Empirical Studies on the Social Structure of Knowledge

Myra Mohnen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Economics

University College London

September 1, 2016

I, Myra Mohnen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis applies micro–econometric techniques to examine the effect of social structure on knowledge. Chapter 2 investigates the role of mass migration in the passage of compulsory schooling laws. It provides qualitative and quantitative evidence that compulsory schooling laws were used as a nation–building tool to homogenise the civic values held by the culturally diverse migrants who moved to America during the "Age of Mass Migration". Our central finding is that the adoption of compulsory schooling by American-born median voters occurs significantly earlier in time in states that host many migrants who had lower exposure to civic values in their home countries and had lower demand for common schooling when in the US. Chapter 3 explores whether, and to what extent, the position in the coauthorship network of medical scientists matters for the productivity of a researcher. I use sudden and unexpected deaths of star scientists as exogenous shocks to the network thus providing a causal identification of the loss of a star on the productivity of a scientist. I characterise the heterogeneity in the impact of the death by exploiting the position of the deceased scientists. Following the death of a star, coauthors suffer on average a 8% decrease in annual publications and this effect can differ by up to 31% depending on the network position. Chapter 4 examines knowledge spillovers by measuring the relative intensity of patent citations in two technological fields for which clean and dirty inventions can be clearly distinguished: energy production (renewables vs. fossil fuel energy generation) and automobiles (electric cars vs. internal combustion engines). We develop a new methodology based on Google's PageRank algorithm to measure the social benefit of knowledge spillover. We find that clean technologies generate 40% higher spillovers than their dirty counterparts.

Acknowledgements

This dissertation is a reflection of the effort of many people besides myself and I am very grateful for all the support that culminated in this work.

First, I wish to thank Uta Schönberg for her steady guidance and supervision. Beyond her advices and encouragement, I appreciate the freedom she gave me in pursuing various projects. I am also indebted to Imran Rasul. In particular, I would like to thank him for the first chapter of this thesis and for his support and helpful comments during the job market process. I am grateful to Antoine Dechezleprêtre for his enthusiasm for research and constructive feedback.

My time at UCL was made enjoyable in large part due to my friends in London. My Ph.D. experience would not have been the same without our coffee breaks, shared lunches, visits to the pub and memorable trips. I owe a special thanks to Andreas for his constant support, patience and insightful comments.

Finally and most importantly, I am deeply grateful to my family for their continuous encouragement and support without which none of this would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis is to shed new light on the social structures generating knowledge. Examining the factors at the foundation of knowledge is difficult in part due to the difficulty of defining what is meant by knowledge and in part due to the fact that it seems impossible to obtain some forms of knowledge without interacting with others. In this thesis, knowledge takes a broad meaning and is examined in different forms: as an institution (i.e. compulsory schooling law) in chapter 2 and in the form of publications and patents in chapters 3 and 4 respectively. Each chapter focuses on different features of the social structure at the foundation of knowledge, ranging from the demography of the population to the structure and position of individuals within a community, and examines how these social arrangements can enhance or impede knowledge generation or transmission.

Educational systems provides an institutional infrastructure to provide knowledge and can be shaped by the composition of the population. Chapter 2 aims to explain the role of mass migration in the creation of compulsory schooling laws in the United States. By the mid-19th century, America was the best educated nation on Earth: significant financial investments in education were being undertaken and the majority of children voluntarily attended public schools. So why did American states start introducing compulsory schooling laws at this point in time? We provide qualitative and quantitative evidence that compulsory schooling laws were used as a nation-building tool to homogenize the civic values held by the tens of millions of culturally diverse

migrants who moved to America during the "Age of Mass Migration" between 1850 and 1914. Using state level data, we show the adoption of compulsory schooling laws occurred significantly earlier in states that hosted a subgroup of European migrants with lower exposure to civic values in their home countries. We present IV estimates based on a Bartik-Card instrument to address concerns over endogenous location choices of migrants. We then use cross-county data to show that these same subgroup European migrant had significantly lower demand for American common schooling pre-compulsion, and so would have been less exposed to the kinds of civic value instilled by the American education system had compulsory schooling not been passed. By providing micro-foundations for schooling laws, our study highlights the link between mass migration and institutional change, where changes are driven by the policy choices of native median-voters in the receiving country rather than migrant settlers themselves.

In the process of creating new ideas, scientists build on knowledge previously found. In fact, scientists are embedded in a network of scientists within which knowledge is shared. It is widely understood that many important bits of information flow through social relations and that certain positions provide better access to such information-flows. Each cluster within the network tends to contain different pools of information, which in turn implies that a scientist's access to information will be increasing in the number of cluster he or she can reach. The fact that certain network positions are more advantageous than others has important implications for the creation of knowledge and consequently, the productivity of a researcher. Chapter 3 empirically explores whether, and to what extent, the position in the coauthorship network matters for the productivity of a researcher. For this purpose, we exploit a comprehensive dataset covering all major medical publications since 1965. As coauthorships are the result of a purposeful matching likely to reflect the quality of a researcher, we consider the sudden and unexpected death of a coauthor as an exogenous shock to the coauthorship network. This framework identifies a causal impact of the loss of a coauthor on the productivity of surviving coauthors. The identification of sudden death in obituary records

creates a bias towards "star" scientists who have been to be mentioned in an obituary. We therefore create a pool of stars through propensity score matching. Through a difference–in–difference, we quantify the change annual publications of a researcher following the sudden and unexpected death of a star–coauthor relative to a matched researcher whose associated star–coauthor is still alive. We then characterise the heterogeneity in these effects by network position. Based on the idea that scientists each embody unique knowledge and that knowledge flows within the network, we propose a measure, called brokerage degree, based on the number of scientists further away one can only reached via a specific coauthor. In other words, brokerage degree measures how much a scientist depends on a coauthor to gain access to scientists further away. The particularly rich dataset allows to condition upon a wealth of factors that influence the productivity of a scientist such as the access to resources or characteristics of the local network. My results reveal heterogeneous peer effects: the death of scientists occupying positions that enable access to non–redundant knowledge leads to a larger decline in the productivity of their coauthors.

Knowledge spillover from innovative activities provide a case for government intervention in the market because private R&D investments are likely too low. An important example are climate change policies that typically try to support so called clean technologies that avoid greenhouse gas pollution and hamper dirty technologies that are associated with polluting emissions. Chapter 4 systematically compares knowledge spillovers in two technological fields for which clean and dirty inventions can be clearly distinguished: energy production (renewables vs. fossil fuel energy generation) and automobiles (electric cars vs. internal combustion engines). We provide consistent evidence that clean inventions generate up to 40% higher level of knowledge spillovers than their dirty counterparts. To establish this finding, we use patent citation and develop a new methodology based on Google's Page Rank. Two explanations stand out as drivers behind this clean advantage: clean technologies have more general applications, and they are radically new compared to more incremental dirty innovation. Our results imply that stronger public support targeted to clean R&D is warranted.

They also suggest that green policies might be able to boost economic growth if the factors leading to an under-provision of knowledge goods are more severe for clean knowledge.

Chapter 2

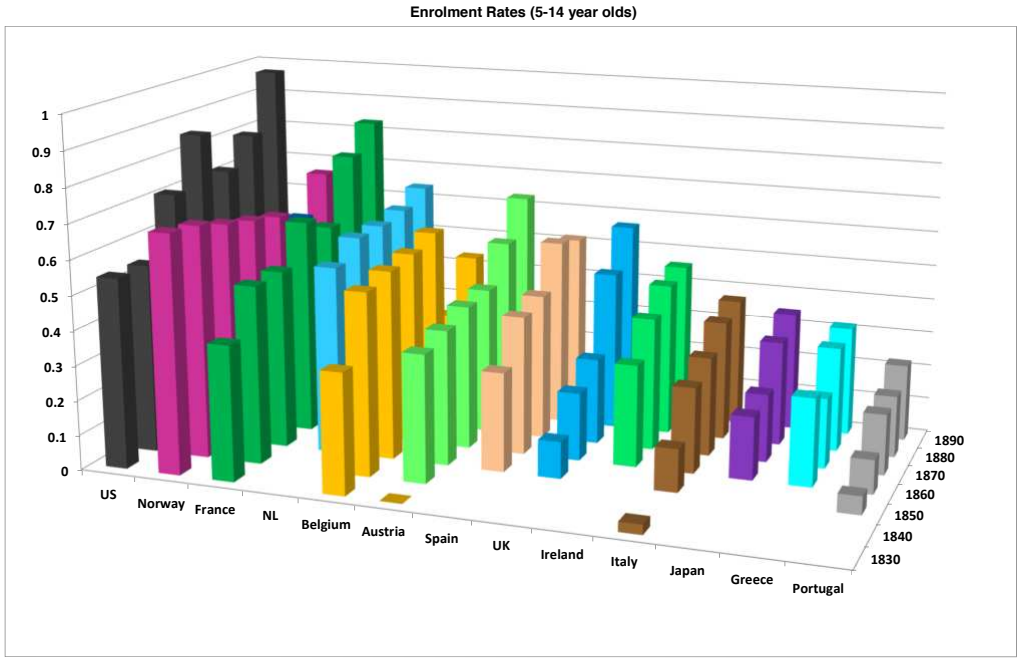
Nation-Building Through Compulsory Schooling During the Age of Mass Migration

2.1 Introduction

By the mid-19th century Americans were the best-educated in the world; financial investments into education were substantial and voluntary attendance was high [Landes and Solomon 1972, Black and Sokoloff 2006, Goldin and Katz 2008]. Figure 2.1 illustrates this point with newly assembled panel data on enrolment rates for 5-14 year olds from 1830 through to 1890 for the US and similarly developed nations. The Figure clearly shows that US rates were always above 50%, trending upwards, and diverging from other countries' from 1850 onwards.

This raises the puzzle that motivates our research question: why did US states start introducing compulsory schooling laws at a time when enrolment rates were high and trending upwards? These laws would not have been binding for the average American child, nor would they be binding for the marginal child and thus the driving force behind 'the educated American' [Goldin and Katz 2003, 2008]. Nor were they meant for blacks, as legislative caveats often effectively excluded them from schools even

Figure 2.1: The Educated American



Notes: Enrollment rates represent students enrolled in public and/or private schools for children aged 5-14. The enrollment rates are extracted from: (i) Lindert [2004] for Austria (1830-1870); Belgium (1830,1840,1860); France (1830,1840); Greece (1860); Ireland (1860); Italy (1830,1850,1860); Japan (1860); the Netherlands (1850, 1860); Norway (1830-1860,1890); Portugal (1850,1880); Spain (1850,1860,1890); the US (1830,1840) (ii) Flora et al. [1983] for Austria-Hungary (1891); Belgium (1850,1869,1881); Ireland (1890); Italy (1890); Norway (1870,1880); the UK (1850,1870-1890); Prussia (1871,1882,1891) (iii) Benavot and Riddle [1988] for Austria (1880); France (1870,1890); Greece (1870,1880); Ireland (1870,1880); Italy (1870,1880); Japan (1870-1890); the Netherlands (1870-1890); Spain (1870); the US (1870-1890). All other rates were calculated using enrollments from Banks and Wilson [2011] and the total population between 5-14 years old from Mitchell [2007a, 2007b] for France (1851,1861,1881); Greece (1889); Portugal (1864,1875,1890); Spain (1877,1887); the UK (1861); the US (1850,1860).

post-compulsion [Black and Sokoloff 2006, Collins and Margo 2006].[1]

The hypothesis we test is that compulsory schooling laws were used to expose the children of migrants who moved to America during the 'Age of Mass Migration' from 1850 to 1914 to American common schools and so instill them with the same civic values held by American-born children, who were voluntarily attending such schools in large numbers.

The idea that underpins this hypothesis is that a state provided formal education system can shape civic values. This idea is central in history studies of why European schooling systems developed at the time they did [Weber 1976, Ramirez and Boli 1987] and of why compulsory schooling was introduced in America [Cubberley 1947, Meyer *et al.* 1979, Engerman and Sokoloff 2005, Brockliss and Sheldon 2012]. The economics literature identifies 'civic values' as the values that: (i) make individuals more likely to take actions to improve the common welfare of their community [Alesina and

[1]A body of work has emphasized Americans became educated because of fiscal decentralization, public funding, public provision, separation of church and state, and gender neutrality [Goldin and Katz 2008]. Goldin and Katz [2003] document that compulsion accounts for at most 5% of the increase in high school enrolment over the period 1910-40, when such laws were being fully enforced.

Reich 2015]; (ii) underpin democratic institutions [Glaeser *et al.* 2007]; (iii) shape the acceptability of welfare transfers [Lott 1999]. Existing empirical evidence supports the idea that schools affects values via the content of curricula [Clots-Figueras and Masella 2013, Cantoni *et al.* 2015], and that those exposed to compulsory schooling are significantly more likely to be registered to vote, to vote, to engage in political discussion with others, to follow political campaigns and attend political meetings, as well as having higher rates of participation in community affairs and trust in government [Milligan *et al.* 2004].

Our research design exploits variation in civic values among European migrants from different countries, generated by differences in compulsory schooling laws in different countries. Since European countries introduced compulsory schooling to teach civic values, migrants exposed to compulsory state schooling in their country of origin were more likely to have been taught to civic values. Of course, the exact way in which compulsory state schooling operated would likely differ between each European country. What we emphasize here is the notion that most state education systems generally instill more values that underpin democratic institutions and trust in the state relative to the counterfactual of a non-state provided compulsory education system: in nineteenth century Europe this would have amounted to either attending a private school, a religious school, or not attending school altogether.

These ideas fix our identification strategy, namely we exploit the fact that the need for American-borns to teach civic values to European migrants was greater in US states where European migrants without historic exposure to compulsory state schooling in their home country, and hence with weaker civic values, were more numerous. We thus exploit differences in the composition of the migrant population, holding constant state characteristics that attract all migrants regardless of the compulsory schooling laws in their country of origin.[2]

[2]Of course, this logical chain requires two further conditions to hold. This first is that migrants transport their values with them, a hypothesis that has empirical support [Guinnane *et al.* 2006, Fernandez 2013, Fernandez and Fogli 2009]. This implies migrants' civic values depend on whether they have been exposed to some form of compulsory state education in their *home country*. The second condition is that parents transmit civic values, and other preferences, to their children. Again, this hypothesis also finds empirical support [Bisin and Verdier 2000, Dohmen *et al.* 2012].

Our analysis proceeds in three stages. We first present qualitative evidence to underpin the hypothesis that American society used compulsory schooling as the key policy tool to nation-build in response to mass migration. We show this was driven by the view that exposure to American public schools would instill the desired civic values among migrants, and a recognition that such values could be transmitted from children to their parents.

Second, we assemble a new data-set on the timing of compulsory schooling laws across European countries and we combine it with US Census data on state population's by country of origin to explain the timing of compulsory schooling laws across US states. We use survival analysis to estimate whether the cross-state timing of compulsory schooling laws is associated with the composition of migrants in the state. Our central finding is that American-born median voters pass compulsory schooling laws significantly earlier in time in US states with a larger share of migrants from European countries without historic exposure to compulsory state schooling in their country of origin: a one standard deviation in the share of these migrants doubles the hazard of compulsory schooling laws being passed in a decade between census years.

We show our core result to be robust to controlling for potentially confounding factors: literacy rates among adult migrants do not predict the cross-state passage of compulsion, and attendance rates of migrant children in some form of school, be they common or parochial school, only weakly impact the timing of compulsory schooling law. We also document that our main result is not driven by other forms of within-migrant diversity – not just differences in human capital but also in the religion, region of origin and English language proficiency of migrants. Finally, we show the main result holds across US regions, including in Southern and Western states.

The nation-building interpretation hinges on the comparison of the differential impact Europeans with and without historic exposure to compulsory state schooling in their home country have on the timing of such legislation in US states. Unobserved state factors that make a location equally attractive to both migrant groups do not bias this comparison. The chief econometric concern is that the process driving the location choices of migrants differ between these groups of European migrants. To address the

endogenous location choices of migrants, we present IV estimates using a control function approach in the non-linear survival model, based on a Bartik-Card instrumentation strategy: these show our main result to be robust to accounting for the endogenous location choices of migrants.

Finally, we set up a horse-race between the nation-building hypothesis and other mechanisms driving compulsory schooling, such as redistributive motives, or due to a complementarity between capital and skilled labor. We find some evidence for these alternatives, but none of them mutes the nation-building channel.

The third part of our analysis provides direct evidence on migrants' demand for American common schooling that underpins the nation-building efforts of American-borns. During the study period, many migrant groups faced a choice between sending their children to parochial schools (so based on religion), or to attend an American common school. Only if migrants' demand for American common schools was sufficiently low would compulsory schooling bind and be required to change migrants' civic values. We develop a probabilistic voting model of schooling provision to pin down the demand for American common schools in various migrant groups and test its predictions using cross-county data from 1890 that contains information on the most important investment into American common schools: teachers.

The revealed demands for American common schooling across migrant groups match up closely with the cross-state analysis. We find that European migrants from countries without long exposure to compulsory state schooling in their country of origin have significantly lower demand for American common schools relative to European migrants from countries with compulsory schooling. Furthermore, we document a significant convergence in demand for, and pupil attendance at, common schools between natives and both groups of European migrants when compulsory schooling laws are introduced. Hence compulsory schooling did indeed lead European migrants to be more exposed to the civic values being taught in American common schools, and this was especially so for Europeans from countries without historic exposure to compulsory state schooling in their country of origin. This cross-county analysis links tightly with the state-level analysis by establishing the counterfactual of what would have been

migrants' exposure to the kinds of civic values instilled through American common schools absent compulsory schooling laws.

Our finding that compulsory schooling laws were driven by the need to foster the assimilation of migrants complements the literature that studies the individual determinants of migrants' economic and cultural assimilation during the Age of Mass Migration [Abramitzky *et al.* 2014, 2016]. It is well recognized that during this period a wider set of educational policies collectively known as the *Americanization Movement*, encompassing language requirements in schools and ultimately citizenship classes targeted towards adult migrants and conducted by the US Bureau of Naturalization [Cubberley 1947, Carter 2009], were introduced primarily to assimilate migrants. While other disciplines have recognized that there have been periods of American history where the schooling system has been used to inculcate values among the foreign-born [Tyack 1976], our analysis contributes to the literature by showing nation-building motives drove the passage of compulsory schooling laws from the 1850s onwards, the first pillar of the *Americanization Movement*, and the legislative bedrock on which all later developments of the American education system have been built.[3]

Most broadly, we contribute to the literature linking the national origins of migrants and institutional change. The seminal work of Acemoglu *et al.* [2001] illustrates how colonial settlers from Europe established institutions that had long lasting impacts on economic development. Our analysis can be seen as 'Acemoglu *et al.* in reverse' as we analyze how the American-born population, from whom the median voter determines state-level policies such as compulsory schooling, best responded in public policy to large migrant flows from a set of culturally diverse countries.

The paper is organized as follows. Section 2 presents qualitative evidence on the use of compulsory schooling as a nation-building tool during the Age of Mass Migration. Section 3 develops a conceptual framework describing how compulsory schooling can be used to nation-build by homogenizing civic values between its native and immi-

[3]For example: (i) Native American children being sent to boarding schools in the early nineteenth century; (ii) the dispatch of American teachers to Puerto Rico and the Philippines after the Spanish-American war; (iii) attempts to democratize Germany and Japan after World War II. In more recent times, Arlington [1991] describes how English became the required language of instruction in Southern US states in 1980s, in response to mass migration from Latin American.

grant members. Section 4 describes the state level data and newly assembled database of compulsory state education laws by European country. Section 5 presents evidence linking the composition of migrant groups and the cross-state passage of compulsory schooling. Section 6 develops and tests a model of investment into education to estimate the relative demand for American common schools across migrant groups using county data. Section 7 concludes. The Appendix provides proofs, data sources and robustness checks.

2.2 Qualitative Evidence

That American society used compulsory schooling as a tool to nation-build during the Age of Mass Migration has been recognized in leading accounts of the development of the American schooling system written by educationalists [Cubberley 1947], sociologists [Meyer *et al.* 1979] and economic historians [Engerman and Sokoloff 2005, Brockliss and Sheldon 2012]. We highlight those pieces of qualitative evidence that inform our research design and presentation of quantitative evidence. We review how long-standing concerns over immigrants' assimilation informed political debate, and how the education system was viewed as the key policy tool to deal with these concerns. This was driven by the view that exposure to American common schools would instill the desired civic values among migrants, and a recognition that such values could then be transmitted from children to parents. We then provide evidence that nation-building motives informed the architects of the common school movement, both as a general principle and to foster the assimilation of migrants in particular. We conclude by providing some evidence of curricula in common schools, as this relates directly to the inculcation of civic values.

2.2.1 Migrants and Compulsory Schooling in the Political Debate

American society's anxieties over immigrant assimilation have been well documented for each wave of large-scale migration. These concerns became politically salient from the 1850s onwards, most famously in 1855 when the *Native American Party* (also referred to as the 'Know Nothing Party') elected six governors and a number of Congressional representatives. The party's core philosophy was one of 'Americanism',

consistently communicating the fear of the "unAmericanness" of immigrants [Higham 1988].

Much of the political debate and concerns of American-borns over migrants' assimilation are crystallized in the Dillingham Report, widely regarded as the most comprehensive legislative study on immigration ever conducted. The Report was drafted over 1907-11 by a Commission of senators, members of the House of Representatives and Presidential appointees. The Commission was established in response to concerns over the assimilation of migrants from Southern and Eastern Europe, and produced a 41-volume report, including a number of volumes solely dedicated to the role of the education system in the assimilation process. Throughout its work, the Commission highlighted the importance of *Americanizing* immigrants where the English language and learning were central to becoming an American citizen.

Moreover, the Commission explicitly recognized the role that children played in the wider long run process of inculcating values in the entire migrant population:[4] *"The most potent influence in promoting the assimilation of the family is the children, who, through contact with American life in the schools, almost invariably act as the unconscious agents in the uplift of their parents. Moreover, as the children grow older and become wage earners, they usually enter some higher occupation than that of their fathers, and in such cases the Americanizing influence upon their parents continues until frequently the whole family is gradually led away from the old surroundings and old standards into those more nearly American. This influence of the children is potent among immigrants in the great cities, as well as in the smaller industrial centers."* [p.42, Volume 29].

2.2.2 Nation Building and the American Common School Movement

The key individuals driving the American common school movement were Horace Mann (1796-1859), Henry Barnard (1811-1900) and Calvin Stowe (1806-1882). They were united in a belief that schooling was the instrument, "*by which the particularities*

[4]This view also matches with historic evidence on the inter-generational transmission of human capital, especially language skills, from children to parents [Ferrie and Kuziemko 2015].

of localism and religious tradition and of national origin would be integrated into a single sustaining identity" and could foster "*goals of equity, social harmony, and national unity*" [p9, p39, Glenn 2002].

Horace Mann is widely regarded as the most prominent figure of the common school movement, becoming the first secretary of the Massachusetts Board of Education in 1837 (the earliest adopter of compulsory schooling). He believed common schools would, *"promote moral education"* and *"unite the country by teaching common values"* [p147, p150, Jeynes 2007]. Like many advocates for the common school movement, he recurrently emphasized the link between education and the civic virtues necessary for effective participation in a democracy.

Henry Barnard was the secretary of the Connecticut Board of Education, and was very much influenced by what he had seen of the European education system. His motives for building the public school system have been described as follows: *"Despite the challenges that Barnard faced, he, like Mann, was tenacious in maintaining the view that the common school cause was for the good of the country. He believed that democracy and education went together "in the cause of truth, justice, liberty, patriotism, religion.""* [p154, Jeynes 2007].

Finally, Calvin Stowe was a key driver of the common school movement in the Midwest. Stowe, like Mann, believed moral education was the most important aspect of schooling and was also heavily influenced by what he saw of European education practices.[5]

It has been argued that all these central figures ultimately saw schools as the *key*

[5]When Calvin Stowe reported back to American education leaders about European practices, he emphasized that "*public education in Europe was having a civilizing effect on that continent because it was bringing Christianity and the teachings of democracy to the most remote parts, where despotism often ruled*" [Jeynes 2007]. Glenn [p100, 2002] writes, *"The influence of foreign models, especially that of Protestant states of the Continent, Prussia and the Netherlands, was of critical importance in shaping the goals and the arguments of he education reformers. It was through the nation-building role of popular schooling in those countries that key ideas of the Enlightenment and the French Revolution of 1789 because central elements of what was virtually a consensus program along elites in the United States throughout the century and a quarter beginning around 1830",* and, *"that the alternative model offered by England, where education remained essentially in the hands of private, ecclesiastical, and charitable enterprise until the 20th century, did not have more appeal suggests how strongly Enlightenment concerns for national unity and uniformity dominated the thinking of the leaders in the common school movement."*

tool for social control and assimilation. Certainly, advocates of common schools came to emphasize their role as an alternative to families to foster the assimilation of immigrant children. As Tyack [p363, 1976] argues, "*Advocates of compulsory schooling often argued that families–or at least some families–like those of the poor or foreign-born–were failing to carry out their traditional functions of moral and vocational training...reformers used the powers of the state to intervene in families to create alternative institutions of socialization*." One of the most noted advocates for common schools in Philadelphia was E.C.Wines best articulated the link between compulsory schooling, immigration and nation-building: *"We refer to that overflowing tide of immigration, which disgorges our shores its annual tens of thousands of Europe's most degraded population–men without knowledge, without virtue, without patriotism, and with nothing to lose in any election..Are these persons fit depositaries of political power? The only practicable antidote to this, the only effectual safe-guard against the other, the only sure palladium of our liberties, is so thorough an education of all our citizens, native and foreign, as shall nullify the dangerous element in immigration."* [p742-3, Wines 1851].

2.2.3 Compulsory Schooling and Civic Values

American educators wanted their schooling system to place relatively more emphasis on the role of schooling in shaping the character, values and loyalties of students as future participants in political and social life. This philosophy is what would have driven the civic values instilled into American-born children voluntarily attending schools in such high numbers (Figure 2.1) and would drive some of the legislative acts that introduced compulsory schooling, to also make explicit references to civic values.[6] In detailing how compulsory schooling laws were implemented to provide insights on the values to be taught, it is important to note that American school districts have always had a high degree of autonomy. This has led to considerable heterogeneity in practices, making it almost impossible to track curriculum changes over time by district [Goldin 1999a]. Subject to this caveat, we highlight the following.

[6]For example, in Connecticut the law states the curriculum must cover "US history and citizenship", and in Colorado it states that instruction "must cover the constitution".

First, the alternative source of education to common schools were parochial and private schools. According to Lindert [2004], 12% of all pupils were enrolled in such schools in 1880. Migrant specific shares are not available but were presumably higher given that the language of instruction in these schools was not necessarily English (and the figure aligns closely with the overall share of migrants in the population). In some cases, compulsory schooling laws required children to be taught in some public school.[7] In other cases, states regulated parochial and private schools by specifying standards they had to comply with to meet compulsory state schooling requirements. For instance, the standards set in Illinois and Wisconsin aroused fierce opposition because of their provisions that private schools teach in the English language and that they be approved by boards of public education [Tyack 1976].

Second, states differed as to whether English should be the main language of instruction. Some states imposed clear English language requirements early on, while in others bilingualism was first accepted and then banned from public schools.[8] Eventually the *Americanization Movement* led to further legislative iterations making language and instruction requirements more explicit [Lleras-Muney and Shertzer 2015]. This was ultimately followed by the introduction of citizenship classes targeted to foreign-born *adults* from 1915-16 onwards, that were in part conducted by the US Bureau

[7]For example, the Massachusetts law of 1952 states that, *"Every person who shall have any child under his control between the ages of eight and fourteen years, shall send such child to some public school within the town or city in which he resides..."*

[8]For example, a 1919 law in Minnesota reads: *"A school, to satisfy the requirements of compulsory attendance, must be one in which all the common branches are taught in the English language, from textbooks written in the English language and taught by teachers qualified to teach in the English language. A foreign language may be taught when such language is an elective or a prescribed subject of the curriculum, not to exceed one hour each day."* [Minnesota, Laws 1919, Ch. 320, amending Gen. Stat. 1914, sec. 2979 as described in Ruppenthal 1920]. Daniels [pp.159-60, 1990] discusses the variation across states: *"Beginning in 1839 a number of states, starting with Pennsylvania and Ohio, passed laws enabling (or in some cases requiring) instruction in German in the public schools when a number of parents, often but not always 50 percent, requested it, and these laws were copied, with inevitable variations, in most states with large blocs of German settlers. The Ohio law authorized the setting up of exclusively German-language schools. In Cincinnati this option was exercised so fully that there were, in effect, two systems, one English, one German, and, in the 1850s, the school board recognized the right of pupils to receive instruction in either German or English. In Saint Louis, on the other hand, the use of bilingualism was a device to attract German American children to the public schools. In 1860 it is estimated that four of five German American children there went to non-public schools; two decades later the proportions had been reversed. In Saint Louis all advanced subjects were taught in English. So successful was the integration that even before the anti-German hysteria of World War I, German instruction as opposed to instruction in the German language was discontinued."*

of Naturalization [Cubberley 1947]. These classes were designed to, "*imbue the immigrant with American ideals of living...and preparing them for citizenship*" [Carter 2009, p23-4]. In short, it is not that nation-building efforts ignored adult immigrants. Rather, as recognized by the Dillingham Report, policies to target immigrant children were prioritized and attempted earlier.

2.3 Conceptual Framework

To bridge between the qualitative and quantitative evidence, we present a framework to make precise the idea of how a society made up of native and migrant groups, with heterogeneity in values across groups, can use compulsory schooling to nation-build. The framework is closely based on Alesina and Reich [2015].

Consider a state comprised of: (i) American-borns, normalized to mass 1; (ii) newly arrived immigrants of mass $\gamma \leq 1$. Individuals have heterogeneous civic *values* represented by a point on the real line. Let $f(j)$ be the density of American-borns with values $j \in \mathbb{R}$, and $g(j)$ be the corresponding density among immigrants. Denote by d_{ij} the 'distance' between values i and j, $d_{ij} = |i-j|$, and let c denote private consumption. An American-born individual with values $i \in \mathbb{R}$ is assumed to have utility:

$$u_i = c - \int_{j\in\mathbb{R}} f(j)d_{ij}dj - \int_{j\in\mathbb{R}} g(j)d_{ij}dj. \tag{2.1}$$

The second term on the RHS of (2.1) measures the difference between her values and those of other American-borns; the third term measures the difference between her values and those of immigrants. American-borns thus prefer to live in a more homogeneous society in which individuals share values. This is an *intrinsic* preference held by natives: homogenizing the population might have other *indirect* benefits, but the underlying nation-building motive of natives is that they prefer to live with others that share their values.

To see how schooling might affect the homogeneity of values held in society, assume first that some voluntary schooling system is in place, attended by American-borns (as described in Figure 2.1). We assume the school curriculum matches the val-

ues of the median American, i_m. Attending school shifts individual values towards i_m by degree λ. Schooling can impact a variety of specific values [Lott 1999, Glaeser *et al.* 2007], and contemporary evidence suggests that the content of school *curricula* do indeed influence beliefs and values held later in life [Milligan *et al.* 2004, Clots-Figueras and Masella 2013, Cantoni *et al.* 2015]. The population then decides by majority rule whether to make this schooling system compulsory.

In line with our empirical setting, γ is sufficiently small so the median voter is an American-born.[9] As American-borns already attend school, the direct effect of implementing compulsory schooling is on the migrant population who are homogenized towards the values of the median American, i_m. Assuming a fixed cost of implementing (and enforcing) compulsory schooling, the policy increases the tax burden for all by an amount T. Hence the utility of an American with median values, i^m, if compulsory schooling were to be introduced is,

$$u_{i^m} = c - \int_{j\in\mathbb{R}} f(j) d_{i^m j} dj - \int_{j\in\mathbb{R}} g(j)(1-\lambda) d_{i^m j} dj - T. \tag{2.2}$$

Proposition 1 *Suppose all immigrants have values $j > i^m$ to the left of the median American, then a majority of Americans vote for compulsory schooling if and only if,*

$$\int_{j\in\mathbb{R}} g(j) d_{i^m j} dj \geq T/\lambda. \tag{2.3}$$

The Proof is in the Appendix.[10]

The framework makes precise that whether a state votes for compulsory schooling depends on: (i) how different the migrant population is from the median American, $d_{i^m j}$; (ii) the size of the migrant group, $g(j)$; (iii) the effectiveness of schooling in

[9] Figure A1 uses IPUMS 1880 census data (a 100% sample) to show that while migrants account for a sizeable share of each state's population, they remain a minority in each state. This fact also holds on subsamples that better reflect those eligible to vote, such as the share of men, those in the labor force, and those residing in urban areas. Hence, even if migrants themselves demanded compulsory schooling, they were not pivotal at the state level in determining the passage of such legislation.

[10] The assumption $j > i^m$ simplifies the algebra and best describes our setting. Allowing for overlapping preferences of Americans and migrants implies that if compulsion is introduced, this moves the values of some immigrant *further* from the preferences of some Americans. The condition under which the majority of Americans then vote for compulsory schooling depends on the entire distribution of preferences among them.

shifting preferences, λ; (iv) the fiscal cost of making schooling compulsory (and its enforcement), T.[11]

Section 4 details how we proxy the key measure, $d_{i^m j}$: pre-held civic values among migrants using their historic exposure to compulsory state schooling in Europe. Section 5 takes this to the data to explain the cross-state timing of compulsory schooling in US states. A necessary condition for natives to prefer to make schooling compulsory is because it binds on immigrants and so exposes them to American civic values. This is at the heart of the analysis in Section 6 that estimates the relative demand for American common schooling among immigrants and natives.

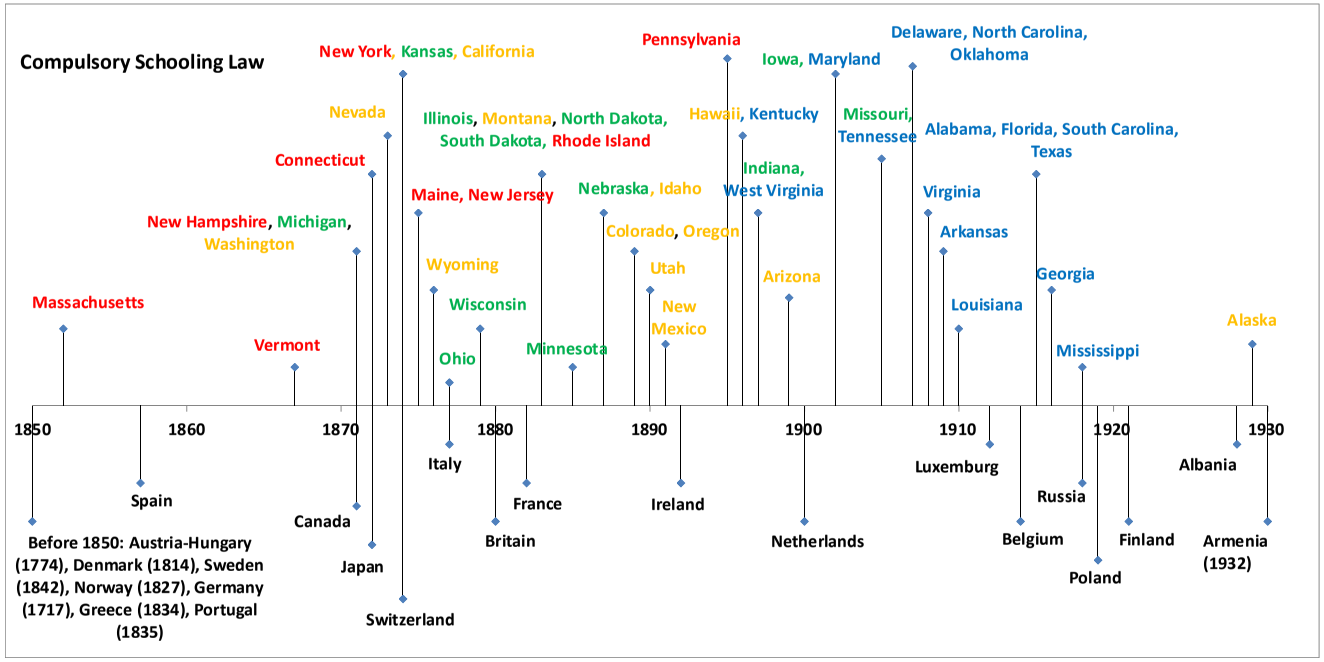
2.4 Data and Methods

The top half of Figure 2.2 illustrates the variation we seek to explain: the timing of compulsory schooling laws by US state, as coded in Landes and Solomon [1972]. This coding is our preferred source because it covers all states from the 1850s. A prominent alternative coding is that provided by Goldin and Katz [2003] (who extend the coding of Lleras-Muney [2002]). The Goldin and Katz [2003] data only covers the period from 1900 onwards, and so does not provide information on the 33 states that introduced compulsory schooling before 1900. For those 15 states that overlap between the Landes and Solomon [1972] and Goldin and Katz [2003] codings, we find the year of passage for compulsory schooling is identical for 13 states, and the differences are minor in the other two cases (Louisiana: 1912 vs. 1910; Tennessee: 1906 vs. 1905).[12]

[11]The costs of compulsory schooling laws can also be interpreted more broadly. For example, with compulsion, immigrant children would have had to reallocate time away from potentially more productive labor market work, to be exposed to the civic values only the state schooling system could provide *en masse*. Second, and related to the evidence in Section 6, there would be greater class sizes as a result for all children including American-borns.

[12]Table A1 shows further details on the passage of key child related legislation by state. There is variation across states in the ages for which compulsory school laws were binding: we do not exploit such variation for our analysis.

Figure 2.2: Timeline for Passage of Compulsory Schooling, by US State and European Country



RED = Northeast, GREEN= Midwest, YELLOW = West, BLUE = South

We focus on understanding what drove the *adoption* of compulsory schooling across states. The existing literature has focused on measuring the *impacts* of this legislation on various outcomes: a question for which the enforcement of compulsory schooling is more first order.[13]

To operationalize the conceptual framework, we need to identify a source of *within-migrant* diversity in values to match $d_{i^m j}$, that is the difference in civic values between Americans and migrants. Our strategy uses the fact that the European state schooling model was itself driven by the promotion of certain civic values and American educators were familiar with this. During the study period, civic values in many European countries and the US were aligned towards instilling values to underpin democratic institutions, to foster trust in the state and to promote the common good. This suggests a natural distinction between two types of European migrant: those from countries that had compulsory state schooling laws in place before the first US state (Massachusetts in 1852) and were thus more likely to be exposed to such civic values in their country of origin, and European migrants from countries that introduced compulsory state schooling after 1850 and were thus less likely to have been inculcated in civic values related to democracy and trust in the state, that were held and valued in American society.

For this purpose we have assembled a novel data-set on the timing of compulsory state schooling laws by European country, shown in the bottom half of Figure 2.2. The Appendix details the data sources underlying this coding. Figure 2.2 shows the European countries defined to have compulsory schooling in place by 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. The adoption of compulsory schooling in Europe is not perfectly explained by geography, language or religion. In particular, within each group of European countries that adopted com-

[13]Economic historians have argued that compulsory schooling laws were initially weakly enforced [Clay *et al.* 2012], and that they become more effective over time. In particular, there were gradual extensions in operation to cover: (i) the period of compulsory schooling each year; (ii) precise age and poverty requirements for children to attend; (iii) the application of schooling laws to private/parochial schools; (iv) increased requirements of cooperation from schools in enforcement; (v) the appointment of attendance officers, and then the institution of state supervision of local enforcement; (vi) and the connection of school-attendance enforcement with the child-labor legislation of States through a system of working permits and state inspection of mills, stores, and factories.

pulsory schooling pre and post 1850, there are countries in Northern, Southern and Eastern Europe, and countries where the main religion is Catholicism or Protestantism. This variation enables us to separately identify the impact on the cross-state passage of compulsory schooling of within-migrant diversity in values from differences along other dimensions.[14]

Of course, the exact way in which compulsory state schooling operated would likely differ between each European country. What we emphasize here is the notion that typical to most state education systems is that they generally instill values: (i) to make individuals likely to take actions to improve the common welfare of their community [Alesina and Reich 2015]; (ii) that underpin democracy and trust in the state [Glaeser *et al.* 2007]; (iii) shape the acceptability welfare transfers [Lott 1999]. We leave for future research a more detailed coding of the specific civil values promoted under each compulsory schooling system, that might then be further exploited in empirical work.

Table A2 also provides the earliest and latest dates by which compulsory schooling might reasonably be argued to have been passed in any country, given the sources cited and ambiguities/regional variations within a country (Table A3 discusses the coding for countries in which there is within-country variation in compulsory schooling). For our main analysis we focus on the dates shown in Figure 2.2. We later provide robustness checks on our results using these lower and upper bound dates of compulsory schooling.[15]

Finally, Table A4 probes the link between compulsory schooling laws and school enrolment rates *in Europe*, exploiting five secondary data sources. In each data set, we compare enrolment rates between countries with and without compulsion. Despite these sources differing in their coverage of countries, years, and the enrolment measure,

[14]This variation also ensures that individuals from both sets of countries arrive in each wave of mass migration to the US (starting with the first waves of migration from Northern Europe, followed by later waves of migration from Southern and Eastern Europe [Bandiera *et al.* 2013]. We also note that European countries without compulsory schooling have higher GDP per capita than those with compulsion, consistent with nation-building rather than economic development driving compulsion in Europe [Ramirez and Boli 1987]. The relative GDP per capita between the two types of European country remains almost fixed over the entire period.

[15]We define countries using pre-1914 borders, that can be matched into US census place of birth codes. Except for Canada and Japan, we were unable to find detailed sources for all non-European countries to accurately divide them into those with and without historic experience of compulsion.

three of them report significantly higher enrolment rates in countries with compulsory schooling than without. This supports the hypothesis that migrants from countries with compulsory state-provided education are likely to have been instilled with the kinds of civic values related to democracy and trust in the state, to a greater extent that children from countries where education would have been various non-state actors: private schools, religious schools or households themselves.[16]

2.4.1 Descriptives

We combine US Census data on state population by country of birth with our coding on the timing of compulsory schooling law by European country to compute for each state-year, the population share of migrants from European countries with and without compulsory schooling before 1850. Data limitations prevent us from dividing non-European migrants between those with and without compulsory schooling at home: thus they are grouped in one category throughout.

Figure A2 shows the share of the state population in each group (Europeans with and without compulsory state schooling in their country of origin, and non-Europeans), averaged across census years before the passage of compulsory schooling laws in each state. There is considerable variation in the size of the groups across US states: the share of Europeans with compulsory schooling ranges from .05% to 18%, the share of Europeans without compulsory schooling from .3% to 29%, the share of non-Europeans from .03% to 32%. Most importantly, the correlation between the migrant shares are positive but not high, allowing us to separately identify the response of American-born median voters to the presence of each group.

Table 2.1 compares the characteristics of the different migrant groups and Amer-

[16]These data make clear that even in European countries with compulsion, enrolment rates remained well below 100% on average, as with US states. Whether these differences in values then translate to differences in values held by Europeans that migrated to the US depends on the nature of migrant selection. The few studies that have examined the question for this period provide somewhat mixed evidence, and highlight that selection varies across entry cohorts and countries. For example, Abramitzky *et al.* [2012] link US and Norwegian census records to provide evidence on the negative selection of Norwegian migrants. At the same time, Abramitzky *et al.* [2014] document that on arrival to the US, the average migrant did not face a substantial occupation based earnings penalty, experienced occupational advancement in the US at the same rate as natives, and those migrants that left the US were negatively selected.

icans in state-census years before compulsory schooling is introduced. The first row describes the relative population share of each group and again highlights the considerable variation in these shares across US states in a given year, and the variation in shares within a state over time. The next two rows in Panel A highlight differences in human capital across groups. Among adults, the share of illiterates is significantly higher among Europeans from countries without compulsory schooling than among European-born adults from countries with compulsory schooling.[17] These differences are significant even conditioning on state fixed effects (Column 6). This is in line with the 'first stage' evidence provided in Table A4 comparing enrolment rates in Europe among countries with and without compulsory schooling. The next row in Table 2.1 shows these patterns persist across generations. Comparing enrolment rates in any type of school in the US (public or parochial) for children aged 8-14 in each group (the cohort for whom compulsory schooling was typically related to), these are significantly *lower* among migrants groups from European countries *without* compulsory schooling than for children from European countries with compulsory schooling in place by 1850. As expected both migrant groups trail behind the enrolment rates of American-borns, and enrolment rates of non-Europeans lie somewhere between the levels of the two European groups.

This evidence suggests that compulsory schooling laws might have been passed by US states to raise the skills of migrant children (that could be acquired through compulsion to attend any school), rather than to instill civic values among them (that could only be acquired through compulsion to attend a common school or requiring other schools to teach elements of the same curriculum). We disentangle these explanations by exploiting variation in enrolment rates within each European group, to see if enrolment rates *per se* drives the passage of compulsion, that would follow from the skills-based rather than values-based nation-building explanation.

The remaining rows of Panel A highlight that the two groups of European migrants

[17] Illiteracy rates among American-born adults are higher than for any of the migrant groups because migrants are much younger on average. This fact combined with the strong upward time trend over the 19th century in the educational attainment of Americans shown in Figure 2.1, means that their adult illiteracy rates of natives are higher than for migrants because older cohorts of American-borns are included.

do not significantly differ from each other on other characteristics including the share of young people in the group (aged 15 or less), labor force participation rates, the share of the group residing on a farm, and an overall measure of the groups economic standing in the US as proxied by an occupational index score available across US census years.[18]

[18]The score is based on the OCCSCORE constructed variable in IPUMS census samples. This assigns each occupation in all years a value representing the median total income (in hundreds of $1950) of all persons with that particular occupation in 1950.

Table 2.1: Characteristics of American-Borns and Immigrant Groups

Sample period for State Descriptives: Census years prior to the introduction of compulsory schooling law
Sample period for County Descriptives: 1880 (based on 100% census sample)
Columns 1 to 4: Mean, overall standard deviation (SD) in parentheses, between SD in brackets, within SD in braces
In Columns 5 and 6, p-values on t-tests are reported in brackets

	(1) American Born	(2) European Born from Countries that did NOT have CSL in 1850	(3) European Born from Countries that had CSL in 1850	(4) Non-European Foreign Born	(5) Test of Equality [Col 2 = Col 3]	(6) Within State Test of Equality [Col 2 = Col 3]
A. State Level						
Population (10,000s)	76.5	4.60	3.14	.862	[.300]	[.333]
	(81.8)	(9.91)	(5.89)	(1.75)		
SD Between States	[70.3]	[10.4]	[5.36]	[1.38]		
SD Within State (over census years)	{45.1}	{2.51}	{2.79}	{1.08}		
Share of Adults (aged 15+) that are Illiterate	.204	.102	.046	.166	[.008]	[.011]
	(.350)	(.074)	(.096)	(.225)		
Enrolment Rate (8-14 year olds)	.570	.297	.441	.331	[.011]	[.016]
	(.245)	(.326)	(.328)	(.368)		
Share Aged 0-15	.445	.081	.065	.156	[.160]	[.188]
	(.097)	(.066)	(.078)	(.162)		
Share in Labor Force	.305	.585	.609	.486	[.345]	[.378]
	(.108)	(.156)	(.200)	(.252)		
Share Residing on a Farm	.501	.225	.243	.261	[.215]	[.246]
	(.189)	(.180)	(.238)	(.274)		
Mean Occupational Score	18.2	21.1	22.2	19.4	[.153]	[.180]
	(2.94)	(3.90)	(7.14)	(7.36)		
B. County Level						
Share of County Population	.894	.041	.040	.025	[.822]	[.335]
	(.136)	(.057)	(.066)	(.072)		
SD Between States	[.121]	[.051]	[.049]	[.048]		
SD Within State (over counties)	{.085}	{.041}	{.043}	{.061}		

Notes: In Panel A, the unit of observation is the state-census year. All variables are constructed from the IPUMS-USA census data using individual weights. For each state, the sample period starts from 1850 and covers all census years prior to the introduction of compulsory schooling laws. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In Panel B, the unit of observation is the county in 1880. All variables are constructed from the IPUMS-USA 100% 1880 census sample. County populations are measured in shares. For both Panels, in Column 1, the American born are those whose recorded nativity is native born. In Column 2, the European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. All other European countries are included in Column 3. In the first row, populations are measured in 10,000s. Adults are defined to be aged 15 and above when defining the share of adults that are illiterate, and enrolment rates for 8-14 year olds are the share of this group that report being in school. The occupational score is a constructed variable from IPUMS-USA that assigns each occupation in all years a value representing the median total income (in hundreds of 1950 dollars) of all persons with that particular occupation in 1950. The occupational score thus provides a continuous measure of occupations, according to the economic rewards enjoyed by people working at them in 1950. Column 5 reports the p-value on a test of the null hypothesis that the values in Columns 2 and 3 are equal – this is derived from an OLS regression allowing standard errors to be clustered by region. Column 6 reports the p-value on the same test where we additionally control for state fixed effects.

