsson, E., McAfee, A., Sorell, M., & Zhu, F. (2008). Scale without mass: business process replication and industry dynamics. *Harvard Business School Technology & Operations Mgt. Unit Research Paper*, (07-016).

Byrne, D., Oliner, S., & Sichel, D. (2013). Is the information technology revolution over? *Available at SSRN 2240961*.

Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: the two faces of R & D. *The economic journal*, 569-596.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly*, 128-152.

Casadesus-Masanell, R., & Llanes, G. (2011). Mixed Source. *Management Science*, *57*(7), 1212–1230.

Cole, S.R., & Hernan, M.A. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168 (6), pp. 656-664.

Daniel, S., Agarwal, R., & Stewart, K. (2006). An absorptive capacity perspective of open source software development group performance. *ICIS 2006 Proceedings*, 59.
Dewan, S., & Min, C. K. (1997). The substitution of information technology for other factors of production: A firm level analysis. *Management Science*, *43*(12), 1660-1675.
Doraszelski, U., & Jaumandreu, J. (2013). R&D and productivity: Estimating endogenous productivity. *The Review of Economic Studies*, *80*(4), 1338-1383.
Fichman, R. G., & Kemerer, C. F. (1997). The assimilation of software process innovations: An organizational learning perspective. *Management Science*, *43*(10), 1345-1363.
Forman, C. (2005). The corporate digital divide: Determinants of Internet adoption. *Management Science*, *51*(4), 641-654.
Forman, C., Goldfarb, A., & Greenstein, S. (2005). How did location affect adoption of the commercial Internet? Global village vs. urban leadership. *Journal of Urban Economics*, *58*(3), 389-420.
Forman, C., Goldfarb, A., & Greenstein, S. (2008). Understanding the inputs into innovation: Do cities substitute for internal firm resources?. *Journal of Economics & Management Strategy*, *17*(2), 295-316.
Fosfuri, A., Giarratana, M. S., & Luzzi, A. (2008). The penguin has entered the building: The commercialization of open source software products. *Organization science*, *19*(2), 292-305.
Garcia-Morales, V. J., Ruiz-Moreno, A., & Llorens-Montes, F. J. (2007). Effects of technology absorptive capacity and technology proactivity on organizational learning, innovation and performance: An empirical examination. *Technology Analysis & Strategic Management*, *19*(4), 527-558.
Gordon, R. J. (2003). *Hi-tech innovation and productivity growth: does supply create its own demand?* (No. w9437). National Bureau of Economic Research.
Greenstein, S., & Nagle, F. (2014). Digital Dark Matter and the Economic Contribution of Apache. *Research Policy* 43, pp.623-631. (Prior version released as National Bureau of Economic Research Working Paper 19507).
Griffith, R., Redding, S., & Van Reenen, J. (2003). R&D and absorptive capacity: Theory and empirical evidence*. *The Scandinavian Journal of Economics*, *105*(1), 99-118.
Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, 92-116.
Harhoff, D., Henkel, J., & Von Hippel, E. (2003). Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. *Research Policy*, *32*(10), 1753-1769.
Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4* (pp. 475-492). NBER.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.
Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, *35*(7), 953-969.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.
Hitt, L. M., & Brynjolfsson, E. (1996). Productivity, business profitability, and consumer surplus: three different measures of information technology value. *MIS Quarterly*, 121-142.

Hogan, J.W., & Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13, pp 17-48.
Horvitz, D.G., & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, pp. 663-685.
Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2013). IT Knowledge Spillovers and Productivity: Evidence from Enterprise Software. *Available at SSRN 2243886*.
Huber, M. (2013). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 2013.
Iacus, S. M., King, G., & Porro, G. (2011a). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, mpr013.
Iacus, S. M., King, G., & Porro, G. (2011b). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.
Iansiti, M., & Clark, K. B. (1994). Integration and dynamic capability: evidence from product development in automobiles and mainframe computers. *Industrial and corporate change*, *3*(3), 557-605
Jorgenson, D. W. (2001). Information technology and the US economy. *The American Economic Review*, *91*(1), 1-32.
Jorgenson, D. W., Ho, M. S., & Stiroh, K. J. (2005). Productivity, Volume 3: Information Technology and the American Growth Resurgence. *MIT Press Books*, *3*.
Knott, A. M. (2008). R&D/returns causality: Absorptive capacity or organizational IQ. *Management Science*, *54*(12), 2054-2067.
Lakhani, K., Lifshitz-Assaf, H., & Tushman, M. (2012). Open innovation and organizational boundaries: the impact of task decomposition and knowledge distribution on the locus of innovation in *Handbook of Economic Organization: Integrating Economic and Organization Theory*, A. Grandori (ed.), Edward Elgar Publishing, Northampton, MA, pp. 355-382.
Lerner, J., Pathak, P. A., & Tirole, J. (2006). The dynamics of open-source contributors. *The American Economic Review*, 114-118.
Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal of Industrial Economics*, *50*(2), 197–234.
MacCormack, A. (2003). Evaluating Total Cost of Ownership for Software Platforms: Comparing Apples, Oranges, and Cucumbers. AEI-Brookings Joint Center for Regulatory Studies Related Publication, April 2003.
MacCormack, A., Rusnak, J., & Baldwin, C. Y. (2006). Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Science*, *52*(7), 1015-1030.
McElheran, K. S. (2014). Delegation in Multi-Establishment Firms: Adaptation vs. Coordination in I.T. Purchasing Authority. *Journal of Economics & Management Strategy*, 23 (2), 225-258.
Nagle, F. (2015). Crowdsourced Digital Goods and Firm Productivity: Evidence from Free and Open Source Software. Harvard Business School Working Paper 15-062.
Tambe, P., & Hitt, L. M. (2012). The Productivity of Information Technology Investments : New Evidence from IT Labor Data. *Information Systems Research*, *23*(3), 599–617.
Tambe, P., Hitt, L. M., & Brynjolfsson, E. (2012). The Extroverted Firm: How External Information Practices Affect Innovation and Productivity. *Management Science*, *58*(5), 843–859.

Von Hippel, E., & Von Krogh, G. (2003). Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science*, *14*(2), 209-223.
West, J., & Gallagher, S. (2006). Challenges of open innovation: the paradox of firm investment in open-source software. *R&d Management*, *36*(3), 319-331.
West, J., & Lakhani, K. R. (2008). Getting clear about communities in open innovation. *Industry and Innovation*, *15*(2), 223-231.
Wooldridge, J. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1, pp. 117-139.
Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281-1301.
Young, R. & Johnson, D. R., (2009). A Comparison of Four Methods for Handling Missing Secondary Respondent Data. *Paper presented at the American Sociological Association Annual Meeting.*

Appendix A: The Shape of the Server Economy

This section details the sample of web servers in this study by examining two questions. First, is this software widely used in the US economy? Is this evidence consistent with a core premise of this study, that server software plays an integral role in the Internet in many locations and at many companies and in many applications? Second, is there evidence that the use of Apache and Microsoft software differ? Do both Apache and Microsoft software appear in many of the same locations and firms?

How do we observe if the use of server software is widespread or concentrated in a small number of locations? To examine these questions, we correlated the IP addresses for our server software against lists of IP locations maintained by MaxMind.com.[106] We computed both state and county numbers. While no state dominates use of server software far out of proportion with its population, for the sake of brevity, we show only one table. This is for server software and market shares for counties in the US.

Table A.1 shows the results for the top 25 counties, where the number of observations are large enough to lend confidence to the results. It lists the 25 counties with the most servers, ranking them in order. It also shows how large a share the servers in that county comprise of the total number of all servers. It then shows that county's rank in terms of Apache servers, and the share of the total of each of the three most common server software platforms, Apache, Microsoft and ngnix.

[106] http://www.maxmind.com/app/geolitecity, accessed December, 2011.