2.4.2 Empirical Method

We use survival analysis to estimate the cross-state timing of the passage of compulsory schooling. We estimate the hazard rate, $h(t)$, namely, the probability of compulsory schooling law being passed in a time interval from census year t until census year $t+10$, conditional on compulsory schooling not having been passed in that state up until census year t. This approach allows for duration dependence in the passage of legislation by states (so that history matters), and corrects for censoring bias without introducing selection bias. The unit of observation is the state-census year where we use census years from 1850 to 1930. In the survival analysis set-up, 'failure' corresponds to the year of passage of compulsory schooling. As that is an absorbing state, state-years after compulsory schooling is passed are not utilized as they provide no information relevant to determine when compulsory schooling is actually passed. We first estimate the following Cox proportional hazard model:

$$h_s(t|\mathbf{x}_{st}) = h_0(t)\exp(\textstyle\sum_j \beta_j N_{st}^j + \sum_j \gamma_j X_{st}^j + \lambda X_{st}), \tag{2.4}$$

where the baseline hazard $h_0(t)$ is unparameterized, and t corresponds to census year. This model scales the baseline hazard by a function of state covariates. In particular, we consider how the composition of various migrant groups j in the state correlate to the passage of compulsory schooling. The division of population groups j we consider is between European migrants in the state from countries with and without historic exposure to compulsory state-provided education systems, as well as non-European migrants. N_{st}^j is the share of the state population that is in group j in year t: this is our key variable of interest; X_{st}^j includes the same group characteristics shown in Table 2.1. X_{st} includes the total population of the state, and the state's occupational index score, a proxy for the state's economic development.

The coefficient of interest is how changes in the composition of the state population group j affect the hazard of passing compulsory schooling laws, $\widehat{\beta}_j$. As population sizes across groups j differ, we convert all population shares N_{st}^j into effect sizes (calculated from pre-adoption state-census years). $\widehat{\beta}_j$ then corresponds to the impact of

a one standard increase in the share of group j in the state on the hazard of passing compulsory schooling law. We test the null that β_j is equal to one, so that a hazard significantly greater (less) that one corresponds to the law being passed significantly earlier (later) in time, all else equal.

The nation-building interpretation hinges on a comparison of $\widehat{\beta}_j$ between Europeans with and without historic exposure to compulsory state-provided education systems. The main econometric concern is that the process driving the endogenous location choices of migrants differs between groups, thus biasing the difference in $\widehat{\beta}_j$'s. We address such concerns using multiple strategies in Section 5.4.

2.5 Results

2.5.1 Baseline Findings

Table 2.2 presents our baseline results. The first specification pools foreign-borns into one group: we find that a one standard deviation increase in the share of the population that is foreign-born significantly increases the hazard rate of compulsory schooling being passed between two Census dates by 24%. Column 2 splits the foreign-born into European and non-Europeans, and the result suggests the presence of European migrants is significantly associated with the passage of compulsory schooling.

While similar results have been noted in the earlier literature studying the passage of compulsory schooling laws, Column 3 splits European migrants along the key margin relevant for the nation-building hypothesis. We find the presence of European migrants from countries that do *not* have historic experience of compulsory state schooling at home significantly brings *forward* in time the passage of compulsory schooling in US states: a one standard deviation increase in the population share of such Europeans is associated with a 64% higher hazard rate. In contrast, the presence of Europeans with a long history of compulsory schooling at home does not influence when compulsory schooling is passed by states. The effect sizes across these types of European migrant are significantly different to each other, as shown at the foot of the Table [p-value=.005].

Column 4 estimates (2.4) in full, so X_{st}^{j} further includes the enrolment rates of

8-14 year olds for American and the three migrant groups j (the age group for whom compulsory schooling in US states was most relevant for), and we present the impacts of these human capital related controls (in effect sizes) in addition to the coefficients of interest, $\widehat{\beta}_j$. Two key results emerge. First, the distinction between the types of European migrant is robust to controlling for other dimensions along which they differ [p-value=.004]. The magnitude of the effect remains large: a one standard deviation increase in the population share of Europeans without compulsory state schooling at home doubles the hazard of a US state passing compulsory schooling. Second, enrollment rates of migrants' children in the US have weak impacts on whether American-born voters introduce compulsory schooling. We note that higher enrollment rates among the children of natives speed up the adoption of the laws, as shown in the literature [Landes and Solomon 1972]. This might reflect the natural complementarity between American enrolment rates, namely, the extent to which American children are instilled in certain civic values in school will inevitably increase the returns to also instill the same values in migrant children using the same common schools.

To further document the link between compulsory schooling and the human capital of adult migrants, Table A5 repeats the specification in Column 4 of Table 2.2 but reports the full set of human capital related coefficients, where all covariates are measured in effect sizes. This highlights that higher illiteracy rates among adults in each group are not associated with the earlier passage of compulsory schooling. Indeed, states with less literate adult populations of American-borns and Europeans with exposure to state compulsory state education systems in their country of origin, adopt compulsory schooling significantly *later* in time, all else equal. This is evidence against the cross-state passage of compulsory schooling being driven predominantly by a desire by American-borns to skill the migrant population.

The nation-building explanation thus remains first order: the conceptual framework highlighted that American-borns have a desire to homogenize those migrants that are more distant from them in values, and the empirical evidence suggests it is the civic values held by migrants, as proxied by their historic exposure to compulsory state-provided education systems at home, rather than migrants' investment in the human

capital of their children in the US, or the skills among adults, that largely drives the cross-state passage of compulsory schooling.

Of course, the American median voter *could* have targeted those with compulsory schooling in their country of origin because state education systems inculcate country-specific identities that are not transportable across locations, and so those individuals are most in need of being re-indoctrinated with American values.[19] Yet, this is strongly rejected by the data. Rather, we find American-borns target those Europeans without historic experience of compulsory schooling in their country of origin (as well as towards non-Europeans who are also unlikely to have compulsory schooling back home). This is consistent with compulsory schooling being a nation-building tool because of its impact on civic values that were *common and transportable* across Europe and America in the nineteenth century.

Such portability of civic values is consistent with ideas that governments have incentives to compel citizens to go through the same state schooling system because, relative to a counterfactual world in which schooling is provided privately, through religious organizations or by households themselves, compulsory state schooling can instill civic values that help underpin democracy [Glaeser *et al.* 2007], trust in government and civic participation [Milligan *et al.* 2004], to shape common interests and goals [Lott 1999, Alesina and Reich 2015], or because state capacity is easier to raise in more homogeneous societies in which the common good is more easily identifiable and political institutions are inclusive [Besley and Persson 2010].

2.5.2 Robustness Checks

The Appendix documents the robustness of our core finding. Specification (2.4) exploits cross-country differences in whether migrants' country of origin had compulsory state schooling laws in place in 1850 or not. The first robustness check explores an alternative specification that exploits *within-country* variation over time, in exposure

[19] This would, for example, be in line with other explanations put forward for why societies compel citizens to go through the same schooling system, such as to build strong national identities in the face of external military conflict [Aghion *et al.* 2012]. The fact that this is rejected is reassuring given that the US context is not one in which the threat of external military conflict is likely to be the driving force behind compulsion.

Table 2.2: Immigrant Groups and the Passage of Compulsory Schooling Laws

Non parametric Cox proportional hazard model estimates, hazard rates reported
Robust standard errors; Populations shares and enrolment rates measured in effect sizes

	(1) Foreign	(2) European	(3) Historic Exposure to Compulsory Schooling	(4) Enrolment Rates
Share of the State Population that is:				
Foreign Born	1.24*			
	(.142)			
European Born		1.43**		
		(.226)		
From European Countries that did NOT have CSL in 1850			1.64***	2.15***
			(.225)	(.509)
From European Countries that had CSL in 1850			.988	.780
			(.122)	(.161)
Non-European Born		.998	.995	1.80***
		(.041)	(.035)	(.409)
Enrolment Rate of American-Borns				2.82**
				(1.39)
Enrolment Rate of Europeans From Countries that did NOT have CSL in 1850				.815*
				(.094)
Enrolment Rate of Europeans From Countries that had CSL in 1850				1.03
				(.153)
Enrolment Rate of Non-European Foreign-Borns				1.18
				(.235)
Group Controls	No	No	No	Yes
State Controls	No	No	No	Yes
European Groups Equal [p-value]			[.005]	[.004]
Euro Without CSL = Non-Euro [p-value]			[.001]	[.505]
Observations (state-census year)	230	230	230	230

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is equal to one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling is passed. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares from census-years prior to the introduction of compulsory schooling law. Robust standard errors are reported. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In Column 4 we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, and the share residing on a farm. We also control for the following state characteristics: the total population and the average occupational score of the population. We also control for the enrolment rate of 8-14 year olds among American borns (in effect sizes), and group specific enrolment rates for all European and non-European groups in the state (in effect sizes). At the foot of Column 3 onwards we report the p-value on the null hypothesis that the hazard coefficients are the same for the two European groups, and the p-value that the hazard coefficients are the same for the non-European immigrant groups and European borns from countries that did not have compulsory schooling in place in 1850.

to compulsory state schooling. To do so, we consider the impact of a rolling window of Europeans' exposure to compulsory schooling by examining whether the American median-voter is differentially sensitive to the presence of European migrants that have passed compulsory schooling at least 30 years ago, versus the presence of Europeans from countries that have either never passed compulsory schooling or passed it less than a generation ago. This highlights how American voters react differently over time to migrants from the same country, as that country becomes exposed to compulsory schooling at home. This helps further pin down that when passing compulsory school-

ing laws, American-born median voters across states are responding to the civic values held by European migrants, rather than some time invariant characteristic of European countries that had compulsion in place in 1850.

The result, in Column 1 of Table A6, demonstrates that with this definition, the sharp contrast between how American-borns react to different types of European migrant becomes even more pronounced: a one standard deviation increase in the population share of European migrants from countries that do *not* have more than a generation of exposure to compulsory schooling at home significantly increases the hazard by 2.31. In contrast, the presence of Europeans with compulsory schooling at home for at least one generation significantly reduces the hazard rate below one. These results highlight how American-born voters appear to react differentially over time to the *same* country of origin as that country's population accumulates experience of compulsory schooling.

Table A6 then shows the robustness of our main finding to additionally controlling for three classes of variable. First, we control for the passage of other legislation in US states, that might be complementary to, or pre-requisites for, compulsory schooling law. For example, child labor laws and the establishment of a birth registration system have been argued to be interlinked with compulsory schooling [Lleras-Muney 2002, Goldin and Katz 2003]. Second, we show the main result survives controlling for proxies for the states' progressivity. Third, we control for additional types of legislation passed in European countries: in particular we show our main result is robust to controlling for the presence of European migrants from countries with and without child labor laws in 1850, to rule out that such policy preferences drive migrants to sort into locations with like-minded Americans, rather than compulsory schooling being introduced as a nation-building tool by American-borns.

Table A5 shows our main result continues to hold using: (i) alternative econometric specifications, including imposing parametric structure on the underlying hazard, $h_0(t)$; (ii) alternative classifications of European countries with and without compulsory schooling, using the lower and upper bound limits of when compulsory schooling could have been introduced, shown in Table A2.

2.5.3 Spatial Variation

Figure 2.2 highlighted a clear spatial pattern in the adoption of compulsory schooling, with Southern and Western states trailing other regions. We thus address whether there could be a very different process driving compulsory schooling law in those regions.

Many Western states were admitted to the Union towards the end of the 19th Century, and passed compulsory schooling laws just before gaining entrance. Such states might have introduced compulsory schooling laws in order to enter the Union, rather than because of nation-building motives. On the other hand, the requirements for entering the Union in the US Constitution (Article IV, Section 3) make no explicit reference to any degree of modernization or institutional complexity that candidate states must have reached, and some educationalists have been explicit that the nation-building hypothesis is as relevant in Western states as others [Meyer *et al.* 1979].

In Southern states there was huge resistance to educating black children (before the Civil War it was illegal in many Southern states to teach slaves to read or write). It is however unclear whether this slowed down the adoption of compulsory schooling laws: typically caveats were included in compulsory schooling laws to ensure blacks did not benefit from compulsion, such as exemptions due to poverty or distance from the nearest public school [Lleras-Muney 2002, Black and Sokoloff 2006, Collins and Margo 2006]. A related concern however arises because during our study period, the Great Migration of Blacks occurred from Southern to urban Northern states (hence more closely matching the spatial patterns in Figure 2.2). However, this is unlikely to be related to the passage of compulsion because the migration of blacks occurred mostly between 1916 and 1930, well after compulsory schooling laws began to be introduced: pre-1910 the net migration of blacks was only .5mn [Collins 1997].[20]

To take these concerns to data, we first limit attention to states that are observed in all census years from 1850 to 1930. These comprise long established states in which the desire to nation-build might be stronger than in states that joined the Union more recently. The result, in Column 1 of Table 2.3, suggests that in long established states,

[20]Chay and Munshi [2013] document that an important pull factor for black migration to start in 1916 was the shutting down of European migration, that left labor supply shortages in Northern states. Prior to 1916 there is little evidence that European and black migration to states was interlinked.

American-born voters remain sensitive to the presence of European migrants from countries without a history of compulsory state schooling. The baseline result is also robust to restricting the sample to the 30 largest states by population (where over 90% of the US population resides): this limits attention to states with weaker incentives to introduce compulsory schooling to attract individuals. The estimated effect size rises because in the most populous states, a one standard deviation increase in European migrant groups corresponds to a far larger change in the absolute group number than in the baseline specification. Column 3 estimates the baseline specification excluding Western states: we continue to find the presence of European migrants from countries without a history of compulsory schooling to be significantly related to the cross-state timing of compulsion across states, and there to be a differential impact from Europeans with historic exposure to compulsory schooling at home [p-value=.000]. Column 4 then estimates (2.4) using only Western and Southern states: even in this subsample the nation-building explanation holds.

Table 2.3: Regional Variation in the Passage of Compulsory Schooling Laws

Non parametric Cox proportional hazard model estimates, hazard rates reported
Robust standard errors; Populations shares and enrolment rates measured in effect sizes

	(1) Established States	(2) Most Populous States	(3) Exclude Western States	(4) Only Western and Southern States
Share of the State Population that is:				
From European Countries that did NOT have CSL in 1850	3.16**	14.6***	5.55***	4.62**
	(1.64)	(14.2)	(2.50)	(2.94)
From European Countries that had CSL in 1850	1.52	.662	.857	.270**
	(.506)	(.205)	(.197)	(.167)
Non-European Born	1.73***	1.66**	1.37	1.60
	(.302)	(.413)	(.337)	(.512)
Group Controls	Yes	Yes	Yes	Yes
State Controls	Yes	Yes	Yes	Yes
European Groups Equal [p-value]	[.094]	[.004]	[.000]	[.016]
Euro Without CSL = Non-Euro [p-value]	[.201]	[.020]	[.004]	[.091]
Observations (state-census year)	187	153	186	141

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is equal to one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling is passed. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. Robust standard errors are reported. In Column 1 the 36 states that are observed in all 8 IPUMS census waves from 1850 to 1930 are included in the sample. These states are Alabama, Arkansas, California, Connecticut, Delaware, Florida, Georgia, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, New Hampshire, New Jersey, New Mexico, New York, North Carolina, Ohio, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Utah, Vermont, Virginia, West Virginia and Wisconsin. In Column 2 the 30 most populous states are included in the sample. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares from census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In all Columns we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the enrolment rate of 8-14 year olds, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, and the share residing on a farm. We also control for the following state characteristics: the total population and the average occupational score of the population. At the foot of each Column we report the p-value on the null hypothesis that the hazard coefficients are the same for the two European groups, and the p-value that the hazard coefficients are the same for the non-European immigrant groups and European borns from countries that did not have compulsory schooling in place in 1850.

2.5.4 Endogenous Location Choices of Migrants

The coefficients of interest $\widehat{\beta}_j$ from (2.4) cannot be interpreted as causal given migrants sort into locations, a process that might be driven by unobserved factors that also drive the passage of compulsory schooling laws. However, endogenous location choices can only drive the core result if the process differs across migrant groups. Specifically, it would have to be that European migrants without long exposure to compulsory state schooling at home are attracted by unobservable state characteristics that correlate with the adoption of schooling laws, while European migrants with long exposure to compulsory schooling at home are not attracted by these same characteristics.

We address the issue instrumenting for the share of the population of group j in state s in census year t using a Bartik-Card strategy, where we use the two-stage residual inclusion (2SRI) method for instrumenting in a non-linear model. The instrument has been much utilized in the immigration literature and is based on the intuition that migrants tend to locate where there are already members of the same group. To construct the instrument for N_{st}^{j} we first calculate the nationwide share of migrant group j (so N_{st}^{j} summed across states s at time t) in states that have not adopted, weighted by state s's share of that migrant group j in the previous census period among states that have not adopted compulsory schooling. We measure population shares in effect sizes and so denote the effect size of migrant group j in state s in census year t by $N_{s,t}^{j,E}$. The instrument is then defined as follows:

$$W_{st}^{j} = \frac{N_{s,t-1}^{j,E}}{\sum_{l \in R(t-1)} N_{l,t-1}^{j,E}} \sum_{k \in R(t)} N_{kt}^{j,E}, \tag{2.5}$$

where $R(t)$ is the set of states that remain at risk of adopting compulsory schooling law in census period t, K is the cardinality of $R(t)$ and L is the cardinality of $R(t-1)$. This instrument can be calculated for all census years except the first.

Table A8 reports the first stage results: for each group j, the instruments correlate with migration shares $N_{st}^{j,E}$: all coefficients lie in the range $.69 - .90$ and all are statistically significant at the 1% level. Column 1 in Table 2.4 shows the second stage results using the 2SRI method. The point estimates for the $\widehat{\beta}_j$'s remain stable, although

each is slightly more imprecise. However, it remains the case that the presence of European migrants from countries that do *not* have historic experience of compulsory state schooling at home significantly brings *forward* in time the passage of compulsory schooling: a one standard deviation increase in the population share of such Europeans is associated with a 65% higher hazard rate. In contrast, the presence of Europeans with a long history of compulsory schooling at home does not influence when compulsory schooling is passed by US states, although the 2SRI estimates are imprecise so we cannot reject the null that these hazards are equal [p-value=.262].

To improve precision, Column 2 presents 2SRI estimates assuming the underlying hazard follows a Log logistic distribution. In this specification the coefficients of interest $\widehat{\beta}_j$ are presented in a time ratio format (rather than a hazard). A time ratio *less* than one has the same interpretation as a hazard greater than one, indicating the covariate is associated with the passage of compulsory schooling *earlier* in time. The second stage results closely align with the baseline findings: the presence of European migrants from countries without historic experience of compulsory schooling at home significantly brings *forward* in time the passage of compulsory schooling. In contrast, the presence of Europeans with a long history of compulsory schooling at home does not influence the timing of compulsory schooling law, and these effect sizes across European migrants are significantly different to each other [p-value=.056].

There is no particular reason to think the first stage relationship between N_{st}^j and W_{st}^j is linear. We therefore consider a non-parametric first stage for N_{st}^j, $N_{st}^j = m(W_{st}^j, Z_{st}^j) + e_{st}^j$,with $m(.)$ unknown.[21] Column 3 shows the result from this more flexible first stage: the passage of compulsory schooling in a state occurs significantly earlier in time in the presence of more European migrants from countries without historic experience of compulsory schooling, and the impacts of the two groups of European migrant are significantly different to each other [p-value=.013].

Finally, Column 4 presents 2SRI estimates from the full model that includes the exogenous variables $Z_{st}^j = (X_{st}^j, X_{st})$. In the first stage, Columns 4-6 in Table A8 show

[21] A consistent estimate of $\hat{e}_{st}^j$ is then obtained as the difference between $\hat{m}(W_{st}^j, Z_{st}^j)$ and N_{st}^j, using local linear regression with Epanechnikov Kernel weights to first obtain $\hat{m}(.)$.

the instrument continues to be highly significantly associated with all three migrant share groups. In the second stage, Column 4 in Table 2.4 shows a pattern of impacts very similar to the baseline estimates from the full model: the findings provide strong support for the nation-building hypothesis. The presence of European migrants without historic exposure to compulsory schooling at home significantly brings forward in time the passage of compulsory schooling law; the presence of European migrants with historic exposure to compulsory schooling has no impact on the timing of compulsory schooling law, and these impacts significantly differ from each other [p-value=.011]. Moreover, we also find a significant impact of the presence of non-Europeans, mirroring the baseline findings.

The Appendix presents additional evidence related to endogenous location choices of individuals, including on: (i) the internal migration of American-borns, to further address the concern the passage of compulsory schooling was an instrument used by states to attract American migrants (or Americans took ideas over compulsory schooling with them as they migrated across states); (ii) the internal migration of the foreign-born, to check if migrants chose to endogenously locate into states after compulsory schooling laws were in place (we find no evidence of trend breaks in migrant population shares in states pre- and post-compulsion).

Table 2.4: Second Stage Estimates for 2SRI Instrumental Variables Method

Non parametric Cox proportional and log logistic hazard model estimates
Robust standard errors; Populations shares and enrolment rates measured in effect sizes

Model:	(1) NP Cox PH	(1) Log logistic (Time Ratio)	(2) Log logistic (Time Ratio)	(3) Log logistic (Time Ratio)
Share of the State Population that is:				
From European Countries that did NOT have CSL in 1850	1.65**	.920***	.906***	.923***
	(.382)	(.022)	(.020)	(.018)
From European Countries that had CSL in 1850	1.15	.098	.098*	.986
	(.152)	(.012)	(.011)	(.015)
Non-European Born	.85	.994	.990	.946***
	(.125)	(.014)	(.012)	(.009)
Includes First Stage Residuals [OLS]	Yes	Yes	No	No
Includes First Stage Residuals [Non-parametric]	No	No	Yes	Yes
Group Controls	No	No	No	Yes
State Controls	No	No	No	Yes
European Groups Equal [p-value]	[.262]	[.056]	[.013]	[.011]
Euro Without CSL = Non-Euro [p-value]	[.019]	[.030]	[.006]	[.217]
Gamma Parameter		.048***	.044***	.017***
		(.007)	(.007)	(.003)
Observations (state-census year)	180	180	180	180

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. In Column 1 a non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. In Columns 2 and 3 a log logistic hazard model is estimated where time ratios are reported. In all cases tests for significance relate to the null that the coefficient is equal to one. The unit of observation is the state-census year, for all census years from 1860. A state drops from the sample once compulsory schooling is passed. Robust standard errors are reported. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares from census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. We control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the enrolment rate of 8-14 year olds, the labor force participation rate, and the share residing on a farm. We also control for the following state characteristics: the total population and the average occupational score of the population. All Columns control for the first stage residuals in the 2SRI method. At the foot of each Column we report the p-value on the null hypothesis that the coefficients are the same for the two European groups, and the p-value that the coefficients are the same for the non-European immigrant groups and European borns from countries that did not have compulsory schooling in place in 1850. At the foot of Columns 2 to 4 the relevant parameters from the parametric hazard and frailty parameters are reported.

2.5.5 Other Forms of Migrant Diversity

The nation-building explanation implies the key source of within-migrant diversity is in their civic values, as proxied by migrants' historic exposure to compulsory state schooling in their origin country. However, American-born voters might actually be sensitive to other correlated sources of within-migrant diversity. Our next set of results establish whether the form of diversity within European migrants we have focused on so far actually proxies for another dimension of heterogeneity across migrants.

The first dimension we consider is religion: during the study period the Catholic church remained the most significant rival to governments in the provision of education [Glenn 2002]. We consider the US as a majority Protestant country, and use the Barro and McCleary [1985] data to group European countries into whether their majority religion is Protestant or Catholic–Other. Column 1 of Table 2.5 shows the result, where the following key points are of note: (i) among European migrants from countries that do not have compulsory state education by 1850, the estimated hazards are above one for both religions, although the hazard for migrants from Catholic–Other countries is

significantly higher than for migrants from Protestant countries [p-value=.013]; (ii) for Europeans with a long history of compulsory state schooling the hazard rate remains below one again for both groups of migrant by religion, and these hazards are not significantly different from each other [p-value=.289]; (iv) within European migrants from Protestant countries, there remain significant differences in the hazard between those with and without long exposure to compulsory schooling in their country of origin [p-value=.052]; (v) within European migrants from Catholic–Other countries, exactly the same source of diversity remains significant [p-value=.000]. In short, while there are important differences in how American voters respond to the presence of European migrants of different religions, being especially sensitive to Europeans from Catholic–Other countries, within religion, historic exposure to compulsory state-provided schooling among European migrants in a state remains a key predictor of when such legislation is passed in each US state.

The Dillingham Report highlighted the divide between "old" (from Northern Europe and Scandinavia) and "new" (from Southern and Eastern Europe) immigrants with respect to their skills, economic conditions at arrival and migratory horizon. Hence the second source of within-migrant diversity we consider is European region of origin. We subdivide European migrants with and without historic exposure to compulsory schooling between these from old and new Europe, so defined. Column 2 shows the result, where we note: (i) among European migrants from countries without compulsory schooling by 1850, the hazards are above one for both subsets of Europeans; (ii) these hazards are not significantly different from each other [p-value=.269]; (iii) for Europeans with a long established history of compulsory schooling the hazard rates remain below one for both groups of European by region of origin, and again these hazards are not significantly different from each other [p-value=.348]; (iv) within European migrants from Northern Europe–Scandinavia, there remain significant differences in the hazard between those with and without long exposure to compulsory state schooling in their country of origin [p-value=.066]; (v) within European migrants from Southern–Eastern Europe, exactly the same source of diversity remains significant in

explaining the cross-state passage of compulsory schooling [p-value=.003]. In short, the evidence suggests while American-born voters are sensitive to the region of origin of European migrants, the over–riding source of diversity the median voter is sensitive to is differences in migrant values.[22]

We next consider English language as the key source of within-migrant diversity. To do so, we subdivide European-born migrants from countries without historic exposure to compulsory schooling in their country of origin, between those from non-English speaking countries and those from English speaking countries. All European migrants from countries with compulsory schooling already in place by 1850 originate from non-English speaking countries. Hence only a three-way division of European migrants is possible when considering English language as the additional source of within-migrant diversity over and above differences in values.

Column 3 shows the result, where the following points are of note: (i) among European migrants from countries that do not have compulsory state schooling in place by 1850, the estimated hazards are above one for both subsets of Europeans; (ii) these hazards are not significantly different from each other [p-value=.555]; (iii) for Europeans with a long established history of compulsory state schooling the hazard rate remains below one; (iv) within European migrants from non-English speaking countries, there remain significant differences in the hazard rate for compulsory schooling between those with and without long exposure to compulsory schooling in their country of origin [p-value=.057]. In short, American-born median voters appear more sensitive to diversity in values among European migrants than diversity in their English speaking

[22]This result reinforces the earlier finding that the human capital or enrolment rates of migrants were not an important factor driving the cross-state adoption of compulsion, as migrants from Southern–Eastern Europe would have had the lowest levels of human capital accumulation. The differences in migrant characteristics between these European regions of origin might capture a host of other factors including: (i) differential propensities to out–migrate [Abramitzky *et al.* 2012, Bandiera *et al.* 2013]; (ii) ties to second generation immigrants in the US (who are then American-born but with foreign born parents). On the first point, we have also taken implied out-migration rates of nationalities from Bandiera *et al.* [2013] and then created a four way classification of European migrants by their historic exposure to compulsory schooling, and whether they have above or below median out–migration rates. The results confirm that within-migrant diversity in values as captured by historic exposure to compulsion remains the key source of variation across migrants. On the second point, in the Appendix we discuss the robustness of our core result to splitting the American-born population between second generation immigrants and those whose parents are both American-born.

abilities. Indeed, the evidence suggests a one standard deviation increase in the population share of English speaking migrants (i.e. British and Irish migrants) significantly increases the hazard of compulsory schooling by 66%, all else equal. As highlighted earlier, this result is most likely picking up the fact that Irish migrants were Catholics, and this was an important divide in values with the median American.

While the evidence points to diversity among migrants along multiple dimensions mattering, all the findings point to the specific targeting of compulsory schooling laws in the US towards European migrants that did not have such values inculcated through a compulsory state education system in their country of origin.

Table 2.5: Other Sources of Diversity Within European Migrants

Non parametric Cox proportional model, hazard rates reported
Robust standard errors; Populations shares measured in effect sizes

	(1) Religion	(2) European Region	(3) Language
Share of the State Population that is From:			
Euro Countries that did NOT have CSL in 1850, Protestant	1.22		
	(.234)		
Euro Countries that did NOT have CSL in 1850, Catholic/Other	2.39***		
	(.596)		
Euro Countries that had CSL in 1850, Protestant	.598*		
	(.176)		
Euro Countries that had CSL in 1850, Catholic/Other	.840***		
	(.044)		
Non-European Born	2.29***	2.08**	1.83***
	(.609)	(.639)	(.227)
Euro Countries that did NOT have CSL in 1850, Northern/Scandinavian		1.89	
		(.837)	
Euro Countries that did NOT have CSL in 1850, Southern/Eastern		1.16*	
		(.099)	
Euro Countries that had CSL in 1850, Northern/Scandinavian		.698	
		(.162)	
Euro Countries that had CSL in 1850, Southern/Eastern		.883***	
		(.038)	
Euro Countries that did NOT have CSL in 1850, English Speaking			1.66*
			(.494)
Euro Countries that did NOT have CSL in 1850, Non English Speaking			1.25
			(.311)
Euro Countries that had CSL in 1850 (all Non English Speaking)			.776
			(.127)
Group and State Controls	Yes	Yes	Yes
With CSL = Without CSL, Protestant	[.052]		
With CSL = Without CSL, Catholic/Other	[.000]		
With CSL = Without CSL, Northern European		[.066]	
With CSL = Without CSL, Southern/Eastern European		[.003]	
With CSL (All Non English) = Without CSL, Non English			[.057]
Observations (state-census year)	230	230	230

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling is passed. Robust standard errors are reported. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares in census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In all Columns we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850, as well as the one additional group defined in each column): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, the enrolment rate of 8-14 year olds and the share residing on a farm. In all Columns we control for the following state characteristics: the total population, and the average occupational score of the population. In Column 1, we use the Barro and McCleary [1985] data to define country religion. The following European countries are then defined to be Protestant: Britain, Denmark, Finland, Germany, Holland, Norway and Switzerland. In Column 2, Northern Europe/Scandinavian countries are defined to be Belgium, Britain, Denmark, Finland, France, Germany, Holland, Iceland, Ireland, Lichtenstein, Luxembourg, Norway, Sweden and Switzerland. In Column 3, English speaking European countries are Britain and Ireland (both without compulsory schooling in 1850). At the foot of each Column we report the p-value on the null hypothesis that the hazard coefficients are the same between various European groups with and without compulsory schooling in 1850.

2.5.6 Alternative Mechanisms

Nation-building motives are not the only reason why the state intervenes in education provision *en masse*. Normative and positive arguments can be used to justify state pro-

vision of education based on efficiency or redistributive concerns, human capital externalities, or complementarity between capital and skilled labor during industrialization. While none of these necessarily require *compulsory* schooling, we now assess whether our core finding remains robust to additionally accounting for the basic predictions of some of these alternative mechanisms.

To examine if redistributive motives drive the passage of compulsory schooling, we estimate (2.4) and additionally control for the standard deviation in the state occupational income score (the mean occupational income score is already in X_{st}). This proxies the redistributive pressures the state faces. Column 1 of Table 2.6 shows that although there is a positive correlation between inequality so measured and the hazard of passing legislation, the coefficient is not significantly different from one. The impacts of the population shares of interest remain almost unchanged from the baseline specification, suggesting the presence of migrant groups and inequality in a state are not correlated.

Column 2 examines the industrialization hypothesis by controlling for the share of workers in the state's labor force working in different occupations: professions, craft and operative. We find that as a greater share of workers are engaged in the middle-skilled craft occupations, the hazard of introducing compulsory schooling significantly increases (the point estimate on the hazard is below one for the least-skilled operative occupations). Hence there is evidence on compulsory schooling being related to industrialization, but this additional mechanism operates in parallel with the nation-building motives embodied in our core finding.[23]

Galor *et al.* [2009] make precise how the industrialization process interacts with land inequality in determining the level of state provision of education. They argue there exists a conflict between the entrenched landed elite (who have little incentive

[23]This is in line with the evidence presented in Galor and Moav [2006] from England, on how members of Parliament voted for the Balfour Act of 1902, the proposed education reform that created a public secondary schooling system. They find Parliamentarians were more likely to vote for the legislation if they represented more skill intensive constituencies (even accounting for their party affiliation). For the US, Goldin and Katz [2001] argue that over 1890-1999 the contribution of human capital accumulation to the US growth process nearly doubled, and Goldin [1999b] describes how the changing industrial structure of the US economy drove changes in the content of what was needing to be taught in secondary schools.

to invest in mass schooling) and the emerging capitalist elite, who do have such incentives given the complementarity between capital and skilled labor. To proxy the relative balance of power in this conflict they propose a measure of land inequality, that is the share of land held by the top 20% of all land holdings. We then additionally control for this same measure in (2.4). The result in Column 3 shows that the effect goes in the expected direction but the ratio is not significantly below one. Moreover, the coefficients relevant for the nation-building hypothesis remain stable, further suggesting the composition of the migrant population is not related to land inequality.[24]

The remaining Columns focus on the explanation that political parties were key to compulsory schooling. Indeed, much has been written about the Republican-Democrat divide over compulsory schooling, with the policy often being seen to be driven by a faction of the Republican party [Provasnik 2006]. In line with this we find that a one standard deviation increase in the vote share for Republicans in Congressional elections significantly increases the hazard rate. Given that significant third parties existed for much of the 19th century, Column 5 repeats the analysis controlling for Democrat party vote shares: as implied by the qualitative evidence, a greater vote share for Democrats does indeed significantly reduce the hazard of passing compulsory schooling law. However, controlling for Republican or Democrat vote shares do not alter the migrant population share coefficients, that remain stable throughout.

[24]This land inequality measure is available for 1880, 1900 and 1920: we linearly interpolate it for other state-census years. Galor *et al.* [2009] show that state schooling expenditures are significantly correlated to land inequality.

Table 2.6: Alternative Mechanisms Driving the Passage of Compulsory Schooling Laws

Non parametric Cox proportional model, hazard rates reported
Robust standard errors; Populations shares measured in effect sizes

	(1) Redistribution	(2) Industrialization	(3) Land Inequality	(4) Republicans	(5) Democrats
Share of the State Population that is From:					
European Countries that did NOT have CSL in 1850	2.14***	2.38***	1.84**	2.62***	3.00***
	(.470)	(.520)	(.461)	(.858)	(1.04)
European Countries that had CSL in 1850	.831	.819	.901	.915	1.02
	(.160)	(.148)	(.196)	(.180)	(.170)
Non-European Countries	1.82***	2.01**	2.14***	1.77**	1.62*
	(.389)	(.554)	(.518)	(.455)	(.459)
SD of Occupational Income Score	1.38				
	(.423)				
Share of Labor Force Engaged in Professional Occupations		1.00			
		(.000)			
Share of Labor Force Engaged in Craft Occupations		2.51*			
		(1.32)			
Share of Labor Force Engaged in Operative Occupations		.550			
		(.296)			
Land Share of Top 20% of Holdings [Galor *et al*. 2009]			.815		
			(.171)		
Republican Party Vote Share in Congressional Elections				1.68*	
				(.455)	
Democratic Party Vote Share in Congressional Elections					.558***
					(.105)
Group and State Controls	Yes	Yes	Yes	Yes	Yes
European Groups Equal (with and without CSL) [p-value]	[.003]	[.000]	[.025]	[.002]	[.003]
Euro Without CSL = Non-Euro [p-value]	[.513]	[.549]	[.591]	[.331]	[.135]
Observations (state-census year)	230	230	216	148	148

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling laws are passed. Robust standard errors are reported. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares in census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In all Columns we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850, as well as the one additional group defined in each column): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, the enrolment rate of 8-14 year olds and the share residing on a farm. In all Columns we control for the following state characteristics: the total population, and the average occupational score of the population. Column 1 controls for the state-year standard deviation in the occupational index score. Column 2 controls for the share of the population defined to be working in craft occupations, and operative occupations (where professional occupations are the omitted category). Column 3 controls for the land share of the largest 20% of farm land holdings, from [Galor *et al*. 2009], to proxy inequality of land holdings. This is available for 1880, 1900 and 1920: we linearly interpolate it for other state-census years. Column 4 (5) controls for the vote share of the Republican (Democratic) party in congressional elections: these are available only in census years from 1860 onwards for a subset of states. At the foot of each Column we report the p-value on the null hypothesis that the hazard coefficients are the same for the two European groups.

2.6 Migrants' Demand for American Common Schooling

The extent to which compulsory schooling is an effective tool by which to expose migrant children to the same kinds of civic value that were being taught to American-born children, depends fundamentally on migrant's underlying demand for American common schooling. Only if their demand for common schooling was sufficiently low would compulsory schooling be required to change the kinds of instruction they were exposed to, and thus shape their civic values. We now exploit detailed information on locally-financed investments into American common schools in the cross-section of counties in 1890 to pin down the relative demands for American common schools of the different migrant groups.

2.6.1 Conceptual Framework

As migrants can form a significant share of the population in jurisdictions that determine investments into common schools, we use a textbook probabilistic voting model [Persson and Tabellini 2000] to derive an empirical specification informative of the relative demands for such schools among migrant groups.[25] A jurisdiction comprises a continuum of citizens. An individual i belongs to group j, where groups are of size N^j, $\sum_j N^j = N$. Within a group, individuals have the same income, y^j. Individual preferences are quasi-linear,

$$u^j(g) = c^j + \alpha^j(.)H(g), \tag{2.6}$$

where c^j is the private consumption of a member of group j, $H(g)$ is concave in the public good, g (common schools), and is assumed twice-differentiable with $H(0) = 0$. The group valuation for American common schools is $\alpha^j(\theta^j, 1(HCSL^j))$: θ^j captures

[25] This is in contrast to the earlier conceptual framework in Section 3, where we utilized a median voter model to understand the passage of compulsory schooling law at the state level. The justification is that: (i) at the state level, migrants never form close to the majority of the electorate (as Figure A1 shows) and so the median voter is American-born; (ii) the outcome studies was a discrete choice of whether to introduce compulsory schooling law or not. In contrast, at the county level, migrant shares are larger, and we study a continuous outcome (investment into common schools) so the probabilistic voting model is more appropriate.

factors that influence the group's demand for common schools (such as the share of young people in the group), and $1(HCSL^j)$ is an indicator for the historic entrenchment of compulsory schooling law (HCSL) in the country of origin for those in migrant group j. In line with our context, the local jurisdiction finances common schools by a local income tax rate τ so individuals face a budget constraint, $c^j = (1-\tau)y^j$, and no group can be excluded. It is because of this local financing that we can map between observed investments into common schools and the underlying demand for those schools.

The probabilistic voting model specifies the following political process that produces a equilibrium level of common schooling: there are two political parties (A, B), whose only motivation is to hold office. The source of within group heterogeneity is a political bias parameter $\sigma^{ij} \sim U[-\frac{1}{2\phi^j}, \frac{1}{2\phi^j}]$: a positive value of σ^{ij} implies that voter i has a bias in favor of party B while voters with $\sigma^{ij} = 0$ are politically neutral. Hence ϕ^j measures the political homogeneity of a group j. Voter i in group j thus prefers candidate A if $u^j(g_A) > u^j(g_B) + \sigma^{ij}$.

The timing of events is as follows. First, parties A and B simultaneously and non-cooperatively announce electoral platforms: g_A, g_B. At this stage, they know the distribution from which σ^{ij} is drawn, but not realized values across voters. Second, elections are held where citizens vote sincerely for a single party. Voters and parties look no further than the next election. Third, the elected party implements her announced policy platform.

Proposition 2 *The political equilibrium is $g^* = g_A = g_B$ where g^* is implicitly defined as,*

$$H_g(g^*) = \frac{\theta \sum_j W^j y^j}{\bar{y} \sum_j W^j \alpha^j(\theta^j, \mathbf{1}(HCSL^j))}. \tag{2.7}$$

$W^j = N^j \phi^j$ is group j's 'political weight', and $\theta = \frac{\sum_j \theta^j N^j}{N}$ is the share of young in the population.

The Proof is in the Appendix.

The group's political weight captures how influential the group is by virtue of its size and how many swing voters are in group j. A key feature of the probabilistic voting model is that all groups have some weight in the determination of commons schooling

g^*. The key comparative static we consider is how the optimal provision of common schooling changes in group-j's size:

$$\frac{\partial H_g(g^*)}{\partial N^j} = \frac{1}{\phi^j}\frac{\partial H_g(g^*)}{\partial W^j} = \frac{\theta y^j}{\phi^j \bar{y}\left(\sum_j W^j \alpha^j(\theta^j, \mathbf{1}(HCSL^j))\right)^2}\left[\sum_{k\neq j} W^k y^k [\alpha^k - \alpha^j]\right] \tag{2.8}$$

Hence the larger is α^j relative to other group α^k's, the more likely is it that $\frac{\partial g^*}{\partial N^j} > 0$. The sign of $\frac{\partial g^*}{\partial N^j}$ can then be informative of $sign(\alpha^j$ relative to $\alpha^k)$. We use this intuition to rank the relative demands for common schools across the j groups. This dovetails with the earlier analysis of what drove the cross-state adoption of compulsory schooling: our results there showed the American-born median voter was especially sensitive to European migrants from countries without historic exposure to compulsory state-provided schooling. Hence they behaved as if,

$$\alpha^j(\theta^j, \mathbf{1}(HCSL^j) = 1) > \alpha^j(\theta^j, \mathbf{1}(HCSL^j) = 0), \tag{2.9}$$

so that absent compulsory schooling in the US, this specific group of European migrants would have demanded less common schooling, and as a result, those migrant children would have been less exposed to the kinds of instruction and shaping of civic values that American-born children were experiencing. We now recover estimates of this relative ranking to understand whether these beliefs were justified. Unlike the earlier cross-state analysis, here it is important that groups have endogenously sorted into counties and so we can recover their equilibrium demand for American common schools.

2.6.2 Empirical Method

We estimate the model using cross-county data from 1890 that were collected as part of the population census, but were the result of a separate report in which the Census Bureau contacted the superintendents of public education in each state. Superintendents were asked to report the race and sex of teachers and enrolled pupils in each county. The data, documented in Haines [2010], details investments into common schools in over 2400 counties in 45 states. We proxy the equilibrium provision of common schooling, g^*, using the number of common school teachers in the county. These are locally

financed and likely comprise the most significant investment into public schooling. As IPUMS 1890 census data is unavailable, we build control variables using 1880 values based on the 100% census sample. The groups considered replicate those in the earlier analysis: the American-born, European migrants from countries with compulsory schooling, European migrants from countries without compulsory schooling and non-European migrants. We then estimate the following OLS specification for county c in state s,

$$\ln(teachers)_{cs} = \sum_j \alpha^j N_{cs}^j + \sum_j \gamma_j X_{cs}^j + \lambda X_c + \delta_s + u_{cs}, \tag{2.10}$$

where N_{cs}^j is the total population size of group j (again measured as an effect size), and X_{cs}^j includes other characteristics of group j (the share aged 0-15, the labor force participation rate, the share residing on a farm, and the average occupational income score).[26]

X_c includes the (log) total population of the county aged below 15, and the county's occupational index score. δ_s is a state fixed effect so the coefficients of interest, α^j, are identified from variation in the composition of migrant populations across counties within the same state. Figure A5 illustrates the cross-county variation in migrant group sizes for four states (one from each census region). Panel B of Table 2.1 provides descriptive evidence on the shares of county populations from each group j and documents the considerable within state variation in these shares. Robust standard errors are reported, and we weight observations by 1880 county population so our coefficients of interest map to the average demand of an individual from group j. Mapping the model to the empirical specification makes clear the relative ranking of $\alpha^j(.)$'s across groups (not their levels) can be identified from the ranking of $\widehat{\alpha}^j$'s estimated from (2.10). As we do not control for the total county population, this allows us to control for the population size and characteristics for *all four* groups j and so measure demands relative to those of the American-born.

[26]The County Yearbook provides information on public education for black and white populations separately. For our analysis, all schooling related variables (teachers and attending pupils) correspond to whites. However, in some states there is expected to be some small bias here as teachers of all races were pooled together. Moreover, there is an imperfect match between true school jurisdictions and counties, and this attenuates our coefficients of interest, α_j.

2.6.3 Results

Table 2.7 presents the results. Column 1 estimates (2.10) only controlling for the populations of each group j. At the foot of the table we report p-values on the equality of these coefficients to establish the ranking of relative demands for common schooling. The results highlight again that a key source of diversity within European migrants in their demand for common schools is whether they have historic exposure to compulsory state schooling in their country of origin. More precisely: (i) a one standard deviation increase in the county population of European migrants with long exposure to compulsory state schooling in their country of origin significantly increases the provision of common school teachers by 5.8%; (ii) a one standard deviation increase in the county population of European migrants without exposure to compulsory schooling in their country of origin significantly decreases the provision of common school teachers by 18%; (iii) these impacts across European migrant groups significantly differ from each other [p-value =.000]; (iii) the presence of non-European migrants is associated with significantly higher investments into common school teachers. This ranking of $\widehat{\alpha}^j$'s is robust to including state fixed effects (Column 2), and group and county controls (X_{cs}^j, X_c) (Column 3).