Table A.1 Top 25 Counties for Server Software Use

Total server Rank	ST	County	Share of servers	# Servers	Rank Apache	Share of Apache	Share MS	Share ngnix
1	OH	Franklin	0.0642	4129	1	0.0775	0.0442	0.003
2	AZ	Maricopa	0.0636	4091	2	0.0576	0.0784	0.0109
3	CO	Arapahoe	0.0529	3407	4	0.0519	0.0588	0.0018
4	IL	Cook	0.0437	2813	3	0.0548	0.0221	0.0573
5	TX	Dallas	0.0432	2778	6	0.0416	0.0414	0.105
6	TX	Harris	0.0397	2553	5	0.0479	0.0252	0.0296
7	CA	Los Angeles	0.0376	2419	7	0.04	0.0316	0.0573
8	WA	King	0.0301	1937	8	0.0291	0.0255	0.1177
9	CA	Orange	0.0256	1649	10	0.0245	0.0216	0.1056
10	GA	Fulton	0.0234	1503	9	0.026	0.0196	0.0091
11	NY	Kings	0.0196	1261	11	0.0226	0.0144	0.0163
12	TX	Bexar	0.017	1097	14	0.0169	0.0148	0.0519
13	CA	Santa Clara	0.0165	1059	16	0.0141	0.0186	0.0465
14	PA	Allegheny	0.0152	980	12	0.0222	0.0035	0.0006
15	DE	New Castle	0.0132	848	30	0.0068	0.0255	0.0066
16	MA	Middlesex	0.0126	811	17	0.0138	0.0101	0.0151
17	MI	Ingham	0.012	774	13	0.0171	0.0033	0.0042
18	CA	San Bernardino	0.0109	699	15	0.0146	0.0039	0.0115
19	VA	Fairfax	0.0106	684	21	0.0105	0.0109	0.0097
20	MO	St. Louis	0.01	645	19	0.0107	0.0083	0.0151
21	PA	Lackawanna	0.0091	587	20	0.0105	0.0033	0.0507
22	FL	Broward	0.0089	575	23	0.0087	0.0099	0.003
23	CA	San Diego	0.0085	547	22	0.0089	0.0083	0.0012
24	UT	Utah	0.008	513	18	0.0119	0.0012	0
25	PA	Delaware	0.007	449	24	0.0082	0.0053	0

The data in Table A.1 show that twenty-five counties account for approximately 60% of servers in the United States. While there is some evidence of concentration of servers in large and populous counties, there is no evidence of concentration in a few locations, such as Seattle, Boston, New York, or Santa Clara. Server software is widely used and in many locations, symptomatic of its importance as an integral piece of the Internet.

Table A.1 also shows the contribution of each county to the total share of Apache, Microsoft and ngnix use, and lists the ranking of the county in terms of Apache software share. These twenty five counties account for approximately 64% of Apache software, 50% of Microsoft server software, and 72% of ngnix software. Once again, this is evidence that server software is widely used and in many locations.

In addition, the ranking for use of Apache is very similar to the ranking for all server use. This is not surprising, since Apache comprises the largest component of total servers in use. Also contributing is another factor. Microsoft and Apache software do not differ tremendously in the extent of deployment within locations. The results for the top 100 counties are positively correlated. The number of servers deployed to Apache and Microsoft are correlated at .86. This last fact also reinforces the observation that arose in the data about counties about the lack of isolation. The market shares for Apache and Microsoft server software are roughly proportionate to one another in different counties. The correlation with ngnix is much lower, .48 for Apache and .42 for Microsoft. This is partial evidence that ngnix differs from the other two.

Now we consider an additional question: Do the data show evidence of isolated use? The presence of such isolation would be evidence that the deployment of Apache and Microsoft software occurs in vastly different locations or companies, which would arise if these were not substitutes for one another.

To address this second set of questions we match the IP addresses with information about the top level and second level domain names. This is obtained using the nslookup tool, which is a

standard feature of Linux. The following tables isolate attention to the three most common servers, Apache, Microsoft and ngnix.

Table A.2 shows the market share for different server software among different types of users, using the top-level domain names. This table shows server software long ago left its academic and government roots. The table shows that the majority of server software is used by organizations that register under TLD com, the most popular TLD, particularly for firms in the hosting business, who are very common users of this software. The second most common TLD is net, reflecting the importance of networking firms as users of server software in the US economy. The two originators of the Internet, the public military network (arpa) and the research network in universities (edu), account for only 9% of Apache and Microsoft server use.

Table A.2 Server use Among Top Level Domain Names

Rank	TLD	Share of Apache	Share MS	Share ngnix
1	com	0.5741	0.5131	0.7398
2	net	0.2320	0.2803	0.1714
3	arpa	0.0609	0.0488	0.0197
4	edu	0.0293	0.0434	0.0047
5	org	0.0254	0.0431	0.0236
6	info	0.0184	0.0065	0.0039

Tables A.3, A.4, and A.5 dig a bit deeper into the market shares for the top deployment of server, ranked by the contribution to the Apache total. As with Table A.1, no single firm dominates the deployment of server software, albeit a few firms have especially large server farms. Each of the tables ranks the listing in terms of the organization's contribution to the Apache total, and in each case it lists the top 15 organizations (16 in edu due to a tie).

Among the top 15 organizations there is only mild evidence of specialization. Many organizations deploy both Microsoft and Apache servers and many use ngnix as well. Some firms only use Apache and nothing else, especially within com, but this is not found in net and edu. This appearance may be a partial artifact of showing only 15 organizations. The correlation between Apache and Microsoft server use for the top 100 users within the com group is .75, which is evidence that users of software from one source tend to be users of both, and roughly in similar scales.

Table A.3 represents 40% of Apache server use, 17% of Microsoft use and 62% of ngnix. That suggests two conclusions. First, ngnix users are disproportionately drawn from Apache users. Second, it also shows that server use is quite spread out.

Table A.3 Server use Among Top 15 Second Level Domain Names Among Com

Rank	SLD	Share of Apache	Share MS	Share ngnix
1	theplanet	0.0855	0.0316	0.0343
2	softlayer	0.076	0.0616	0.1143
3	amazonaws	0.0559	0.0159	0.2069
4	dreamhost	0.0284	0	0.1154
5	cloud-ips	0.0244	0.0116	0.0766
6	bluehost	0.0229	0	0
7	ubiquityservers	0.0205	0.0057	0.0034
8	Rr	0.0161	0.0451	0.0011
9	myhostcenter	0.0134	0	0.0011
10	Linode	0.0132	0	0.0731
11	ecommerce	0.0132	0	0
12	mailengine1	0.0077	0	0
13	hostmonster	0.0073	0	0
14	nocdirect	0.007	0.0001	0
15	gridserver	0.0065	0	0

Tables A.4 and A.5 show the results from a similar exercise, now for net and edu. The results are very similar to those found in Table A.3. The top 15 organizations among net users account for 52% of the deployed Apache software within that group, and 43% and 50% of Microsoft and ngnix users within that group. Once again, there is little evidence of specialization. Among the top 100 users the correlations in the deployment of Apache and Microsoft server software is 78%.

Table A.4 Server use Among Top 15 Second Level Domain Names Among Net

Rank	SLD	Share of Apache	Share MS	Share ngnix
1	Secureserver	0.2964	0.1266	0.0143
2	Comcast	0.0441	0.0496	0.0095
3	Hostnoc	0.0393	0.0086	0.3429
4	comcastbusiness	0.0241	0.0918	0.0095
5	Verizon	0.0219	0.0476	0.0048
6	Sbcglobal	0.0169	0.0376	0.0048
7	Carpathiahost	0.0146	0.0015	0
8	Lstn	0.0102	0.0031	0.019
9	turnkeyinternet	0.0099	0.0007	0.0143
10	Cox	0.0092	0.0352	0.0048
11	Steadfastdns	0.0089	0.0011	0.0048
12	Securesites	0.0083	0.0013	0
13	Qwest	0.008	0.019	0
14	Scent	0.0075	0.0018	0.0333
15	Slicehost	0.0067	0	0.0429

The top 16 organizations among the edu users account for just 28% and 26% of Apache and Microsoft server software use respectively, reflecting the widespread use among many universities, albeit, universities are not a large fraction of server use in the United States. This group represents 50% of ngnix use, however, once again, showing that ngnix use is more concentrated, and largely drawn from large Apache users.