Mapping the marginal impacts from the specification in Column 3 back to the model then implies the following ranking of quasi-linear demand parameters from (2.6):

$$\alpha^{Euro}_{\mathbf{1}(HCSL^j)=1} = \alpha^{Am-born} > \alpha^{NonEuro} > \alpha^{Euro}_{\mathbf{1}(HCSL^j)=0}. \tag{2.11}$$

This links directly to the earlier analysis on how the composition of migrants drove the cross-state timing of compulsory schooling: there we found the American-born median voter was especially sensitive to the presence of migrants from European countries without historic exposure to compulsory schooling. The implied ranking of $\widehat{\alpha}^j$'s across European migrant groups closely matches up across the two sets of analysis, despite the two sets of quantitative evidence using entirely different data sources, econometric methods and identification strategies. Fundamentally, it suggests European migrants from countries without historic exposure to compulsory schooling would have

invested less in American common schools ($\alpha^{Euro}_{\mathbf{1}(HCSL^j)=1} > \alpha^{Euro}_{\mathbf{1}(HCSL^j)=0}$). As such, the American-born median voter held correct beliefs in bringing forward in time compulsory schooling laws in those states where such migrants were more numerous.[27]

Table 2.7: Migrants and County Investments in Common Schools

OLS estimates, robust standard errors
Dependent variable: Log common school teachers in county
County populations measured in effect sizes

	(1) Immigrant Groups	(2) State FE	(3) Controls
County Population that is:			
American Born	.298***	.239***	.029**
	(.060)	(.042)	(.011)
European Born from Countries that did NOT have CSL in 1850	-.180***	-.176***	-.040***
	(.032)	(.024)	(.011)
European Born from Countries that had CSL in 1850	.058*	.076***	.036***
	(.034)	(.025)	(.007)
Non-European Born	.120***	.078***	.017***
	(.018)	(.012)	(.005)
Mean of Dependent Variable (in levels)		133	
State Fixed Effects	No	Yes	Yes
Group and County Controls	No	No	Yes
American = European Born without CSL [p-value]	[.000]	[.000]	[.002]
European Groups Equal (with and without CSL) [p-value]	[.000]	[.000]	[.000]
Observations (county)	2472	2472	2472

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. The unit of observation is a county, and the sample covers counties from 45 states. The dependent variable is the log of the number of white teachers in the county. All outcomes are measured in 1890. All right hand side controls are measured in 1880, and derived from the 100% IPUMS-USA census sample. OLS regression estimates are shown, where robust standard errors are estimated, and observations are weighted by the county population. In all Columns population groupings are all defined in effect sizes, where this is calculated from population numbers in the cross section of counties in 1890. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. Column 2 onwards includes state fixed effects. In Column 3 we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the labor force participation rate, the share residing on a farm, and the average occupational income score. At the foot of each Column we report the p-value on the null hypothesis that the coefficients are the same for various pairs of groups.

Given investments into common school are measured in the cross-section of counties in 1890, and that by then half of all states had passed compulsory schooling, we next estimate a modified version of (2.10) that allows for the demand for common schools to vary within the same migrant group depending on whether or not they reside in a state with compulsory schooling. This allows us to establish whether the

[27] One disconnect between the cross-state and cross-county evidence relates to non-Europeans. This disconnect can stem from two sources: (i) the selection of non-European migrants into the US differs from that for European migrants; (ii) American-borns were less informed about the preferences of non-European migrants, that is plausible given the long history of anti-Chinese discrimination in the US, culminating in the Chinese Exclusion Act of 1882, that banned all immigration of Chinese laborers.

compulsory schooling laws had the intended effect of increasing migrants' exposure to American civic values in common schools. Defining a dummy D_s equal to one if state s has passed compulsory schooling in 1890, we estimate the following specification:

$$\ln(teachers)_{cs} = \sum_j \alpha^{j0} N_{cs}^j + \sum_j \alpha^{j1} \left[D_s \times N_{cs}^j \right] + \sum_j \gamma_j X_{cs}^j + \delta_s + u_{cs}, \qquad (2.12)$$

where $\widehat{\alpha}^{j0}$ and $(\widehat{\alpha}^{j0} + \widehat{\alpha}^{j1})$ measure the relative demand for common schools pre and post-compulsory schooling respectively, for the same migrant group j. The corresponding estimates are shown graphically in Figure 2.3. We focus first on Panel A: the left hand side shows the $\widehat{\alpha}^{j0}$'s for each group j (and their corresponding 95% confidence interval): the y-axis shows the magnitude of each estimate, but as only relative demands for common schools are identified from (2.12), we centre the point estimates on the value for American-borns. This shows that pre-compulsory schooling, a key source of diversity in values for common schools was between European migrants with and without historic exposure to compulsory state schooling in their country of origin. Indeed, pre-compulsory schooling, European-born migrants from countries with compulsory schooling already in place by 1850 have significantly higher demands for common schooling than other European migrants and the American-born.[28]

The right hand side of Panel A in Figure 2.3 shows the change in demand for common schooling for the same groups j: these $\widehat{\alpha}^{j1}$ estimates show there is a significant convergence in demands for common schooling with compulsory schooling. The change in demand for common schools is significantly greater among Europeans without historic exposure to compulsory schooling than among Europeans with such exposure to compulsory state schooling. This evidence suggests the introduction of compulsory schooling did indeed lead European migrants to be significantly more exposed to the American common schooling system, as measured by this willingness to invest in such schools. Moreover, this was especially so for Europeans from countries without historic exposure to compulsory schooling in their country of origin and hence most in need of homogenizing their civic values towards those being instilled among

[28]It is well recognized that compulsory schooling laws necessitated no supply side response, so that the supply of teachers would not have been directly impacted [Margo and Finegan 1996].

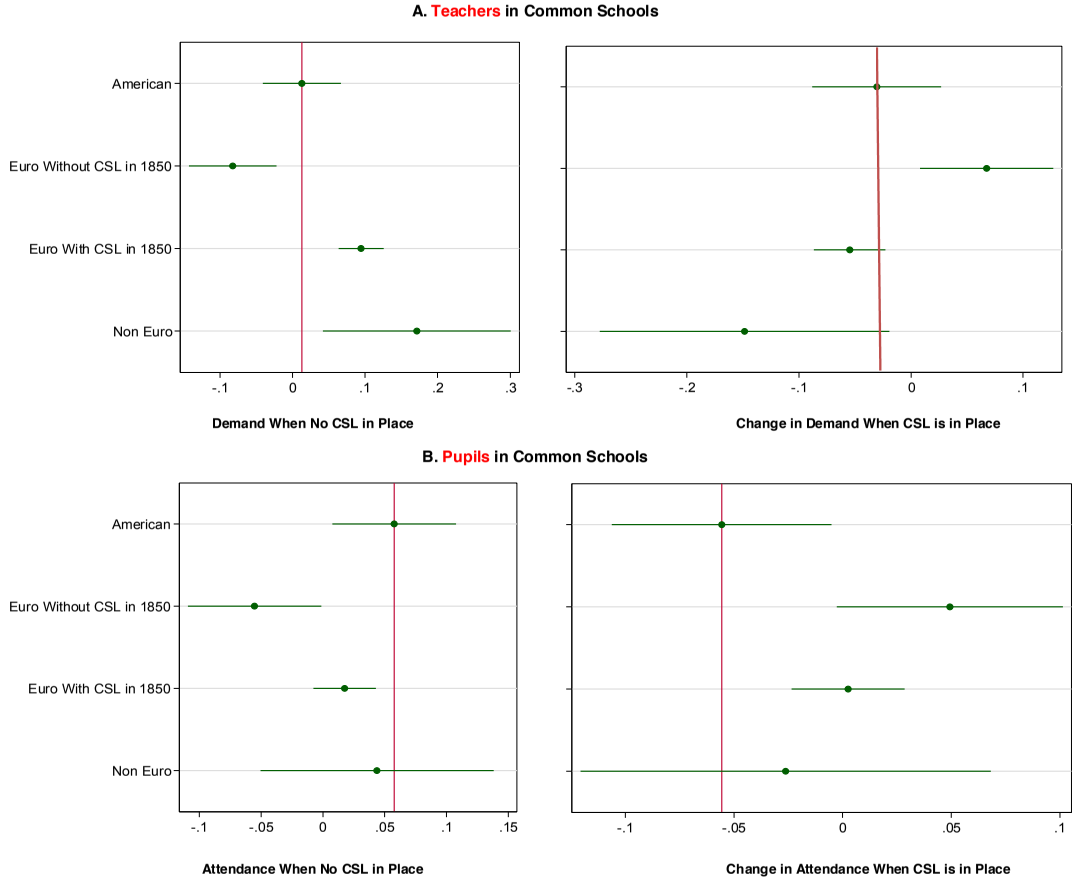
American-born children.

The data complied by Superintendants also allows us to re-estimate (2.12) but considering pupil attendance as a county level outcome. We can thus assess how pupil attendance various with migrant shares in the county, and how this relationship alters under compulsory schooling. The evidence is summarized in Panel B of Figure 2.3. The results highlight that pre-compulsory schooling, counties with more migrants from European countries without historic exposure to compulsory schooling in their country of origin, had lower attendance in American common schools. As shown above, compulsory schooling led to a significant degree of convergence in demands for American common schools between migrant groups and American-borns. These implied changes in demand for common schooling are in line with evidence in Panel B on the impact of compulsory schooling on actual pupil attendance in common schools.

In line with this evidence, Lleras-Muney and Shertzer [2015] show how compulsory schooling laws significant increased enrolment rates of migrant children by 5%, with smaller impacts on American-born children. Ultimately, this will have impacted the instruction migrant children were exposed to (relative to the counterfactual absent compulsory schooling) and so shaped the civic values that were instilled into them. Our evidence thus links closely to the findings of Milligan *et al.* [2004], who show using NES and CPS data, that those exposed to compulsory schooling are later in life, significantly more likely to be registered to vote, to vote, to engage in political discussion with others, to follow political campaigns and attend political meetings, as well as having higher rates of participation in community affairs and trust in government. These are the kinds of changes in values emphasized in Glaeser *et al.* [2007] as being inculcated through compulsory schooling. Indeed, our findings and these related papers all suggest that the original architects of the common school system, as discussed in Section 2, all of whom linked education with inculcating the civic values necessary for effective participation in American democracy, achieved their aim.[29]

[29]Recent evidence also highlights cases in which assimilation policies lead to a backlash among migrants: Fouka [2014] presents evidence showing that Germans that faced restrictions on the use of the German language in primary schools (introduced over the period 1917-23) are less likely to volunteer during the Second World War, more likely to marry within their ethnic group, and be more likely to give German sounding names to their children.

Figure 2.3: Demand for Common Schooling in 1890, by Population Groups and Compulsory Schooling Law



Notes: The Panels show coefficient estimates and robust standard errors from an OLS regression in which the unit of observation is a county, and the sample covers counties from 45 states. The dependent variable in Panel A is the log of the number of white teachers in the county. The dependent variable in Panel B is the log of the number of enrolled white pupils in the county. All outcomes are measured in 1890. All controls in the regressions are measured in 1880, and derived from the 100% IPUMS-USA census sample. Observations are weighted by the county population. In all Panels, the four population groups are controlled for, as well as an interaction between each group and whether compulsory schooling laws are in place in the state prior to and including 1890 (the other controls in each regression are state fixed effects, the average occupational score of the county population, the log of the county population aged 0 to 15, and the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the labor force participation rate, the share residing on a farm, and the average occupational income score). Population groupings are all defined in effect sizes, where this is calculated from population numbers in the cross section of counties in 1890. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In each Panel, the left hand side figure shows the coefficient on the population grouping in the pre-compulsion period. The right hand side figure shows the coefficient on the interaction between the population grouping and the compulsory schooling law dummy.

2.7 Discussion

Many great figures in political and economic history including Napoleon and Adam Smith, have emphasized the central role of a state's education system in nation-building [Milligan *et al.* 2004, Clots-Figueras and Masella 2013]. In this paper we have examined the hypothesis that nation-building efforts were part of the policy response of American voters to the large and diverse waves of migrant inflows during the Age of Mass Migration. In particular, we have provided qualitative and quantitative evidence that compulsory schooling was used by Americans as a nation-building tool to homogenize the *civic values* of migrants who moved to America during the nineteenth century, with these laws being targeting first towards European migrants without exposure to a compulsory state schooling system in their country of origin. Our work adds to the broad literature emphasizing that the national origins of migrants matters [La Porta *et al.* 1998, Acemoglu *et al.* 2001], where we show the importance of national origins for long run outcomes through a new mechanism: the policy response of natives.

By providing micro-foundations for compulsory schooling, our findings also have implications for the large literature examining the impacts of compulsion on the human capital of American-borns. As summarized in Stephens and Yang [2014], this literature has found rather mixed evidence. Our results suggests this is partly because American-borns were not the intended marginal beneficiary, and that the core purpose of compulsion was to instill civic values among the children of migrants. Indeed, our findings build on and complement Lleras-Muney and Shertzer [2015] who show that compulsory schooling laws had significant impacts on the enrolment rates of migrant children (increasing them by around 5% overall), with smaller impacts on native children.

We conclude by highlighting two further directions for research. First, a wide set of public policies might have been impacted by large and diverse inflows during the Age of Mass Migration. The most natural policy dimension to study next would be cross-jurisdiction variations in tax rates used to finance local public goods, but variations observed in the regulation and operation of financial and legal markets, say, might also originate from differences in patterns of mass migration into those states during

the 19th century [Burchardi *et al.* 2016, Fulford *et al.* 2015].[30] It also remains important to understand other policies specifically targeted towards immigrants during the study period. For example, during the early 20th century some states introduced citizenship requirements for foreigners to be able to vote. Such policies presumably held back immigrant assimilation and sustained greater heterogeneity in values among the population. Hence there remains a need to understand the political economy trade-offs involved that led to the simultaneous use of both nation-building efforts towards foreigners as well as their political exclusion.

A second direction for future research is to combine the ideas underpinning this analysis with earlier work that documented high rates of out-migration from the US by Europeans during the Age of Mass Migration [Bandiera *et al.* 2013]. This opens up an agenda examining whether returning Europeans drove institutional and legal change in their home country after having been exposed to American society.

[30]This emerging body of work indeed suggests that migration during the Age of Mass migration is causally linked to: (i) FDI sent and received by firms across US counties [Burchardi *et al.* 2016]; (ii) the evolution of county level income for a century later [Fulford *et al.* 2015].

Chapter 3

Stars and Brokers: Knowledge Spillovers in Medical Science

3.1 Introduction

The importance of technological progress for economic growth is well understood, yet the process of knowledge creation remains elusive. Opening the black box of innovation requires an understanding of the creativity process itself. Endogenous growth models suggest that the production function of new knowledge depends on new (re–)combinations of the existing stock of knowledge by researchers (Jones, 2009, Mokyr, 2002, Romer, 1990a, Weitzman, 1998, Wuchty et al., 2007).

If creativity requires a fresh perspective on existing knowledge, then access to a diverse pool of knowledge plays a crucial role for the productivity of a researcher. Being aware of breakthroughs will alert original and creative individuals to gaps and opportunities in the existing stock of knowledge. As highlighted by Mokyr (2005), the progress in exploiting the existing stock of knowledge will depend first and foremost on the ease and cost of access to knowledge. In sociology, it has long been recognised that the structure of network ties and one's position therein are important determinants of a researcher's access knowledge and the diversity of it (Borgatti, 2005, Wellman and Berkowitz, 1988). Scientists are embedded in a larger scientific community within which knowledge is shared. The structure of the community shape the connections

between scientists, each embodying specialised skill and knowledge.

This paper investigates the importance of network position for the productivity of scientists. Specifically, I study the influence a star has on the productivity of a given scientist and allow for heterogeneity in the effect by network positions. To capture the idea that network positions provide access to different knowledge embodied in researchers, I measure to what extent a scientist is near a so–called "structural hole" (Burt, 2004). Structural holes are gaps in the network at which people on either side of the hole have access to different information flows. Brokers occupy an exclusive intermediate position acting as a "bridge" across structural holes. Through such intermediation, they reduce the cost of accessing new knowledge and provide research opportunities by brokering the flow of information between contacts who differ in their skills and knowledge. Put differently, brokers are potential amplifiers for innovation since they are well–positioned to synthesize ideas arising from different groups (Burt, 2004, Granovetter, 1973).

I propose a new measure of brokerage motivated by the idea that scientists embedded in a collaborative network meet and freely share ideas. Scientists embody unique specialized information and research skills. Knowledge flows along coauthorship links; scientists can get access to the knowledge of their immediate coauthors and scientists further away through mutual contacts. The connections received via a coauthor are beneficial to the extent that they provide access to *new* knowledge that differs from previously available information. This means that the relative position of coauthors to one another will determine how much they rely on the other to provide non–redundant information. Brokerage degree is a pair–specific and asymmetric measure defined for a local neighborhood (up to three links away) that quantifies this dependency. More precisely, the brokerage degree of B to A is defined as the share of scientist B's links that offer a unique and exclusive access to *new* scientists (or *new* knowledge) for A.

The analysis is based on stars in medical science for whom collaboration is the norm

and information on quantity of output is easily measured by publications. Moreover, knowledge, at least when it is new, often remains tacit and confined to tightly-knit groups. Therefore scientific collaborations play an important role as they involve the exchange of ideas and opinions and facilitate the generation of new ideas. The focus on star scientists is driven by the observation that high performers account for the generation of a large share of total research output(Lotka, 1926, Narin and Breitzman, 1995, Rosen, 1981, Zucker and Darby, 1996). Moreover, recent evidence suggests that only colleagues of very high quality affect the productivity of scientists (Azoulay et al., 2010, Borjas and Doran, 2013).

One of the main empirical challenges to establishing the presence of spillovers is the sorting of individuals into collaborations. A scientist's decision to coauthor and therefore his position in the overall network is often strategic with the consequence that networks are endogenously determined. Given that people appreciate that brokerage advantages are possible, they implicitly seek out opportunities to realize them. This is exacerbated by the presence of unobservable factors that affect a researcher's productivity but also the productivity of his peers. I use a well–established identification strategy exploiting the sudden and unexpected deaths of scientists to obtain exogenous shocks to the collaborative network allowing one to uncover the causal impact of the loss of a coauthor on the productivity of a scientist (Aizenman and Kletzer, 2008, Azoulay et al., 2014, 2010, Becker and Hvide, 2013, Bennedsen et al., 2007, Fadlon and Nielsen, 2015, Fracassi and Tate, 2012, Isen, 2015, Jaravel et al., 2015, Jones and Olken, 2005, Nguyen and Nielsen, 2010, Oettl, 2012, Patnam, 2011). For this purpose, I search for the deaths of medical star scientists from obituaries and memoires and gather 1,111 deaths of star scientists, 127 of which were sudden and unexpected. I build the coauthorship network from a panel of 9 million medical scientists. This dataset derives from the combination of MedLine, a comprehensive database on biomedical publications covering the period from 1965 to 2013, and Author-ity (2009), a database resulting from a name disambiguity algorithm that estimates the probability that two articles in MedLine that share the same author name were actually written by

the same individual.

To examine the performance of scientists had they not lost a star–coauthor, I create an appropriate control group through a propensity score matching procedure. I match on the characteristics of the star, the coauthor and their relationships up to the year of the death. In a difference-in-difference regression, I examine the change in research output of scientists following the sudden death of a star–coauthor. To assess the heterogeneity in the effect by network position, I exploit the variation in the brokerage degree between the deceased stars and their coauthors. It is important to emphasize that I do not model the formation of the network. I take the initial configuration of scientists in the network as given. Given that scientists are not randomly assigned to network positions, this means that results cannot be interpreted as the causal effect of losing a broker. The variation in the network position allows me to quantify how the effect of the treatment (i.e. the death of a star–coauthor) varies by brokerage degree.

The main findings are as follows. First, the sudden and unexpected death of a coauthor leads to decrease in output quantity and quality. The loss of a coauthor represents an 8% decrease in annual publications. The number of publications where the star is *not* a coauthor are also affected confirming that knowledge spillovers might be at play. The effect is long–lasting and mainly affects younger coauthors who are still acquiring skills. Second, network position matters. The higher the brokerage degree (i.e. the greater the dependence of the coauthor on his star to provide access to scientists further away) the greater the decline in productivity following his death. The loss of a star providing only non–redundant links (i.e. broker) compared to a star providing only redundant links (i.e. non–broker) is associated with a 31% decrease in the yearly number of publications. Third, the decline in productivity is consistent with an increase in the cost of accessing knowledge embodied in scientists further away. To examine the access to knowledge as a primary explanatory mechanism, brokerage degree is measured based on the number of new topics access only via the star (rather than coauthors further away), taking care to remove any topic known to the scientists

and/or the star. The death of a star who provided access to only non–redundant topics is associated with a 16% decrease in the yearly number of publications. The importance of brokerage degree remains remarkably robust to alternative mechanisms. The fact that coauthors with a high brokerage degree to their star are significantly more severely affected by their death can be an indication that brokerage degree is correlated with some unobserved characteristics of the star, the coauthor, the local network and the type of coauthorship. The death of a star involves the loss of knowledge and access to resources provided by the star himself. More able or better connected coauthors may have an easier time to recover from the loss. The local network, measured by other centrality measures, may in fact serves as a conduit of knowledge. Finally, certain ties yield better or more output.

From a general perspective, these results are indicative, in both significance and magnitude, of the potential importance of network externalities in productivity and provide insight into the mechanisms behind the creation of knowledge. Uncovering the influence of network position is important given the policy implications for the optimal allocation of public research and development funds, and the design of research incentives that foster innovation and potentially economic growth. Biomedical research receives large public subsidies. The United States allocated $30.4 billion in 2015 to research conducted by the National Institute of Health (NIH). Public funding bodies, like the NIH, whose aim is the overall advancement of science and the diffusion of knowledge may want target public funds towards "key scientists", i.e. scientists who generate the highest possible aggregate scientific output in a network. Given the role of a scientist's position, key players are not necessarily the scientists with the highest number of publications (Ballester et al., 2006). Moreover, policy measures aimed at fostering partnerships, networking and knowledge sharing are generally assumed to increase the effectiveness of the system.

This paper contributes to the growing empirical research on peer effects in scientific research. A number of recent empirical studies exploit supply shocks generated

by a variety of natural experiments to single out a causal relationship. The papers that are closest to the research presented here are Azoulay et al. (2010) and Oettl (2012). Both investigate medical scientists and, utilising unexpected deaths of stars as a natural experiment, report significant reduction in output among those who remain. Azoulay et al. (2010) finds that coauthors suffer a 5 to 8% decline in their quality-adjusted publication rates after the death of a superstar coauthor. Oettl (2012) builds on Azoulay et al. (2010) and distinguishes between helpfulness and productivity of a star. Coauthors of highly helpful (not highly productive) scientists that die experience a decrease in output quality but not output quantity. Using the same research design, I present another dimension of heterogeneity in the loss of a star by providing evidence for the role of network position. Although the death of high–productivity stars has a negative impact on the performance of their coauthors, the differential effect between brokers and non–brokers is stark: brokers negatively affect the performance of their coauthors when they die whereas non–brokers do not. Using the the dismissal of scientists in Nazi Germany as an exogenous change in the peer group, Waldinger (2010, 2012) find that the quality of Ph.D. students declined in affected departments while the productivity of colleagues left behind was unaffected. Borjas and Doran (2012b) examine the impact of the large influx of Soviet mathematicians into the United States after the collapse of the Soviet Union and find that the output of American mathematicians with the most Soviet-like research programs fell dramatically. In considering the peer effects in science, most studies have focused their attention on one-degree, egocentric coauthorship or colleague network without specifically examining the wider network these scientists are embedded in. Peer effects are generally conceived as an homogenous dependence across members, and correspond to an average intra–group influence. I extend these studies by studying peer effects through the lens of network position. Despite the growing consensus that social interactions as encoded by a network of relationships matter, the specific effect of different elements of network structure on innovation remains unclear.

This paper also fits into the literature on how information flows through a social

network, how different nodes can play structurally distinct roles in this process, and how these structural considerations shape the evolution of the network itself over time. The sociology literature on social network is well–established (Borgatti, 2005, Wasserman and Faust, 1994). Many studies show the powerful impact of social structure and networks on the extent and source of innovation and its diffusion as reviewed by Rogers (2003). In particular, the notion of structural holes (Burt, 2009), which provides brokerage opportunities and brings social capital, draws on network concepts that emerged in sociology during the 1970s; most notably Granovetter (1973) on the strength of weak ties, Freeman (1977) on betweenness centrality, Cook and Emerson (1978) on the benefits of having exclusive exchange partners, and Burt (1980) on the structural autonomy created by complex networks. Unlike other centrality measures, brokerage is pair–specific and asymmetric and captures flows which are local in nature.

In economics, most empirical studies find that better connected or better positioned individuals benefit from their network position in a variety of settings.[1] Calvo-Armengol and Jackson (2004) describe a network model of information exchange about job opportunities. They show that peer effects in drop–out decisions vary in equilibrium with network location. Ballester et al. (2010), Calvó-Armengol and Zenou (2004), Patacchini and Zenou (2012) embed criminal activities in a social network model. They study the effect of the structure of the network on crime and show that the location in the social network of each criminal not only affects his/her direct friends but also friends of friends, etc. Banerjee et al. (2013a) highlight the role of entry points and document that in villages where leaders occupy central positions in the village network, adoption of a micro finance product is higher. In the context of scientific research, Ductor et al. (2014) show that an important determinant in the prediction of future research

[1]Centrality is important in explaining job opportunities Granovetter (1995), Hellerstein et al. (2015), exchange networks Cook et al. (1983), Marsden (1982), peer effects in education and crime Calvó-Armengol et al. (2009), Hahn et al. (2015), Haynie (2001), power in organisations Brass (1984), the adoption of innovation Coleman et al. (1966), the creativity of workers Perry-Smith and Shalley (2003), the diffusion of micro finance programs Banerjee et al. (2013b), the flow of information Borgatti (2005), the formation and performance of R&D collaborating firms and inter-organisational networks Uzzi (1997), the success of open-source projects Grewal et al. (2006) as well as workers' performance Mehra et al. (2001).

output over and above the information contained in past performance is a researcher's coauthorship links. My results confirm the importance of network position, quantified through a new measure of brokerage, on scientific productivity.

The remainder of the paper is organised as follows. The next section presents the brokerage measure and the intuition behind its importance on the productivity of scientists. The construction of the dataset is described in section 3.3. Section 3.4 outlines the identification strategy. Section 3.5 presents results on the productivity of scientists following the sudden death of a coauthor and then examines the heterogeneity in the effect across the network position of the deceased scientist. To establish the loss of network spillovers as the primary explanatory mechanism, I first rule out alternative mechanisms by looking at characteristics of the coauthor, the star himself, the network and the coauthorship. I then show that the decline in productivity following the death of a co-author is driven by the fact the scientists lost a coauthor who provided access to knowledge embodied in scientists further away. Section 3.6 concludes. Details of the dataset construction, additional descriptives statistics, results and robustness checks are deferred to the Appendices.

3.2 Motivating and Defining Brokerage

To help motivate the importance of network position within a scientific community, consider the discovery of the double helix structure of DNA. Rosalind Franklin, an English biophysicist and X–ray crystallographer, was studying DNA at King's College London. Her experimental work proved to be crucial for the work done by Francis Crick and James Watson, two molecular biologist at the University of Cambridge. Their work resulted in the correct molecular model for the structure of DNA –the double helix– for which they won the Nobel Prize in 1962. By all accounts, Crick and Watson had little interaction with Franklin. They exploited some of Franklin's unpublished work, which was shown to them, without her knowledge, by her colleague, Maurice Wilkins (Watson, 2012).

The controversy surrounding the discovery of the double helix illustrates three key features of the nature of scientific research. First, scientists do not work in isolation. They are part of a wider network of scientists within which knowledge is shared. Second, despite the importance of written communication in science, the majority of scientific communication still takes place through private conversation as knowledge remains tacit long after discoveries are made (Newman, 2001c). In this way, knowledge, at least when it is new, is embodied in particular individuals; it cannot diffuse rapidly, and it cannot be easily-duplicated. Therefore personal scientific collaborations play an important role in the exchange of ideas and opinions and facilitates the generation of new ideas. Third, innovations can arise from the unexpected synthesis of multiple ideas, each of them on their own perhaps well–known, but well–known in distinct and unrelated bodies of expertise.

Now consider these observations in the context of a simple network illustrated in the figure 3.1. The network represents a community of scientists consisting of nodes (i.e. the scientists) and edges (i.e. coauthorships). Each scientist embodies unique ideas, knowledge and specialized skills. Knowledge circulates from one scientist to the next along the edges. Scientists can thus receive knowledge not only from direct coauthors but also from scientists further away through common coauthors. The node B is connected to four other nodes: A, C, D, and E but the link between B and A is different from B's other links. The links to C, D and E connect her to a tightly–knit group. Given that knowledge is more homogeneous within a tightly–knit group than in a more open group with few overlapping links, these links expose B to similar opinions and ideas. The link to A however reaches into different parts of the network, offering access to knowledge B would not otherwise hear about. Given the opportunity to generate novel ideas by combining these disparate sources of information in new ways, the more non–redundant knowledge a scientist receives, the more productive he becomes. This means that the link to A is especially important for B's productivity. My aim is to quantify the special dependency between a scientist and his neighbor in providing access to knowledge embodied in scientists further away.

The edge joining nodes $A - B$ is defined as a *local bridge*. Deleting the edge between A and B would cause the distance between them, as measured by minimum number of co–authors indirectly linking A to B, to a value strictly more than two. They have no common coauthor. However, the $A - B$ edge isn't the only path connecting A and B. They are also connected by a longer path through E and F. The *span* of a local bridge is the distance its endpoints would be from each other if the edge were deleted. Here the $A - B$ edge is a local bridge with span three. An edge joining two nodes is a *bridge* if deleting the edge would cause the two nodes to lie in two different components. In other words, this edge is literally the only route between them. This is an extreme case of a local bridge. If the cost of accessing knowledge from two nodes is a function of the distance between them, then the dependency of one scientist to his neighbor should be a function of the span.

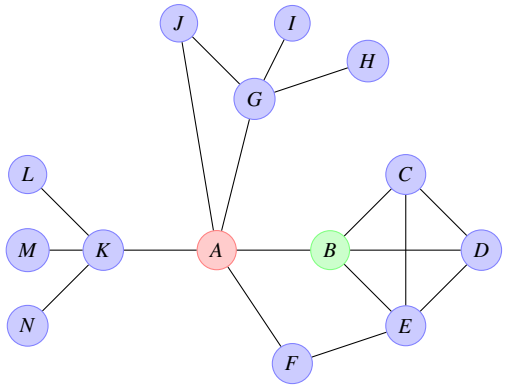


Figure 3.1: Example: Local bridge, neighborhood overlap and brokerage degree

The notion of local bridges is closely related to Granovetter (1973)'s "weak ties". Local bridges tend to be weak ties because if they weren't, triadic closure[2] would tend to produce short-cuts that would eliminate the local bridge. Weak ties serve to link different tightly–knit communities that each contain a large number of stronger ties. This means that more novel information flows to individuals through weak than through strong ties. Burt (2004, 2009) extended and reformulated the weak ties argument by

[2]Triadic closure is the property among three nodes A, B, and C, such that if a strong tie exists between A-B and A-C, there is an increased likelihood that B-C will form a link – a structure called *triangle*. The terms "triadic closure" comes from the fact that the B-C edge has the effect of "closing" the third side of the triangle.(Easley and Kleinberg, 2010)

emphasizing that what is of central importance is not the quality of any particular tie but rather the way different parts of networks are bridged. He emphasizes the strategic advantage that may be enjoyed by individuals with ties into multiple networks that are largely separated from one another. In his words, node A, with her multiple local bridges, spans a *structural hole* – the "empty space" in the network between two sets of nodes that do not otherwise interact closely.

A well–known measure used to measure the local bridges is the one of neighborhood overlap.

Definition 1 *Neighborhood overlap of an edge connecting A and B is the ratio*

$$NO_{AB} = \frac{\#\textit{ nodes who are neighbors of } \text{both } A \textit{ and } B}{\#\textit{ nodes who are neighbors of } \text{at least one of } A \textit{ or } B} \tag{3.1}$$

where in the denominator we don't count A or B themselves.[3]

In the example 3.1, the neighbourhood overlap between A and G (NO_{AG}) is 1/6 as only J is a common neighbour of both A and G. If the bridge between A and G were to disappear, the distance between A and G would be two. Therefore the neighborhood overlap is the continuous measure of a local bridge of span of two. A has no common coauthors with her other neighbors ($NO_{AB} = NO_{AK} = NO_{AF} = 0$). This means that these edges acts more as a local bridge than the edge between A and G.

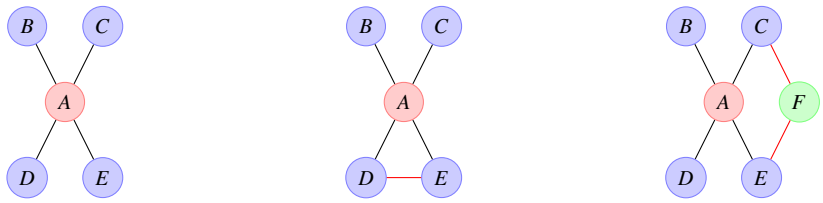
Neighborhood overlap is a pair–specific and symmetric measure which takes into account the network up to two links away. It measures how much common knowledge they share. However, my interest lies in how much one scientist depends on a coauthor to provide him access to *new* knowledge embodied in scientists further away. To quantify this dependency, I propose a pair–specific and asymmetric measure, called *brokerage degree*, which takes into account a larger local network than neighborhood overlap.

[3]The numerator of the neighborhood overlap ratio is known as the *embeddedness* of a edge (i.e. the number of common neighbors the two nodes share).

Definition 2 *The brokerage degree* $0 \leq b_{AB} \leq 1$ *is a continuous measure defined at the coauthorship level.*

$$b_{AB} = \frac{\#\ nodes\ who\ are\ neighbors\ of\ B \text{ but not } A}{\#\ links\ of\ B\ -\ 1} \tag{3.2}$$

where the numerator will be called *# of non–redundant nodes*. Dividing the number of non–redundant nodes scientist B offers to A by the number of coauthors of B allows us to take into account the local neighborhood of B, making it a scale–free measure. Brokerage degree therefore represents the share of *non–redundant* nodes scientist B offers to A. Specifically, it counts the number of new coauthors a specific coauthor offers access to, taking care to eliminate coauthors with direct links to one another and coauthors with a middleman linking them (as illustrated by the three examples in the figure 3.2). Unlike neighborhood overlap, brokerage degree is based on the network up to three links away whereas neighborhood overlap is based on a local network up to two links away. In the example, $b_{AB} = b_{AG} = \frac{2}{3}$, $b_{AK} = \frac{3}{3}$, $b_{AF} = 0$, $b_{GA} = b_{BA} = \frac{3}{4}$. K only provides non–redundant links to A whereas F provides only a redundant one.



$b_{BA} = 1$
$b_{CA} = 1$
$b_{DA} = 1$
$b_{EA} = 1$

$b_{BA} = 1$
$b_{CA} = 1$
$b_{DA} = \frac{2}{3}$
$b_{EA} = \frac{2}{3}$

$b_{BA} = 1$
$b_{CA} = \frac{2}{3}$
$b_{DA} = 1$
$b_{EA} = \frac{2}{3}$

Figure 3.2: Degree of brokerage

This measure of brokerage is closely related to the the density of a "local" network (close or loose–knit network) (Harary, 1969). Counting the number of ties observed in the network formed by a scientist and his immediate coauthors and dividing it by the ratio of possible ones. It rests on the fact that the denser the network, the more alternative paths there are along which information, ideas and influence can travel between

any two nodes. Pairs of scientists with high degree of brokerage tend to form in less dense networks. Yet another closely related measure is betweenness (Freeman, 1979). This measure counts the number of times a given scientist falls along the shortest path between two others scientists. Although closely related to these measures, brokerage degree is pair–specific and measures a feature of the relationship between two nodes within a very local network.[4]

Scientists with high brokerage degree to their coauthors benefit in a number of ways. The first benefit is an informational one: they receive earlier access to a valuable pieces of information originating from multiple and non–interacting parts of the network (Burt, 1997, 2004). Moreover, scientists can benefit from referrals. It is easier for two scientists to meet and connect if they have a mutual coauthor. A long line of research in sociology has argued that if two individual are connected by an embedded edge, then this makes it easier for them to trust one another, and to have confidence in the integrity of the transaction (Coleman, 1988, Granovetter, 1985, Uzzi, 1996). In section 3.5, I examine the importance of brokerage degree on the productivity of a scientists, the new links formed and new topic explored through potential referrals.

3.3 Data and Descriptives

3.3.1 Publications

Publication information is taken from the MedLine, the National Library of Medicine's bibliographic database. It contains over 20 million publications in leading biomedical journals over a period covering 1965 to 2013. The following information on each publication is provided: title of article, abstract, journal name, author list, medical subject heading (MeSH)[5], language, and affiliation. In order to capture the quality of scientific output, I also examine the number of publications weighted by the journal

[4]Formal definitions can be found in Appendix B

[5]Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the United States National Library of Medicine to index all the articles in MedLine and consists of 24,767 terms arranged in a hierarchical structure.

impact factor and the number of citations received.[6]

A well-known problem with this dataset is that authors are not uniquely identified. In the case of MedLine, almost 2/3 of authors have an ambiguous name (where their last name and first initial is shared with one or more other authors) (Torvik and Smalheiser, 2009). The "John Smith" issue has been the topic of many disambiguate algorithm (i.e. the identification of individuals from author names, or equivalently, the identification of author oeuvres within a literature).[7] I rely on the disambiguate algorithm "Author-ity" developed by Torvik et al. (2005). It is based on the idea that different articles written by the same individual will tend to share certain characteristic article attributes, much more so than pairs of articles authored by different individuals.[8] The resulting database, Author-ity 2009, based on a snapshot of PubMed[9] taken in July 2009, assigns each article to one of 6.7 million inferred author-individual clusters.[10]

Combining the MedLine and Author-ity datasets, I trace the publication record of 9 million scientists over their entire career and extract the coauthorship network from

[6]I use journal ranking developed by SCImago which is based on the Google PageRank algorithm. The indicator shows the visibility or prestige of the journals contained in the Scopus database from 1996 by the frequency the average article in the journal has been cited in a particular year. Subject field, quality and reputation of the journal have a direct effect on the value of a citation. SJR also normalizes for differences in citation behavior between subject fields. The number of citations received per publications is retrieved from Scopus. One drawback is that the journals covered in the Scopus and MedLine do not perfectly overlap. Therefore there are publications in journals without any impact factor or any forward citations.

[7]Author name disambiguation is one of the great unresolved issues in bibliometrics. Kang et al. (2009) provide an excellent review of the problem and of the literature that explores approaches for dealing with it.

[8]The algorithm is based on an unsupervised machine learning algorithm, fed with information extracted from each record in PubMed using, among others, text processing on several fields of the record. They present a probabilistic model that describes how two articles in MedLine, sharing the same author name (last name and first initial), were written by the same individual given how similar the two articles are across 8 dimensions: middle initial match, suffix match, journal name, language of article match, number of coauthor names in common, number of title words in common after preprocessing and removing title-stopwords, number of affiliation words in common after preprocessing and removing affiliation-stopwords, and the number of MeSH words in common after preprocessing and removing mesh-stopwords. The method is applied to the entire PubMed/Medline database of roughly 15 million records, and is computationally expensive.

[9]PubMed includes MedLine and PubMed-not-records.

[10]The Author-ity 2009 database is available for nonprofit academic research, and can be freely queried via http://arrowsmith.psych.uic.edu

the joint publications. I consider two scientists to be linked if they have ever published together. Links can never be broken. To guarantee that the sample represents actual researchers and not practitioners with an "accidental" publication, I keep authors who have at least five years of career, a minimum of two publications and at least one coauthor.

The general descriptives of the publication and patent datasets are presented in B.5 in the Appendix. The patterns described have been widely documented. The average number of authors on a publication and the specialisation of scientists has increased over time. These patterns hint at the rising burden of knowledge as documented by Jones (2009). The number of publications per scientists has remained stable over time. The number of clusters has increased and so has the assortative coefficient. There are few very large collaborations in the database, and yet there are a small number of individuals with very large numbers of collaborators. One possibility is that it is the result of the practice in the biomedical research community of laboratory directors signing their name to all (or most) papers emerging from their laboratories. One can well imagine that, with some individuals directing very large laboratories, this could generate authors with a very high number of collaborators.

3.3.2 Obituary Records

In an attempt to collect as many obituary records as possible, manual searches and web scraping of medical obituaries and memoires are performed. Information on the deceased, the year of birth and death, and the cause of death are gathered.[11] I categorise the cause of death in "sudden" and "anticipated" groups.

I then identify the deceased scientists in the Author-ity publications through first and last names, and the publication years relative to the age (proxied by the year of first publication) and the year of the death. The matching is by no means trivial and its

[11] Obituary sources are detailed in the Appendix B. In contrast to Azoulay et al. (2010), who start from a pool of scientists and search for the cause of death, I gather obituary records relevant to medical scientists and match these records to publication.

validity is essential to the credibility of this exercise. I initially eliminate the following: (1) all individuals younger than 18 years old the year of their first publication, (2) individuals whose first publication occurs within five years preceding their death if they were 50 years old or older at the time of their death and, (3) individuals whose last publications are more than 10 years before or after their death. I then proceed by iterations until a unique identifier is found for an obituary record. First, I match the first and last name. Second, I restrict the sample to individuals with a first publication between the ages of 25 and 40 and a last publication within 5 years pre– or post–death. Finally, I select the scientists with the largest number of publications. Details of the procedure to link the can be found in Appendix B.[12]

In Appendix B, I present an example of a deceased scientist and the match to a publication record. "Jane Doe" and her husband, both researchers at the Johns Hopkins University School of Hygiene and Public Health, died in a plane crash in 1998. They were both leaders in the research on AIDS. The New York Times and the Gazette of the Johns Hopkins University both feature news of the accident. "Jane Doe" published in May of the year of her death in the Journal of Infectious Disease as a lead author. All coauthors of this articles are the focus of my analysis. For instance, Weinhold K., the second author if this article, later published in 2011 in the PLOS pathogens.

The final dataset contains 1111 deaths of star scientists and among the ones with a cause of death, 127 are sudden and 224 are anticipated. Researchers recorded in the obituaries die on average at the age of 56 due to some form of cancer. The leading causes of sudden death are heart attacks and car accidents (see tables B.3 and B.4 in the Appendix).[13]

[12]The fuzzy string matching may lead to false positives. These measurement errors relate to the name matching of the obituary records to the MedLine publication database. To alleviate these concerns, I also examine a subsample only including deceased star scientists for which the quality of the string matching was perfect. Reassuringly, the results do not seem to be driven by the quality of the string matching (see table B.9 in the Appendix).

[13]The deaths collected represent approximately 1% of deaths in my sample assuming that medical scientists are as likely to die as the average American.

3.4 Empirical Strategy

3.4.1 Estimating Knowledge Spillovers

My aim is to observe whether, and if so to what extent, the relative position of two neighboring scientists in the local co–authorship network affects their productivity through knowledge spillovers. The evolution of a coauthorship (i.e. the links formed or broken) are endogenous. Given that there tends to be strong similarities across linked individuals, the endogeneity could generate correlations in their productivities even when there are no spillovers. The sudden and unexpected death of a coauthor provides a quasi–natural experiment which exogenously breaks network links and changes the local network structure for reasons unrelated to the productivity of surviving scientists.This allows us to examine the causal effect of the loss of a coauthor on the productivity of a scientists in a difference–in–differences strategy.[14] I estimate the following equation for a scientist i who coauthored with scientist j,

$$y_{i,t} = \alpha \text{ post death}_{j,t} + \beta \text{ post death}_{j,t} * b_{ij,death} + \text{ age}_{i,t} + \rho_t + \tau_{ij} + \varepsilon_{i,t} \qquad (3.3)$$

where $y_{i,t}$ is the outcome variable of scientist i at time t. The variables of interest are *post death* which is equal to one for coauthors of scientists who die for years after the death and the variable $b_{ij,death}$ is the brokerage degree from j to i at the time of death as defined in section 3.2. Because brokerage degree is time–invariant, I only identify brokerage degree through the interaction with *post death*. The coefficient α captures the causal effect of coauthor j's death on the net change in scientist i's productivity. The coefficient β captures the change in productivity of scientist i if the brokerage degree of j to i is one at the time of his death (i.e. all of j's links are non–redundant for i) relative to the productivity change when the brokerage degree is zero (i.e. all of j's links are redundant for j).

[14] I take the initial network as given and do not model the strategic formation of the coauthorship network. There has been an active area of research on strategic network formation (see e.g. Aumann and Myerson (1988), Bala and Goyal (1999), Bramoullé et al. (2014), Dutta et al. (2005), Jackson and Wolinsky (1996)) including papers that explicitly incorporate the notion of structural holes (Buskens and Van de Rijt, 2008, Goyal and Vega-Redondo, 2007). These theoretical papers are highly stylized, and hence difficult to directly apply to empirical work.

I evaluate a researcher's productivity through several measures. First, the annual number of publications[15] is used to capture the quantity of scientific output. Second, I examine the quality of scientific output by weighting the number of publications by the journal impact factor and the number of citations received. Both these measures are widely employed as tools for evaluating performance. The citation weighted publication has the advantage over the journal impact factors to differentiate between the quality of publications within the same journal. However, citations as a proxy for quality is not without problems (Coupé et al., 2010, Torgler and Piatti, 2011). Some fields attract more citations (Arrow et al., 2011, Cole and Cole, 1971) and citations can be driven by fashion (Van Dalen and Klamer, 2005). Moreover, due to censoring publications near the end of my period each publication does not have the same opportunity of being cited. Nevertheless, citations are highly correlated with the assessed quality of papers and remain a valuable metric for evaluating the significance of a publication through peer–ratings. Lastly, for a subset of the scientists the number of patents are used as a measure of innovative output. In addition to the productivity of scientists, I look at the effect of the death on other margins, for instance the number of new coauthorships formed and the number of new MeSH codes explored.

A set of career age brackets dummies are included to control for life–cycle changes in productivity. The career age is measured by the number of years since the first publication.[16] The calendar year fixed effect ρ_t is included to take care of yearly fluctuation in publications that affect all researchers in the same way. Research fields may exhibit rapid growth or decay which has a direct impact on the probability of publishing or collaborating (Levin and Stephan, 1991). τ_{ij} is a coauthorship fixed effect which soaks up all unobserved time-invariant characteristics such as complementarity and social proximity effects (age, gender, ethnicity, month tongue, place of education, stable research

[15]The number of publications per author is highly skewed with a few outliers. MedLine contains many journals which do not have any journal impact factor. Therefore the number of publications is winsorized at the 1% level. The results are robust to winsorizing at the 5% level.

[16]I do not observe the start of a career (end of Ph.D. or first job). I will examine how this assumption affects results in section 3.5.1.1.

interest, mutual empathy, complementarity in skills etc.). In particular, it accounts for the level of brokerage degree as it is measured at the time of death.

I estimate equation 3.3 using OLS with robust standard errors clustered at the star level. This takes into account the possible serial correlation in a scientist's outcomes over time and the possible correlation in the outcomes of scientists linked to the same star. [17]

3.4.2 Appropriate Control Group

The aim of my analysis is to examine what would have happened to the performance of scientists had they not experienced the sudden and unexpected death of a coauthor. One possible strategy is to estimate equation 3.3 for the treated group only (i.e. scientists who are affected by the death of a star–coauthor). In this specification, the control group for scientists who have experienced the death of a star–coauthor consists of scientists who have experienced the loss of their star in the past or will in the future. The life–cycle patterns of control scientists would not be appropriately captured, especially if there is indeed a negative effect on the productivity following the death.[18]

An alternative specification is estimating equation 3.3 over the entire sample of scientists. However this could potentially lead to biased estimates due to selection. Figures B.1 and B.2 in the Appendix show that the deceased star scientists are at the top end of the distribution in terms of the number of publications and the number of coauthors.[19] These figures point to the lack of balance and only partial overlap between the distribution of the deceased stars and the general population of scientists. Having been

[17] Given the nature of the outcome variables which are highly skewed and nonnegative, one may prefer to perform a conditional quasi-maximum likelihood Poisson specification (Hausman et al., 1984). Results are robust to this specification

[18] Results of this specification are similar to the baseline ones and can be found in table B.13 in Appendix.

[19] The distribution of the number of publications and coauthors are highly skewed. One possible explanation is that it is the result of the practice in the biomedical research community of laboratory directors signing their name to all (or most) papers emerging from their laboratories. One can well imagine that, with some individuals directing very large laboratories, this could generate authors with a very high apparent number of collaborators.

archived in an obituary or a memoire, it is likely that the group of deceased scientists are a highly selected group of "star" scientists, meaning that they were at the time of their death of relatively higher productivity or had promising careers. Moreover, this selection bias is likely to spill over to the coauthors of the deceased scientists due to homophily. This is problematic as the full population of scientists in the data may not be the appropriate counterfactual for the coauthors of deceased star scientists.

One method for dealing with this selection issue is to find other stars who exhibit no systematic difference in output trends to the treated group up to the time of death. Matching will discard observations so that the remaining data show good balance and overlap. I match deceased scientists to five scientists that are as likely to be archived in an obituary when they die, conditional on a vector of observable characteristics.[20] For every year of death and cohort (proxied by the year of first publication), a nearest neighbour propensity score matching without replacement is implemented. First, I collected the pool of potential controls consisting of all researchers who have started their career around the same time as the deceased star: their first publications is within a three years window of the deceased star's. From this set, I rule out (1) all deceased scientists, (2) scientists that have already been matched to deceased stars in previous rounds of matching, and (3) all scientists who are coauthors of the deceased scientist. Given the organisation of research with large components and coauthorship networks often being tight-knit, the last criteria creates a "buffer" between the treated and controls. For the remaining potential control pairs, I then measure their productivity up to the year of death. Last, I estimate the propensity of a scientist to be treated (i.e. to be mentioned in an obituary), using a logistic regression of the likelihood of a scientist being recorded in an obituary record as a function of the characteristics measured at the year of death for the deceased star and the pseudo–year of death for the potential controls. The variables include the cohort, the cumulative number of JIF–weighted publications, and the cumulative number of coauthors. In choosing the covariates I

[20]The decision to match each deceased star to *five* scientists is motivated by the fact that there will be a second matching procedure for the coauthors of the deceased and matched stars. Having five matched stars allows for a trade–off between the two matching procedures.

control for systematic differences between stars and scientists on dimensions that are possible sources of status. The matching procedure results in a pool of stars: each of the 701 deceased star scientists is matched to five other star scientists.[21]

A series of QQ–plots are presented in figures B.3, B.4, and B.5 in the Appendix. They compare the empirical distribution of treated and matched pairs on a few key variables used to construct the match. The quantities of the treated and control samples are calculated and plotted against each other. If the empirical distributions are the same in the treated and control groups, the points in the QQ–plots should all lie on the 45 degree line. We can see that the central points seem to agree fairly well for all three variables of the first matching, but there are some discrepancies at the tails of the distributions for the number of publications. I also examine the QQ–plots on other variables for which I do not match on in figures B.6, B.7, B.8, B.9, B.10, and B.11) of the Appendix. They show that the matching procedure appears to have improved on the fit on other network centrality measures.

All scientists who have coauthored with these stars (the deceased stars and the matched stars) are gathered. I remove all coauthors who are associated with two or more stars. I also eliminate from the sample all coauthors who are both treated and control. However, simply comparing the two groups of coauthors might still lead to erroneous conclusions. First, there are many reasons to believe that there is a life-cycle to collaboration, with their productive potential first increasing over time, eventually peaking, and thereafter slowly declining (Levin and Stephan, 1991). Second, coauthorship pairs are embedded in local communities of researchers which can be more or less dense. In a dense community, all researchers are connected to one another and the degree of brokerage is likely to be very low. Since I have not controlled for any local network

[21] I have assumed that the star status is a time–independent characteristic of a scientist. That is, a deceased scientist is considered a star throughout the analysis if s/he is recorded in an obituary record. The "star" status and health conditions or early death may be correlated. For instance, stars may be more likely to experience stress leading to adverse health conditions and eventually early death. The issues of immortality time bias and healthy survivor bias have been examined by Han et al. (2011), Redelmeier and Singh (2001) and, Rablen and Oswald (2008). Work in medical science has shown that, on the contrary, measures of socio–economic status are associated with better health and longer life.

characteristics, the brokerage relationship may be very different between coauthors of deceased star scientists (the treated group) and coauthors of the matched scientists (the control group). Finally, a more able coauthor is probably better able to turn the ideas accessed through their associated star into publications than a less able researcher and a well–connected coauthor can more easily replace the lost coauthor.

Taking these possible biases into consideration, I perform a second matching procedure on coauthorship pairs within the pool of coauthors of the deceased and matched stars. Ideally, the two separate matching procedures would be implemented directly in one step based on coauthorships characteristics and characteristics of individual researchers within the team. However, this is computationally infeasible due to the size of the dataset. Matching each deceased star to five matched stars in the first matching allows for a trade–off between the two matching procedures. The second matching is based on the characteristics of the coauthors (the cumulative number of quality–weighted publications, the cumulative number of coauthors at the time of the death of their associated star, the cohort and the age at the time of death) and the characteristics of the coauthorship itself (the brokerage relationship, recency of the coauthorship measured as the number of years since the last coauthorship, and the strength of the coauthorship measured as the number of joint publications). For each year of death, a one–to–one nearest neighbour propensity score matching is implemented. Calipers are applied to ensure that no matched group is too dissimilar (all matches not equal to or within 0.25 standard deviations are dropped). Out of the 304,703 coauthors of stars, I match 80,284 (40,142 treated and 40,142 control). Figures B.12 to B.17 in the Appendix show that the second matching procedure resulted in a good balance and overlap.

Table 1 presents summary statistics for the variables of interest in the samples build in the first and second matching for sudden deaths (table D2 in the Appendix presents the same statistics for sudden deaths only). All variables are computed based on years prior to the year of death (or pseudo–year of death for the controls). I define a scientist's

cohort as the first year he appears in the dataset (i.e. the year of his first publication). Deceased star scientists have 62 publications and 82 coauthors by the time of their death. Coauthors are approximately 8 years younger than their associated stars and lag behind them both in terms of publications and number of coauthors. Overall, the matched group is not systematically different from the treated group on aspects relevant to their publishing behaviour, except for the death of a coauthor.

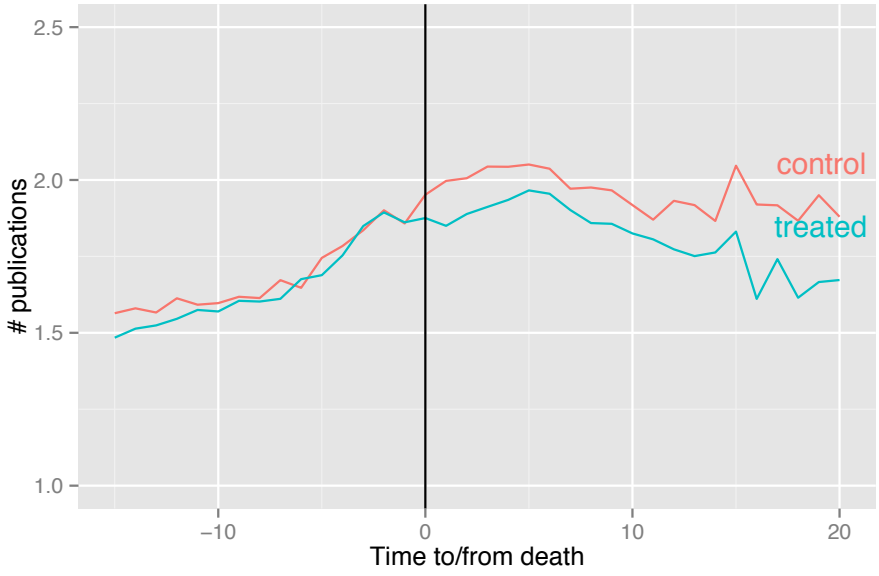
Table 3.1: Descriptives

	Mean	Median	Std. dev.	Mean	Median	Std. dev.
A. Star Level	Deceased stars (701 Obs.)			Matched stars (3,505 Obs.)		
Age at death	55.87	57.00	8.49			
Cohort	1976.05	1973	9.77	1976.02	1973	9.89
Cum. # publ.	61.72	26.00	84.21	61.56	20.00	98.02
Cum. # JIF-weighted publ.	120.17	28.59	210.90	118.83	25.58	214.44
Cum. # cit.-weighted publ.	2910.04	629.50	5586.48	2563.94	462.50	5277.22
Cum. # grants	11.51	1.00	19.01	8.91	1.00	17.49
Cum. # R01 grants	0.47	0.00	1.10	0.32	0.00	1.09
Cum. # coauthors	82.42	40.00	112.22	81.07	30.00	127.68
Closeness	1.83E-06	2.02E-06	9.26E-07	1.85E-06	2.11E-06	9.27E-07
Betweenness	3.84E+05	4.68E04	1.14E06	5.76E+05	6.28E04	2.99E06
Eigenvector	4.68E-03	1.98E-07	5.41E-02	5.55E-03	9.43E-07	6.17E-02
Clustering coefficient	0.18	0.11	0.20	0.15	0.10	0.19
Triangle	3.54E+03	1.68E02	3.20E04	3.38E+03	1.98E02	3.29E04
B. Coauthor Level	Treated coauthors (40,142 Obs.)			Matched coauthors (40,142 Obs.)		
Cohort	1984.74	1985	10.92	1984.62	1985.00	11.18
Cum. # publ.	19.64	7.00	34.09	19.83	7.00	33.79
Cum. # JIF-weighted publ.	30.27	9.77	58.92	30.98	9.16	62.36
Cum. # cit.-weighted publ.	702.31	189.00	1630.84	673.95	173.00	1582.42
Cum. # MeSH codes	153.62	102.00	156.87	161.52	108.00	163.42
Cum. # coauthors	43.57	18.00	83.57	44.60	19.00	85.84
C. Coauthorship Level	Treated pairs (40,142 Obs.)			Control pairs (40,142 Obs.)		
Strength of coauthorship	2.51	1.00	5.34	2.53	1.00	5.80
Recency of coauthorship	8.75	6.00	7.72	8.81	6.00	7.82
# non-redundant nodes	29.75	10.00	51.45	30.43	11.00	52.00
Brokerage degree	0.21	0.07	0.30	0.20	0.06	0.29
Embeddedness	10.84	4.00	55.86	11.21	4.00	56.61

Notes: The unit of observation is the star (panel A), the coauthor of a star (panel B) and the star-coauthor pair (panel C). The cohort is the year of first publication. All variables apart from the cohort are defined at the time of the death of the star or pseudo-death of the matched star. The R01 grants are NIH grants awarded to individual researchers. The strength of the coauthorship is the number of joint publications between the star and the coauthor, the recency of the coauthorship is the number of years since the last joint publication before the death. The brokerage degree is the fraction of non-redundant nodes offered by the star to his coauthor over all the links of the star as defined by equation 2.

The most intuitive way to view the results is by plotting the output trend before and after the death for the treated and control group without any adjustment for age and calendar time effects. Figures 3.3 plots such a graph (figure B.18 can be found in the appendix for sudden deaths only). The pattern of the treated and controls appear similar up to the year of death from which point the treated exhibit a decrease in their productivity compared to the control group. Ten years prior to the death/pseudo–death, both groups of coauthors publish approximately 1.5 publications per year. The publication rate of both treated and control groups increases steadily over the next few years reaching 1.9 right before the death/pseudo–death. Scientists who suffer the loss of a star–coauthor experience a decrease in productivity in comparison to those who do not suffer this lost. There is no evidence of recovery.

Figure 3.3: Publication Trends for Treated and Control Coauthors



Notes: Mean number of publications around the time of the death. The solid red line corresponds to treated group (coauthors of star scientists who suddenly die) and the green line corresponds to the control group (matched coauthors of star scientists).

3.5 Results

3.5.1 Main Results

Table 3.2 presents the main results on the productivity of scientists both in terms of quantity (columns 1 and 2) and in terms of quality (columns 3 to 6). The quantity is measured by the annual number of publications. The quality is captured by the annual number of publications weighted either by the journal impact factor or by the number of citations received. I focus on sudden deaths only to remove any possible anticipation effects.[22] Columns 1, 3 and 5 first show the direct effect on productivity following the death of a coauthor. Consistent with previous results (Azoulay et al., 2010, Jaravel et al., 2015, Oettl, 2012), coauthors suffer significantly both in terms of the quantity and quality of their productivity following the death of their associated star. The coefficient on post death is -0.136, which translates into an 8% decrease in the yearly number of publications coauthors produce after the star dies.

In the other columns of table 3.2, I examine the heterogeneity in the effect of the death of a coauthor by brokerage degree. The baseline death coefficient becomes insignificant in column 2, indicating that the death of a star providing only redundant links (i.e. brokerage degree equal to 0, also called "star non–brokers") have no negative effect on the number of publications of their coauthors, whereas the death of a star providing only non–redundant links (i.e. brokerage degree equal to 1, referred to as "star–brokers") appears to have a large negative impact. A scientist experiencing the loss of a broker instead of a non–broker will experience a significant relative loss of 0.87 publications per year. This represents a 31% decrease in the yearly number of publications for scientist associated to a star with a brokerage degree of 1. Columns 4 and 6 show that the loss of any star, broker or non–broker, is detrimental to the number of quality–weighted publications. However, coauthors linked to stars providing greater

[22] Unlike anticipated death, sudden deaths do not leave time for the coauthors to react in order to soften the blow by starting new projects or seeking out new coauthors to replace him/her. Therefore I have focused the main results on scientists who have experienced the sudden loss of a coauthor. Column 8 of table B.9 and figure B.22 present the results for anticipated deaths separately and show similar results as in the case of sudden deaths.

brokerage degree suffer an even greater decrease in the quality of publications following the death. These results confirms that the heterogeneity in the treatment effects is relevant for both the quantity and quality of scientific output.[23] [24]

Table 3.2: Main Results

Sample: Sudden deaths						
	(1)	(2)	(3)	(4)	(5)	(6)
	# publ. per year		# JIF-publ. per year		# citations-publ. per year	
post death	-0.136***	0.036	-0.271***	-0.031	-13.221***	-8.221**
	(0.023)	(0.027)	(0.065)	(0.084)	(2.242)	(2.596)
*post death*b*$_{ij,death}$		-0.866***		-1.208***		-25.171**
		(0.115)		(0.237)		(8.055)
R^2	0.066	0.068	0.036	0.037	0.031	0.031
Nb. of obs.	503,748	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications includes a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications (winsorized at the 1% level) in columns 1 and 2, the annual number of publication weighted by their journal impact factors in columns 3 and 4 and the number of publications weighted by the number of citations received in columns 5 and 6. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death.

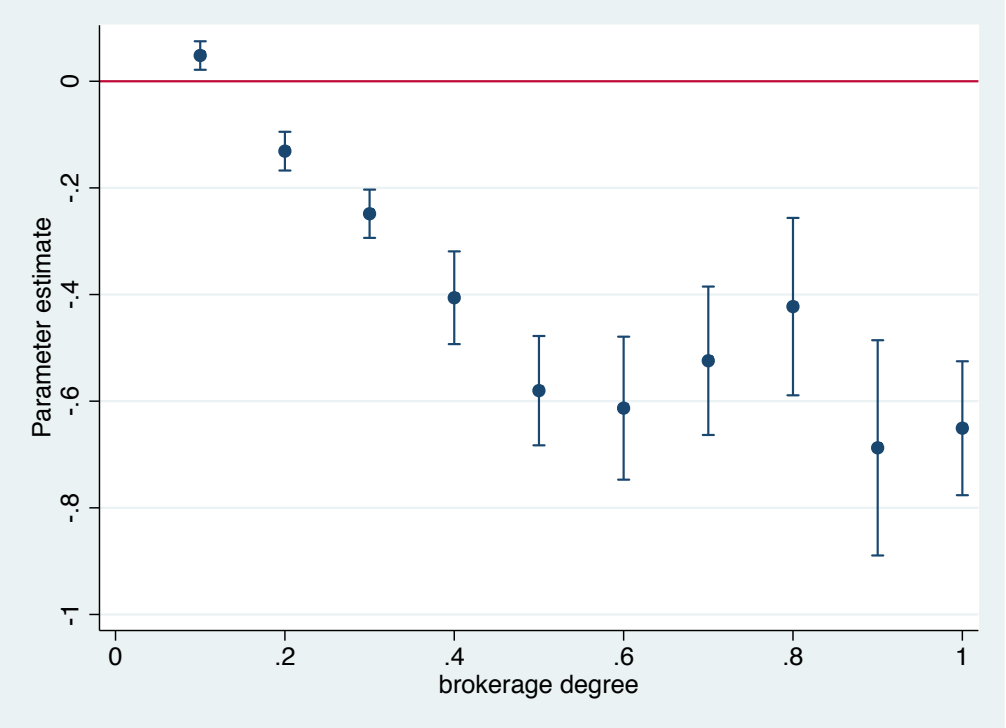
Figure 3.4 plots the interaction between a set of indicator variables corresponding to different levels of brokerage and the treatment status. For brokerage degree ranging from 0 to 0.5, this figure clearly show that the higher the brokerage degrees the larger the negative effect on output. For higher brokerage degree, the coefficients are not significantly different from one another.

Given the significant and negative effect of the death, it is interesting to investigate

[23] In table B.9 in the Appendix, I replicate the main specification using Azoulay et al. (2010)'s death and my data. The death of a superstar is associated with a 6% decrease in the performance of their coauthors. This point estimate is only slightly smaller than the estimate provided in Azoulay et al. (2010), who report a decrease of 8% in their base specification. Azoulay et al. (2010) examines the hypothesis that their superstars broker relationships by computing the betweenness centrality of the deceased superstars. They do not find a significant differential effect between coauthors of centrally located superstars with high betweenness centrality and coauthors of less central superstars.

[24] Instead of using the brokerage degree (as defined in equation 2), one could use the number of non–redundant nodes offered exclusively by the star (i.e. the numerator of brokerage degree). The number of coauthors of the star at the year of the death is included in order to control for the size of the local network. The loss of a star offering an additional non–redundant link is associated with a significant decrease of -0.007 in annual publication (see table B.12 in the Appendix).

Figure 3.4: Results by Brokerage Degree



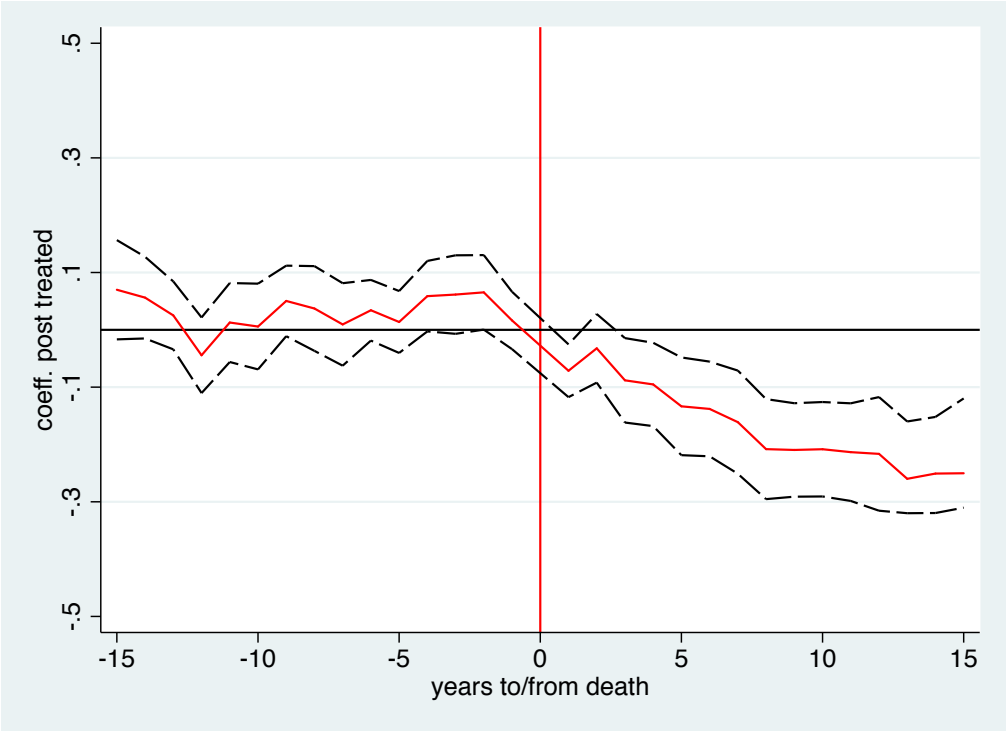
Notes: The solid lines in the above plots correspond to coefficient estimates of OLS specifications in which the number of publication (winsorized at the 1% level) of a scientist is regressed onto year effects, indicator variables corresponding to different career age brackets, and interactions of the *post death* with 10 indicator variables corresponding to different levels of brokerage degree. The bars around these estimated correspond to the 95% confidence interval (corresponding to robust standard errors, clustered around the stars).

how persistent this effect is over time. Figure 3.5 plots the interaction between a set of indicator variables corresponding to a particular year relative to the coauthor's death and the treatment status (having a coauthor who eventually dies) along with the 95% confidence interval.[25] This shows that the estimated causal effects of the death of a star is significant only after the year of death, which alleviates any concerns that death was not really exogenous of collaboration. There is no evidence of recovery. Researchers who are still acquiring skills are presumably more sensitive to changes in peer quality. I therefore explore the time pattern of young and experienced coauthors in figures B.23 and B.24. The cutoff between young and experienced coauthors is defined at 27 years

[25]This is based on a regression which include a full set of leads and lags around the death interacted with the treatment status (i.e. experiencing the death of a coauthor)

of career. Young coauthors suffer a persistent effect whereas experienced coauthors do not appear to be significantly affected by the death.

Figure 3.5: Dynamics of the Treatment Effect



Notes: The solid lines in the above plots correspond to coefficient estimates of OLS specifications in which the number of publication (winsorized at the 1% level) of scientist is regressed onto year effects, seventeen indicator variables corresponding to different age brackets, and interactions of the treatment effect with 30 indicator variables corresponding to years around the year of the death. The 95% confidence interval (corresponding to robust standard errors, clustered around superstars) around these estimates is plotted with dashed lines.

In table 3.3, I look at the impact of the death of the star taking into account the input of other coauthors on each publications. The death of any coauthor generally leads to a mechanical decrease in publications because the coauthor was a direct input in the joint publications. Star scientists, regardless of their degree of brokerage to their coauthors, provide input in terms of time and effort into their joint publications. The number of publications where the star is not a coauthor removes the direct input of the star in the productivity of his coauthors. Examining publications without the star (columns 1 and 2) also sheds light onto the ability to substitute toward new coauthorships or place more time on existing ones upon the death of the star. The death significantly decreases

these publications as well suggesting that coauthors provide more than a direct input into the joint research and actually transfers knowledge useful for other projects and are not easily replaced. The death of star who provided only redundant links actually significantly *increases* the number of publications not involving the star. One explanation is that coauthors of non–broker stars worked in teams involving the star at the head of the lab. Such stars might have been particularly demanding in terms of time. After the death, coauthors now have more time to dedicate to other projects. The loss of a star–broker is associated with a significant decrease in publications without the star. This result hints to the fact that knowledge spillovers might be at play.

The increasing tendency across scientific disciplines to write multi–coauthored papers makes the issues of the sequence of contributors' names a topic of research in terms of reflecting actual contributions (Baerlocher et al., 2007, Laurance, 2006). To control for co-author influence (not only the star's input as in columns 1 and 2), I divide each publications by the total number of authors to reflect the fact that a single author's contribution is smaller in larger teams (Hollis, 2001, Lindsey, 1980, Long and McGinnis, 1982). Columns 3 and 4 confirm the previous results. The loss of a star non–broker has a significant and positive effect while the loss of a star–broker has a significant and negative effect on the number of publication even when controlling for the team size.

Finally, I examine the number of publications for which the scientist appears as a first and/or last authors in the author list. Traditionally, the first author contributes the most and also receives most of the credit. In biomedical sciences, the last author generally gets as much credit as the first author because he or she is assumed to be the driving force, both intellectually and financially, behind the research.[26] I find that the loss of a star–brokers has a significant impact on both these measures. In all remaining specifications, I will include the interaction between post death and brokerage as

[26]This practice is unofficial and hence not always followed, meaning that sometimes last authors mistakenly benefit when they are not the principal investigators.

explained in equation 3.3.

Table 3.3: Alternative Productivity Measures

Sample: Sudden deaths								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	# publ. w/o star per year		# publ. weighted by team size		# publ. as first author		# publ. as last author	
post death	-0.090***	0.088***	-0.027***	0.035***	-0.026**	0.068***	-0.027*	0.003
	(0.024)	(0.026)	(0.008)	(0.009)	(0.009)	(0.011)	(0.011)	(0.011)
*post death*b*$_{ij,death}$		-0.896***		-0.313***		-0.476***		-0.154**
		(0.113)		(0.037)		(0.040)		(0.056)
R^2	0.047	0.050	0.040	0.043	0.037	0.040	0.025	0.025
Nb. of obs.	503,748	503,748	503,748	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications includes a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications without the star in columns 1 and 2, the annual number of publication weighted by the number of authors in columns 3 and 4, the number of publications as a first author in columns 5 and 6, and the number of publications as a last author in columns 7 and 8. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death.

3.5.1.1 Robustness Check

I conduct a number of robustness checks which are reported in table B.8 in the Appendix. I first point two potential sources of biases: the data is right–censored and truncated. Then, I check whether the empirical specification correctly capture the life–cycle pattern of the career of a scientist and the common trends around the year of the death. I finally examine whether the results are sensitive to medical subfields and period of time.

The data is right–censored. I do not know whether the scientists matched to the deceased ones are in fact still alive and active as their year of death or retirement is unknown. This can lead to a possible contamination of the control group. In columns 1 to 3 of table B.8 I examine sudden deaths occurring at a young age for which the matched star scientist is less likely to be deceased. The effects are still presents albeit small in magnitude. The fact that I do not know the end of the career of scientists also has implications for the coauthors of stars. When a scientist is no longer publishing I cannot determine whether it is because he quit research or because he is in fact still an active researcher but has been unsuccessful in publishing. I assume that scientists can have up to 45 years of career (years from their first year of publication). There should be no reason for this assumption to lead to a differential bias between treated and control coauthors. Nevertheless, I examine different cutoffs for the length of a career (see columns 4 and 5 of table B.8) and the main results stay significant.

Due to the limited window of observation (1965–2013), I have to deal with a truncation problem as I do not witness individuals' publications outside this period. For example, recent deaths have many pre–death observations but few post–death observations while the opposite holds for early deaths in the sample. The dynamic specification can confound true dynamics with the changes in the composition of the sample. Column 6 of table B.8 confirms that the results are robust to restricting attention to a balanced panel, focusing on scientists whose star passed away between 1985 and 1993.

A key concern is to correctly model the career of a scientists which may follow a life–cycle pattern. However, the nature of collaboration (Agrawal et al., 2013) and the life–cycle of a scientists (Jones et al., 2014, Jones, 2010) have changed over time. To control for such differences across cohorts, I include the interaction between cohort and decade for both the treated and control groups in column 8 of table B.8. The point estimates are very similar to those obtained when not these interaction effects.

An important assumption for using the death of a scientist as an exogenous shock to the network is that the publication trends of treated (i.e. scientists who experience the loss of a star-coauthor) and the control (i.e. scientists who remain linked to their star-coauthor) would have followed the same trend in the absence of the death. To investigate this identification assumption, I estimate a placebo experiment using placebo years of death for the control group, where those years are drawn at random from the empirical distribution of death across years for the deceased scientists. The results of column 7 of table B.8 report a non–significant coefficient close to zero.

I examine whether my results are robust by decades and fields.The average number of authors per paper and consequently the average number of coauthors per researcher have been increasing over time as can be seen in B.5. Different fields may also differ in their propensity to collaborate. These patterns may have different implications for the importance of brokers. On the one hand, advances in communication and transportation technologies have decreased the distance between scientists. It imply that cost of meeting new scientists further away and accessing their knowledge is now easier (Rosenblat and Mobius, 2004). Therefore, the importance of brokerage may have diminished over time. On the other, research has become increasingly complex and research costs may have been rising over time (Jones, 1995). This in turn implies that brokerage may have a greater role now than before. I find that the loss of a broker (i.e. coauthor who provided only non–redundant links) leads to a decrease in productivity in all decades although it is only significant from 1990 onwards (see table B.11 in the Appendix). The magnitude of the broker effect is particularly large in the period

between 1990 to 1995 and decreases afterwards. I also examine the eight most popular topics among star scientists (see table B.10 in the Appendix) and find that effect of brokers are not driven by any specific field of research despite the fact that research in different subfields is conducted very differently.

3.5.2 Mechanisms

In this section, I explore potential explanations behind the decrease in productivity following the sudden and unexpected death of a star and, in particular, the role of network position. I first rule out alternative mechanisms by examining characteristics of the coauthors, stars and local network measures. I then present evidence in support of the hypothesis that brokers offer access to knowledge by provided exclusive access to new scientists and the ideas they embody.

3.5.2.1 Coauthor Characteristics

If network position is correlated with personal characteristics of scientists, then the results found for brokerage might simply reflect some unobserved dimension of a researcher's ability. More able scientists might have an easier time to recover from the loss of a coauthor. It still takes ability and effort to turn a new idea into a publication. Better–connected scientists might have an easier time to connect to new coauthors. Researchers who are still acquiring skills are presumably more sensitive to changes in their peer quality. Finally, the level of specialization of a researcher might reflect his dependency on the star.

I control for the productivity, connectedness, experience and specialization based on the number of publications, the number of coauthors, the number of years since the first publication and the number of MeSH codes measured at the time of the death of their associated star respectively. *Less productive*, *isolated* and*young* are indicator variables equal to one if the coauthor is among the bottom 25% of the distribution. *Generalist* is an indicator variable equal to one if the coauthor . In table 3.4, I control for these characteristics and find that the death of a star–broker remains significant and negative as in the baseline specification. Moreover, the coefficient remains of similar

magnitude. Younger and more generalist scientists are more affected by the death of a star. This is unsurprising given the fact that researchers who are starting their career or who branch out to many different fields are more depended on their coauthors. On the contrary, less productive and isolated researchers actually benefit from the death of the star. This might be explained by the fact less productive and isolated scientists may be trapped in a relationship. If the search costs for finding a new coauthor are very high, it may be rational to stay in a relationship with the star although coauthors would actually benefit from severing all ties to the star. The death in such cases provide an opportunity to form more profitable collaborations.

Table 3.4: Coauthor characteristics

Sample: Sudden deaths Dep.Var.: Number of publications per year					
	(1)	(2)	(3)	(4)	(5)
post death	-0.054	-0.051	0.097***	0.121***	0.082**
	(0.031)	(0.032)	(0.026)	(0.029)	(0.032)
*post death*b*$_{ij,death}$	-0.717***	-0.722***	-0.925***	-0.892***	-0.733***
	(0.118)	(0.120)	(0.117)	(0.0113)	(0.117)
*post death*less productive*$_{i,death}$	0.281***				0.305***
	(0.026)				(0.025)
*post death*isolated*$_{i,death}$		0.264***			0.142***
		(0.028)			(0.024)
*post death*young*$_{i,death}$			-0.273***		-0.323***
			(0.043)		(0.038)
*post death*generalist*$_{i,death}$				-0.195***	-0.272***
				(0.036)	(0.036)
R^2	0.069	0.069	0.069	0.069	0.070
Nb. of obs.	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications includes a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. This variables *less productive, isolated, young* and *generalist* are equal to one if the coauthor is has fewer than 2 publications, 7 coauthors, less than 5 years since the first publication, and more than 6 MeSH codes.

3.5.2.2 Star Characteristics

Upon the death of a coauthor, a scientist loses the coauthor and all his links. This means that the loss in access to knowledge may come from the star himself rather than from the loss of connection to part of the network. To investigate this possibility, I control for the quality of the star using different definitions of stars: the number of publications (top pub), the number of citations received (top cite), and the career age (experience). Medical science often requires labs and expensive equipment and access to funding is crucial for research. If a coauthor provides resources, the death of such a coauthor should lead to a negative effect on joint productions only. I control for the access to resources of a star through the number of grants (top grant) and the share of publication as a last author. For each of these dimensions, I use an indicator to identify stars in the top 25% of the distribution of these characteristics.[27] In columns 2 to 6 of 3.5, we see that that brokerage degree remains significant even after controlling for these different characteristics. Moreover, the magnitude of the coefficient is quite stable across all specifications and even increases when all these different star characteristics are included. The loss of a star who had many grants is particularly detrimental, confirming the fact that access to resources is important in medical science. It is interesting to point out that, although not highly significant, the loss of a star who is experienced and is a last author in many publications *positively* affects the productivity of scientists. These types of stars might have been particularly demanding in terms of time. Why would a coauthor remain in a coauthorship when severing the link would lead to an increase his productivity? It is still economically rational to continue a collaboration if the cost of rematching outweighs the output gain.

[27] This represents more than 206 publications, 995 citations, 32 years of career, 3 grants, and 60% of publications as last author at the year of death.

Table 3.5: Star characteristics

Sample: Sudden deaths Dep.Var.: Number of publications per year							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Baseline		Knowledge		Access to resources		All
post death	0.036	0.047	0.056*	-0.002	0.077*	0.003	0.057
	(0.027)	(0.036)	(0.028)	(0.036)	(0.030)	(0.036)	(0.033)
$post\ death * b_{ij,death}$	-0.866***	-0.864***	-0.869***	-0.829***	-0.881***	-0.832***	-0.889***
	(0.115)	(0.128)	(0.128)	(0.126)	(0.125)	(0.127)	(0.128)
$post\ death * top\ pub_{j,death}$		-0.100					-0.096
		(0.057)					(0.056)
$post\ death * top\ cite_{j,death}$			-0.088				-0.071
			(0.069)				(0.061)
$post\ death * experienced_{j,death}$				0.075			0.093*
				(0.049)			(0.046)
$post\ death * top\ grants_{j,death}$					-0.156**		-0.153**
					(0.058)		(0.051)
$post\ death * lead\ author_{j,death}$						0.059	0.100
						(0.053)	(0.055)
R^2	0.068	0.055	0.055	0.055	0.055	0.055	0.055
Nb. of obs.	503,748	503,748	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variable is the annual number of publications (winsorized at 1%). The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. All other interaction variables are defined at the star level at the time of his death. The variable top pub, top cite, experienced, top grant, and lead author are indicator variables equal to one if the star is at the top 25% of the distribution of in terms of number of publications, number of citations, years since their first publication, number of grants and share of publications as a first or last author in the author list respectively.

3.5.2.3 Network Characteristics

Brokerage degree tries to capture the dependency of a coauthor on his associate star to provide him with new knowledge. If brokerage degree simply captures an aspect of the local network structure or the structural importance of the star (power, influence, or prestige), there are many other standard centrality measures which could potentially also account for the decrease in productivity. For instance, less connected parts of the network create opportunities for brokerage. So, the number and size of local networks, and their "connection topology" may be consequential for predicting both the opportunities of scientists.

Degree approaches are based on the idea that having more ties means being more important. Closeness measures go slightly further and assumes that actors who are able to reach other actors at shorter path lengths, or who are more reachable by others at shorter path lengths, are in favoured positions. In an information flow context, we can interpret closeness as an index of the expected time until the arrival of new information. Nodes with low closeness have short distances from others, and will receive information sooner. When using betweenness approaches (Freeman, 1977, 1979), it is being an intermediary between many other actors that makes an actor central. Betweenness can be thought of as measuring the volume of traffic moving from any one node to every other node that would pass through a given node. It measures the amount of network flow that a given node controls in the sense of being able to shut it down if necessary. Eigenvector centrality captures indirect reach so that being well-connected to other well-connected nodes makes one more central. Triangles counts how many times a node is part of a triangle.The clustering coefficient goes one step further and measures the probability that a node a part of a triangle. It is the ratio of existing links connecting the node's neighbours to each other, to the maximum possible number of such links. The clustering coefficient is a measure of the extent to which the friends of my friends are my friends. Finally, neighborhood overlap is a pair–specific measure of the number of common neighbors as defined in Section 3.2. All these measures are detailed in the Appendix B along with a correlation matrix in table B.7. The negative

correlation between the mean brokerage degree and degree can be explained by the fact that it is "easier" for a node with two neighbors to get a score of 1 (only one tie is need) than for a node with 10 neighbors.[28]

Table 3.6 reports the estimation results obtained when using these alternative network measures as an additional explanatory variable in the regression. With the exception of clustering coefficient, all centrality measures show a significant negative effect on the productivity after the star dies. In particular, once we control for the degree of the star (i.e. the denominator of the brokerage degree), brokerage degree remains negative and significant and even larger in magnitude. This means that the result is not only driven by the size of the network of the star or the centrality of the star. The loss of a star–coauthor with a high number of triangles is particular detrimental to productivity. When all these network characteristics are included, brokerage degree remains significant (column 9) and of similar magnitude. More generally, these results validate the importance of *localized* network measures.

As discussed in Bloch et al. (2016), Borgatti (2005), Wasserman and Faust (1994), these different measures each make implicit assumptions about the manner in which information flows through the network. For instance, closeness and betweenness are based on the assumption that flows moves along the shortest path, taking one path or the other. Eigenvector centrality counts walks which assume that the trajectories can not only be circuitous, but also revisit nodes and lines multiple times along the way. It assumes that the information can take multiple paths simultaneously. Neighborhood overlap and clustering coefficients are the measures closest in spirit to brokerage degree. However neighborhood overlap restricts the neighborhood of interest to first degree links while clustering coefficient is defined at the node level. Brokerage degree not only captures flows which are local in nature, but also specific to a coauthor. It

[28] All network measures are computed each year based on the network of coauthorship in the last five years. In other words, the network measure at the time of death takes into account all joint projects published in the last five years prior to the death. The betweenness centrality measures have a cutoff of 4 links away.

assumes that information need not take the shortest path and can travel simultaneously through different paths.[29]

[29]For computational reasons, I limit myself to nodes that are up to two links away. An implicit assumption is that flows can only be reached up to two links away. However, conceptually one could easily extend the brokerage degree definition to include flows from further network distances.

Table 3.6: Network characteristics

Sample: Sudden deaths Dep.Var.: Number of publications per year									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$post\ death$	-0.145***	-0.187***	-0.113***	-0.144***	-0.116***	-0.141***	-0.773***	-0.229***	0.057
	(0.029)	(0.032)	(0.024)	(0.027)	(0.026)	(0.029)	(0.159)	(0.055)	(0.185)
$post\ death * b_{ij,death}$	-0.247***	-0.301***	-0.297***	-0.260***	-0.279***	-0.251***	-0.271***	-0.227***	-0.299***
	(0.037)	(0.039)	(0.037)	(0.037)	(0.040)	(0.037)	(0.037)	(0.034)	(0.036)
$post\ death * degree_{j,death}$		-0.266***							-0.151
		(0.063)							(0.062)
$post\ death * closeness_{j,death}$			-0.177***						-0.148***
			(0.022)						(0.037)
$post\ death * betweenness_{j,death}$				-0.131***					0.007
				(0.033)					(0.019)
$post\ death * clustering_{j,death}$					0.150***				0.002
					(0.034)				(0.038)
$post\ death * eigenvector_{j,death}$						-0.024***			-0.016
						(0.003)			(0.011)
$post\ death * triangle_{j,death}$							-5.970***		2.384
							(1.438)		(1.780)
$post\ death * overlap_{ij,death}$								-0.828*	-0.539
								(0.377)	(0.364)
R^2	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.056
Nb. of obs.	503,748	503,748	503,748	503,748	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variable is the annual number of publications. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. The variables *degree, closeness, betweenness, clustering coefficient, eigenvector centrality, triangle* and *neighbourhood overlap* are defined in appendix A. All variables are measured for the stars at the time of death (and pseudo-death).

3.5.2.4 Coauthorship Characteristics

If a key feature of brokerage degree is that it is pair–specific, it may be confounded by other characteristics of the tie between the star and the coauthor. I distinguish between different type of ties using the frequency and recency of the joint publications and the overlap in research interest. The recency and strength of coauthorships are defined by the number of years since the last publication and the length of exposure to the star prior to the death. The variables *recent* and *weak* are indicator variables equal to one when the last joint publications was less than 2 years ago and they have fewer than 2 years of joint collaboration. I measure whether the coauthors were specializing in the same research area at the time of the death.

Table 3.7 presents results controlling for these tie characteristics. The loss of a star–coauthor with recent and weak ties significantly decreases publications rate of the surviving coauthors. The loss of a star–coauthor working in the same field also significantly decreases publications rate. In all specification, the coefficient on the interaction between the treatment and brokerage degree remains negative and significant. The magnitude of the effect of losing a star–broker is even larger when controlling for the characteristics of the tie. These results are suggestive that recent ties, which can be considered as active ones, are especially important because the star still has knowledge which hasn't already been shared to his coauthor. The fact that the loss of a star with weak ties has a negative effect goes against a possible grief mechanism. Under this mechanism, we would expect scientists with long–standing collaboration to a deceased coauthor to be more grief stricken than scientists with more distant ties to a deceased coauthor. The importance of weak ties for brokerage is consistent with Granovetter (1973)'s argument that weak ties act as bridges which are particularly valuable for the transmission diverse information. Brokers are also especially important when coauthors have common research focus. One explanation is that coauthors sharing common research interests are more likely to transmit relevant information and information that is understand by the recipient.

Table 3.7: Coauthorship characteristics

Sample: Sudden deaths Dep.Var.: Number of publications per year				
	(1)	(2)	(3)	(4)
post death	0.052*	1.574***	0.065	1.664***
	(0.026)	(0.056)	(0.034)	(0.059)
post death $*$ $b_{ij,death}$	-0.853***	-1.044***	-0.881***	-1.035***
	(0.114)	(0.117)	(0.117)	(0.116)
post death $*$ *recent tie*$_{ij,death}$	-0.089*			-0.193***
	(0.045)			(0.044)
post death $*$ *weak ties*$_{ij,death}$		-1.606***		-1.632***
		(0.056)		(0.056)
post death $*$ *same research*$_{ij,death}$			-0.111*	-0.115**
			(0.045)	(0.044)
R^2	0.068	0.079	0.069	0.080
Nb. of obs.	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variable is the annual number of publications (winsorized at 1%). The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. The variables interacted with post death are indicator variables measured at the time of the death: *recent ties* takes the value of one when the last joint publication was within 2 year of the death, *weak ties* is equal to one when there has been fewer than 2 years of collaboration, and *same research interest* is equal to one if the star and his coauthor wok in the same field. The specialty is defined as the modal topic (defined through MeSH codes) a scientist has published in.

3.5.2.5 Knowledge transmission via star-brokers

Stars in a brokerage position connect coauthors to other parts of the network thereby providing access to knowledge spillovers from further away. The broker may serve as information conduit through which knowledge and ideas is transmitted. To establish the access to knowledge flow via the star as a primary explanatory mechanism, I create a brokerage degree measure based on the new *topics* (instead of the number of new coauthors as in equation 2) a coauthor provides unique and exclusive access to.

Given that research is generally done in teams, determining an individual scientist's

field of expertise or knowledge, which may evolve over the career, poses a challenge. Moreover, fields are assigned to publication and not to individual researchers. I use all topics associated to the MeSH codes assigned to publications as a measure of a scientist's knowledge.[30] A scientist can access knowledge directly from his immediate coauthors and indirectly from his coauthors' coauthors (i.e. coauthors two links away). The pool of knowledge a scientist has access to is therefore defined as all the topics scientists up to two links away have worked in. From this set, I remove all the topics of the scientist in question and his direct coauthors. This eliminates any *direct* access to knowledge either through joint publications or knowledge embodied in immediate coauthors. This also removes any confounding factors associated to the characteristics of a collaboration. Two coauthors who share many common topics (i.e. low brokerage degree in terms of topics) may be a sign of a close and fruitful collaboration rather than a potential knowledge transmission. Two coauthors with many joint publications will inevitably have a large overlap in their fields of work.

The equivalent to the brokerage degree in section 2 using the number of topics (rather than coauthors) is defined as the share of new topics a star provides unique and exclusive access to. Specifically, from the pool of indirect knowledge offered by a star j to i, one can count the topics j is the only one among i's coauthors who provides this indirect link. As previously done, brokerage degree in terms of topics is measured at the time of death for the star (and the pseudo time of death for the matched stars). The histogram of brokerage degree in terms of topics can be found in figure B.21 in the appendix.

I find a significant and sizeable decrease in the number of publications after the death of a broker in terms of topics. The death of a star who provided only non–redundant indirect links to topics at the time of his death is associated with a 16% decrease in annual publications compared to a star who provided only redundant links to topics. Once we

[30] One could use the MeSH codes directly, but for computational reasons I use subfields. There is a total of 107 fields.

control for both measures of brokerages, the magnitude of the brokerage degree based on the number of coauthors is smaller and no longer significant. Brokerage degree based on topics on the other hand is significant and remains of similar magnitude. This suggests that brokerage degree previously used indeed captured knowledge embodied in scientists and that access to topics further away is important for the productivity of scientists.[31]

Table 3.8: Brokerage in terms of topic

Sample: Sudden deaths		
	(1)	(2)
$post\ death$	0.409***	0.399***
	(0.036)	(0.034)
$post\ death * b^{topic}_{ij,death}$	-0.603***	-0.559***
	(0.072)	(0.067)
$post\ death * b_{ij,death}$		-0.136
		(0.202)
R^2	0.069	0.070
Nb. of obs.	248,231	248,231
Nb. of coauthors	9,376	9,376
Nb. of deceased stars	107	107

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. The variable $b^{topic}_{ij,death}$ is the number of non–redundant topics offered by the star "j" to his coauthor "i" at the time of his death.

[31]The number of observations differ from previous tables because some scientists do not have any MeSH codes associated to their publications leaving them with a missing $b^{topic}_{ij,death}$

3.6 Conclusion

This paper is built on the argument that brokers – individuals in network positions that connect otherwise unconnected network members – benefit from the access to non-redundant information (Burt, 2008, Reagans and Zuckerman, 2008). Focusing on star medical scientists and their coauthors, I provide empirical evidence that brokerage relationships, measured by the number of scientists who can be reached solely through a given coauthor, induces heterogeneity in peer effects on publication rates. To circumvent the identification problems caused by endogenous collaborative link formation, I use the sudden and unexpected deaths of coauthors which induces changes in the network structure for reasons that are unrelated to his abilities as a researcher and/or network position.

My results validate the importance of star scientists and produce new evidence that the network position plays an important role in the productivity of scientists. Researchers suffer a 8% decrease in publication following the death of a coauthor. Losing a star–coauthor who offered only non–redundant links (i.e. a "broker") instead of a star–coauthor who offered only redundant links (i.e. a "non–broker") is associated with a significant loss of 0.24 publications per year. This represent a 31% decrease in the yearly number of publications for the average scientist. Measuring brokerage degree based on the number of topics (rather than the number of links to coauthors) reveals that the loss of a star who provided access to only non–redundant topics is associated with a 16% decrease in annual publication. Taken together, my results support the hypothesis that access to novel knowledge embodied in researchers is an important determinant of the productivity of a researcher. Moreover, network position and the access to flows in the network can be captured by a measure which is is local in nature.

It is important to emphasize that the focus of this paper has been to characterize the heterogeneity in the causal impact of the loss of a coauthor in terms of network position. The causal effect of network position (i.e. what would have happened to a coauthor's productivity after the death of a coauthor if he had been located in another

network position) would need to take into account the endogeneity in network position. It is hard to think of a natural experiment which would provide exogenous variation in the network position. Alternatively, the formation of network would need to be directly modeled. However, given the current state of theoretical research on network formation such an exercise

Uncovering the influence of immediate peers, and more generally, the network structure on productivity has important implications, not just for the creation of scientific knowledge, but also for long–run growth potential. The creation of knowledge appears to be a "social" production. The role of peer learning in promoting innovation and potentially growth has been featured in the endogenous growth literature (Aghion and Howitt, 1992a, Romer, 1990a). Growth is driven by technological change that results from research and development effort. This leads to the prediction that an increase in the level of resources devoted to R&D, as measured by the number of scientists engaged in research, should lead to a proportional increase in the per capita growth rate of output. However, this "scale effect" is known to be at odds with observed trends (Jones, 1995, Segerstrom, 1998). Using the number of scientists engaged in research may be misleading if their configuration is not taken into account. An increase in scientific output may not follow from an increase in the number of scientists if the additional scientists are isolated or not well–positioned. Therefore, any policy which increases the level of resources devoted to R&D should take the network topology into account.