The evidence for specialization is stronger for this special group than for either com or net. Among the top 50 edu users the correlations in the deployment of Apache and Microsoft server software is only 17%. This arises because many universities tend to be small (the fiftieth ranked university in this data is CUNY and it has only 19 servers). Most universities tend to have large investments in either one or another server, albeit it often is no more than a few dozen.

Table A.5 Server use Among Top 15 Second Level Domain Names Among Edu

Rank	SLD	Share of Apache	Share MS	Share ngnix
1	utpa	0.0549	0.0162	0
2	utexas	0.0224	0.0068	0
3	mit	0.0191	0.0014	0.1667
4	wisc	0.0179	0.0122	0
5	stanford	0.0168	0.0054	0.1667
6	psu	0.0157	0.1664	0
7	northwestern	0.0157	0.0027	0
8	columbia	0.0146	0.0041	0.1667
9	vt	0.0135	0.0041	0
10	umn	0.0135	0.0054	0
11	duke	0.0135	0.0014	0
12	umich	0.0123	0.0081	0
13	harvard	0.0123	0.0054	0
14	uchicago	0.0112	0.0027	0
15	ucsd	0.0112	0.0041	0
16	usc	0.0112	0.0149	0

In summary, server software is widely used in the US economy, as one would expect if it plays an integral role in the Internet in many locations and at many companies and in many applications. In addition, both Apache and Microsoft software appear in many of the same locations and at the same using organizations. The absence of evidence showing isolated use is consistent with the premise that the two are substitutes for one another.

Appendix B: Substitutability of Apache and IIS

The insights into the data in Appendix A do not end the discussion about substitution between Apache and IIS. When considering substitution, it is also important to compare the boundaries and functionality of the products.

When facing a decision to utilize a web server other than the Apache HTTP Server, businesses must consider a number of other costs associated with this substitution. Such costs often result from any switching between open and closed systems (Scacchi, 2002, Zhu, Kraemer, Gurbaxani, and Xu, 2006), but are especially relevant for a technology as important as a web server. Although there are other free options for web servers, the Apache community is by far the largest community supporting any of the open source web servers (and one of the most widely used open source projects after Linux). Substituting a different open source web server for Apache HTTP Server alters the ecosystem that comes along with the software. A change in the software results in a loss of the large network of users and contributors who can be called upon for support. Additionally, because web server products exhibit network effects and Apache has already gained dominance in the web server market, most system engineers are only familiar with the Apache HTTP Server, and utilizing a different open source product can lead to a need to retrain engineers.

Another difference between the two is that IIS only runs on Microsoft Windows, while Apache HTTP Server can run on a variety of different operating systems, including Windows. This results in the added expense of purchasing the Windows operating system, as discussed above, to run IIS, whereas HTTP Server can be run on any operating system.

Compatibility with development languages is another area of differentiation for the two web servers. Active Server Pages on the .NET Framework (ASP.NET) is a web application framework produced by Microsoft that allows for the development of dynamic web sites and applications. It is integrated by design into IIS, whereas it can be run on HTTP Server via an add-on module called PHP: Hypertext Preprocessor (PHP), an open source web application framework used for developing dynamic web sites and applications. PHP is designed to be easily integrated with an Apache server, but it can also be run on IIS. ASP.NET and PHP have different pros and cons as well; however, the choice of a web server often depends on the preferences of the web application developer, with ASP.NET being optimized for IIS and PHP being optimized for Apache.

Additionally, IIS is generally considered easier to use due to its graphical interface when compared to the command line interface of Apache. However, the graphical interface also utilizes a greater deal of system resources than a command line interface, and therefore it is difficult to configure a Windows system running IIS to run in a very lean fashion, while it is very easy to do this for a Linux system running Apache. Therefore, a large percentage of Windows system resources and power are often devoted to tasks other than serving web pages, whereas a Linux/Apache system can be configured to spend the majority of resources and power serving web pages.