The death of a star has wider repercussion than the performance of immediate coauthors. Stars influence colleagues, competitors (Borjas and Doran, 2013), PhD students (Waldinger, 2010), new hires (Agrawal et al., 2013), scientists further away in the network (Jaravel et al., 2015), and can shape entire fields (Azoulay et al., 2014). An interesting extension to this paper would be to examine how the death of stars affect the *community* of scientists. Network–level characteristics, and particularly the sorting pattern of scientists, influence both individual and collective outcomes (Uzzi and Spiro,

2005). Different elements of the network structure such as its size, density, and average path length can aid or hinder the creation and the diffusion of knowledge. Certain network structures may be more resilient to such productivity shocks and offer an environment fostering the creativity of its members and consequently promote social capital.

Chapter 4

Knowledge Spillovers from Clean and Dirty Technologies

4.1 Introduction

It is commonly recognized that knowledge spillovers from innovative activities provide a case for government intervention in the market because private R&D investments are likely too low. It has also been recognized that not all innovations create spillovers to the same extent. In particular more basic research is assumed to create stronger spillovers and therefore should attract more government support. However, for better or worse governments often champion specific technological areas – rather than types – such as defence, IT, aerospace, bio–technology etc. Often this is because a certain area promises auxiliary (i.e. not necessarily economic) benefits such as security, health or simply prestige. If the level of spillovers generated by these different areas are the same, then the choice to target a specific technological area does not matter. However, if the spillovers vary substantially across areas then the distribution of government intervention can affect the level and growth of economic well being. To the best of our knowledge, this study is the first to systematically compare spillovers between different technology areas. Our main focus is what we have dubbed dirty and clean technologies; i.e. technologies that are associated with GHG gas pollution and alternative technologies that can replace them. We also examine other emerging technologies and develop a methodology that will be relevant for comparing spillovers between

technological areas more generally. We focus on clean and dirty technologies because they are an important example of deliberate differential treatment of technology areas by government policies. Increasingly, governments are deployment carbon pricing policies which incentivize clean and hamper dirty technology development. This includes carbon and energy taxes (Aghion et al., 2016) but also direct subsidies for clean innovation. In 2012, OECD countries spent over 3 billion euros to support the development of new clean technologies such as renewable energy or hydrogen cars. This is motivated by the desire to mitigate climate change in the long run. However, many policy makers – often in an effort to make climate change policy attractive to the public – have suggested that this could also have a beneficial impact on economic outcomes such as growth or employment in the short run. Theoretically, this can only be the case if clean technology innovations lead to larger spillovers than the dirty technology innovations that it replaces. Hence, the main objective of this paper is to measure and compare the amount of knowledge of spillovers from clean and dirty technologies.

Following a long tradition in the literature, we derive our measure of knowledge spillovers from patent citation data (Caballero and Jaffe, 1993, Hall et al., 2005, Jaffe and Trajtenberg, 1999, Trajtenberg, 1990). Patent documents offer a paper trail of knowledge flows as inventors are required to reference previous patents which have been useful for developing the new knowledge described in the patent. Patent citations are not without limitations, but an important advantage of our dataset is that it allows us to deal with most of the problems usually associated with their use. For example, we can identify (and discard) self-citations by inventors, as well as citations added by patent examiners, which might not capture external knowledge spillovers. We rely on the PATSTAT database, a new dataset assembled by the European Patent Office in collaboration with the OECD. It provides information on nearly all patents filed worldwide in almost all national patent offices. It also provides information on patent families; i.e. when the same innovation is filed repeatedly in different jurisdictions. This allows us to use an innovation, rather than a patent as the unit of analysis avoiding any double counting. Our main analysis focuses on two technology fields – cars

and energy generation – and within each field on two main areas: fossil fuel based technologies (dirty) and alternative (clean) technologies. Cars and power generation account for about 40% of global carbon emissions (IPCC, 2007). PATSTAT also allow an easy distinction between dirty - i.e. everything related to fossil fuel combustion - and clean - i.e. alternative technologies such as electric vehicles and solar power generation. As an extension we also consider "grey" technologies; i.e. innovation to improve the pollution efficiency of fossil technologies.

There are a variety of confounding factors that might lead to differences in citations between technology areas such as clean and dirty that are unrelated to spillovers in an economic sense. Citations in patent documents are driven by legally binding definitions on what constitutes prior art. These differ over time and between jurisdictions. Clean and dirty innovations are not uniformly distributed across space or time. Hence, average citation counts for the two technological areas can differ because one technology tends to file more in patent offices that require more citations. We address a wide range of such concerns by including a set of control variables including patent office by year fixed effects. However, this will not deal with variation in citation practices between different technological areas; e.g. suppose that in some technological areas, it is customary to cite more by referring to more remote underlying ideas. Because most innovations receive their citations from within their own technological area, this could lead to differences in citation counts that reflect "cultural" differences between technological fields rather than economically meaningful spillovers. We address this in two ways. Firstly, we examine spillovers differences relying only on citations outside an innovation's technological area. Secondly, we use a new measure, called PatentRank, rather than the mere citation count. PatentRank is derived from the Page-rank algorithm developed by Google's Larry Page to rank the relevance of webpages on the basis of how they are hyperlinked; i.e. cited. It is recursively computed as the weighted average of all citing patent page ranks weighted by the inverse ratio of citations in a citing patent. Hence, a patent receives a high page rank if it is cited by many other patents that are themselves cited a lot but do not cite many others themselves. This not

only deals with potential variation in citation culture between technological areas but also considers indirect spillovers; i.e. an innovation can create spillovers because it is cited a lot by itself or because it is cited by another innovation that is cited a lot. We are one of the first to apply this to patent data.

Our results suggest that clean innovations generate significantly more knowledge spillovers than their dirty counterparts. All other things being equal, clean patented inventions receive 43% more citations than dirty inventions. The gap is larger in the electricity production sector (49%) than in the transportation sector (35%). Interestingly, the gap between clean and dirty technologies has been constantly increasing during the past 50 years. We show that clean patents are not only cited more often, they are also cited by patents that are themselves cited more often (irrespective of their technological area). When considering our new PatentRank index, we also find strong evidence of larger spillovers from clean technologies. Our conclusions are robust to a large number of sensitivity tests. These include discarding citations added by patent examiners, correcting for self-citations at the applicant level, including inventor fixed effects, looking at different subsamples and including additional control variables.

How can we account for the larger knowledge spillovers from clean technologies? One explanation stands out from our investigation: clean technologies seem to benefit from steep learning curves associated with new technological fields.[1] When we control for the age of the technology, the clean premium decreases by 14%. We then compare knowledge spillovers between clean, grey and "truly dirty" innovations. The analysis suggests a clear ranking: clean technologies exhibit significantly higher levels of spillovers than grey technologies, which themselves outperform truly dirty technologies. We also compare clean inventions with other emerging technologies such as biotechs, IT, nanotechnology, robot and 3D, and find that clean patents appear much closer in terms of knowledge spillovers to these radically new fields than to the dirty

[1] We partially control for the novelty of a technological field by including a measure of previous patenting within the technology class of a given patent in our regressions. However, the number of patents might not capture novelty entirely.

technologies they replace. Interestingly knowledge spillovers from clean technologies appear comparable in scope to those in the IT sector, which has been the driver behind the third industrial revolution. When comparing clean, dirty and emerging technologies to all other inventions patented in the economy, we find a clear ranking in terms of knowledge spillovers: dirty technologies have lower knowledge spillovers than the average invention, while clean and other emerging technologies exhibit larger knowledge spillovers. With the exception of biotechs, all other emerging technologies (IT, nanotechnology, robots and 3D) show larger knowledge spillovers over the average invention than clean inventions. Taken together, these pieces of evidence suggest that the clean advantage might be a feature of the radical novelty of the field.

Our results have a number of immediate implications. Firstly, our results highlight the large and economically relevant spillover differences between technology areas. This means that any meaningfully growth policy design should take these spillovers into account. Secondly, with respect to climate change policy, our findings provide support for the idea that pollution pricing should be complemented with specific support for clean innovation that goes beyond standard policies in place to internalize knowledge externalities. Indeed, the spillover advantage of clean innovations compared to dirty innovations (including "grey" energy efficiency technologies) uncovered in this paper justify higher subsidies to clean R&D in a first best policy setting. Radically new clean technologies should receive higher public support than research activities targeted at improving on the existing dirty technologies. However, such specific support could equally be justified for a range of other emerging areas, such as nanotechnologies or IT. Therefore our results go some way into supporting the recommendation by Acemoglu et al. (2012) that only clean (and not dirty) technologies should receive R&D subsidies.[2]

Thirdly, our results lend support to the idea that a redirection of innovation from

[2]Acemoglu et al. (2012) do not assume different spillovers from clean and dirty technologies. The crucial assumption on which the results hold is that patents last only for one period. Greaker and Heggedal (2012) show that it is possible to obtain similar results when relaxing this assumption if one assumes that clean technologies exhibit larger knowledge spillovers than dirty technologies.

dirty to clean technologies not only reduces the net cost of environmental policies but can also lead to higher economic growth in the short run, if the benefits from higher spillovers exceed these costs. Indeed, if the factors leading to an under-provision of knowledge are more severe for clean technologies and if new clean technologies are induced by environmental regulation, environmental policies could generate growth by un–intendedly correcting a market failure that has been hampering the economy, irrespective of the environmental problem (Neuhoff, 2005). In fact, the presence of a market failure associated with R&D spillovers from clean innovations is one of the possible theoretical foundations for the Porter hypothesis (Porter and Van der Linde, 1995) according to which environmental regulations may enhance firms' profits and competitiveness (see Ambec et al. (2013) and Ambec and Barla (2006), for a recent review). For example, in Mohr (2002), the existence of knowledge spillovers prevents the replacement of an old polluting technology by a new, cleaner and more productive technology, as firms have a second-mover advantage if they wait for someone else to first adopt. The introduction of an environmental regulation may thus induces firms to switch to the new, cleaner technology. This simultaneously improves environmental quality and eventually increases productivity. Our results however suggest that the potential growth effects of environmental policies very much depend on the type of displacement being induced by increasing support for clean technologies. If this leads to less investment in dirty technologies, as evidenced by Aghion et al. (2012a), there seems to be scope for medium run growth effects. If innovation in other emerging areas is crowded out, such effects are less likely.

Finally, our results also have implications for the modelling of climate change policy. For example, Fischer and Newell (2008), Fischer et al. (2014) assess different policies for reducing carbon dioxide emissions and promoting innovation and diffusion of renewable energy, with an application to the electricity sector. They model R&D investments and learning-by-doing, but assume that knowledge spillovers have the same intensity across clean and dirty technologies. Our paper suggests that this assumption does not hold in practice and provides estimated parameters that can be used to more

precisely model the difference between clean and dirty technologies.

Our paper relates to three main strands of the literature. First, our work draws on the extensive empirical literature that has used patent data to analyze the determinants and the effects of knowledge spillovers. Pioneers of patent citation data as a measure of knowledge spillovers include Scherer (1965) and Schmookler (1966). Griliches (1992), Griliches et al. (1991) survey this earlier literature. Since then, a large number of papers have used this method to investigate knowledge diffusion (Caballero and Jaffe, 1993, Hall et al., 2001, Trajtenberg, 1990). In particular, many papers have focused on the geography of knowledge spillovers (Jaffe and Trajtenberg, 1996, 1999, Jaffe et al., 1993, Thompson and Fox-Kean, 2005).

Second, in the energy literature some papers have recently attempted to compare knowledge spillovers from energy technologies with those of non-energy technologies. Bjørner and Mackenhauer (2013) compare the spillover effects of private energy research with those of other (non-energy) private research. They find that spillover effects of energy research may be lower than for other types of private research. Popp and Newell (2012) use US patent citation data to compare the social value of alternative energy patents to that of other patents filed by the same firms. They find that alternative energy patents are cited more frequently by subsequent patents, and by a wider range of technologies, than other patents filed by the same firms. However, none of these papers distinguishes between clean and dirty technologies within energy technologies nor do they go beyond comparing simple citation counts.

Third, our paper is closely related to the literature on the impact of environmental policies on economic growth, which is itself rooted in the endogenous growth literature (for seminal contributions, see Aghion and Howitt (1992b, 1996, 1998), Grossman and Helpman (1991), Romer (1990b)). Smulders and De Nooij (2003) introduce a difference in spillovers from the clean and the dirty sector into a model in which both the rate and direction of technological change are endogenous. They discuss the implication

of this difference for growth in the long run. In a Schumpeterian growth model where new technologies are both more productive and more environmentally-friendly, Hart (2004) shows that environmental policy can stimulate economic growth (see also Hart (2007), Ricci (2007b), for similar types of models, and Ricci (2007a), for a review of this literature).

The remainder of the paper is organized as follows. In the next section we present the datasets, explain how we measure knowledge spillovers and conduct some preliminary data exploration. In section 3, we discuss our empirical strategy in greater detail. Section 4 reports our main results. In section 5, we estimate the market value of clean knowledge spillovers. We discuss the implications of our findings in the final section.

4.2 Data and Descriptive Statistics

4.2.1 The Patent Database

We use data from the World Patent Statistical Database (PATSTAT), maintained by the European Patent Office (EPO). PATSTAT includes close to 70 million patent documents from 107 patent offices. We identify clean and dirty patents using the International Patent Classification (IPC) and the European Patent Classification (ECLA). For this purpose we rely heavily on work carried out at the OECD and the EPO, which has recently developed a patent classification scheme for "Technologies related to climate change mitigation and adaptation" (see Veefkind et al. (2012) for more information on how this scheme was constructed).[3]

We focus on two sectors where we can precisely distinguish between clean and dirty patents: electricity production (renewables vs. fossil fuel energy generation) and automotive (electric and hydrogen cars vs. internal combustion engines). Our paper rests primarily on a distinction between radically clean innovations (electric cars, so-

[3]This new scheme was defined with the help of experts in the field, both from within and outside the EPO, including from the Intergovernmental Panel on Climate Change (IPCC). It brings together technologies related to climate change that are scattered across many IPC sections and includes around 1,000 classification entries and nearly 1,500,000 patent documents.

Table 4.1: Number of clean and dirty inventions by sector

Sector	Clean	Grey	True Dirty	Total
Transport	74,877	133,083	212,193	420,153
Electricity	103,659	19,827	627,590	751,076
Total	178,536	152,910	839,783	1,171,229

lar energy...) and their dirty counterparts (gasoline–fueled cars, coal-based electricity generation...). However, an important feature of the dirty category is that some patents included in this group aim at improving the efficiency of dirty technologies (for example motor vehicle fuel efficiency technologies), making the dirty technology less dirty. We refer to these energy-efficiency patents as "grey" inventions. The list of patent classification codes used to identify clean, dirty and grey inventions is shown in table C.1 and C.2.

Given that the same invention may be patented in several countries, our level of observation is the patent family (the set of patents covering the same invention in several countries). In other words, we treat multiple filings of an invention as one invention and count citations by patent family instead of individual patents.[4] In total, our sample spans from 1950 to 2005[5] and includes over 1 million inventions with approximately 3 million citations made to these inventions. Table 4.1 shows a breakdown of the number of inventions in each sector. Clean inventions represent around 15% of our sample.

4.2.2 Citation Counts as Knowledge Spillovers

Patent data has a number of attractive features. First, patents are available at a highly technologically disaggregated level. This allows us to distinguish between clean and dirty innovations in several sectors, in particular electricity production and transportation. In comparison, R&D expenditures of a car company cannot usually be broken down into clean and dirty innovations. Second, patent documents contain citations to "prior art" as inventors are required to reference previous patents that have been used

[4] A patent family is considered clean if at least one patent within the family is clean

[5] We stop in 2005 to allow at least five years for patent to get cited. The majority of citations occur during the first five years of a patent.

to develop the new technology described in the patent. Citations are a response to the legal requirement to determine the scope of an inventor's claim to novelty and thus represent a link to the pre-existing knowledge upon which the invention is built.[6]. In other words, a citation indicates that the knowledge contained in the cited document has been useful in the development of the new knowledge laid out in the citing patent and thus represents a knowledge flow (Collins and Wyatt, 1988). It is therefore not surprising that patent data have been widely used in empirical studies of knowledge spillovers (Caballero and Jaffe, 1993, Jaffe and Trajtenberg, 1996, 1999, Jaffe et al., 1993, Keller, 2004).

To give a concrete example of knowledge spillovers, take the patent entitled "X'Ray Apparatus" (US8036340B2, see figure C.3). It was applied for in 2008, published in 2011 and belongs to the H05K class of electric techniques. The patent documents the inventor(s), and the applicant of the invention as well as their addresses. It also lists the claims of the invention and references other patents which will be useful in the making of the invention, including whether these citations were added by the examiner or not. Among its references, it lists a patent US6727670B1 entitled "Battery Current Limiter for a High Voltage Battery Pack in a Hybrid Electric Vehicle Power train" (see figure C.2) which was published in 2004. It belongs to the "electric motor" class (H02P). The citation received represents a transfer of knowledge. Looking in turn at the list of ref-

[6]US patent law 37 C.F.R 156 establishes that 'each individual associated with the filing and prosecution of a patent application has a duty of candour and good faith in dealing with the (US Patent) Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability [...] no patent will be granted on an application in connection with which fraud on the Office was practiced or attempted or the duty of disclosure was violated through bad faith or intentional misconduct'. In contrast, the EPO has no requirement similar to the duty of candour. Rule 42 of the European Patent Convention requires that the description in a European patent application should 'indicate the background art which, as far as is known to the applicant, can be regarded as useful to understand the invention, draw up the European search report and examine the European patent application, and, preferably, cite the documents reflecting such art'. The different legal requirements of the two systems have implications both in terms of who adds the citations and in the number of citations in the patents. For EPO patents, it is the patent office's examiner rather than the inventors or applicants who adds the majority of patent citations. This implies that in the EPO system, inventors are more likely to be unaware of the patents that are (ultimately) cited in their patents. However, citations in EPO patents may be less 'noisy' than USPTO citations, since it can be assumed that they have been scrutinised and chosen by the patent examiner, and citing-cited patent pairs might be 'closer' both in time and technological content than those extracted from the USPTO (Breschi and Lissoni, 2005, Michel and Bettels, 2001)

erence, it cites the patent US6026921A ("Hybrid Vehicule Employing Parallel Hybrid System, using both Internal Combustion Engine and Electric Motor for Propulsion", see figure C.1) which was published in 2000 is classified as B60K which falls under our clean transport category. This represents a clean knowledge spillover.

For each patent family in our dataset, we compile all the citations received regardless of their field and whether or not they are clean. Nevertheless, there are a few drawbacks to bear in mind. Patent citations are an incomplete measure of knowledge flows because they only capture flows that result in a novel and patentable technology. For this reason Griliches (1992) refers to citations as "pure knowledge spillovers". Since not all inventions are patented, patent citations underestimate the actual extent of knowledge spillovers. Other channels of knowledge transfers, such as non-codified knowledge and embodied know-how (inter-firm transfer of knowledge embodied in skilled labor, knowledge flows between customers and suppliers, knowledge exchange at conferences and trade fairs, etc.) are not captured by patent citations. It is however reasonable to assume that knowledge spillovers within and outside the patent system are correlated. Furthermore, there is a consensus that patent citations are a noisy measure of knowledge flows (Jaffe et al., 2000). First, citations made to patents by the same inventor (referred to as self–citations) represent transfers of knowledge that are mostly internalized, whereas citations to patents by other inventors are closer to the true notion of diffused spillovers. This problem can be (at least partly) resolved by excluding self-citations by the inventor. Second, some citations are added by patent examiners during the examination process (see Cockburn et al. (2003), Lemley and Sampat (2012) for an overview of the process). In a survey of inventors, Jaffe et al. (2000) show that the influence of examiners on citations is considerable, and that inventors were fully aware of less than one-third of the citations on their patents. Alcacer and Gittelman (2006) find that examiners are responsible for 63% of citations on the

average patent, and that 40% of patents have all citations added by the examiners.[7] These types of citations might not capture pure knowledge spillovers if the inventor was genuinely unaware of that invention.[8] Fortunately, our patent data indicate whether the citations was included by the applicant or the patent examiner. We can thus check the robustness of our results to excluding citations added by patent examiners.[9] Third, inventors and applicants might be strategically referencing prior art. Citing more prior art will make a patent more valuable in litigation, as it is much harder to prove a patent is invalid if the patent office has already considered it and rejected the relevant prior art (Allison et al., 2003). Most firms employ patent attorneys - many of whom were formerly patent examiners - to maximise the chances of approval by the examiner in order to avoid potential infringement and costly holdups. However, inventors have an incentive not to cite patents unnecessarily as it may reduce their claims to novelty and therefore affect the scope of the monopoly rights granted by the patent (Hegde and Sampat, 2009, Sampat et al., 2005). Moreover, not properly referencing priori art can lead to the invalidation of the patent and is therefore a dangerous strategy.[10]

[7] Alcacer et al. (2009) utilize a change in the reporting of US patent data that allows to separate citations added by the inventor and the examiners to examine the examiners' behaviour with respect to inventor citations. In the first case, the patent examiner might add citations that differ in nature from the inventor/applicant citations ('gap-filling'). Statistically, the gap-filling scenario would bias estimates of inventor knowledge. In the second case, the examiner might add similar citations ('tracking'). Tracking does not lead to any bias but it may cause standard errors in statistical estimations to be inflated. This raises doubts about patent citations as good indicators of knowledge flows. If examiner and inventor citations resemble each other closely, this suggests that firms and inventors choose their citations with respect to potential infringement and holdup threats and anticipate with some error citations most likely to be added by examiners. Moreover, examiners and inventors might exchange information during the application process, and examiners themselves are prone to biases in favour of citing particular patents. Using the EPO data which allows to identify the source of the citations since 1979, Criscuolo (2006) attempt to identify the factors that influence whether an observed patent-to-patent citation was added by the applicant/examiner.

[8] Of course, if the inventor has deliberately omitted to cite a relevant invention, then citations added by patent examiners actually capture true knowledge spillovers.

[9] Note that even if the citations was added by the inventor, s/he might have learnt about the cited invention only after the development of the invention. We have no way to control for this potential issue.

[10] "Failure of a person who is involved in the preparation or prosecution of a United States patent application to disclose material prior art can result in the patent not issuing, or if issued, being held unenforceable or invalid. As in many instances, the issue of whether prior art is material to patentability can be quite subjective; it is critical that inventors, assignees, and attorneys be acquainted with the obligations to disclose such prior art." (Silverman, 2003)

4.2.3 A New Measure of Spillovers: PatentRank

A potential concern with citation counts is that a citation from an obscure patent is given the same weight as a citation from a highly-cited work. Hence it is possible that some patents receive less citations than others but are cited by patents that are themselves more influential (i.e., more cited themselves). In particular many ground-breaking patents are modestly cited due to the small size of the scientific community in their area at the time of the publication, but subsequent patents are themselves increasingly cited (Maslov and Redner, 2008).

In order to take into account the whole network of patent citations, we apply the random surfer PageRank algorithm (Page et al., 1999) to our patent dataset. This algorithm was originally used by the web search engine Google to help determine the relevance or importance of a webpage. It does so by analyzing the network of hyperlinks of web pages. The basic idea is that a webpage is considered important if many other webpages point to it, or if many webpages point to the webpages that point to it (or both), and so on. To date, a handful papers have applied this method to rank the importance of patent documents (Lukach and Lukach, 2007, Shaffer, 2011). The resulting PatentRank has the advantage to readily identify patents that are modestly cited but nevertheless contain ground-breaking results. It also normalizes the impact of patents from different areas allowing for a more objective comparison (Maslov and Redner, 2008).

The PatentRank of a patent i is defined as the weighted sum of PatentRanks of all patents citing i, where the weights depend on the number of citations *made* by these citing patents. Therefore, a patent has a high rank if it is cited by many patents with a high rank, and it is better to be cited by a patent that cites only one patent than by a patent that has a long list of references. The PatentRank $r(i)$ of patent i is defined

according to the following formula and is computed recursively:[11]

$$r(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j \varepsilon F(i)} \frac{r(j)}{B(j)}$$

where N is the total number of patents, $F(i)$ is the set of patents that cite patent i (i.e. patent i's "forward citations"), and $B(j)$ is the number of citations made by patent j (i.e. patent j's number of "backward" citations). The parameter α, the damping factor, is used to avoid sink patents (i.e. patents that are never cited) because sink patents will lead to an endless loop.[12]

When constructing the PatentRank, we use the entire population of inventions and their citations correcting for self-citations by the inventor. We give inventions that are never cited the smallest PatentRank and rank these PatentRanks to create a PatentRank index. Thus the higher the PatentRank the greater impact or relevance of the invention. Figure 4.1 shows that there is a positive correlation overall between the citation count and the PatentRank but also a vast heterogeneity: many patents have few citations but a high PatentRank and vice versa. As opposed to citation counts, PatentRank allows us to capture the network centrality and in particular the influence of a patent. Hence, both indicators are complementary measures of the intensity of knowledge spillovers.

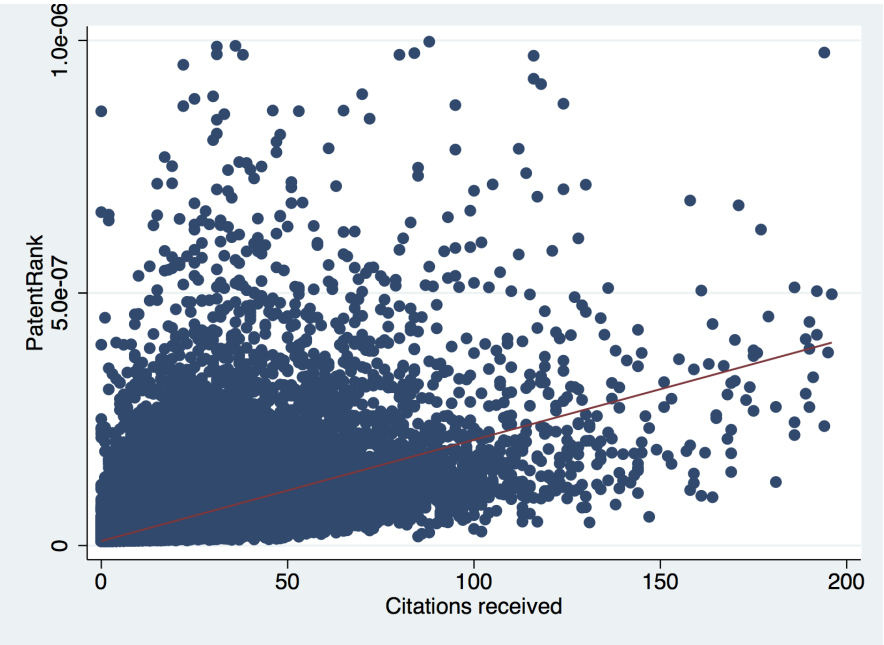
4.2.4 Exploratory Data Analysis

The objective of this paper is to compare the extent of knowledge spillovers that arise from clean and dirty innovations. As shown in table 4.2, aggregating both sectors together, clean inventions receive on average 3.40 citations throughout their life time while dirty inventions receive on average 2.30 citations. This difference is highly statistically significant (see column 3). An obvious problem with this simple comparison

[11] The process converges very quickly. In practice we use 50 iterations but the process converges after just a few iterations.

[12] The mechanism behind the ranking is equivalent to the random-surfer behavior, a person who surfs the web by randomly clicking links on the visited pages but periodically gets bored and jumps to a random page altogether. Therefore, when a user is on a web page, she will select one output link randomly with probability α or will jump to other webpages with probability $1-\alpha$. It can be understood as a Markov process in which the states are web pages, and the transitions are all equally probable and are the links between webpages.

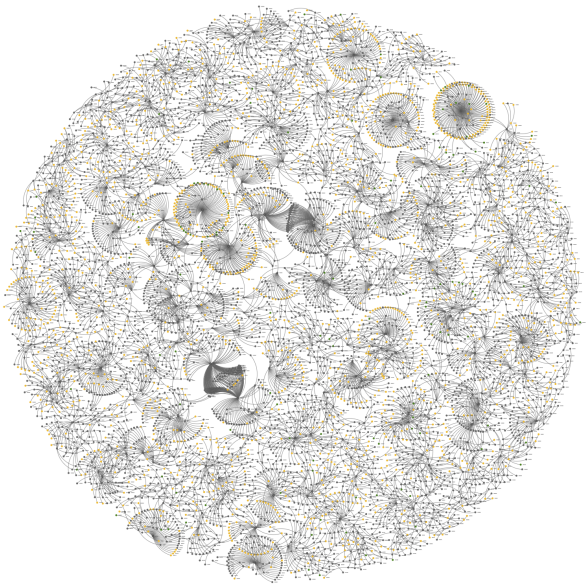
Figure 4.1: Citation counts and PatentRank



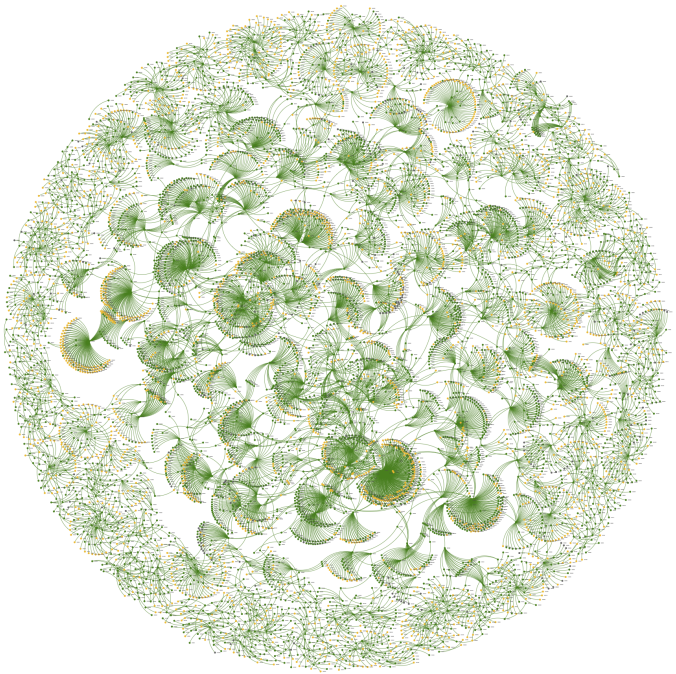
is that clean patents are relatively newer, and hence have had less time to be cited. The average age of clean patents (the time between the publication year and today) is 22 years as opposed to 27 years for dirty patents. In order to partly deal with this truncation issue, we look at the number of citations received within the first five years of the patents' publication (Hall et al., 2001). The difference between the number of citations received by clean and dirty inventions increases: clean patents receive 74% more citations than dirty patents within their first five years. Clean inventions also have a significantly higher PatentRank index than dirty inventions. Looking separately at each technological field, we find that the mean number of citations and the differences between clean and dirty patents vary across sectors. Inventions in the transportation sector are more cited overall and have a higher PatentRank index. Nevertheless, clean inventions are more cited and have higher PatentRank than dirty ones in both sectors and this difference is always significant. The "innovation flowers" in figure 4.2 show a network diagram for a random sample of 1000 clean and 1000 dirty innovations where the edges represent citations. This visual representation of PatentRank highlights the greater PatentRank of clean inventions.

Figure 4.2: Innovation Flowers

(a) Dirty



(b) Clean



Notes: The figures visualize innovation spillovoers. We draw a random sample of 1000 dirty and 1000 clean innovations corresponding to the nodes in the figures. The edges correspond to backwards citations. An interactive version is under http://www.eeclab.org.uk/forcedirect_arx.html?tojson_dirlinks0_1995_15_1000_0.json and http://www.eeclab.org.uk/forcedirect_arx.html?tojson_dirlinks0_1995_15_1000_2.json.

Table 4.2: Mean number of citations and PatentRank

	Clean	Dirty	Diff.
	Transport and Electricity		
Citations received	3.399	2.295	1.104***
	(8.256)	(5.921)	[0.016]
Citations received within 5-years	1.807	1.066	0.741***
	(4.754)	(3.109)	[0.009]
PatentRank index	2,335,270	1,920,395	414,874.3***
	(3,019,924)	(2,813,827)	[7,354.756]
	Transport		
Citations received	4.275	3.215	1.060***
	(9.626)	(7.185)	[0.031]
Citations received within 5-years	2.572	1.65	0.920***
	(5.903)	(4.174)	[0.018]
PatentRank index	2,645,597	2,429,006	216,591.2***
	(3,081,718)	(3,126,471)	[12,455.71]
	Electricity production		
Citations received	2.800	1.839	0.961***
	(7.092)	(5.091)	[0.018]
Citations received within 5-years	1.281	0.767	0.514***
	(3.681)	(2.312)	[0.009]
PatentRank index	2,119,068	1,666,122	452,945.3***
	(2,922,871)	(2,633,157)	[8,948.939]

Notes: The first two columns report the mean values with standard deviation in parentheses. The last column reports a t-test for the difference in means with standard error in parentheses. *** indicates significance at 0.1% level.

4.3 Econometric Analysis

Results from the exploratory data analysis point to larger knowledge spillovers from clean technologies. The results from this exploratory analysis can however be driven by some unobserved shocks to citation patterns disproportionately affecting clean patents. For example, the number of citations received by patents have increased recently due to the development of online patent search engines which facilitate identification of previous patents. Since clean patents are on average younger, they are likely to have been disproportionately affected by changes in the IT system. Moreover, the truncation issue is exacerbated for patents of older vintage. Even if each patent have the same amount of time to be cited, the increase in the universe of citing patents would increase

the total number of citations made. Econometric methods allow us to control for these potential confounding factors.

Our strategy is to estimate a simple count data model of the type

$$C_i = exp(\beta Clean_i + \gamma X_i + \varepsilon_i) \tag{4.1}$$

where C_i is the number of citations received by invention i (excluding self-citations) or the PatentRank index associated to invention i, $Clean_i$ is a dummy variable indicating whether invention i is clean, X_i are controls and ε_i is the error term. Our sample is the population of clean and dirty patents. Hence, the main coefficient of interest, β, captures the percentage difference between the number of citations received by clean and dirty patents, all other things being equal. Given the count data nature of the dependent variable, we estimate equation 4.1 by Poisson pseudo-maximum likelihood. We condition out the patent office-by-year-by-sector fixed effects using the method introduced by Hausman et al. (1984), which is the count data equivalent to the within groups estimator for OLS.[13]

We include a number of control variables to purge the estimates from as many potential confounding factors as possible. First, as explained above, the average number of citations *received* and *made* has been rising over time (Hall et al. (2001)). Moreover, differences in patent office practices across time and technological areas may produce artificial differences in citations intensities. We therefore include a full range of patent

[13]This is implemented by the xtpoisson, fe command in STATA. Note that Poisson models estimated by pseudo-maximum likelihood can deal with over-dispersion (see Silva and Tenreyro (2006)), so that negative binomial models offer no particular advantage. In particular, we find the pseudo-fixed effects negative binomial estimator available in stata (xtnbreg, fe) untrustable, since it does not truly conditions out the fixed effects (only the overdispersion coefficient is assumed to vary across units - see Allison and Waterman (2002), Greene (2007), for more information on this issue). However, as a robustness check we also estimated equation 4.1 using an unconditional negative binomial estimator with patent office, year, month and sector dummies (including a whole range of sector by year by patent office dummies is computationally infeasible) and find very similar results. The coefficient obtained for the clean dummy variable is 0.508***. The standard error varies from 0.041 when we cluster at the patent office and sector level, 0.023 when we cluster at the patent office level only and 0.093 when we cluster at the sector level only.

office-by-year-by-sector fixed effects. Practically speaking, this means that we effectively compare for example clean energy patents filed at the USPTO in 2000 with dirty energy patents filed at the USPTO that same year. To account for seasonality effects, we also include dummy variables for each month using the publication date.[14]

Second, the main problem we face is the fact that clean technologies are relatively newer, which makes them intrinsically different from dirty technologies. Note that the direction of the potential bias is not obvious. On the one hand, inventors start from a lower knowledge base which may lead to greater opportunities for big breakthroughs and larger positive spillovers than more mature technologies. On the other hand, the number of opportunities to be cited is smaller for clean technologies because we only know about citations received so far. As a result, we might be overestimating or underestimating spillovers effects from clean patents, depending on which effect dominates. In order to make a first attempt at controlling for this issue, we include the stock of past patents from the same technological field (defined on the basis of 4-digit IPC code) in the regressions.[15] Clearly, the stock of past patents might not perfectly capture the level of development of the technology and we come back to this point later.

Finally, citations might not exclusively capture knowledge flows, but also the commercial value of the patent. In order to control for this problem and focus on the part of the patent's value that is not appropriated by the inventor we include three measures of patent value: the patent's family size, a dummy variable indicating a "triadic" patent, and a dummy variable indicating the grant status. Family size is the number of patent offices where the invention has been filed. Family size has been used widely as a measure of patent value (Harhoff et al., 2003, Lanjouw and Mody, 1996, Lanjouw and Schankerman, 2004). Triadic patents are patents which have been filed in the US, European, and Japanese patent offices. Triadic patents have also been used extensively as

[14]Remember our unit of observation is the patent family. We use the earliest publication date within the family as the invention publication date.

[15]We also tried including higher-order polynomial terms of the past patent stock. This does not alter the results in any way.

a way to identify highest-value patents (Dernis and Khan, 2004, Grupp, 1998, Grupp et al., 1996, Guellec and Van Pottelsberghe de la Potterie, 2004, Van Pottelsberghe et al., 2001). The grant status of an invention indicates whether the patent has been granted by the patent office yet and obviously indicates a higher quality patent.

4.4 Results

Results from equation 4.1 can be found in table 4.3. The results from the econometric analysis confirm those of the exploratory data analysis: conditional on sector, patent office, publication year, commercial value and level of technology development, clean inventions appear to give rise to larger knowledge spillovers than dirty inventions. On average across the two technological fields, we find that clean inventions receive between 40% and 43% more citations than dirty ones depending on the specification. The coefficient is highly statistically significant across all models ($p<0.001$).[16] We get the strongest effect when adding all three measures of value as controls, but there is little variation across specifications. Given that these value measures all enter with a highly statistically significant coefficient, column 4 is our preferred specification. Notably, the number of past patents is always negative and significant, indicating that the latest patents in a field receive a decreasing number of citations as the field grows over time.[17] Results based on PatentRank confirm the results found with citations counts. Clean inventions have a significantly higher PatentRank across all sectors. Hence, when considering the whole citation network, knowledge spillovers from clean technologies are still larger than those generated by dirty technologies. Moreover, PatentRank has the advantage of taking into account the possible effect of different citing behavior of inventors citing clean and dirty patents. For instance, if inventors citing clean patents generally cite more patents than inventors citing dirty patents, this would translate into higher citations received for clean innovations but not in a higher PatentRank. Recall the PatentRank normalizes the number of citations received by the number of citations

[16] We cluster standard errors at the sector by patent office by year level. To check the robustness of the results we cluster-bootstrap the standard errors instead. The standard error increases slightly from .0137 to .0146 with the associated p-value still <0.001.

[17] Including the squared stock in the regression leads to a clean invention coefficient of 0.404*** (with a coefficient of -0.613*** for the stock and 0.028*** for the squared stock)

made by a citing patent.

Table 4.3: Basic results

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.	Citations received			PatentRank		
Clean invention	0.398***	0.392***	0.430***	0.267***	0.264***	0.292***
	(0.015)	(0.015)	(0.014)	(0.013)	(0.014)	(0.014)
Number of patents		-0.092***	-0.057***		-0.052***	-0.031***
		(0.008)	(0.007)		(0.006)	(0.005)
Family size			0.073***			0.067***
			(0.004)			(0.003)
Triadic			0.456***			0.241***
			(0.036)			(0.025)
Granted			0.947***			0.491***
			(0.031)			(0.021)
Patent office-by-year-by-sector	yes	yes	yes	yes	yes	yes
Month fixed effect	yes	yes	yes	yes	yes	yes
Obs.	1,149,988	1,149,988	1,149,988	1,149,988	1,149,988	1,149,988

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors (columns 1 to 3) and the PatentRank after 20 iterations (columns 4 to 6). All columns are estimated by fixed-effects Poisson pseudo-maximum likelihood.

We conducted a number of robustness checks on the basic specification. First, we use variations over our dependent variable to measure knowledge spillovers: the number of citations received within a five-years window (table C.21), the number of citations discarding citations added by the patent examiner (table C.22), and the number of citations excluding self-citations at the applicant level on top of excluding citations at the inventor-level (table C.23). Second, we add various controls: the number of claims, the number of 3-digit IPC codes, the number of citations made, the number of inventors, and the number of applicants (table C.24). Finally, we focus on various subsamples including patents that received at least one citations, triadic patents, patents from the US and European patent office (table C.26). None of these tests modifies our results.

In order to investigate the evolution of the relative intensity of spillovers across time, we run our estimation for each five years period between 1950 and 2005 and plot the coefficients obtained for clean invention along with their 95% confidence intervals in figures 4.3 and 4.4. We find that there has been a clear increase in the clean premium

over time.

Figure 4.3: Clean coefficient between 1950 to 2005 using citations received

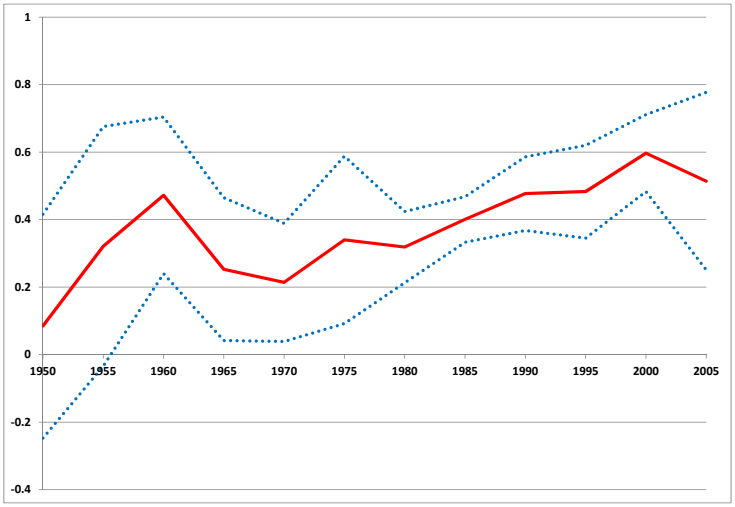
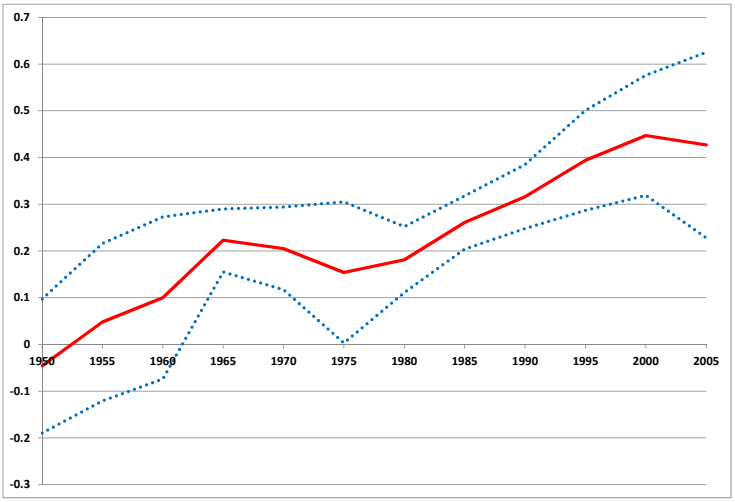


Figure 4.4: Clean coefficient between 1950 and 2005 using PatentRank



In table 4.4 we present the regressions results for each technology separately. The results are robust across both sectors, but we find some heterogeneity in the clean coefficient. Clean inventions in the transportation sector receive 35% more citations than

dirty inventions, while the clean premium in the electricity is larger (49%).

Table 4.4: Results by sector

	(1)	(2)	(3)	(4)
Sector	Transport	Electricity	Transport	Electricity
Dep. var.	Citation count		PatentRank	
Clean invention	0.347***	0.488***	0.219***	0.333***
	(0.018)	(0.023)	(0.014)	(0.023)
Number of patents	-0.068***	-0.047***	-0.048***	-0.019**
	(0.008)	(0.009)	(0.006)	(0.007)
Family size	0.070***	0.067***	0.062***	0.060***
	(0.008)	(0.004)	(0.007)	(0.004)
Triadic	0.512***	0.432***	0.279***	0.252***
	(0.056)	(0.050)	(0.045)	(0.041)
Granted	1.134***	0.725***	0.620***	0.381***
	(0.034)	(0.024)	(0.027)	(0.017)
Obs.	419,959	748,918	419,959	748,918

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variables are the total number of citations received excluding self-citations by inventors in columns 1 and 2 and the PatentRank index in columns 3 and 4. The regressions are all estimated by Poisson pseudo-maximum likelihood. The sample includes inventions from the transport (columns 1 and 3) and electricity (columns 2 and 4) sectors. All columns include a patent office-by-year and month fixed effects.

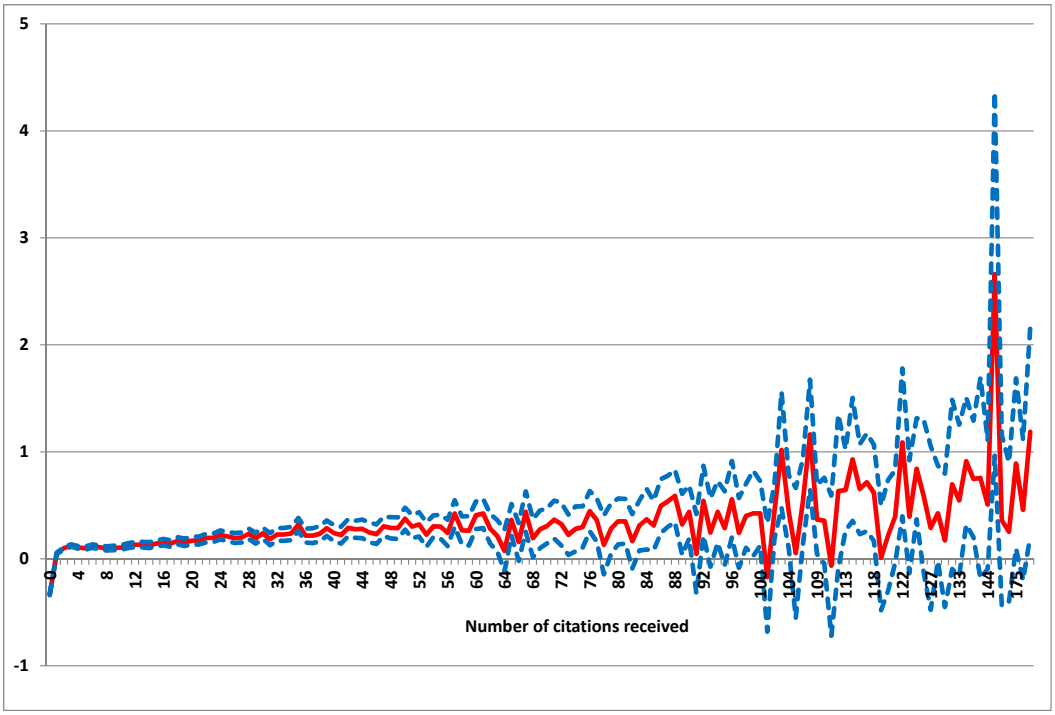
So far we have focused on the average effect of being a clean invention on the citation outcome. We now investigate the heterogeneity of the clean premium across the distribution of citations. Quantile regression techniques are not readily available for count data models, but we bypass this issue by estimating probit models of the likelihood that a patent falls within a given percentile of the patent citation distribution (see Chernozhukov et al. (2013) for a discussion of this issue). We run the following model:

$$Prob(Cite_i^j = 1) = \alpha + \beta Clean_i + \gamma X_i + \varepsilon_i \tag{4.2}$$

where $Cite^j{}_i$ equals one if invention i receives j citations where j varies between 0 (56% of inventions are never cited) and 479 (the most highly cited invention). $Clean_i$ and X_i are identical to the previous section. Hence the coefficient obtained for $Clean_i$ captures the difference between clean and dirty inventions in the probability of inven-

tion i to receive j citations. Figure 4.5 shows the coefficient obtained for $Clean_i$ and the associated 95% confidence interval on the number of citations received. We conclude from these results that (i) clean inventions are *always* more likely to have a positive citation count than dirty inventions at all levels of the distribution and (ii) the higher intensity of knowledge spillovers from clean technologies is even more pronounced for most highly cited patents.

Figure 4.5: Heterogeneity



4.4.1 Localized Knowledge Spillovers

The existence of localized knowledge spillovers has been widely documented (see Audretsch and Feldman (2004) for an overview). In one of the earliest papers on this subject, Jaffe et al. (1993) show that spillovers from research to firms are more intense when the firm is closer to the institution that generated the research. Jaffe and Trajtenberg (1996, 1999) show that patent citations tend to occur initially between firms that are close to each other, and later on spread to a larger geographical area and other countries. Using European patent data, Maurseth and Verspagen (2002) show that

patent citations occur more often between regions which belong to the same country, same linguistic group and geographical proximity (see also Peri (2005)). Similar results have been found for energy technologies (see Braun et al. (2010), Verdolini and Galeotti (2011)).

In our case, clean technologies could generate larger knowledge spillovers than dirty technologies simply because the clean industry might be more clustered geographically than the dirty industry. Although we do not have detailed information on the exact localization of inventors, we do have extensive information on their country of residence. We use this information to distinguish between national (within-border) and international (cross-border) citations. We then separately run regressions on these two sets of citation counts.[18] For the PatentRank, we compute a new PatentRank on the pool of national citations and international separately. We find that clean inventions exhibit larger national (column 2) and international (column 3) spillovers. For the remainder of the paper, all results will be presented for citations. The PatentRank index results can be found in the appendix. This suggests that clean inventor community transcend country borders. The clean advantage is larger in terms of domestic spillovers are larger than international ones.

[18] In the case of collaboration, we weight each citations by the number of inventors from each country involved in the invention. For example, three inventors working together, one in country A and two in country B, will count as 1/3 of a citation for country A and 2/3 of a citation for country B.

Table 4.5: Within vs. across-country spillovers

	(1)	(2)	(3)
Dep. var.	Citations received	Citations received within country	Citations received across country
Clean invention	0.430***	0.423***	0.247***
	(0.014)	(0.017)	(0.019)
Number of patents	-0.057***	-0.057***	-0.081***
	(0.007)	(0.008)	(0.006)
Family size	0.073***	0.062***	0.066***
	(0.004)	(0.003)	(0.004)
Triadic	0.456***	0.363***	0.212***
	(0.036)	(0.028)	(0.040)
Granted	0.947***	0.757***	0.829***
	(0.031)	(0.029)	(0.030)
Obs.	1,149,988	1,149,988	1,149,988

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variables are the total number of citations received (column 1), the total number of citations received from the inventor's country (column 2), the total number of citations received from all countries except the invention's (column 3) corrected for self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

4.4.2 Public Support for R&D

With many clean technologies dependent on policy support of one form or another, the expansion of clean technologies and its spillovers could be due in part to public investment. For instance, in 2011 OECD countries spent over 3 billion euros on R&D support to renewable energy technologies. To control for the government spending level, we include in the first two columns of table 4.6 the government spending in clean and dirty technologies within the transport and electricity sectors. Since we only have information on R&D spending for 28 countries from 1974 onwards, we run the baseline regression for this sample in columns 1, 3 and 5 and the include the government spending in columns 2, 4 and 6. On average, clean inventions exhibit even larger spillovers than dirty inventions after controlling for government spending. This effect is driven by the electricity production sector.

Another related concern is that research in clean technologies might come disproportionately from universities rather than private firms. If this is the case, the clean

premium might come from the fact that university patents are more highly cited and more general (Henderson et al., 1998). Moreover, the incentive and reward structure within the university system induce scientists to invest in their reputation by making research publicly available (openness of the academic community) and make them more willing to recognize the influence of their predecessors. We control for whether the patent was filed by a university or a firm in the last two columns of table 4.6 with private individuals being the baseline and still find that clean inventions receive 42% more citations than their dirty counterpart. Taken together, these results suggest that public support for R&D is not the driving force behind the clean premium.

Table 4.6: Public spending

	(1)	(2)	(3)	(4)
Dep. var.	Citations received			
	Government Spending		University	
Clean invention	0.493***	0.507***	0.421***	0.423***
	(0.026)	(0.026)	(0.014)	(0.015)
Number of patents	-0.007	-0.006	-0.047***	-0.050***
	(0.009)	(0.009)	(0.006)	(0.006)
Family size	0.067***	0.067***	0.070***	0.067***
	(0.004)	(0.004)	(0.003)	(0.003)
Triadic	0.452***	0.450***	0.450***	0.432***
	(0.046)	(0.046)	(0.034)	(0.034)
Granted	0.689***	0.688***	1.005***	0.992***
	(0.025)	(0.025)	(0.031)	(0.032)
Government spending		0.034***		
		(0.007)		
University				0.429***
				(0.022)
Firms				0.271***
				(0.018)
Obs.	496,788	496,788	826,078	826,078

Source: International Energy Agency (2013): Energy Technology Research and Development Database (Edition: 2013). Mimas, University of Manchester

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects. The samples of columns 1 and 2 include patent families for which we have government spending, where column 1 is the baseline and column 2 add a control for government spending. The sample of the last two columns include the patent families for which we have university or firm, where column 1 is the baseline and the column 2 add a control for university and firms.

4.4.3 Network Effects

Whether guided by "norms of science" (Merton, 1957, Small and Griffith, 1974) or self-interest including personal connections (Case and Higgins, 2000, Leopold, 1973), one might be concerned that inventors working on clean innovation behave systematically differently from inventors working on dirty innovations. The community of researchers working on clean technologies could perhaps be smaller and more close-knit. Stuart and Podolny (1996) for instance argue that there is also a strong social

component to a citation. The clean premium would then represent inventors' networks rather than true knowledge spillovers. To address this issue we restrict our sample to inventors who have been working both on clean and dirty technologies and include inventor fixed effects in our baseline estimations. Our data includes 41,713 such inventors (representing 2.92% of total inventors). Results are presented in table 4.7. We similarly introduce applicant fixed effects and the results do not change either. The clean premium remains significant albeit of slightly smaller magnitude. However, this is due to the different sample as can be seen by comparing columns 1 and 2 and columns 3 and 4 respectively.

Table 4.7: Adding inventor and applicant fixed effect

	(1)	(2)	(3)	(4)
Dep. var.	Citations received			
Clean invention	0.274***	0.336***	0.400***	0.380***
	(0.007)	(0.011)	(0.019)	(0.040)
Number of patents	-0.096***	-0.081***	-0.038***	-0.067***
	(0.004)	(0.006)	(0.008)	(0.010)
Family size	0.038***	0.094***	0.091***	0.100***
	(0.002)	(0.006)	(0.007)	(0.011)
Triadic	0.866***	0.644***	0.461***	0.444***
	(0.012)	(0.026)	(0.056)	(0.089)
Granted	1.234***	1.008***	1.022***	1.000***
	(0.007)	(0.011)	(0.033)	(0.046)
Inventor fixed effect	no	yes	no	no
Applicant fixed effect	no	no	no	yes
Obs.	697,192	697,192	435,584	435,584

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office, sector, year and month fixed effects.

4.4.4 Nature of the Citations

There are two important types of citations: references to patent documents that are particularly close to the new invention, which restrict the claims of the inventor, and references related to the technological background of the new invention. Therefore citations may reflect the similarity of inventions rather than the cumulative nature of

innovation Packalen and Bhattacharya (2012). To account for the heterogeneous nature of citations, we distinguish between citations received from inventions in the same technological sector (defined using the 3-digit IPC code as assigned by the patent examiner) and citations received from inventions in a different technological sector.[19] While the former include citations which might merely reflect similarities between patents, the latter should be closer to true knowledge spillovers. We then run our baseline regression separately on these two types of citations. Table 4.8 shows that clean inventions receive more citations both within and across technological fields, suggesting they do generate larger knowledge spillovers in the economy. The PatentRank index is computed on the pool of intrasectoral and inter-sectoral citations separately.

Table 4.8: Intra vs. inter-sectoral spillovers

	(1)	(2)	(3)
Dep. var.	Citations received	Intra-sectoral citations	Inter-sectoral citations
Clean invention	0.430***	0.457***	0.247***
	(0.014)	(0.015)	(0.019)
Number of patents	-0.057***	-0.053***	-0.081***
	(0.007)	(0.007)	(0.006)
Family size	0.073***	0.074***	0.066***
	(0.004)	(0.004)	(0.003)
Triadic	0.456***	0.487***	0.212***
	(0.036)	(0.036)	(0.040)
Granted	0.947***	0.963***	0.829***
	(0.031)	(0.032)	(0.030)
Obs.	1,149,988	1,149,988	1,149,988

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001).The dependent variables are the total number of citations (column 1), within a technological field (based on IPC 3 digit code) (column 2), across technological field (column 3) corrected for self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

[19] An important difference between the EPO and the USPTO systems is that in European search reports, cited documents are classified by the patent examiner within a particular citation category according to their relevance. When assessing the novelty of patent applications the examiner searches for earlier documents which have the same or almost the same features as the patent concerned (Schmoch, 1993).

4.4.5 Generality and Originality

Clean technologies, being relatively newer, might have more opportunities for "fundamental" research while older dirty technologies might instead be focused on the development of new applications. If clean technologies have more general applications, this might explain why they receive more citations and appear to induce larger knowledge spillovers.

In the previous section, clean inventions were found to be more likely to be cited both within or across their originating technological field. To further investigate the generality of clean and dirty inventions, we construct a measure of generality based on the Herfindahl index of concentration introduced by Trajtenberg et al. (1997). It measures the extent to which the follow-up technical advances (i.e. the citations) are spread across different technological fields, rather than being concentrated in just a few of them (i.e., they are more likely to have the characteristics of a General Purpose Technology, see Bresnahan and Trajtenberg (1995), Popp and Newell (2012)). The generality of a patent is defined in the following way:

$$Generality_i = 1 - \sum_{j}^{n_i} s_{ij}^2 \tag{4.3}$$

where s_{ij} is the percentage of patent citations *received* by patent family i that belong to patent class j (defined at 3-digit IPC code), out of n_i patent classes.[20] An originating patent with generality approaching one receives citations that are very widely dispersed across patent classes; a generality equal to zero corresponds to the case where all citations fall into a single class.

Similarly, one might suspect that clean technologies are more *original* than their dirty counterparts because they are relatively newer. We construct an originality measure using the same approach as in equation 4.3 but replacing s_{ij} by the percentage of ci-

[20]Specifically, we count the number of citations made by a *patent* and received by a *patent family*. This way we are only capturing citations directly made to an invention as oppose to citations made from one patent family to another.

tations *made* (instead of received) by invention i that belong to patent class j (defined again at 3-digit IPC code).[21] Thus, if a patent cites previous patents that belong to a narrow set of technologies the originality score will be low, whereas citing patents in a wide range of fields would render a high score.

We carry out regressions using this generality measure as a new outcome variable. Clean technologies are significantly more general and original in the transport industry while the opposite in true for the electricity production industry (see Table 4.9).[22] Adding generality (column 2), originality (column 3) and finally both measures (column 4) as control in Table C.18 confirms the finding of greater knowledge spillovers from clean inventions. Interestingly, the coefficient is slightly smaller when adding these controls than under the baseline specification (column 1). This suggests that these measures, particularly the generality measure, explain (a small) part of the clean premium.

Table 4.9: Generality and Originality

	(1)	(2)	(3)	(4)	(5)	(6)
Sector	All	Transport	Electricity	All	Transport	Electricity
Dep. var.	Generality measure			Originality measure		
Clean invention	0.008*	0.047***	-0.034***	-0.003	0.049***	-0.054***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)
Number of patents	-0.047***	-0.081***	-0.024***	-0.050***	-0.086***	-0.027***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
Family size	0.012***	0.011***	0.012***	0.008***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)	(0.0004)	(0.001)	(0.001)
Triadic	0.035***	0.028***	0.046***	0.026***	0.017***	0.037***
	(0.003)	(0.004)	(0.005)	(0.003)	(0.003)	(0.005)
Granted	0.047***	0.053***	0.039***	0.024***	0.024***	0.022***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Obs.	515,217	227,678	291,989	382,236	162,919	222,538

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is a generality measure (columns 1 to 3) and a originality measure (columns 4 to 6) based on Herfindahl index of concentration. The sample includes patents in the transport sectors only (colunns 2 and 5), in the electricity sector only (column 3 and 6), and in both sectors (columns 1 and 4). All columns are estimated by OLS and include patent office-by-year-by-sector fixed effects, and month fixed effects.

[21] These measures depend upon the classification system: a finer classification would render higher measures, and conversely for a coarser system. We use 3-digit IPC code as used in Hall et al. (2001)

[22] Note that there is a potential selection bias here, as patents that have never been cited have no generality measure and are therefore left out of the sample.

4.4.6 Clean Technologies Versus Other Emerging Fields

Technologies that contain a high degree of new knowledge (radical innovations) are likely to exhibit higher spillover effects than technologies that contain a low degree of new knowledge (incremental innovations). Clean technologies are new and rather under-developed technologies. In contrast, the dirty technologies they replace are much more mature and developed. Therefore research in clean technologies might yield spillovers that are completely different in scope from research in dirty technologies because they can be considered as radically new innovations. In order to investigate this assumption, we use several strategies.

First, we control for the age of the invention's technological field defined as the time elapsed since the date of the first appearance of this technological field (defined at the 15-digit IPC code) in any patent. Results are reported in column 2 of table 4.10. Controlling for the age of the technology decreases the coefficient obtained for the clean dummy variable. In order to account for potential non-linearities we further add squared age (column 3) and a whole range of dummy variables for each percentile of the age distribution (column 4). This exercise further diminishes the clean coefficient from 0.430 to 0.353, indicating that part of the clean premium is explained by the relative novelty of the field.

Table 4.10: Controlling for age of technological field

	(1)	(2)	(3)	(4)
Dep. var.	Citations received			
Clean invention	0.410***	0.381***	0.363***	0.354***
	(0.013)	(0.013)	(0.013)	(0.013)
Number of patents	-0.094***	-0.052***	-0.043***	-0.046***
	(0.004)	(0.005)	(0.005)	(0.005)
Family size	0.070***	0.067***	0.068***	0.068***
	(0.004)	(0.003)	(0.003)	(0.003)
Triadic	0.448***	0.431***	0.406***	0.397***
	(0.035)	(0.035)	(0.034)	(0.034)
Granted	0.939***	0.929***	0.917***	0.912***
	(0.031)	(0.030)	(0.030)	(0.030)
Age of tech field		-0.177***	0.194***	
		(0.009)	(0.034)	
Age of tech field^2			-0.023***	
			(0.002)	
Age of tech dummies	no	no	no	yes
Obs.	1,149,237	1,149,237	1,149,237	1,149,237

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Second, we distinguish between inventions which are radically clean from those which are related to energy efficiency improvements that make the dirty technology less dirty. So far, our paper revolves mostly around a distinction between radically clean innovations (e.g. electric cars, wind turbines) and dirty innovations (e.g. combustion engines, coal power plants). In the results presented thus far we have included grey innovations in the "dirty" category. We now identify these inventions and label these "grey" innovations. In tables 4.11 and C.11, we compare clean inventions with grey inventions (column 2), grey and truly dirty inventions (column 3), and finally clean with truly dirty inventions only (column 4). As a benchmark, column 1 simply reproduces the results from table4.4 where grey innovations are included in the dirty category. This analysis suggests a clear ranking in citations counts: clean technologies exhibit significantly higher levels of spillovers than grey technologies, which themselves outperform

truly dirty technologies. From a policy perspective, this result implies that radically clean technologies should receive higher public support than incremental innovation in dirty technologies.

Table 4.11: Clean, Grey and True Dirty

	(1)	(2)	(3)	(4)
Sample	Clean vs. Grey and true Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs. True Dirty
Dep. var.	Citations received			
Clean/Grey invention	0.430***	0.191***	0.307***	0.502***
	(0.014)	(0.016)	(0.016)	(0.015)
Number of patents	-0.057***	-0.051***	-0.114***	-0.060***
	(0.007)	(0.009)	(0.005)	(0.007)
Family size	0.073***	0.069***	0.072***	0.071***
	(0.004)	(0.007)	(0.004)	(0.004)
Triadic	0.456***	0.481***	0.454***	0.441***
	(0.036)	(0.055)	(0.037)	(0.035)
Granted	0.947***	0.997***	0.977***	0.868***
	(0.031)	(0.035)	(0.033)	(0.027)
Obs.	1,149,988	326,942	978,179	1,006,996

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. The sample includes clean, grey and truly dirty (column 1), clean and grey (column 2), grey and truly dirty (column 3), and clean and truly dirty (column 4) inventions. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Third, we compare knowledge spillovers between clean inventions in the transport and electricity technologies to other radically new technologies, namely IT, biotechnologies, nanotechnologies, robots and 3D (see Table C.3 for the list of related IPC codes). Results in table 4.12 show that clean inventions receive 41% more citations than biotech inventions. However, clean inventions receive significantly fewer citations than inventions in the IT, nanotechnology, robot and 3D industries. In tables C.18 and C.20, we find that clean inventions are less general and less original than all new technologies apart from nanotechnologies. Taken together, these results suggest that the relative novelty of clean technologies might explain why they exhibit larger spillovers. Looking at the coefficients obtained for the clean invention variable, it is interesting to

note that knowledge spillovers from clean technologies appear comparable to those in the IT sector, which has been behind the third industrial revolution.

Table 4.12: Spillovers from clean and other new technologies

	(1)	(2)	(3)	(4)	(5)
Baseline sector	IT	Biotechs	Nano	Robot	3D
Dep. var.	Citations received				
Clean invention	-0.153***	0.408***	-0.337***	-0.127***	-0.278***
	(0.029)	(0.033)	(0.062)	(0.042)	(0.036)
Number of patents	-0.013	-0.160***	-0.031***	-0.039***	-0.037***
	(0.008)	(0.014)	(0.008)	(0.008)	(0.008)
Family size	0.020***	0.033***	0.063***	0.063***	0.062***
	(0.003)	(0.005)	(0.007)	(0.007)	(0.007)
Triadic	0.574***	0.663***	0.525***	0.550***	0.528***
	(0.057)	(0.053)	(0.070)	(0.069)	(0.068)
Granted	1.181***	0.806***	0.862***	0.877***	0.882***
	(0.065)	(0.023)	(0.038)	(0.036)	(0.037)
Obs.	1,445,552	403,294	180,441	198,602	185,726

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. The sample includes all clean patents (transport and electricity) and patents from the following technologies: IT (column 1), bioechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Fourth, we compare the previous sample (clean transport, clean electricity, IT, biotech, nano, robots and 3D) to all other inventions. Figure 4.6 plots the coefficient of the dirty (in black), grey (in grey), clean (in green) and radically new technologies (in orange). Clean transport and electricity exhibit larger spillovers than the average invention. In terms of relative ranking, the clean transport and clean electricity are positioned between their dirty counterparts and radically new technologies.

Fifth, we restrict the sample of radically new technologies (IT, biotechs, nano, and robots) and compare clean and dirty inventions within these technologies. While clean inventions within the IT and the biotechs technologies still exhibit larger knowledge spillovers, there is no clean advantage within the nano and robot sectors.

Figure 4.6: Clean, grey, dirty, and radically new technologies vs. all other technologies- Citations count

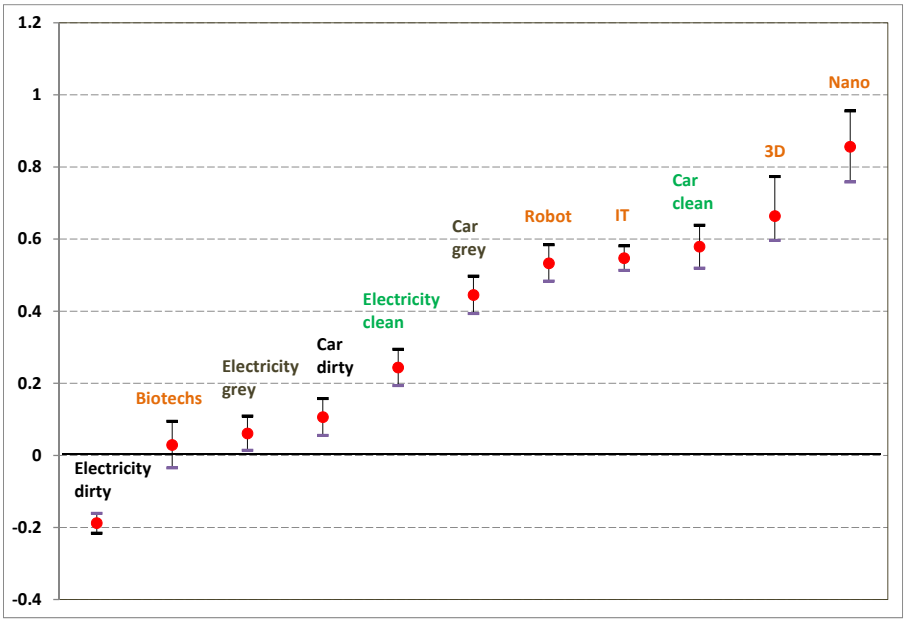


Table 4.13: Comparing spillovers from clean and dirty within new technologies

	(1)	(2)	(3)	(4)
Sector	IT	Biotechs	Nano	Robot
Dep. var.	Citations received			
Clean invention	0.222*	0.609**	0.313	0.677
	(0.091)	(0.053)	(0.211)	(0.525)
Number of patents	-0.012	-0.257***	-0.169***	-0.051
	(0.008)	(0.016)	(0.044)	(0.047)
Family size	0.020***	0.033***	0.109***	0.104***
	(0.003)	(0.005)	(0.018)	(0.014)
Triadic	0.547***	0.583***	0.268*	0.387***
	(0.055)	(0.056)	(0.136)	(0.113)
Granted	1.220***	0.699***	0.961***	1.005***
	(0.072)	(0.031)	(0.145)	(0.053)
Obs.	1,270,842	227,100	1,481	22,266

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. The sample includes patents from the following technologies: IT (column 1), bioechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Finally, in an attempt to find a dirty yet radically new technology, we compare knowledge spillovers between clean electricity production technologies and carbon capture and storage technologies (CCS) in table 4.14. The clean advantage disappears when considering simple patent counts and PatentRank, suggesting it is not because they are clean that clean technologies generate larger knowledge spillovers.

4.5 Discussion and Conclusion

In this paper we compare the relative intensity of knowledge spillovers from clean and dirty technologies. To measure knowledge spillovers, we use a rich dataset of 3 million citations received by over a million inventions patented in the automobile and electricity production sectors. This analysis is crucial to answer the question of whether clean technologies warrant higher subsidies than dirty ones. Our results unambiguously show that clean technologies induce larger knowledge spillovers than their dirty counterparts. We conduct a large number of sensitivity tests and the findings are remarkably robust.

Table 4.14: Spillovers from clean and CCS technologies

Dep. var.	Citations received
Clean invention	-0.083*
	(0.034)
Number of patents	0.037***
	(0.010)
Family size	0.065***
	(0.006)
Triadic	0.477***
	(0.062)
Granted	0.681***
	(0.030)
Obs.	106,700

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. The sample includes clean electricity production inventions and CO2 Capture and Storage technology. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

In particular, as depicted by the innovation flowers, this result is confirmed when using a completely novel methodology to measure knowledge spillovers that does not only count immediate forward citations but takes into account the whole network of patent citations.

We explore five potential explanations for our findings. First, we find no evidence that the clean industry is more geographically clustered. Second, differential citations behaviors among scientists involved in clean technologies cannot fully explain the clean advantage. Third, we find no evidence that government spending cannot account for clean premium. Fourth, we examine the generality and originality features of clean inventions. We find that clean inventions in the automobile industry are more general (i.e. they are cited by a wider range of technological fields) and more original. However, clean inventions in the electricity production industry are less general and less original. Finally, we compare clean inventions to other radically new inventions

such as IT, biotechnologies and nanotechnologies. We conclude that clean inventions seem to benefit from early returns to scale and steep learning curves. Interestingly we observe that knowledge spillovers from clean technologies appear comparable in scope to those in the IT sector.

Our results have two important policy implications. Firstly, the larger knowledge spillovers from clean technologies uncovered in this study justify higher subsidies for clean R&D or specific R&D programs for clean technologies, in addition to implicit support for clean R&D trough climate policies such as carbon taxation. Radically new clean technologies should receive higher public support than research activities targeted at improving on the existing dirty technologies.[23] However, such specific support could equally be justified for a range of other emerging areas, such as nanotechnologies or IT. This recommendation has been made in the past, for instance by Hart (2008) or Acemoglu et al. (2012) but it is the first time to our knowledge that it is substantiated by robust empirical evidence.[24] While a first best policy scenario would suggest a combination of emissions pricing and R&D subsidies *specifically* targeted at clean technologies, in times of tight government budgets it might be difficult to achieve the necessary subsidy levels. There might also be concerns over governments' ability to channel funds to R&D projects with the highest potential either because of information asymmetry or because of political interference. In this case our results would support a second best policy with more stringent emission pricing and regulation that would otherwise be the case (see for exampleGerlagh et al. (2009), Hart (2008), Kverndokk and Rosendahl (2007), Kverndokk et al. (2004)).

[23]Importantly, our results suggest that the relative support to clean R&D should grow over time. Incidentally, in a recent working paper Daubanes et al. (2013) show that gradual rise in subsidies to clean R&D activities causes a less rapid extraction of fossil resources, because it enhances the long-run resource productivity.

[24]Interestingly, statistics in OECD countries show that there is higher public R&D spending in clean technologies than in dirty ones. A look at the International Energy Agency's R&D expenditures data reveals that between 2000 and 2012, OECD countries have spent 198 million euros on dirty cars and 18 billion euros on dirty energy but 327 million euros on clean cars (65% more than dirty cars) and 25 billion euros on clean energy (35% more than dirty energy). However, these numbers do not include subsidies to *private* clean R&D, which is also warranted in a first best policy setting.

Secondly, our results lend support to the idea that a redirection of innovation from dirty to clean technologies induced by environmental or climate policies can lead to higher growth in the short and medium run. This can happen if the larger spillover effects from clean technologies exceed any negative growth effects from more stringent regulation. Our results however suggest that the potential growth effects of environmental policies very much depend on the type of displacement being induced by increasing support for clean technologies. If clean innovation crowds out dirty innovation, as shown by Aghion et al. (2012a) for the transport industry, there is scope for medium run growth effects. If innovation in other emerging areas is crowded out, such effects are less likely. At any rate, one should keep in mind that higher spillovers are only a necessary but not a sufficient condition for growth effects from green policies.

Our work can be extended in several directions. First, it would be interesting to investigate how knowledge spillovers affect firms' decisions to invest in radical innovation (clean technologies) or in incremental innovation (less dirty technologies), and how they respond to R&D subsidies targeted at clean technologies. Second, an interesting direction is to understand the spatial pattern of knowledge diffusion for clean technologies, including the transfer of knowledge across borders, in particular between developed and developing countries. Third, we could use micro data to estimate the impact of knowledge spillovers from clean and dirty technologies on firms' productivity. These parameters are crucial to empirically validate the potential impact of green policies on economic growth.

Appendix A

Appendix to Chapter 2

Additional Tables

Table A.1: Year of Passage of Laws, by US States

State	Territory Joined the Union[1]	State Joined the Union[2]	Introduction of Compulsory Schooling[3]	Age Groups Compulsory Schooling Laws Applied to[4]	Introduction of Child Labor Laws[5]	Introduction of Birth Registration Proof[6]
Alabama	1817	1819	1915	8 - 14	1910	1908
Alaska		1959	1929			
Arizona	1863	1912	1899	8 - 14	after 1910	1909
Arkansas	1819	1836	1909	8 - 14	1910	1914
California		1850	1874	8 - 14	1890	1905
Colorado	1861	1876	1889	8 - 14	1890	1907
Connecticut		1788	1872	7 - 14	1890	1897
Delaware		1787	1907	7 - 14	after 1910	1881
Florida	1822	1845	1915	8 - 12	1910	1899
Georgia		1788	1916	8 - 12	1910	1919
Hawaii		1959	1896			
Idaho	1863	1890	1887	8 - 14	1910	1911
Illinois	1809	1818	1883	7 - 14	1900	1916
Indiana	1800	1816	1897	7 - 14	1890	1908
Iowa	1838	1846	1902	7 - 14	1910	1880
Kansas	1854	1861	1874	8 - 14	1910	1911
Kentucky		1792	1896	7 - 14	1910	1911
Louisiana	1804	1812	1910	- 14	1890	1918
Maine		1820	1875	7 - 14	1890	1892
Maryland		1788	1902	8 - 12	1900	1898
Massachusetts		1788	1852	7 - 14	before 1880	1841
Michigan	1805	1837	1871	7 - 14	1890	1906
Minnesota		1858	1885	8 - 14	1900	1872
Mississippi	1798	1817	1918	7 - 12	1910	1912
Missouri		1821	1905	8 - 14	1900	1910
Montana	1864	1889	1883	8 - 14	1910	1907
Nebraska		1867	1887	7 - 14	1890	1904
Nevada	1861	1864	1873	8 - 14	after 1910	1911
New Hampshire		1788	1871	8 - 14	before 1880	1883
New Jersey		1787	1875	7 - 14	before 1880	1878
New Mexico	1850	1912	1891	7 -	after 1910	1920
New York		1788	1874	7 - 14	1890	1880
North Carolina		1789	1907	8 - 12	1910	1914
North Dakota	1861	1889	1883	8 - 14	1900	1907
Ohio		1803	1877	8 - 14	1890	1909
Oklahoma	1890	1907	1907	8 - 14	1910	1917
Oregon	1848	1859	1889	9 - 14	1910	1903
Pennsylvania		1787	1895	8 - 14	before 1880	1906
Rhode Island		1790	1883	7 - 14	before 1880	1896
South Carolina		1788	1915	8 - 14	1910	1915
South Dakota	1861	1889	1883	8 - 14	1910	1905
Tennessee	1790	1796	1905	8 - 14	1900	1914
Texas		1845	1915	8 - 12	1910	1903
Utah	1850	1896	1890	8 -	after 1910	1905
Vermont		1791	1867	8 - 12	before 1880	
Virginia		1788	1908	8 - 12	1910	1912
Washington	1853	1889	1871	8 - 14	1910	1907
West Virginia		1863	1897	8 - 12	1900	1925
Wisconsin	1836	1848	1879	7 - 12	before 1880	1908
Wyoming	1868	1890	1876	7 -	after 1910	1909

Notes and Sources:
* The District of Columbia is not included as it is a federal district.
[1] Year when the territory joined the Union [extracted from Braun and Kvasnicka 2013]
[2] Year when the state joined the Union [extracted from US Census Office]
[3] Year of introduction of compulsory school attendance laws [extracted from Landes and Solomon 1972]
[4] Year of introduction of child labor laws for manufacturing employment [extracted from Moehling 1999]
[5] Age groups that compulsory schooling laws applied to when the laws were introduced (i.e., the closest year available) [extracted from Lleras-Muney and Shertzer 2015]
[6] Year of introduction of birth certificate as official proof of a child's age [extracted from Fagernäs 2014]

Table A.2: Compulsory Schooling Laws, by Country

Country	Introduction of CSL: Preferred Year	Lower Bound	Upper Bound	Sources	Legislation Introducing Compulsory Schooling	Notes
Albania	1928	1928	1928	Hörner et al. (2007), Sefa and Lushnje (2012)	Fundamental Statute of the Kingdom of Albania (Constitution)	
Armenia	1932	1932	1932	Hörner et al. (2007), EFA (2000)		
Austria-Hungary	**1774**	**1774**	1869	Melton (1988), Slaje (2009), Schneider (1982), Donnermair (2010), Fort (2006), Ramirez and Boli (1987), Flora et al. (1983), Cohen (1996)		In Austria, the principle of compulsory education was introduced in 1774 by Joseph II but met with opposition (Flora et al. (1983), p.555). Six years of compulsory schooling were introduced in 1774 together with state-controlled public schools (Fort (2006), p.20). Maria Theresa and Joseph II reformed the education the education system in pursuit of pragmatic goals for the state. In 1781 Joseph II established the principle of mandatory primary education for all children aged 6-12, although in practice it took decades to realize this in many crown lands (Cohen (1996), p.15). As attendance was still not satisfactory a century later, the law was re-iterated with the 1869 Reichsvolksschulgesetz. Complete separation of schools from the Church was achieved in 1868 (Ramirez and Boli 1987, p.5). In Hungary, compulsory schooling was introduced in 1777 with the "Ratio Educationis". The 1869 Reichsvolksschulgesetz (the upper bound) applied to all the countries of the Empire
Belgium	1914	1914	1914	Wielemans (1991), Gathmann et al. (2012), Flora et al. (1983), Colle-Michel (2007), Ramirez and Boli (1987)	Loi Poullet (Loi du 19 mai 1914)	Compulsory education was introduced in 1914 but implemented only after World War I (Flora et al. (1983), p.561)
Britain	1880	1872	1880	Soysal and Strang (1989), Flora et al. (1983), Ritter (1986), Salimova and Dodde (eds.) (2000), Anderson (1995)		Compulsory education of eight years was introduced with exceptions in England and Wales in 1880 (Flora et al. (1983), p.623). School became compulsory in 1881 and free in 1891. However, the legislation was not implemented in the same way in every community. That is, some communities continued to depend on voluntary schooling or under the control of religious groups (Salimova and Dodde (eds.) (2000), p.108). In Scotland, compulsory schooling was already introduced in 1872 (lower bound) with the "Education (Scotland) Act"
Canada	1871	1871	1943	Oreopoulos (2005)		In the case of Canada, schooling was made compulsory at different points in time in different Canadian states. The first state to introduce a CSL was Ontario (1871), the last one was Quebec (1943) (Oreopoulos 2005). The first date (1871) was chosen as the CSL enactment date for Canada
Denmark	**1814**	**1739**	**1814**	Bandle et al. (2005), Gathmann et al. (2012), Simola (2002), Schneider (1982), Flora et al. (1983)	Education Act	Compulsory education was first enacted in 1739, but consisted only of religious education and the reading of certain familiar texts. In 1814, writing was added to the curriculum. Compulsory education covered only three days a week. Starting from 1869 compulsory education was extended to cover six days a week (Flora et al. (1983), p.567)
Finland	1921	1921	1921	Hörner et al. (2007), Simola (2002), Flora et al. (1983), Salimova and Dodde (eds.) (2000)	Compulsory School Attendance Act	Finland became an independent state in 1917; the primary school institution was established in 1866, but only became compulsory in 1921 (Simola 2002, p.212) with the introduction of eight years of compulsory schooling (Flora et al. (1983), p.572). The Parliament passed the law on compulsory education in 1921. The law entitled everyone to receive education free of charge, regardless of sex, language, or class. […] Towns were given five years to enforce the law and rural municipalities fifteen. In other words, the elementary schools were not functioning properly until the late 1930s (Salimova and Dodde (eds.) (2000), p.136)
France	1882	1882	1882	Soysal and Strang (1989), Cubberley (1920), Schriewer (1985), Schneider (1982), Flora et al. (1983), Salimova and Dodde (eds.) (2000)	Lois Jules Ferry (Loi n° 11 696 du 28 Mars 1882 (Article 4))	The Jules Ferry Laws established free education (1881) and laic and compulsory education (1882) (Garnier et al. 1989, p.291)
Germany	**1717**	**1592**	1871	Ramirez and Boli (1987), Stolze (1911), Salimova and Dodde (eds.) (2000), Flora et al. (1983), Oelkers (2009)		The first German state to introduce compulsory schooling was Palatinate-Zweibrücken in 1592. In Prussia, compulsory schooling was introduced by Frederick William in 1717, and reiterated by Frederick II in 1763. The general law of the land (Allgemeines Landrecht) of 1794 makes instruction - as opposed to attendance - mandatory, a fact that had consequences for school attendance and organization. In this system the state only regulates the minimum for those parents who cannot provide for their children's attendance. [...] Elementarschulen became unavoidable but actually only for the poorer classes of the population, who could not afford a better form of education (Salimova and Dodde (eds.) (2000), pp.179-180). Upon unification of the German Empire in 1871, compulsory schooling (which existed in Prussia) was extended to all states. Eight years of compulsory education were introduced in the German Empire with the exception of Wurtemberg and Bavaria where only seven years were introduced (Flora et al. (1983), p.584). Most states already had compulsory schooling before 1871 (detailed information on all states was not available). As Prussia was the largest and dominant state at the time of unification, we use the date of its first CSL enactment (1717) as the reference date for Germany

Italy	1877	1859	1877	Cubberley (1920), Schneider (1982), Ramirez and Boli (1987)		In the Kingdom of Sardinia, compulsory education was introduced in 1859 (2 years in all communes, 4 years in communes over 4,000 population) (Flora et al. (1983), p.598). Upon unification, compulsory school attendance was extended to all Italian provinces. This process was completed in 1877. The education system was quite effective in some of the Northern regions by 1880 and in Southern regions by 1900 (Ramirez and Boli 1987, p.7)
Japan	1872	1872	1872	Duke (2009), Loomis (1962), Burnett and Wada (2007), Salimova and Dodde (eds.) (2000)	Gakusei (Fundamental Code of Education)	The Fundamental Code of Education - the Gakusei - was announced in 1872. […] They declared their intention to spread education and mentioned that educational opportunity should be available for all people […] they emphasized parents' responsibility for education, every guardian shall bring up his children with tender care, never failing to have them attend school (Salimova and Dodde (eds.) (2000), p.275)
Luxembourg	1912	1912	1912	Soysal and Strang (1989), UNESCO (2007), European Commission (2010)	Loi du 10 août 1912 sur l'organisation de l'enseignement primaire	
Netherlands	1900	1900	1900	Soysal and Strang (1989), Gathmann et al. (2012), Schneider (1982), Flora et al. (1986), Salimova and Dodde (eds.) (2000)	De Leerplichtwet (July 7, 1900, Staatsblad No. 111)	Introduction of six years of compulsory education (Flora et al. (1983), p.603). When compulsory education was introduced in 1900, about 90% of children was already attending a primary school (Salimova and Dodde (eds.) (2000), p.315)
Norway	**1827**	**1739**	1860	Soysal and Strang (1989), Bandle et al. (2005), Hove (1967), Einhorn (2005), Rust (1990)	Primary School Act	
Poland	1919	**1825**	1919	Karsten and Majoor (1994), Slaje (2009), Biskup (1983), Salimova and Dodde (eds.) (2000)	Decree On Compulsory Schooling (O obowiazku skolnym) (February 7, 1919)	In the Prussian part of partitioned Poland, compulsory schooling was introduced in 1825. Shortly after the reunification, compulsory schooling was extended to the entire country in 1919. School systems inherited from Russia, Prussia and Austria were different and school traditions varied […] the young country's most important task in the field of education policy was to adopt a uniform school system (Salimova and Dodde (eds.) (2000), p.340) […] The Constitution of 1921 failed to provide the rural population with guarantees of any rights to education of the same quality as that provided to urban areas (Salimova and Dodde (eds.) (2000), p.341)
Portugal	**1835**	**1835**	**1835**	Ministro dos Negocios do Reino (1835)	Regulamento Geral Da Instrucção Primaria	
Russia	1918	1918	1918	Decree of October 16, 1918, on the Comprehensive Labor School of the Russian Socialist Federative Soviet Republic	Decree of October 16, 1918, on the Comprehensive Labor School of the Russian Socialist Federative Soviet Republic	
Spain	1857	1857	1857	Gathmann et al. (2012), De Maeyer et al. (2005), Ministerio de Fomento (1857)	Ley Moyano de Instrucción Pública de 1857	
Sweden	**1842**	**1842**	**1842**	Soysal and Strang (1989), Simola (2002), Schneider (1982)	Folkskolestadgan (SFS 1842:19)	The 1842 law was followed in later decades by other bills that made the system entirely universal (Ramirez and Boli 1987, p.6)
Switzerland	1874	1874	1874	Bundesverfassung (Federal Constitution)	Bundesverfassung (Federal Constitution)	Sources contradict each other with respect to introduction of compulsory schooling in different cantons. After the constitutional change of 1874, age of entry still varied according to cantonal law which also governed the duration of the primary school course (Flora et al. (1983), p.618). It was the radical new arrangement of society that made first attempt in 1798, but in a permanent manner only in the 19th century led to the establishment of the compulsory state school (Salimova and Dodde (eds.) (2000), p.433)

Table A.3: Compulsory Schooling Laws, for European Countries With Potential for Within-Country Regional Variation

Country	Region	Year of Introduction of Compulsory Schooling	Lower Bound	Upper Bound	Sources	Legislation Introducing Compulsory Schooling	Notes
Austria-Hungary	**Austria**	1774	1774	1869	Melton (1988), Slaje (2009), Schneider (1982), Donnermair (2010), Fort (2006), Ramirez and Boli (1987), Flora et al. (1983), Cohen (1996)	Allgemeine Schulordnung für die deutschen Normal-, Haupt- und Trivialschulen in sämmtlichen Kaiserlich-Königlichen Erbländern (General School Ordinance)	In Austria, the principle of compulsory education was introduced in 1774 by Joseph II but met with opposition (Flora et al. (1983), p.555). Six years of compulsory schooling were introduced in 1774 together with state-controlled public schools (Fort (2006), p.20). Maria Theresa and Joseph II reformed the education the education system in pursuit of pragmatic goals for the state. In 1781 Joseph II established the principle of mandatory primary education for all children aged 6-12, although in practice it took decades to realize this in many crown lands (Cohen (1996), p.15). As attendance was still not satisfactory a century later, the law was re-iterated with the 1869 Reichsvolksschulgesetz. Complete separation of schools from the Church was achieved in 1868 (Ramirez and Boli 1987, p.5). In Hungary, compulsory schooling was introduced in 1777 with the "Ratio Educationis". The 1869 Reichsvolksschulgesetz (the upper bound) applied to all the countries of the Empire.
	Hungary	1777	1777	1869		Ratio Educationis	
Britain	**England**	1880	1880	1880	Soysal and Strang (1989), Flora et al. (1983), Ritter (1986), Salimova and Dodde (eds.) (2000), Anderson (1995)	Elementary Education Act 1870	Compulsory education of eight years was introduced with exceptions in England and Wales in 1880 (Flora et al. (1983), p.623). School became compulsory in 1881 and free in 1891. However, the legislation was not implemented in the same way in every community. That is, some communities continued to depend on voluntary schooling or under the control of religious groups (Salimova and Dodde (eds.) (2000), p.108). In Scotland, compulsory schooling was already introduced in 1872 (lower bound) with the "Education (Scotland) Act".
	Scotland	1872	1872	1872		Education (Scotland) Act	
	Wales	1880	1880	1880		Elementary Education Act 1870	
Germany*	**Prussia**	1717	1717	1763	Ramirez and Boli (1987), Stolze (1911), Salimova and Dodde (eds.) (2000), Flora et al. (1983), Oelkers (2009)	Schuledikt (Schools Edict, September 28, 1717)	The first German state to introduce compulsory schooling was Palatinate-Zweibrücken in 1592. In Prussia, compulsory schooling was introduced by Frederick William in 1717, and reiterated by Frederick II in 1763. The general law of the land (Allgemeines Landrecht) of 1794 makes instruction - as opposed to attendance - mandatory, a fact that had consequences for school attendance and organization. In this system the state only regulates the minimum for those parents who cannot provide for their children's attendance. [...] Elementarschulen became unavoidable but actually only for the poorer classes of the population, who could not afford a better form of education (Salimova and Dodde (eds.) (2000), pp.179-180). Upon unification of the German Empire in 1871, compulsory schooling (which existed in Prussia) was extended to all states. Eight years of compulsory education were introduced in the German Empire with the exception of Wurtemberg and Bavaria where only seven years were introduced (Flora et al. (1983), p.584). Most states already had compulsory schooling before 1871 (detailed information on all states was not available)
	Palatinate-Zweibrücken	1592	1592	1592			
	German Empire	1871	1871	1871			
Italy	**Kingdom of Sardinia**	1859	1859	1859	Cubberley (1920), Schneider (1982), Ramirez and Boli (1987)	Legge Casati	In the Kingdom of Sardinia, compulsory education was introduced in 1859 (2 years in all communes, 4 years in communes over 4,000 population) (Flora et al. (1983), p.598). Upon unification, compulsory school attendance was extended to all Italian provinces. This process was completed in 1877. The education system was quite effective in some of the Northern regions by 1880 and in Southern regions by 1900 (Ramirez and Boli 1987, p.7)
	Kingdom of Italy	1877	1877	1877		Legge Coppino	

Notes: ** The data for Germany is not exhaustive as we were unable to locate information for all regions. Only Prussia (the largest state) and Palatinate-Zweibrücken (the earliest state to enact compulsory schooling) are included here.

Table A.4: Compulsory Schooling Laws and European Enrolment Rates

Source, Enrolment Measure	CSL pre-1850	No CSL pre-1850	Difference (t-test)	Sample	Notes
Lindert [2004]: Primary enrolment rate, 5-14 year olds					
Public+private	60.71	57.28	3.43	Austria, Belgium, England and Wales, Finland, France, Ireland, Italy, Netherlands, Norway, Scotland.	The data from Lindert (2004) *available at lindert.econ.ucdavis.edu. The main data sources used* for Europe are Flora et al. (1983) and Mitchell (2007). He discusses problems with these data provides an alternative estimate based on educational censuses, inspections data and school attendance rates. The exact measure of enrolment in Lindert's data differs between countries. For some countries, he provides public plus private enrollments, for others, only public enrollments; and for others, the exact measure is not specified. For some countries, more than one measure is provided. Therefore, comparisons cannot be made between all countries, but only between those for which a common measure is available. Our initial dataset compiled from Lindert 250 observations from 20 countries. Out of these, 84 from 10 countries are used in the public plus private comparison, 111 from 14 countries in the public comparison, and 30 from 50 countries in the not specified comparison.
Public	57.46	55.9	1.56	Austria, Belgium, Canada, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Scotland, Sweden, Switzerland.	
Not specified	51.42	43.31	8.11	Denmark, Greece, Japan, Russia, Spain.	
Mitchell [2007]: Primary enrolment rate, 5-14 year olds					
Primary	65.24	56.7	8.54**	Austria, Belgium, Denmark, England and Wales, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Poland, Scotland, Spain, Sweden, Switzerland.	Mitchell (2007) compiles data from a large number of sources, mainly the official publications of European governments. He provides yearly data on the number of pupils in primary and secondary school and the size of certain age groups in the population. Age groups provided are however not uniform across countries, and population data only exists for few years (while enrollment numbers are very complete). The data exhibits a number of breaks, at which enrollment "jumps" due to changes in measurement or the school system. These are, however, well-documented. As highlighted by Morrisson and Murtin [2009], the definition of primary and secondary schooling in adopted in Mitchell (2007) is unknown, and information on the school systems of the countries in the sample "can hardly be recovered from other sources" (p. 9). Our initial dataset compiled from Mitchell contains 1274 observations from 19 countries (20 after the partition of Northern and Southern Ireland in 1921). Out of these, 98 from 17 countries could be u84). Most states already had compulsory schooling before 1871 (detailed information on all states was not available)hare residing on a farm, and the average occupational income score). Population groupings are all defined in effect sizes, where this is calculated from population numbers in the cross section of co
Banks and Wilson [2012], CNTS: Number of 5-14 year olds enrolled divided by total population					
Primary	11.66	9.73	1.94***	Albania, Austria (Austria-Hungary until 1913), Belgium, Canada, Denmark, Finland, France, Germany (Prussia until 1866), Greece, Ireland, Italy, Japan, Luxembourg, Netherlands, Norway, Poland, Portugal, USSR (Russia until 1913), Spain, Sweden, Switzerland, United Kingdom.	The data from Banks and Wilson [2012] is available on the CNTS website (http://www.databanksinternational.com/71.html). They adopt the UNESCO definitions of primary and secondary schooling: "First level: Education whose main function is to provide basic instruction in the tools of learning (e.g., at elementary school, primary school). Its length may vary from 4 to 9 years, depending on the organization of the school system in each country; Second level: Education based upon at least four years of previous instruction at the first level, and providing general or specialized instruction, or both (e.g., at middle school, secondary school, high school…)". Furthermore, they aim to omit "data on preprimary, vocational or technical, part-time, and adult education students". Their main data sources are The Statesman's Yearbook and a dataset compiled by Zapf and Flora [1973]. In addition, they use a number of official national government sources and own estimates. Enrolment rates are not measured in terms of a particular age group, but in terms of the entire population. Our initial dataset compiled from CNTS et al. contains 2061 observations from 22 countries. Out of these, 1522 are used in the primary, 1456 in the secondary, and 1455 in the primary plus secondary comparison test.
Secondary	0.76	0.61	0.15***		
Primary + secondary	12.18	10.41	1.77***		
Flora et al. [1983]: Primary enrolment rate, 5-14 year olds					
Primary	67.5	62.3	5.2***	Austria, Belgium, Denmark, England and Wales, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Prussia, Scotland, Sweden, Switzerland.	Flora et al. (1983) use data from the Western European Data Archive to compile their dataset on education, which contains yearly data on primary and secondary school enrollment. It is an unbalanced panel, in which the most common distance between two observations is 5 years, with a mean distance of 4.4. The data is characterized by a large number of missing values. For primary school enrollment, data on the total number of pupils and on their percentage in the 5-14 age group is provided, both for public plus private enrollment and for public enrollment only. Percentages are often missing, as data for the population distribution is not available. As noted by Benavot and Riddle (1988), "it is not always clear whether counts of pupils refer to total enrollments, average yearly enrollments, average yearly enrollments, average daily attendance, or attendance on inspection or examination days" (p. 196). For secondary school enrollment, the data is more complex, reflecting the diversity of schooling systems across cular age group, but in terms of the entire population. Our initial dataset compiled from CNTS et al. contains 2061 observations from 22 countries. Out of these, 1522 are used in the primary, 1456 in the secondary, and 1455 in the primary plus secondary comparison test.population numbers in the cross section of counties in 1890. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In each Panel, the left hand side figure shows the coefficient on the population grouping in the pre-compulsion period. The right hand side figure shows the coefficient on the interaction between the population grouping and the compulsory schooling law dummy.holdings, from [Galor et al. 20
Benavot and Riddle [1988]: Primary enrolment rate, 5-14 year olds, by decade					
Primary	56.15	58.92	-2.77	Austria, Belgium, Canada, Denmark, England and Wales, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Russia, Scotland, Spain, Sweden, Switzerland.	Benavot and Riddle [1988] provide primary enrollment rates for age groups 5-14 for a large number of countries. The data is per decade and spans from 1870 to 1940. It is compiled from several sources; the main source for Western Europe being Flora et al. [1983]. When data gaps were small, they base estimates for the proportion of 5-14 year olds in a country's population on observations from adjacent years. When data gaps are large, estimates are based on a country's level of development. The percentage of estimated values among all values for a decade ranges between 47% and 67%. A particular drawback of this dataset is that it is relatively coarse and provides only few data points for estimations, as it is measured by decade. Our initial dataset compiled from Benavot and Riddle contains 176 observations from 21 countries. In the comparison table, 154 observations are used and no country has to be dropped entirely.

Table A.5: Full Baseline Specification

Non parametric Cox proportional hazard model estimates, hazard rates reported
Robust standard errors; All covariates measured in effect sizes

	(1) Baseline
Share of the State Population that is:	
From European Countries that did NOT have CSL in 1850	2.15***
	(.509)
From European Countries that had CSL in 1850	.780
	(.161)
Non-European Born	1.80***
	(.409)
Enrolment Rate of American-Borns	2.82**
	(1.39)
Enrolment Rate of Europeans From Countries that did NOT have CSL in 1850	.815*
	(.094)
Enrolment Rate of Europeans From Countries that had CSL in 1850	1.03
	(.153)
Enrolment Rate of Migrants From Non-European Countries	1.18
	(.235)
Illiteracy Rate of Adult American-Borns	.155**
	(.134)
Illiteracy Rate of Adult Europeans From Countries that did NOT have CSL in 1850	1.12
	(.197)
Illiteracy Rate of Adult Europeans From Countries that had CSL in 1850	.256***
	(.088)
Illiteracy Rate of Adult Migrants From Non-European Countries	.753
	(.186)
Group Controls	Yes
State Controls	Yes
European Groups Equal [p-value]	[.004]
Euro Without CSL = Non-Euro [p-value]	[.505]
Observations (state-census year)	230

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is equal to one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling is passed. Robust standard errors are reported. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. All coefficients are defined in effect sizes, where this is calculated using census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. We control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the enrolment rate of 8-14 year olds, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, and the share residing on a farm. We also control for the following state characteristics: the total population and the average occupational score of the population. At the foot of the Column we report the p-value on the null hypothesis that the hazard coefficients are the same for the two European groups, and the p-value that the hazard coefficients are the same for the non-European immigrant groups and European borns from countries that did not have compulsory schooling in place in 1850.

Table A.6: Robustness Checks

Non parametric Cox proportional model, hazard rates reported
Robust standard errors; Populations shares measured in effect sizes

	(1) Rolling Window	(2) Americans	(3) Child Labor and Birth Registration Laws in Place	(4) Universal Suffrage and Women's Property Rights	(5) European Child Labor Laws
Share of the State Population that is From:					
European Countries that did NOT have CSL introduced in the past 30 years	2.31*				
	(.995)				
European Countries that had CSL introduced sometime in the past 30 years	.628*				
	(.170)				
American-Born, Second Generation		.777			
		(.213)			
European Countries that did NOT have CSL in 1850		1.62*	2.22***	2.20***	2.58***
		(.447)	(.533)	(.528)	(.851)
European Countries that had CSL in 1850		1.07	.836	.819	.856
		(.244)	(.195)	(.198)	(.161)
Non-European Countries	1.08	1.56**	1.77***	1.76***	1.85***
	(.262)	(.304)	(.377)	(.386)	(.434)
European Countries that had Child Labor Law in 1850					.693
					(.317)
Child Labor Laws in Place			1.19	1.19	
			(.366)	(.360)	
Birth Registration Law in Place			.707	.716	
			(.283)	(.293)	
Universal Suffrage for Men and Women				.904	
				(.199)	
Women Have Right to Property and their Own Earnings				1.15	
				(.356)	
Group and State Controls	Yes	Yes	Yes	Yes	Yes
European Groups Equal (with and without CSL) [p-value]	[.049]	[.241]	[.005]	[.004]	[.004]
Euro Without CSL = Non-Euro [p-value]	[.218]	[.894]	[.386]	[.382]	[.316]
Observations (state-census year)	230	230	230	230	230

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. A non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. Hence tests for significance relate to the null that the coefficient is one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling laws are passed. Robust standard errors are reported. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares in census-years prior to the introduction of compulsory schooling law. From Columns 3 onwards, the European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In all Columns we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850, as well as the one additional group defined in each column): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, the enrolment rate of 8-14 year olds and the share residing on a farm. In all Columns we control for the following state characteristics: the total population, and the average occupational score of the population. In Column 2 we split the American-born population into those with and without foreign-born parents. In Column 3, the child labor laws are derived from Moehling [1999, Table 1], and the year of introduction of birth certificate as official proof of a child's age is extracted from Fagernäs [2014]. In Column 4 the coding for whether the US state has universal suffrage for men is derived from multiple sources, and the state coding for whether women have the right to property and their own earnings is extracted from Geddes *et al.* [2012]. In Column 5 the following European countries are defined to have child labor laws in place in 1850: Britain, France, Germany and Switzerland. At the foot of each Column we report the p-value on the null hypothesis that the hazard coefficients are the same for the two European groups, and the p-value that the hazard coefficients are the same for the non-European immigrant groups and European borns from countries that did not have CSL in place in 1850.

Table A.7: Alternative Estimation Methods and Alternative Coding of CSL in Europe

Robust standard errors; Populations shares measured in effect sizes

	Estimation Method: Parametric: Log Logistic		OLS	Non Parametric: Cox Proportional	
	Coefficients Reported: Time Ratio		LPM	Hazard Rate	
	(1) Log Logistic Time Ratio	(2) Log Logistic Time Ratio and Frailty Parameter	(3) OLS	(4) Lower Bound Definition of CSL	(5) Upper Bound Definition of CSL
Share of the State Population that is From:					
European Countries that did NOT have CSL in 1850	.940***	.944**	.019	1.59**	1.20
	(.020)	(.021)	(.036)	(.343)	(.257)
European Countries that had CSL in 1850	1.02	1.01	.017	.821	.759***
	(.026)	(.015)	(.042)	(.151)	(.076)
Non-European Born Country	.953***	.970*	.050*	2.08***	1.73**
	(.017)	(.016)	(.030)	(.478)	(.433)
State and Group Controls	Yes	Yes	Yes + State and Year FE	Yes	Yes
European Groups Equal [p-value]	[.012]	[.006]	[.967]	[.004]	[.005]
Euro Without CSL = Non-Euro [p-value]	[.520]	[.078]	[.543]	[.332]	[.251]
Gamma Parameter	.025***	.016***			
	(.004)	(.005)			
Theta Parameter		.324			
		(.270)			
Observations (state-census year)	230	230	371	230	230

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. In Columns 1 to 5 a non-parametric Cox proportional hazard model is estimated, where hazard rates are reported. robust standard errors are reported. In Columns 1 and 2 a parametric hazard model is estimated, where the baseline hazard is assumed to follow a log logistic distribution: the time to failure is then reported, and in Column 2 we also allow for a frailty parameter to be estimated. At the foot of Columns 1 and 2 the relevant parameters from the parametric hazard and frailty parameters are reported. In all Columns except 3 tests for coefficient significance relate to the null that the coefficient is one. The unit of observation is the state-census year, for all census years from 1850. A state drops from the sample once compulsory schooling laws are passed. In Column 3 an OLS panel data model is estimated (controlling for state and year fixed effects) where the dependent variable is equal to one if compulsory schooling laws are in place. The year of passage of compulsory school attendance laws is extracted from Landes and Solomon [1972]. In all Columns population share groupings are defined in effect sizes, where this is calculated using population shares in census-years prior to the introduction of compulsory schooling law. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In all Columns we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the labor force participation rate, the enrolment rate of 8-14 year olds and the share residing on a farm. In all Columns we control for the following state characteristics: the total population, and the average occupational score of the population.

Table A.8: First Stage Estimates for 2SRI Instrumental Variables Method

OLS and Nonparametric First Stage Estimates
Standard errors clustered by state in Columns 1 to 3

	Share of the State Population that is:					
	(1) From European Countries that did NOT have CSL in 1850	(2) From European Countries that had CSL in 1850	(3) Non-European Born	(4) From European Countries that did NOT have CSL in 1850	(5) From European Countries that had CSL in 1850	(6) Non-European Born
Bartik-Card Instrument	.807***	.898***	.687***	.484***	.708***	.564***
	(.050)	(.072)	(.151)	(.057)	(.078)	(.160)
Group Controls	No	No	No	Yes	Yes	Yes
State Controls	No	No	No	Yes	Yes	Yes
Observations (state-census year)	180	180	180	180	180	180

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. In Columns 1 to 3 an OLS regression model is used. In Columns 3 to 6 a local linear regression is estimated with Epanechnikov Kernel weights and (constant) optimal cross-validated bandwidth selection based on the leave-one-out Kernel. The outcome variable is the share of state s's population from each migrant group (measured as an effect size). The unit of observation is the state-census year, for all census years from 1860 (the first census year in 1850 is dropped because the Bartik-Card Instrument cannot be constructed for that first period). Standard errors are clustered by state in the OLS specifications in Columns 1 to 3. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden. In Column 4 onwards we control for the following characteristics of each group (American born, non-European, European with and without compulsory schooling laws in 1850): the share aged 0-15, the share of adults (aged 15 and over) that are illiterate, the enrolment rate of 8-14 year olds, the labor force participation rate, and the share residing on a farm. We also control for the following state characteristics: the total population and the average occupational score of the population.

Table A.9: Population and the Passage of Compulsory Schooling Laws by US State

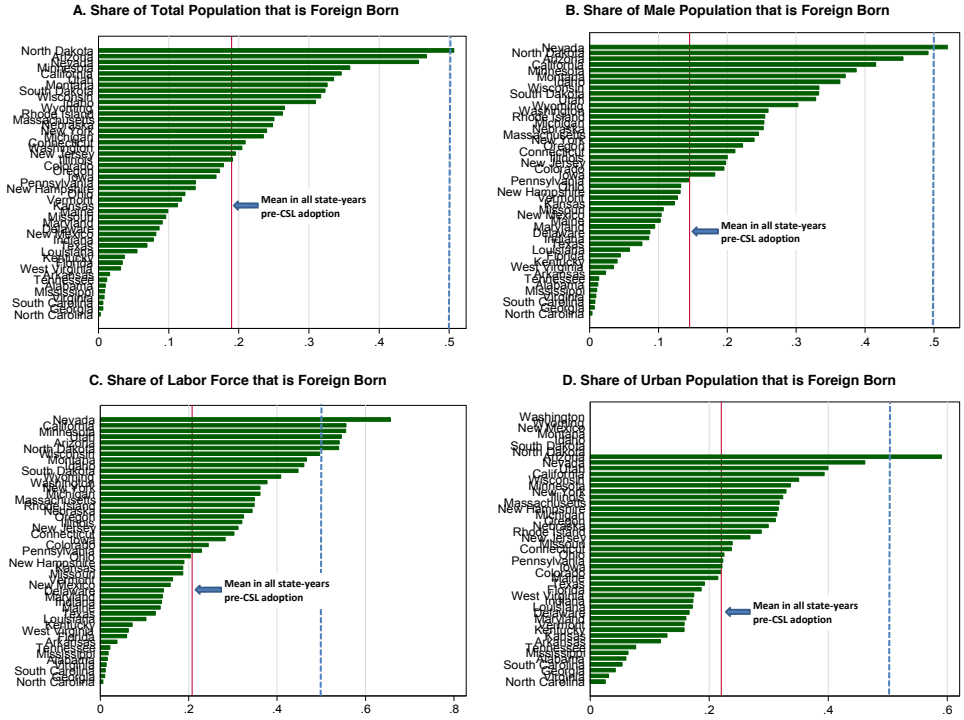
OLS estimates, standard errors clustered by region

	Log (State Population)			Foreign Born Population			
	(1) Unconditional	(2) Fixed Effects	(3) Mean Reversion	(4) Foreign Born Population	(5) European Born from Countries that had CSL in 1850	(6) European Born from Countries that did NOT have CSL in 1850	(7) Ratio of Europeans from Countries without CSL in 1850 to Those that had CSL in 1850
A. Mean Reversion Model							
CSL Passed [yes=1]	1.04***	-.112*	-.074	.113	.098	.063	-2.96
	(.174)	(.056)	(.062)	(.078)	(.106)	(.103)	(2.43)
State Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Census Year x 1850 Population Interactions	No	No	Yes	Yes	Yes	Yes	Yes
Census Year x 1850 Occ Score Interactions	No	No	No	No	No	No	No
Observations (state-census year)	288	288	288	288	286	288	286
B. Trend Break Model							
Post CSL Passage Trend Break	-.003	-.013*	-	-.001	.008	.001	-.251
	(.009)	(.016)	-	(.005)	(.005)	(.004)	(.216)
1850-1930 Trend	.025***	.030***	-	.020***	.017***	.018***	-.032
	(.004)	(.004)	-	(.005)	(.003)	(.003)	(.040)
State Fixed Effects	No	Yes	-	Yes	Yes	Yes	Yes
Observations (state-census year)	288	288	-	288	286	288	286

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. The unit of observation is a state-census year from 1850 to 1930. The dependent variable varies across columns: in Columns 1 to 3 it is the log of the total state population, and in Columns 4 to 7 it relates to various migrant populations. All variables are derived from the IPUMS-USA census samples. OLS regression estimates are shown with standard errors clustered by census region. In Panel A, a mean reversion model is estimated (allowing for state and year effects, as well as a linear time effect of the outcome in 1850) and in Panel B a trend break model is estimated (including state fixed effects and a linear time trend). The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden.

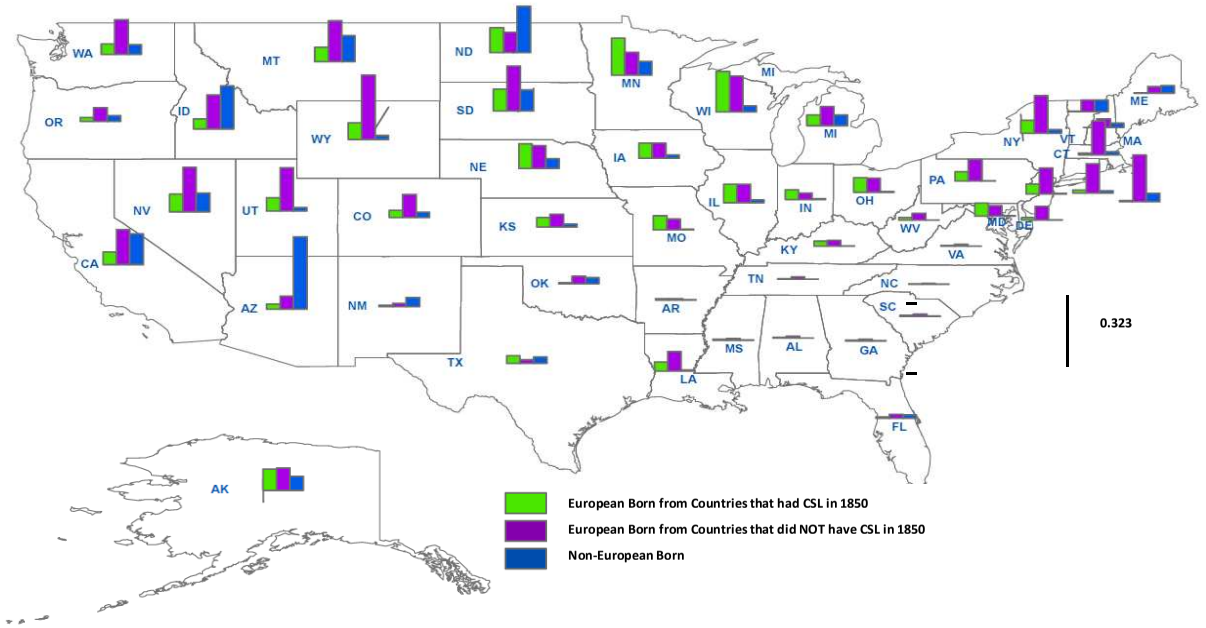
Additional Figures

Figure A.1: Foreign Population by US State, 1880



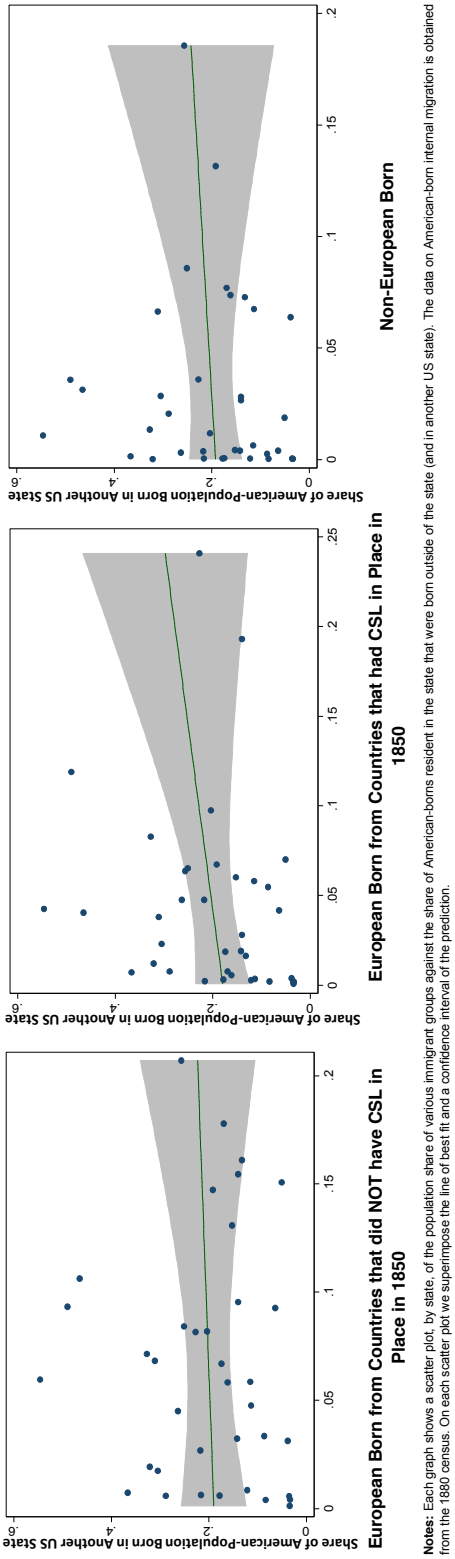
Notes: All variables are derived from the 100% IPUMS-USA 1880 census sample. In Figure D, there are some states in which none of the foreign-born population resides in urban areas. The solid line shows the mean of each variable in all state-census years prior to the adoption of compulsory schooling laws. The dashed line shows the .5 population share.

Figure A.2: Migrant Groups Population Shares, Averaged Across pre-Compulsory Schooling Census Years



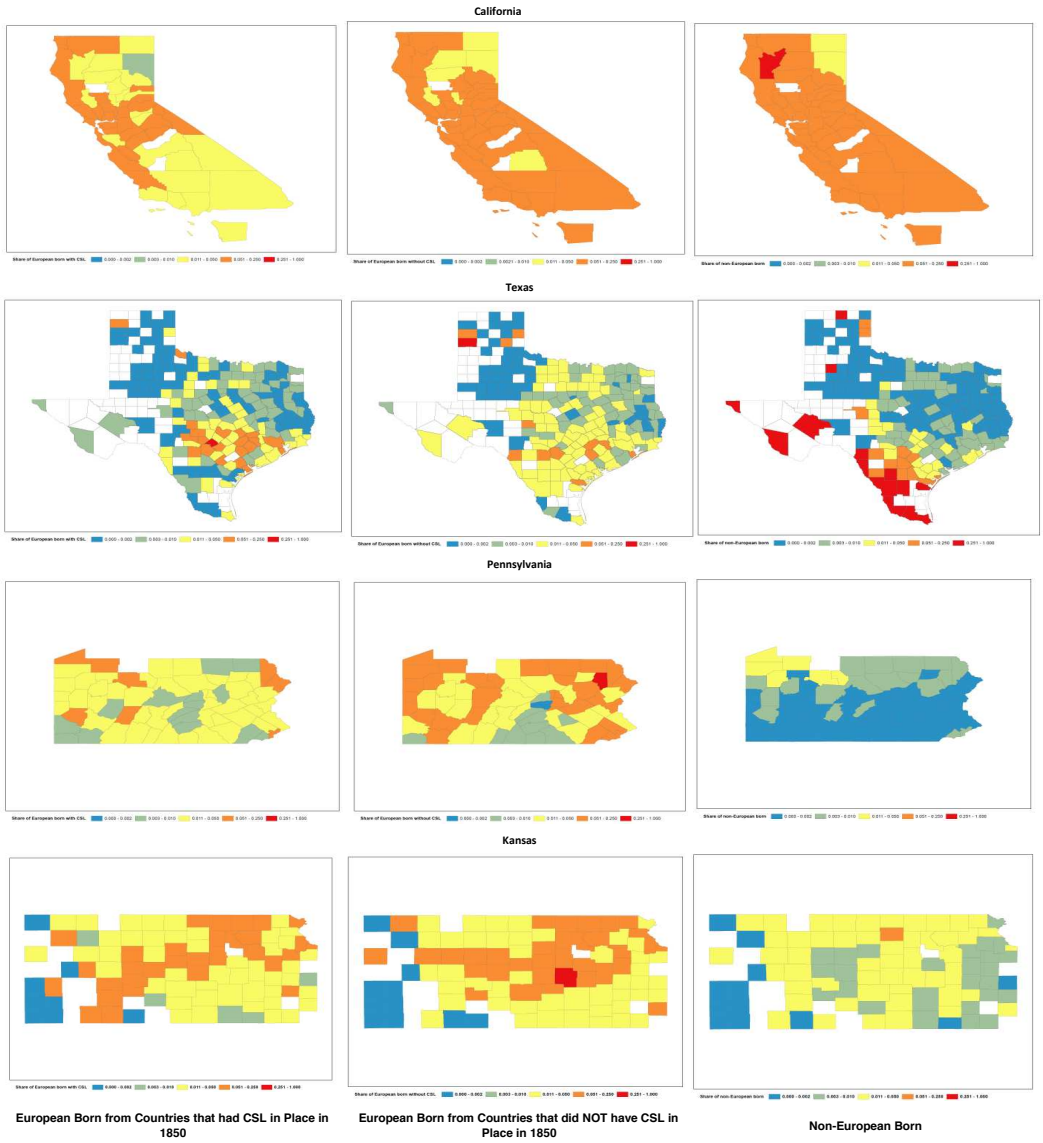
Notes: The bars represents the mean population share of immigrants by group for each US state prior to the passage of compulsory schooling laws in the state. The year of passage of compulsory school attendance laws are extracted from Landes and Solomon [1972]. The European countries defined to have had compulsory schooling laws in place in 1850 are Austria-Hungary, Denmark, Germany, Greece, Norway, Portugal and Sweden.

Figure A.3: Internal Migration by American-Borns and Immigrant Groups



Notes: Each graph shows a scatter plot, by state, of the population share of various immigrant groups against the share of American-borns resident in the state that were born outside of the state (and in another US state). The data on American-born internal migration is obtained from the 1880 census. On each scatter plot we superimpose the line of best fit and a confidence interval of the prediction.

Figure A.4: Foreign Population by US County, 1880



Proofs

Proof of Proposition 1: For any $i \leq i^m$ and for any $j \in \mathbb{R}$ where $j > i^m$ we can rewrite $d_{ij} = d_{ii^m} + d_{i^m j}$. Schooling shifts migrant values towards i^m by λ. So for $i \leq i^m$, as all migrants have values $j > i^m$ this distance becomes $d_{ij} = d_{ii^m} + (1-\lambda)d_{i^m j}$. Introducing compulsory schooling then gives an American-born individual $i \leq i^m$ utility,

$$\begin{aligned} u_{i^m} &= c - \int_{j\in\mathbb{R}} f(j) d_{i^m j} dj - \int_{j\in\mathbb{R}} g(j)[d_{ii^m} + (1-\lambda)d_{i^m j}]dj - T \qquad \text{(A.1)} \\ &= c - \int_{j\in\mathbb{R}} f(j) d_{i^m j} dj - \int_{j\in\mathbb{R}} g(j) d_{ii^m} dj - \int_{j\in\mathbb{R}} g(j) d_{i^m j} dj + \int_{j\in\mathbb{R}} g(j)\lambda d_{i^m j} dj - T \\ &= c - \int_{j\in\mathbb{R}} f(j) d_{i^m j} dj - \int_{j\in\mathbb{R}} g(j)[d_{ii^m} + d_{i^m j}]dj + \int_{j\in\mathbb{R}} g(j)\lambda d_{i^m j} dj - T \end{aligned}$$

Hence the American-born individual $i \leq i^m$ votes for compulsory schooling if $\int_{j\in\mathbb{R}} g(j)\lambda d_{i^m j} dj \geq T$, that can be re-written as (2.3). As this inequality is the same for all American-borns with values $i \leq i^m$, a majority of American-borns vote for compulsory schooling if (2.3) is satisfied and a majority vote against otherwise.■

Proof of Proposition 2: The voter in group j indifferent between voting for party A or B is given by,

$$\begin{aligned} \sigma^{j*} &= u^j(g_A) - u^j(g_B) \qquad \text{(A.2)} \\ &= (g_B - g_A)\frac{y^j\theta}{\bar{y}} + \alpha^j(\theta^j, \mathbf{1}(HCSL^j))(H(g_A) - H(g_B)). \qquad \text{(A.3)} \end{aligned}$$

All voters i in group j with $\sigma^{ij} \leq \sigma^{j*}$ prefer party A. Therefore, the share of the electorate that vote for party A is,

$$\begin{aligned} \pi_A &= \sum_j W^j \phi^j (\sigma^{j*} + \frac{1}{2\phi^j}) \qquad \text{(A.4)} \\ &= \sum_j W^j ((g_B - g_A)\frac{y^j\theta}{\bar{y}} + \alpha^j(\theta^j, \mathbf{1}(HCSL^j))(H(g_A) - H(g_B)) + \frac{1}{2\phi^j}) \qquad \text{(A.5)} \end{aligned}$$

where $W^j = N^j\phi^j$ is group j's political weight. Party A wins the election if $\pi_A > 1/2$. As both parties facing the same optimization problem, in equilibrium they announce the same policy. The equilibrium amount of common schooling is then derived by taking

the first order condition of π_A with respect to g_A and using the fact that $g_A = g_B = g^*$. Solving gives (2.7).■

Coding Compulsory Schooling Laws

US States

The data on the year of enactment of compulsory schooling laws (CSL) across US states was extracted from Landes and Solomon [1972], whose original source was Steinhilber and Sokolowski [1966]. The Landes and Solomon [1972] data has been compared to alternative sources including Katz [1976], Leddon [2010], and the Workers' Compensation Project of Fishback [2000]. Katz [1976] mentions the dates of CSL enactment for a number of states: they are all in accordance with the Landes and Solomon data. Leddon [2010] provides a table with the enactment years of CSL, which correspond exactly to those in Landes and Solomon [1972]. Finally, the Workers Compensation Project Data does not include Alaska and Hawaii, but coincides with Landes and Solomon [1972] for all other available states.

European Countries

Our coding of the introduction of compulsory state schooling laws across European countries relies on primary sources (original laws were consulted whenever possible) and secondary sources of a scientific and official nature (monographs and papers, mostly written by historians, and information provided by governments or the European Union). We focus on the first establishment of general compulsory education in the respective territory of interest. We do not explicitly differentiate between compulsory school attendance and compulsory education, as some countries allow for home schooling. It should be noted that sources on the history of compulsory education in different countries sometimes contradict each other: this is a particular concern for countries with federal systems (such as Switzerland) and for territories which belonged to different national entities over the 19th and 20th century (such as today's Poland and Germany).

Albania Compulsory schooling was introduced when the country became a monarchy in 1928. Article 206 of the Royal Constitution, adopted in 1928, states, "The primary

education of all Albanian subjects is obligatory, and the State schools are free" [Hörner *et al.* 2007, Sefa and Lushnje 2012].

Armenia Compulsory primary schooling was introduced in 1932 [EFA 2000, Hörner *et al.* 2007].

Austria-Hungary As part of a comprehensive schooling reform, Maria Theresia signed the General School Ordinance (Allgemeine Schulordnung) in 1774, which made schooling compulsory for children of both genders between 6 and 12 throughout most of the Austro-Hungarian territory. Article 12 of the ordinance states, "children of both sexes whose parents or guardians do not have the will or the means to support a tutor should go to school without exception (...) as soon as they have entered their 6th year". In order to be allowed to leave school before the age of 12, children needed to "prove in public exams, and provide a written certificate by the superintendent, that they had learnt all the necessary".[1] The ordinance further stipulates that municipal authorities in the city and teachers in the country should keep a list of children who have to attend school and admonish parents to send their children to school. This regulation did not apply to Hungary, where schooling was however made compulsory in 1777 with the Ratio Educationis [Melton 1988]. The 1774 law could not be fully enforced, such that analphabetism remained a widespread phenomenon in Austria in the 19th century. To increase school attendance, Maria Theresia's son and successor Joseph II established punishments for non-compliance in 1781. In 1869, a comprehensive new schooling law (the Reichsvolksschulgesetz) was enacted. It restated the compulsory character of schooling (Art. II.20) and increased years of compulsory attendance from 6 to 8

[1] "Kinder, beiderlei Geschlechts, deren Ueltern, oder Vormünder in Städten eigene Hauslehrer zu unterhalten nicht den Willen, oder nicht das Vermögen haben, gehören ohne Ausnahme in die Schule, und zwar sobald sie das 6te Jahr angetreten haben, von welchem an sie bis zu vollständiger Erlernung der für ihren künftigen Stand, und Lebensart erforderlichen Gegenstände die deutschen Schulen besuchen müssen; welches sie wohl schwerlich vor dem 12ten Jahr ihres Lebens, wenn sie im 6ten, oder nach dem 6ten angefangen haben, gründlich werden vollbringen können; daher es denn gerne gesehen wird, daß Ueltern ihre Kinder wenigstens durch 6 oder 7 Jahre in den deutschen Schulen liessen (...) Wenn aber einige vor dem 12ten Jahre zu dem Studiren übergehen, oder aus der Schule entlassen sein wollen; so müssen sie in den öffentlichen Prüfungen beweisen, und von dem Schulaufseher ein schriftliches Zeugnis erhalten, daß sie alles Nöthige wohl erlernet haben".

(Art II.21) [Slaje 2009, Donnermair 2010].[2,3] According to Schneider [1982], the 1869 Reichsvolksschulgesetz achieved compulsory schooling even in rural areas.

Belgium Primary schooling was made compulsory in 1914 with the Loi Poullet [Flora *et al.* 1983, Wielemans 1991, Colle-Michel 2007, Gathmann *et al.* 2012].

Denmark Education was first made compulsory in Denmark-Norway in 1739, to prepare children for confirmation. Under those provisions, education consisted of the basics of religion and the reading of familiar texts. In Denmark, writing was added to the curriculum with the 1814 Education Act, when compulsory primary schools were established [Schneider 1982, Flora *et al.* 1983, Simola 2002, Bandle *et al.* 2005, Gathmann *et al.* 2012].

Finland Primary schools were established in 1866 and became compulsory in 1921 with the Compulsory School Attendance Act. However, universal primary school attendance was only achieved at the time of the Second World War [Flora *et al.* 1983, Simola 2002].

France In France, law no. 11 696 of March 28, 1882 (Loi Jules Ferry), made primary education compulsory for children of both sexes aged 6-13 years [Cubberley 1920, Schneider 1982, Flora *et al.* 1983, Schriewer 1985]. Its Article 4 states, "primary instruction is compulsory for children of both sexes from 6 to 13 years of age".[4] Children were allowed to leave school at age 11 if they passed the public examination for the "certificate of primary studies". A municipal commission was set up to monitor and encourage school attendance by keeping lists of school-aged children and taking different types of measures in case of non-compliance.

Germany Education was made compulsory in Prussia in 1717 with the School Edict (Schuledikt) enacted by Frederick William I, who "made attendance at village schools compulsory for all children not otherwise provided with instruction" [p4, Ramirez and Boli 1987]. According to Stolze, this was the first time Frederick William proclaimed

[2] "Die Eltern oder deren Stellvertreter dürfen ihre Kinder oder Pflegebefohlenen nicht ohne den Unterricht lassen, welcher für die öffentlichen Volksschulen vorgeschrieben ist."

[3] "Die Schulpflichtigkeit beginnt mit dem vollendeten sechsten, und dauert bis zum vollendeted vierzehnten Lebensjahre."

[4] "L'instruction primaire est obligatoire pour les enfants des deux sexes âgés des six ans révolus à treize ans révolus."

schooling to be compulsory in all Prussian provinces [Stolze 1911]. This regulation was reiterated by his son Frederick II in his 1763 "General Regulations for Village Schools" (General-Landschul-Reglement), which decreed compulsory schooling for the entire Prussian monarchy. Article 1 of the general regulations stipulates that "all subjects sent both their own children and children entrusted to them, boys or girls, from their fifth year of age on, to school".[5] The regulation stated the school fees to be paid. For those too poor to afford them, they should be financed through church or village donations. The responsibility to enforce attendance lay with the local preacher and court authorities, who were able to sanction fines for non-compliance. The General-Landschul-Reglement did not apply to Catholics and urban residents. However, a separate edict was promulgated in 1765 for Silesian Catholic schools. Given widespread opposition, compulsory schooling only became effective over a long period [Ramirez and Boli 1987, Melton 1988]. In the German Empire, education became compulsory upon unification in 1871, but precise regulations differed between states (in Bavaria and Wurtemberg, school was compulsory for children between 7 and 14, whereas in the rest of the Empire, it was for those aged between 6 and 14) [Flora *et al.* 1983]. Not only Prussia, but also most of the other German territories had already introduced compulsory schooling before unification. The first state to do so was Palatinate-Zweibrücken in 1592 [Oelkers 2009]. The state of Weimar introduced compulsory education in 1619 according to Ramirez and Boli [1987], and the Kingdom of Bavaria in 1802 according to De Maeyer [2005], a date which is, however, contradicted by other sources.

Great Britain In England and Wales, the 1870 Elementary Education Act (Forster's Education Act) established state responsibility for primary education. Schooling was made compulsory for children aged between 5 and 13 ten years later, in the Education Act of 1880 [Flora *et al.* 1983, Ritter 1986]. In Scotland, education became compulsory

[5] "Zuvörderst wollen Wir, daß alle Unsere Unterthanen, es mögen denn Eltern, Vormünder oder Herrschaften, denen die Erziehung der Jugend oblieget, ihre eigene sowol als ihrer Pflege anvertraute Kinder, Knaben oder Mädchen, wo nicht eher doch höchstens vom Fünften Jahre ihres Alters in die Schule schicken, auch damit ordentlich bis ins Dreyzehente und Vierzehente Jahr continuiren und sie so lange zur Schule halten sollen, bis sie nicht nur das Nöthigste vom Christenthum gefasset haben und fertig lesen und schreiben, sondern auch von demjenigen Red und Antwort geben können, was ihnen nach den von Unsern Confistoriis verordneten und approbirten Lehrbüchern beygebracht werden soll."

for all children between 5 and 13 in 1872 with the Education (Scotland) Act [Flora *et al.* 1983, Anderson 1995].

Greece Education was made compulsory in a 1834 decree on elementary education, which was part of the so-called "Bavarian Plan", an educational reform which took place under the reign of King Otto, a Prince of Bavaria. [Gkolia and Brundrett 2008, Cowen and Kazamias 2009].

Ireland Schooling was made compulsory in 1892 by the Irish Education Act [Akenson 1970, Schneider 1982, Flora *et al.* 1983]. Children were excused from compulsory attendance during harvest and other seasons during which their labor was needed. Furthermore, children aged between 11 and 14 could obtain a work permit if they had a "certificate of proficiency in reading, writing and arithmetic". School attendance committees were in charge of enforcing the legislation, and courts could impose modest fines on parents who refused to comply. Nonetheless, the law appeared to have little impact on school attendance during the 19th century [Akenson 1970].

Italy Compulsory schooling in Italy is based on the Legge Casati, enacted in 1859 in the Kingdom of Sardinia. This law defined elementary schooling to consist of two grades, inferior and superior, each of which takes two years. Article 326 states that "[p]arents, and those who act as their substitutes, are obliged to procure, in the way they believe most convenient, to their children of both sexes in the age of attending public elementary school of the inferior grade, the instruction which is given in those".[6] Elementary education was provided free of charge. The law became effective in 1860, and was extended to all Italian provinces upon unification. The legal framework was completed in 1877 with the Legge Coppino, which reiterates the compulsory character of education in its first article: "Boys and girls who have completed the age of six years, and to those parents or those acting as their substitutes have no procured the necessary instruction (...) have to be sent to the local public school".[7] However, it did not result in

[6] "I padri, e coloro che ne fanno le veci, hanno obbligo di procacciare, nel modo che crederanno più conveniente, ai loro figli dei du sessi in età di frequentare le scuole pubbliche elementari del grado inferiore, l'istruzione che vien data nelle medesime."

[7] "I fanciulli e le fanciulle che abbiano compiuta l'età di sei anni, e ai quali i genitori o quelli che ne tengono il luogo non procaccino la necessaria istruzione (...) dovranno essere inviati alla scuola elementare del comune."

universal school attendance everywhere. Additional laws were hence enacted in 1904 and 1911, which made more stringent provisions for school attendance and increased state aid for elementary schools [Cubberley 1920, Schneider 1982, Ramirez and Boli 1987].

Luxembourg Compulsory schooling was introduced in Luxembourg through the 1881 law on the organisation of primary education [European Commission 2010]. Article 5 of this law states that "every child of either sex, having completed six years of age at the beginning of the school year, has to receive during six consecutive years instruction in the subjects listed...".[8] However, the compulsory character of schooling is reflected in earlier laws as well. Article 23 of the 1843 law on primary instruction (which is bilingual) defines "children of school-age" ("schulpflichtige Kinder" in its German, "enfans susceptibles de fréquenter l'école" in its French version) as those between 6 and 12 years of age.[9] While the French wording is less explicit, the German wording "Schulpflicht" clearly implies an obligation to attend school. Article 56 of the same law even specifies sanctions for non-compliance. For example, "indigent parents who habitually neglect sending their children to school, can be prived from public support."[10,11]

Netherlands Compulsory education was introduced in 1900, with "De Leerplichtwet" [Schneider 1982, Flora *et al.* 1983, Gathmann *et al.* 2012].

---

[8] "Tout enfant de l'un ou de l'autre sexe, âgé de six ans révolus au commencement de l'année scolaire, doit recevoir pendant six années consécutives l'instruction dans les matières énumérées (...)" / "Jedes Kind beiderlei Geschlechts, welches bei Beginn des Schuljahres das sechste Lebensjahr zurückgelegt hat, muß während sechs aufeinander folgender Jahre in den (...) angegebenen Lehrgegenständen unterrichtet werden."

[9] Sont considérés comme tels, les enfans qui, á partir du premier octobre de chaque année, ont six ans révolus et moins de douze ans accomplis (...)" / "Als solche werden diejenigen Kinder betrachtet, welche vom 1. October jedes Jahres an sechs Jahre zurückgelegt haben und noch nicht volle 12 Jahre alt sind (...)".

[10] "Les parens indigens qui négligeront habituellement 'envoyer leurs enfans aux écoles, pourront être privés des secours publics." / "Die dürftigen Eltern, die gewohnheitlich unterlassen, ihre Kinder in die Schule zu schicken, können von den öffentlichen Unterstützungen ausgeschlossen werden."

[11] Earlier administrative documents, in particular a circular from 1842 and an ordinance from 1840, refer to a school regulation from 1828. The original text of the 1828 regulation could not be accessed, which is why we could not determine whether schooling was made first made compulsory in 1828 or in 1843.

Norway Education was first made compulsory in Denmark-Norway in 1739, to prepare children for confirmation. Under those provisions, education consisted of the basics of religion and the reading of familiar texts. In Norway, writing was added to the curriculum in 1827 with a new primary school law, but children were typically unable to write more than their name and the letters of the alphabet. Several authors regard the 1827 Primary School Act as the first compulsory schooling law of Norway [Hove 1967, Einhorn 2005]. Still in 1857, 80% of rural children only had access to ambulant schooling, as there were no schools in their parishes. This changed after the 1860 School Law, which provided for permanent schools instead [Rust 1990]. In 1889, a stricter compulsory schooling law was enacted, requiring "a more demanding mother tongue subject" and 7 years of primary school attendance [Hove 1967, Bandle *et al.* 2005].

Poland During the 19th century Poland was partitioned between Prussia, Russia and Austria-Hungary on three occasions. Education in Poland was, on the one hand, largely determined by the respective occupier, but reflected, on the other hand, the efforts of the Polish to upheld their cultural heritage [Slaje 2009]. In the Prussian part of Poland, compulsory schooling was introduced in 1825 [Biskup 1983]. Sources are contradictory on whether there was corresponding legislation in the Austrian and Russian parts during the partition. Shortly after re-obtaining its independence in 1918, Poland enacted a decree "On Compulsory Schooling" (O obowiazku skolnym) which made school attendance compulsory for children between 7 and 14 in 1919 [Slaje 2009].

Portugal Compulsory schooling was first introduced in Portugal in 1835, with the Regulamento Geral da Instrucção Primaria. In Title VII, Article 1, it states that "To the obligation imposed, by the constitution, on the government to provide all citizens with primary education, corresponds the obligation of parents to send their children to public schools, as soon as the pass 7 years (...) if they don't have the means to educate them otherwise".[12] The responsibility for enforcement rested on municipal authorities and

[12]"A obrigação imposta, pela Carta Constitucional, ao Governo de proporcionar a todos os Cidadãos a Instrucção Primaria, corresponde a obrigação dos Pais de familia de enviar seus filhos às Escòlas Publicas, logo de passem de 7 annos, (...), se meios não tiverem de o fazer construir de outro modo."

priests.[13]

Russia Compulsory education for children between 6 and 17 years of age was introduced shortly after the success of the October Revolution, with the Dekret ot "ob Edinoy Trudovoy Shkole Rossiyskoy Sozialisticheskoy Federativnoy Sovetskoy Respubliki (Polojenie)" (Decree on the Unified Labour School of the Russian Soviet Federative Socialist Republic) of October 16, 1918 [Presidential Library 2013].

Spain The first law to regulate education in Spain was the 1838 Law of Primary Instruction (Ley de Instrucción Primaria). It was accompanied by a Plan of Primary Instruction (Plan de Instrucción Primaria), which stipulates the obligation of villages and cities to provide primary schools (Art. 7-10). Furthermore, its Article 26 states that "[a]s it is an obligation of parents to procure for their children, and for guardians to procure for the persons under their responsibility, the amount of instruction which can make them useful for society and for themselves, the local commissions will assure by the means their prudence dictates them to stimulate parents and guardians to comply with this important duty, applying at the same time all their enlightenment and zeal to the removal of obstacles which would impede it,", remaining thus highly vague with respect to the content and form of such an instruction.[14]

Compulsory education was introduced with the Law of Public Instruction of September 9, 1857 [De Maeyer 2005, Gathmann *et al.* 2012]. Article 7 states that "Elementary primary education is compulsory for all Spanish. The parents and guardians must send their children and wards to public schools from the age of six to nine years; unless they provide them sufficiently with this type of instruction in their homes or in private establishments".[15]

[13] "A's Camaras Municipaes, e aos Parochos incumbe o procurar mover por todos os meios de que poderem usar, os Pais de familia a cumprir com esta importante obrigação..."

[14] Siendo una obligacion de los padres procurar á sus hijos, y lo mismo los tutores y curadores á las personas confiadas á su cuidado, aquel grado de instruccion que pueda hacerlos útiles á la sociedad y á si mismos, las Comisiones locales procurarán por cuantos medios les dicte su prudencia estimular á los padres y tutores al cumplimiento de este deber importante, aplicando al propio tiempo toda su ilustracion y su celo á la remocion de los obstáculos que lo impidan."

[15] "La primera enseñanza elemental es obligatoria para todos los españoles. Los padres y tutores o encargados enviarán a las Escuelas públicas a sus hijos y pupilos desde la edad de seis años hasta la de nueve; a no ser que les proporcionen suficientemente esta clase de instrucción en sus casas o en establecimiento particular".

Sweden Compulsory education was introduced in 1842 with the Folkskolestadgan [Schneider 1982, Soysal and Strang 1989, Simola 2002].

Switzerland With the adoption of the Swiss Federal Constitution (Bundesverfassung) of 1874, primary schooling became mandatory in all Swiss cantons [Schweizerische Eidgenossenschaft 1874, Muller 2007]. Article 27.2 states that "Cantons provide sufficient primary education, which shall be exclusively under the control of the state. It is compulsory and, in public schools, free of charge."[16] However, compulsory schooling had been introduced previously by different cantons at different points in time. Sources contradict each other in terms of the dates of introduction. For example, Forster [2008] dates the introduction of compulsory schooling in Geneva in 1536, whereas Muller [2007] sets it at 1872.

Robustness Checks

Our first robustness check explores a specification exploiting *within-country* variation over time, in exposure to compulsory state schooling. To do so, we consider the impact of a rolling window of Europeans' exposure to compulsory schooling by examining whether the American median-voter is differentially sensitive to the presence of European migrants that have passed compulsory schooling at least 30. Figure 2 makes clear that using a rolling window for Europeans' exposure to compulsory schooling adds in a number of significant countries that pass compulsory schooling between 1850 and 1880 (Spain, Switzerland, Italy and Britain) and so might impact the cross-state passage of compulsory schooling in the US from 1910 onwards. Column 1 of Table A5 shows that with this definition the sharp contrast between how American-borns react to different types of European migrant becomes even more pronounced.

Another way to examine differential responses over time of American voters to individuals with the same country of origin is to focus in on second generation migrants. They are American-born and coded as such, but the next specification splits American-borns between those with American-born parents and those with at least one foreign-born parent. This latter group of individuals form an additional group j that

[16] "Die Kantone sorgen für genügenden Primarunterricht, welcher ausschliesslich unter staatlicher Leitung stehen soll. Derselbe ist obligatorisch und in den öffentlichen Schulen unentgeltlich."

can then also be controlled for (we then also control for the group characteristics of second generation immigrants in X_{st}^{j}). Column 2 in Table A5 shows the result: the passage of compulsory schooling is not significantly impacted by the presence of second generation migrants, rather it is the composition of more *recent* foreign-born migrants that drives the policy response of US states.

Other Legislation

The next set of robustness checks include additional controls in (2.4). First, we consider the passage of *other* pieces of state legislation, that might be complementary to, or pre-requisites for, the passage of compulsory schooling. For example, the passage of child labor laws and the establishment of a birth registration system have been argued to be interlinked with compulsory schooling [Lleras-Muney 2002, Goldin and Katz 2003]. Column 3 of Table A5 shows the baseline results to be unchanged if we additionally control for whether a state has child labor laws or a system of birth registration. Given the stability of our coefficients of interest, this finding further implies migrant groups were not differentially attracted to states based on these legislative and regulatory characteristics.[17]

A second concern is that some states might be more progressive than others, in that they are more likely to pass compulsory schooling, but also be more likely to universal suffrage or to allow women property rights and over their own earnings. If migrants from European countries are differentially likely to locate to such progressive states (as a function of their country of origin's own legislative history), our earlier result would be spurious. To check for this we then additionally control for both state characteristics. Column 4 shows that neither having universal suffrage nor property rights for women have significant impacts on the passage of compulsory schooling in the state (neither hazard significantly differs from one). Moreover, the impacts of the presence of different migrant groups replicate the baseline findings.

Finally, we consider additionally controlling for the presence of European mi-

[17]The coding for child labor laws are extracted from Moehling [1999, Table 1] as these extend back to the mid-1800s (an updating coding is also provided in Lleras-Muney and Shertzer.[2015] for the 1910-39 period); the coding for the introduction of birth registration proofs is extracted from Fagernas [2014].

grants from *countries* that have passed other pieces of legislation, apart from compulsory schooling, that might relate to migrant values. For example, we consider whether the American-born median voter responds to the presence of Europeans from countries with child labor laws in place since 1850. Column 5 shows there is no impact of having migrants in the state from European countries with a long history of child labor laws, that might otherwise have reflected the passage of compulsory schooling as being driven by the child-related preferences of migrants (and natives), rather than compulsory schooling being driven by the desire of the American-born median voter to homogenize certain incoming migrants.

Alternative Econometric Specifications

A next set of robustness checks relate to using alternative econometric specifications. We impose more parametric structure on the underlying hazard, $h_0(t)$, using a log logistic model. When estimating this model, time ratios are reported.[18] Recall that a time ratio *less* than one has the same interpretation as a hazard greater than one, indicating the covariate is associated with the passage of compulsory schooling *earlier* in time. Column 1 in Table A6 shows that imposing this parametric structure leaves our core findings unchanged: (i) the passage of compulsory schooling occurs significantly earlier in time when a greater share of the population comprises European migrants without historic exposure to compulsory schooling; (ii) the time ratio on Europeans with historic exposure to compulsory schooling is above one and these time ratios are significantly different between the European migrant groups; (iii) compulsory schooling is passed significantly earlier in time when a greater share of the population is non-European born. All these findings to continue to hold when we allow for there to be cross-state heterogeneity in hazard rates as captured by a frailty parameter (Column 2).

We next move away from survival models and use a linear probability regression, following some of the earlier literature examining the passage of compulsory school-

[18] In the log logistic model the hazard rate is characterized as $h(t,X) = \frac{\lambda^{\frac{1}{\gamma}} t^{(\frac{1}{\gamma}-1)}}{\gamma[1+(\lambda t)^{\frac{1}{\gamma}}]}$, where $\lambda = \exp -(X\beta)$. This has two parameters: λ is the location parameter and γ is the shape parameter, allowing for non-monotonic hazards.

ing. Such models use *all* state-years (not just those pre-adoption) to essentially estimate the probability that state s has compulsory schooling in place, and are equivalent to a survival model assuming duration *independence* in the passage of legislation. Column 3 shows the result: using a regression model we find no significant partial correlation between the population shares of either European migrant grouping and the likelihood compulsory schooling is passed, although an increase in the population share of non-Europeans does have a positive and significant impact, consistent with earlier work [Landes and Solomon 1972, Lleras-Muney and Shertzer 2015]. The reason why the OLS and survival results differ is that the assumption of duration *independence* is strongly rejected in our data: history does matter and so the hazard of passing legislation, $h_0(t)$, varies over census years t, a result demonstrated in the unparameterized Cox proportional hazard model, and the parametric log logistic specification.

Alternative Classifications

We now consider alternative ways to group European countries by their exposure to compulsory schooling. We first regroup European countries using the lower and upper bound definitions of the introduction of compulsory schooling (shown in Table A2). The results are in Columns 4 and 5 of Table A6: our baseline result is robust to using the lower bound definition and so narrowing down the focus on those European countries that have the longest exposure to compulsory schooling at home. Using the upper bound definitions, the results suggest compulsory schooling is significantly less likely to be passed in the presence of European migrants with exposure to compulsory schooling at home, and the hazard of compulsory schooling being passed across US states remains significantly differently related to the two groups of European migrant, with and without compulsory schooling at home [p-value= .005].

Internal Migration

American-borns

If the passage of compulsory schooling was an instrument used by states to attract American migrants (or Americans took ideas over compulsory schooling with them as they migrated across states), and that the location of the foreign-born groups we

focus on in Table 2 is interlinked with the internal migration of white American-borns, this would generate a spurious correlation between the presence of these foreign-born groups and the cross-state passage of compulsory schooling. To check for this, we use data on the internal migration of Americans from the 1880 census to plot the cross-state variation in Americans born out-of-state (but in the US) and the foreign-born population group shares core to our analysis ($N^{j}_{s,1880}$). Figure A3 shows the result (and line of best fit): we find no significant relationship between the population share of out-of-state American-borns, with the population shares of Europeans with and without long exposure to compulsory schooling at home, or non-Europeans. This suggests our findings are not merely picking up the internal migration of white American-borns.[19]

Foreign-borns

We can further check whether the passage of compulsory schooling in state s by census year t, is associated with subsequent changes in the composition of the migrant population within the state. This sheds light on the narrower issue of whether any process by which natives and migrants sort into states is significantly altered by the introduction of compulsory schooling law. We use two specifications to check for whether population trends shift in response to compulsory schooling:

$$N^{j}_{st} = \mu \mathbf{1}(CSL_{st} = 1) + \delta_s + \delta_t + \sum_t \theta_t (N^{j}_{s1850}.\delta_t) + u_{st}, \tag{A.6}$$

$$N^{j}_{st} = \delta t + \kappa [(t - CSL_{st})\mathbf{1}(CSL_{st} = 1)] + \delta_s + \varepsilon_{st}, \tag{A.7}$$

where N^{j}_{st} corresponds to measures of the state-year population, and $1(CSL_{st} = 1)$ is a dummy for whether compulsory schooling law has been adopted in state s by census year t. Specification (A.6) allows for a complete set of state and year fixed effects (δ_s, δ_t), and also allows for there to be long run reversion to the mean in populations across states, as captured in the $N^{j}_{s1850}.\delta_t$ term. Specification (A.7) is a standard trend break model, that allows for state fixed effects, but assumes population follows a linear

[19] Rocha *et al.* [2015] provide long run evidence on the economic/industrial development of Brazilian municipalities that explicitly used settlement policies to attract high skilled migrants into them in the late 19th and early 20th century.

time trend (δt) and then tests for a break in this linear trend in the years after compulsory schooling law has been adopted in state s.

Table A8 presents the results: Panel A shows estimates of μ from (A.6), and Panel B shows estimates of κ and δ from (A.7). In Columns 1 to 3 we focus on the partial correlation between the passage of compulsory schooling in a state on the subsequent total state population ($N_{st} = \sum_j N_{st}^j$). Examining Panel A, we see that unconditionally, states with compulsory schooling subsequently have significantly larger populations, but this result is not robust: including state fixed effects reduces the magnitude of the partial correlation by 90%, and allowing for reversion to the mean eliminates any significant correlation between the total population and the earlier passage of compulsory schooling. Columns 4 to 7 focus on the composition of the foreign-born population in the state. We find no evidence that after compulsory schooling laws are passed, the foreign born population, European migrants from countries with a long history of compulsory schooling, European migrants from countries without a long history of compulsory schooling, or the ratio of the two groups of European migrant, are significantly different. These results go firmly against the idea that native or migrant population movements are endogenously driven by the earlier passage of compulsory schooling in a state. Equally, the results suggests migrant groups were not resisting the civic values being imparted onto them via compulsory schooling by moving to other states. These conclusions are reinforced if we move to Panel B where (A.7) is estimated: we again find little evidence of native or migrant populations being responsive to the earlier passage of compulsory schooling ($\widehat{\kappa} = 0$ in five out of six specifications).

IV Method

We use a control function (CF) approach to implement an instrumental variables strategy based on a Bartik-Card style instrument for migrant shares. The non-linear hazard model in (2.4) is a special case of a generalized regression model: $y_i = D.F(x_i\beta, u)$ for $D : \mathbb{R} \to \mathbb{R}$ a known non-degenerate and monotonic function and $F : \mathbb{R}^2 \to \mathbb{R}$ monotonic in each variable [Han 1987].[20] To overcome potential endogeneity of one of the regres-

[20]For the Cox proportional hazard model, $y_i = F^{-1}(x_i b + u_i)$ with $F(\zeta) = \log \int_0^\zeta h(\tau) d\tau$, and h being the hazard function, $y_i > 0$, $h() > 0$, and $u_i \sim EV(1)$ [Han 1987].

sors in such generalized regression models, the CF approach can be adopted where the unobservable covariate is directly controlled for (rather than instrumenting the endogenous variable as for 2SLS linear models). Terza *et al.* [2008a, 2008b] and Wooldridge [2010] show the consistency of such a two-stage residual inclusion (2SRI) methods for non-linear models.

To make explicit the nature of the endogeneity problem, we first let Z_{st}^{j} denote the exogenous variables (X_{st}^{j}, X_{st}) and add a state-migrant-specific unobservable to the empirical specification in (2.4), denoted V_{st}^{j}, with V_{st} an $S \times J$ matrix of state-migrant unobservables. These unobservables enter additively in the proportional hazard model, that can be written in the regression form,

$$H(t) = \exp(-N_{st}\beta - Z_{st}\psi - V_{st}) + U, \tag{A.8}$$

where $H(t) = \int_0^t h(s)ds$ is the integrated hazard function, $U \sim Exp(1)$, with $U \perp (N_{st}, Z_{st}, V_{st})$, $Z_{st} \perp V_{st}$ but $N_{st} \not\perp V_{st}$. Hence the migrant shares are endogenous in that they correlate with unobservable determinants of compulsory schooling law. The endogenous migration shares N_{st}^{j} are assumed to relate to some instrument W_{st}^{j} according to the following parametric model,

$$N_{st}^{j} = \alpha_j W_{st}^{j} + \delta_j Z_{st}^{j} + e_{st}^{j}, \tag{A.9}$$

where e_{st}^{j} is an error term. We assume the rank condition holds, that the instruments are exogenous ($W_{st}^{j} \perp e_{st}^{j}, \varepsilon_{st}^{j}$) and that $\mathbb{E}[e_{st}^{j} | Z_{st}^{j}, W_{st}^{j}] = 0$. The unobserved V_{st}^{j} component can be decomposed into a term that is potentially correlated with N_{st}^{j} and a residual,

$$V_{st}^{j} = e_{st}^{j\prime}\rho_j + \varepsilon_{st}^{j}, \tag{A.10}$$

where $\varepsilon_{st}^{j} \perp e_{st}^{j}$, and wlog, $\mathbb{E}[\exp(\varepsilon_{st}^{j})] = 1$. The key to the CF approach is to obtain the population expectation conditional on V_{st}^{j}, which under the above assumptions is,

$$\mathbb{E}[H(t) | N_{st}, Z_{st}, V_{st}] = \exp(-N_{st}\beta - Z_{st}\psi - e_{st}\rho), \tag{A.11}$$

where e_{st} is a $S \times J$ matrix of residuals from (A.9). In the first stage, consistent estimates of $(\hat{\alpha}_j, \hat{\delta}_j)$ are obtained by OLS, and predicted values of the residuals are obtained as $\hat{e}^j_{st} = N^j_{st} - \hat{N^j_{st}}$. In the second stage, $\hat{e}_{st} = (\hat{e}^1_{st}, ..., \hat{e}^J_{st})$ is then included in (A.11),

$$\mathbb{E}[H(t)|N_{st}, Z_{st}, \hat{e_{st}}] = \exp(-N_{st}\beta - Z_{st}\psi - \hat{e}_{st}\rho). \tag{A.12}$$

If the first stage is correctly specified, estimating this exponential regression model conditioning on $\hat{e_{st}}$ gives consistent estimates of (β, ψ) [Wooldridge 2010]. The need to include additional covariates when estimating the second stage equation is demanding given our data dimensions: hence we first present result from the most parsimonious model that excludes the exogenous covariates $Z_{st} = (X^j_{st}, X_{st})$ from both stages.

Appendix B

Appendix to Chapter 3

Centrality measures

In this section, I provide the brief description of some terms and variables for a non–valued and undirected network.[1]

Some Basic Notions

Consider a graph $G = (V, E)$ with $|V|$ nodes (also called vertices) and $|E|$ edges, encoded using $|V| \times |V|$ adjacency matrix $\mathbf{A}$:

$$a_{uv} = \begin{cases} 1 & \text{if } u, v \in E \\ 0 & \text{otherwise} \end{cases}$$

A *path* in a graph is simply a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge.

The *distance* between two scientists is the minimum number of edges that takes to go from one to another. This is also known as the *geodesic distance*. If two nodes are neighbours or adjacent, the distance between them is 1. If A links to B, and B links to C (and A does not link to C), then actors A and C are at a distance of 2. The *shortest path length* between nodes v and u, $dist(u, v)$, is defined as the minimum distance (i.e.

[1] For more detailed definitions and descriptions, see Easley and Kleinberg (2010), Jackson et al. (2008)

number of edges) from node v to reach node u:

$$dist(v,u) = \min(a_{vx} + ... + a_{yu}) \tag{B.1}$$

Measures at the node level

The *degree* of node v is the number of edges connected to node v, which is the cardinality of the node's neighbourhood:

$$Degree(v) = \sum_{i=1}^{|V|} a_{iv} \tag{B.2}$$

If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future. This principle is captured by "triadic closure" and its prevalence is measured by the *clustering coefficient*. It is defines as the proportion of neighbours of v that are also connected to each other. This corresponds to dividing the amount of "triangles" of which node v is part of, $\tau_\Delta(v)$, to the amount of possible triangles of which v could be part of, $\tau_3(v)$:

$$cl(v) = \frac{\tau_\Delta(v)}{\tau_3(v)} \tag{B.3}$$

In general, the clustering coefficient of a node ranges from 0 (when none of the node's friends are friends with each other) to 1 (when all of the node's friends are friends with each other).

Closeness defines that a node is central if it is close to other nodes. This takes the inverse of the sum of all path lengths going from node v to all other nodes:

$$Closeness(v) = \frac{1}{\sum_{i=1}^{|V|} dist(v,i)} \tag{B.4}$$

Betweenness of node v is defined as the sum of proportions of the number of shortest

paths between all pairs of nodes that go through node v:

$$Betweenness(i) = \sum_{i \neq j \neq k \in V}^{n} \frac{\sigma(i,j|v)}{\sigma(i,j)} \tag{B.5}$$

where $\sigma(i,j)$ is the total number of shortest paths between any two nodes and $\sigma(i,j|v)$ the amount of those paths that go through v.

Eigenvector centrality states that a node is central if its neighbours are centrals through an iterative paths. Let $\mathbf{A} = (a_{ij})$ be the adjacency matrix. The eigenvector centrality of node v is given by

$$Eigenvector(v) = \frac{1}{\lambda} \sum_{\{u,v\} \in E} eigenvector(u) \tag{B.6}$$

where $\lambda \neq 0$ is a constant. In the matrix form, we have:

$$\lambda eigenvector = eigenvectorA \tag{B.7}$$

Measured at the network level

The *diameter* is the length of the largest geodesic path in a network.

The *average path length* is the mean length of the shortest path between any two nodes. Shorter average path lengths means information has to travel less (on average) to get from randomly–selected node i to j.

The *assortativity coefficient* is a correlation coefficient between the degrees of all nodes on two opposite ends of a link. A positive assortativity coefficient indicates that nodes tend to link to other nodes with the same or similar degree.

Network components are maximally connected networks, that satisfy two conditions. First, two scientists in a network component g_c are either directly linked, or indirectly

linked through a sequence of agents in g_c (i.e. there is a path between the two of them). Second, two scientists in different network components g_c and $g_{c'}$ cannot be connected through any such sequence. The *number of clusters* is the maximal connected components of a graph.

Obituary sources

Manually searches

In addition to obituary records in Azoulay et al. (2010) and Azoulay et al. (2015), other obituary records were manually collected from American National Biography, Australian Dictionary of Biography, National Academy of Science Memoires, Oxford Dictionary of National Biography, Plarr's Lives of the Fellows Online, Royal Society, Who was who, and World Biography Information System. When possible, searches were targeted to the medical profession and death occurring before the age of 68.

Webscrapping

As a second collection method, a webscrapping algorithm was used to query the ProQuest database for the words "Dr." and "MD." along with the surname of medical scientists within obituaries between to 2009. The ProQuest database offers access to obituaries in newspapers since the 18th century. Newspapers include The Guardian (1821-2003), The Observer (1791-2003), The Atlanta Constitution (1868-1945), The Baltimore Sun (1837-1986), Boston Globe (1872 - 1979), Chicago Defender (1910-1975), Chicago Tribune (1849-1988), The Christian Science Monitor (1908-1998), Hartford Courant (1764-1986), Los Angeles Times (1881-1988), The New York Times (1851-2008), New York Tribune (1841-1922), San Francisco Chronicle (1865-1922), The Wall Street Journal (1889-1994), and The Washington Post (1877-1995).

Linking obituary records to "Author-ity"

Linking obituary records to the 9 million individuals in the "Author-ity" database can be a challenge. For each obituary record there are several possible authors with similar names. Using the first and last names of authors is an obvious first step to match records. The "Author-ity" database provide all the different versions of the names of the authors. For a large fraction of the authors, no first name is provided. Given these limitations, I construct four groups: group A consists of individuals with exact matches of the first and last names of authors, group B matches the last name and the

Table B.1: Number of obituary record by source

Obituary source	Count	With a cause of death	Sudden death
American National Biography	380	4	3
Australian Dictionary of Biography	535	0	0
Azoulay et al. (2010)	137	137	57
Azoulay et al. (2015)	451	394	34
International Who is who	102	0	0
National Academy of Science Memoires	216	12	4
Oxford Dictionary of National Biography	709	0	0
Plarr's Lives of the Fellows Online	5,523	3	1
ProQuest (webscrapping)	301	177	55
Royal Society	1,149	1	0
Who was who	626	2	0
World Biography Information System	4,522	9	3
Total	6,186	852	337

first initial, group C fuzzy matches the first and last name[2], and group D fuzzy matches on last name and exactly matches on the initial. All matches regardless of the group must satisfy the following conditions: I eliminate individuals younger than 18 years old the year of their first publication, individuals whose first publication occurs in the last five years to their death if they were 50 years old or older the year of their death and individuals whose last publication was more than ten years before or after their death. The assumptions behind these restrictions is that individuals should be active researchers rather than practitioners. I also eliminate all scientists who have died after 2008. Following these basic restrictions, each group contains a set of potential matches. I then proceed by iterations: at each step, I collect the number of unique matches and continue to the next step with the obituary records for which there are still multiple matches until I find a unique author for each scientist found in the obituary record.

Iteration 1 I first individually match individuals with the same institutional affilia-

[2]For the fuzzy string matching, I compute pairwise string distances between obituary names and author names based on the Levenshtein distance. This algorithm counts the number of deletions, insertions and substitutions necessary to turn one name into another.

tion in the "Author-ity" and the death records.[3]

Iteration 2 For the remaining multiple matches, I further restrict the sample to (1) individuals with first publications between the ages of 25 and 40, and (2) last publication within 5 years pre-post death.

Iteration 3 For those remaining, I choose the individual with the highest number of publication. This is based on the assumption that scientists who have obituary records are most likely to be written for well known scientists.

The following table summarises the number of unique matches found by iteration. The sample used in this study includes all uniquely identified obituary records for each step. Out of the 6186 scientists found in obituary records, I successfully match 1111.

Table B.2: Fuzzy name matching by group

Number of obituary record with unique author ID				
Group	Iteration 1[a]	Iteration 2[b]	Iteration 3[c]	Total
A: Perfect match first and last name	259	41	30	330
B: Perfect match last name and initial	301	80	150	531
C: Fuzzy match first and last name	126	19	20	165
D: Fuzzy match last name and perfect match initial	62	12	11	85
Total	748	152	211	1111

[a] Affliation between Author-ity and death records match.
[b] Among the remaining individuals (i.e. those with multiple matches), I increasing the restriction: first publication below age 40 and last publication can be up to 5 years pre/post death.
[c] Finally, the individual with the most publication is chosen.

[3] Until 2014, only the affiliation and the address of the affiliation of the lead author was included. The obituaries generally record the last institutional affiliation of the deceased. Therefore potential unique match at this stage only include scientists within this subset of lead authors and recorded affiliations in the obituary.

Table B.3: Sudden deaths

Cause of death	Count
heart attack	51
“died suddenly”	23
car accident	14
airplane crash	6
heart failure	5
stroke	5
murder	4
pulmonary embolism	3
scuba-diving accident	2
“died while travelling”	2
accidental fall	2
bike accident	1
aneurysm	1
cardiac arrest	1
cerebral aneurysm	1
brain haemorrhage	1
complications from routine surgery	1
drowned	1
hit in the head by a stone	1
adverse drug reaction/ multi-organ fail	1
trekking accident in Nepal	1
TOTAL	127

Table B.4: Anticipated deaths

Cause of death	Count
cancer	50
brain cancer	16
cancer of the lung	14
breast cancer	12
diabetes	11
heart disease	11
“long illness”	11
cancer of the pancreas	10
cancer of the prostate	7
leukaemia	7
“anticipated”	5
cancer of the colon	5
melanoma	4
bone cancer	3
kidney cancer	3
lymphoma	3
AIDS	2
Alzheimers disease	2
aortic aneurysm	2
esophageal cancer	2
gastric cancer	2
glioblastoma	2
heart attack	2
Hodgkins disease	2
infection of the heart	2
multiple myeloma	2
ocular melanoma	2
Parkinsons disease	2
suicide	2
tumor	2

Anticipated deaths (continued)

Cause of death	Count
aortic dissection	1
asthma attack	1
brain tumour	1
cancer of the chest or abdomen	1
cancer of the stomach	1
carcinoma of nasal sinus	1
cervical cancer	1
complication from kidney transplant	1
complications following surgery	1
complications from brain surgery	1
complications of liver surgery	1
Creutzfeldt Jacob disease	1
embolism	1
glioma	1
hemorrhagic dementia	1
liver cancer	1
Lou Gehrig disease	1
lymphatic cancer	1
malignant melanoma	1
metastatic disease	1
ovarian cancer	1
pneumonia	1
postpolio complications	1
renal cancer	1
renal cell carcinoma	1
sarcoma of the lung	1
suddenly	1
T cell lymphoma	1
uterine cancer	1
Total	224

Example

September 4, 1998

[redacted], 51, An Expert on AIDS Vaccines

By WOLFGANG SAXON

Dr. [redacted], an epidemiologist and vaccine expert who headed the center for immunization research at the Johns Hopkins University School of Hygiene and Public Health in Baltimore, died with her husband in Wednesday night's plane crash off Nova Scotia.

Like her husband, Dr. [redacted], she was 51 and lived in Columbia, Md.

The couple were on their way to meetings of the World Health Organization in Geneva. Dr. [redacted] was to have attended a United Nations AIDS conference dealing with research into vaccines against H.I.V. She was a leading researcher on AIDS vaccines and was considered an expert on the assessment of vaccines.

September 8, 1998
VOL. 27, NO. 42

OBITUARY

[redacted] and [redacted] of Public Health, die in Swissair Jetliner Crash

The Gazette Online
THE NEWSPAPER OF THE JOHNS HOPKINS UNIVERSITY

[redacted], longtime head of the Division of Vaccine Sciences in the Department of International Health at the School of Public Health, and her husband, [redacted], a visiting professor at Hopkins and dean of the School of Public Health at Allegheny University of the Health Sciences, died Sept. 2 in the Swissair jetliner crash off the coast of Canada.

The couple, both leaders in the world's fight against AIDS, were en route to a U.N. AIDS vaccine conference in Geneva. Both were 51.

"[redacted] and [redacted] were distinguished and celebrated members of the faculty," said Alfred Sommer, dean of the School of Public Health. "Their loss will be felt deeply by all of us here and by the thousands, if not millions, of people who live better lives today because of their work."

J Infect Dis. 1998 May;177(5):1230-46.

Immune responses to human immunodeficiency virus (HIV) type 1 induced by canarypox expressing HIV-1MN gp120, HIV-1SF2 recombinant gp120, or both vaccines in seronegative adults. NIAID AIDS Vaccine Evaluation Group.

[redacted][1], Weinhold K, Matthews TJ, Graham BS, Gorse GJ, Keefer MC, McElrath MJ, Hsieh RH, Mestecky J, Zolla-Pazner S, Mascola J, Schwartz D, Siliciano R, Corey L, Wright PF, Belshe R, Dolin R, Jackson S, Xu S, Fast P, Walker MC, Stablein D, Excler JL, Tartaglia J, Paoletti E, et al.

Author information

[1]Johns Hopkins University School of Public Health, Baltimore, Maryland, USA. mclement@jhsph.edu

Abstract

A safety and immunogenicity trial was conducted in vaccinia-immune and vaccinia-naive human immunodeficiency virus (HIV)-uninfected adults who were randomized to receive 10(6) or 10(7) TCID50 of canarypox (ALVAC) vector expressing HIV-1MN gp160 or 10(5.5) TCID50 of ALVAC-rabies virus glycoprotein control at 0 and 1 or 2 months and ALVAC-gp160 or 50 microg of HIV-1SF2 recombinant (r) gp120 in microfluidized emulsion at 9 and 12 months; others received rgp120 at 0, 1, 6, and 12 months. All vaccines were well-tolerated. Neither vaccinia-immune status before vaccination nor ALVAC dose affected HIV immune responses. HIV-1MN and HIV-1SF2 neutralizing antibodies were detected more often (100%) in ALVAC-gp160/rgp120 recipients than in recipients of ALVAC-gp160 (<65%) or rgp120 (89%) alone. ALVAC-gp160/rgp120 also elicited more frequent HIV V3-specific and fusion-inhibition antibodies, antibody-dependent cellular cytotoxicity, lymphoproliferation, and cytotoxic CD8+ T cell activity than did either vaccine alone. Trials with ALVAC expressing additional HIV components and rgp120 are underway.

PMID: **9593008** [PubMed - indexed for MEDLINE] **Free full text**

PLoS Pathog. 2011 Feb 10;7(2):e1001273. doi: 10.1371/journal.ppat.1001273.

Relationship between functional profile of HIV-1 specific CD8 T cells and epitope variability with the selection of escape mutants in acute HIV-1 infection.

Ferrari G[1], Korber B, Goonetilleke N, Liu MK, Turnbull EL, Salazar-Gonzalez JF, Hawkins N, Self S, Watson S, Betts MR, Gay C, McGhee K, Pellegrino P, Williams I, Tomaras GD, Haynes BF, Gray CM, Borrow P, Roederer M, McMichael AJ, Weinhold KJ.

Author information

[1]Department of Surgery, Duke University Medical Center, Durham, North Carolina, United States of America. gflmp@duke.edu

Abstract

In the present study, we analyzed the functional profile of CD8+ T-cell responses directed against autologous transmitted/founder HIV-1 isolates during acute and early infection, and examined whether multifunctionality is required for selection of virus escape mutations. Seven anti-retroviral therapy-naïve subjects were studied in detail between 1 and 87 weeks following onset of symptoms of acute HIV-1 infection. Synthetic peptides representing the autologous transmitted/founder HIV-1 sequences were used in multiparameter flow cytometry assays to determine the functionality of HIV-1-specific CD8+ T memory cells. In all seven patients, the earliest T cell responses were predominantly oligofunctional, although the relative contribution of multifunctional cell responses increased significantly with time from infection. Interestingly, only the magnitude of the total and not of the poly-functional T-cell responses was significantly associated with the selection of escape mutants. However, the high contribution of MIP-1β-producing CD8+ T-cells to the total response suggests that mechanisms not limited to cytotoxicity could be exerting immune pressure during acute infection. Lastly, we show that epitope entropy, reflecting the capacity of the epitope to tolerate mutational change and defined as the diversity of epitope sequences at the population level, was also correlated with rate of emergence of escape mutants.

PMID: 21347345 [PubMed - indexed for MEDLINE] PMCID: PMC3037354 Free PMC Article

Additional Descriptives

Table B.5: Descriptives by decade

	1970	1980	1990	2000
A. Publications Level				
Number of publ.	5,424,571	9,213,862	15,902,121	25,377,269
Number of single authored publ.	820,376	801,511	1,223,776	1,563,419
Avg. number of authors per publ.	3,23	4.16	4.91	11.37
B. Author Level				
Number of authors	3,036,674	4,845,869	7,877,119	12,391,948
Mean publ. per year	1.79	1.90	2.02	2.05
Std. dev.	1.72	1.98	2.32	2.43
Max.	88	139	181	279
Mean number of coauthors per year	3.79	5.28	6.75	10.53
Std. dev.	3.57	5.32	7.34	31.66
Max.	119	289	504	1,863
C. Network Level				
Number of vertices	209,785	336,606	584,270	969,803
Number of edges	712,704	1,481,102	3,477,344	7,250,480
Diameter	55	45	40	44
Assortative coefficient	0.35	0.29	0.17	0.66
Number of clusters	34,265	40,293	43,419	63,236

Notes: All variables are constructed from the coauthorship network in 1970, 1980, 1990 and 2000. The unit of observation is the publication (panel A), the author (panel B) and the entire network (panel C). Definitions for the network level variables can be found in the appendix A.

Figure B.1: Total number of publications

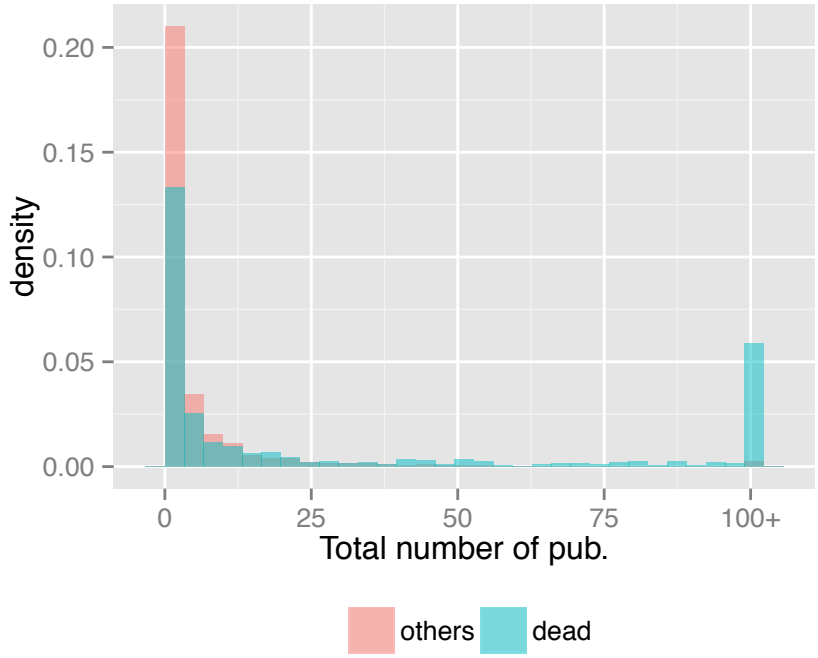


Figure B.2: Total number of coauthors

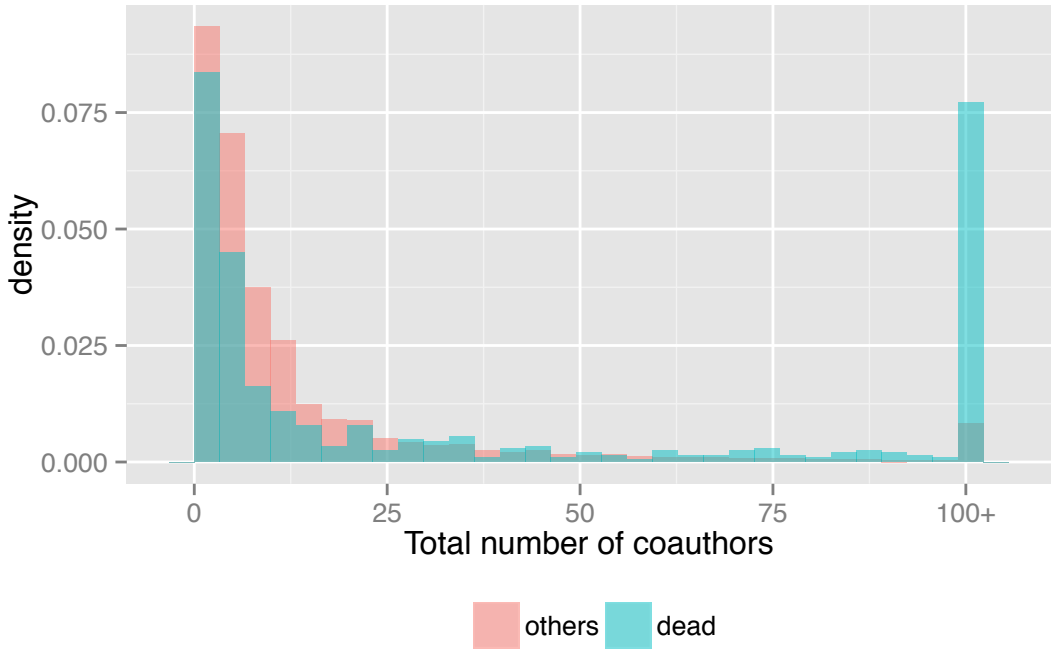


Table B.6: Descriptives - Sudden death

	Mean	Median	Std. dev.	Mean	Median	Std. dev.
A. Star Level	Deceased stars (112 Obs.)			Matched stars (569 Obs.)		
Age at death	56.44	57.00	6.44			
Cohort	1975.66	1973	9.30	1975.50	1973	9.46
Cum. # publ.	85.51	68.50	79.33	89.17	60.00	102.10
Cum. # JIF-weighted publ.	179.20	87.16	225.41	184.73	106.15	234.02
Cum. # citations received	4794.02	2214.00	8027.73	3583.83	1671.00	5281.60
Cum. # grants	21.01	12.00	22.60	13.71	3.00	21.06
Cum. # R01 grants	1.07	0.00	1.53	0.56	0.00	1.36
Cum. # coauthors	114.63	74.50	105.75	112.44	72.00	130.66
Closeness	2.21E-06	2.26E-06	7.37E-07	2.14E-06	2.24E-06	7.05E-07
Betweenness	6.25E05	1.31E05	1.62E06	6.59E05	1.25E05	1.71E06
Eigenvector	1.50E-02	4.68E-07	1.13E-01	3.08E-04	2.15E-06	1.52E-03
Clustering coefficient	0.15	0.11	0.16	0.14	0.10	0.15
Triangle	5.06E02	2.72E02	6.54E02	6.47E02	2.83E02	1.09E03
B. Coauthor Level	Treated coauthors (8,636 Obs.)			Matched coauthors (8,636 Obs.)		
Cohort	1984.79	1986	10.13	1984.54	1985.00	10.64
Cum. # publ.	18.44	7.00	30.09	19.35	7.00	34.78
Cum. # JIF-weighted publ.	31.43	11.15	58.36	30.87	9.37	62.92
Cum. # citations received	693.39	211.00	1615.93	675.42	174.00	1508.51
Cum. # MeSH codes	131.93	86.00	133.65	146.84	92.00	159.05
Cum. # coauthors	35.23	18.00	46.26	36.20	18.00	50.11
C. Coauthorship Level	Treated pairs (8,635 Obs.)			Control pairs (8,636 Obs.)		
Strength of coauthorship	2.34	1.00	3.86	2.30	1.00	5.74
Recency of coauthorship	8.03	6.00	7.15	8.20	5.00	7.42
# non-redundant nodes	27.31	11.00	43.90	28.38	10.00	47.35
Brokerage degree	0.18	0.06	0.26	0.17	0.05	0.26
Neighborhood overlap	5.92	5.00	4.99	5.95	5.00	5.67

Notes: The unit of observation is the star (panel A), the coauthor of a star (panel B) and the star-coauthor pair (panel C). The cohort is the year of first publication. All variables apart from the cohort are defined at the time of the death of the star or pseudo-death of the matched star. The R01 grants are NIH grants awarded to individual researchers. The strength of the coauthorship is the number of joint publications between the star and the coauthor, the recency of the coauthorship is the number of years since the last joint publication before the death. The brokerage degree is the fraction of non-redundant nodes offered by the star to his coauthor over all the links of the star as defined by equation 2.

Figure B.3: Star Matching – Cohort

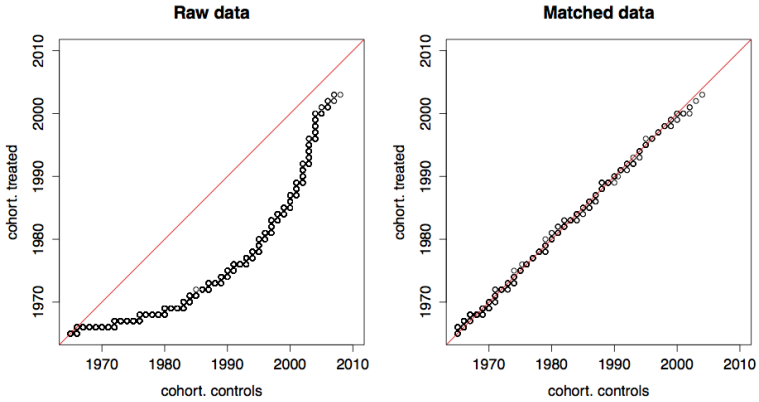


Figure B.4: Star Matching – Productivity

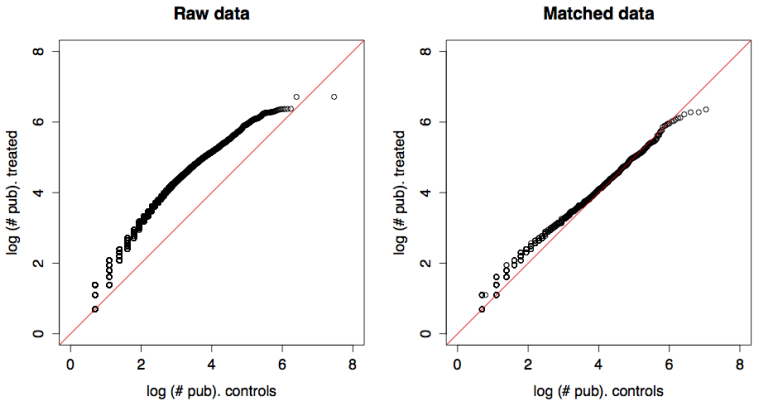


Figure B.5: Star Matching – Connectedness

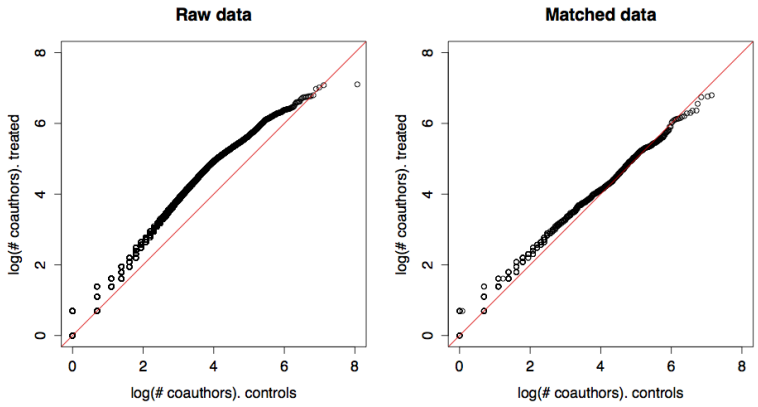


Figure B.6: Star matching – Grant

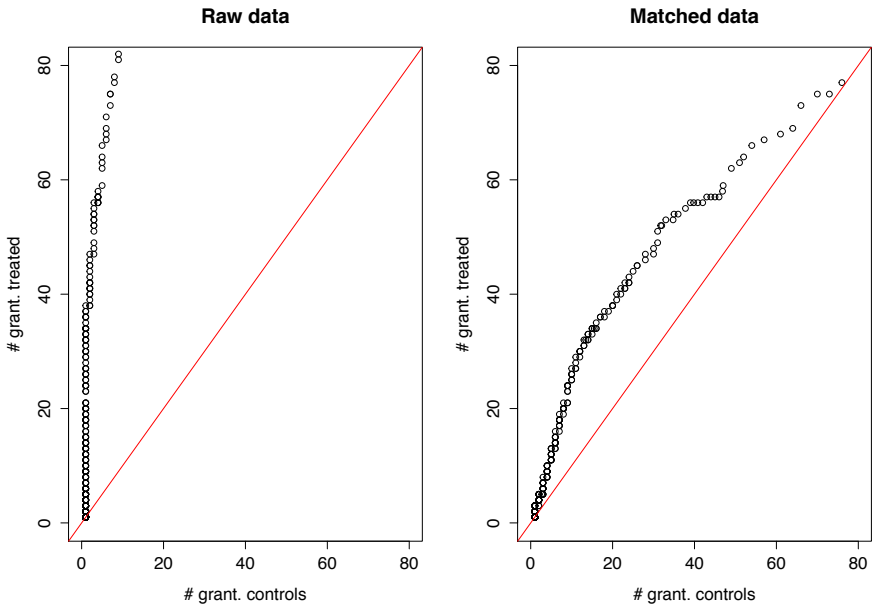


Figure B.7: Star matching – Closeness

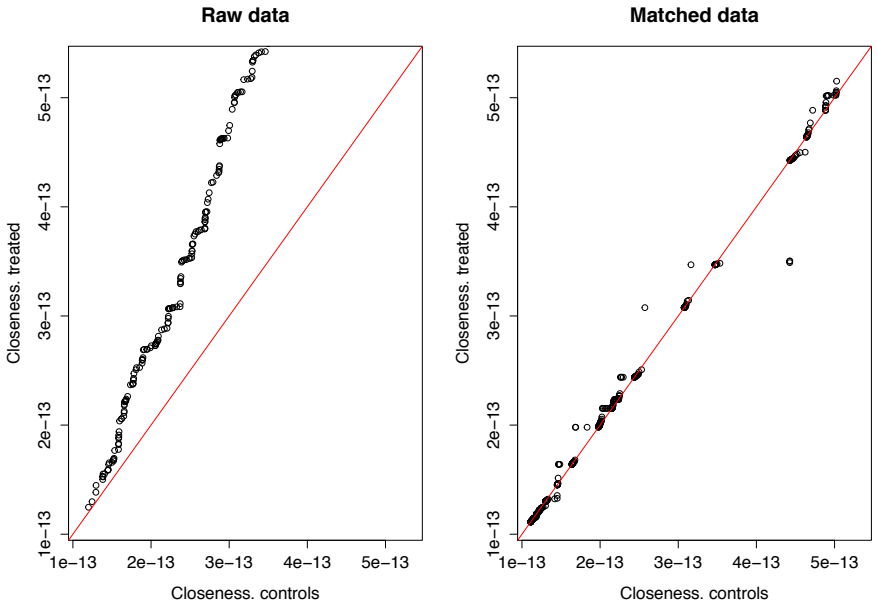


Figure B.8: Star matching – Betweenness

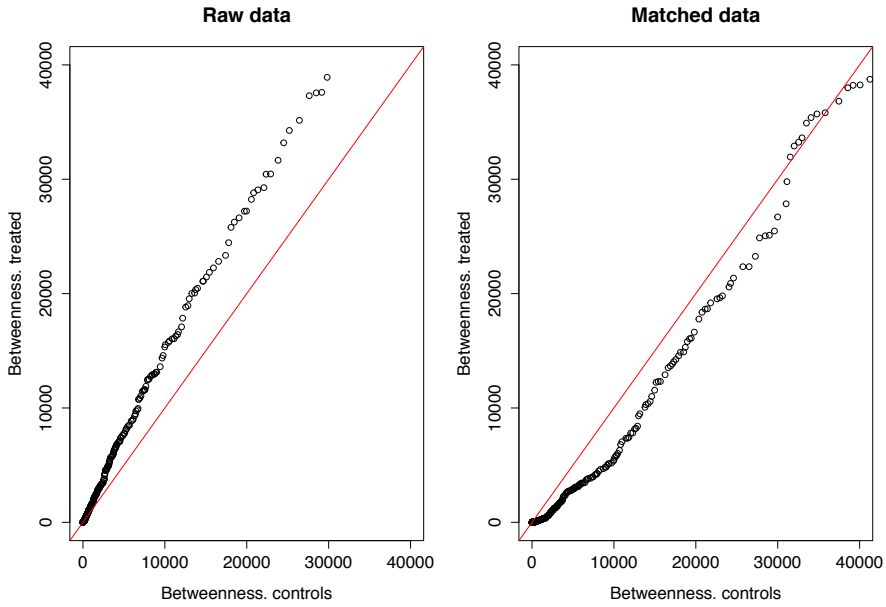


Figure B.9: Star matching – Eigenvector Centrality

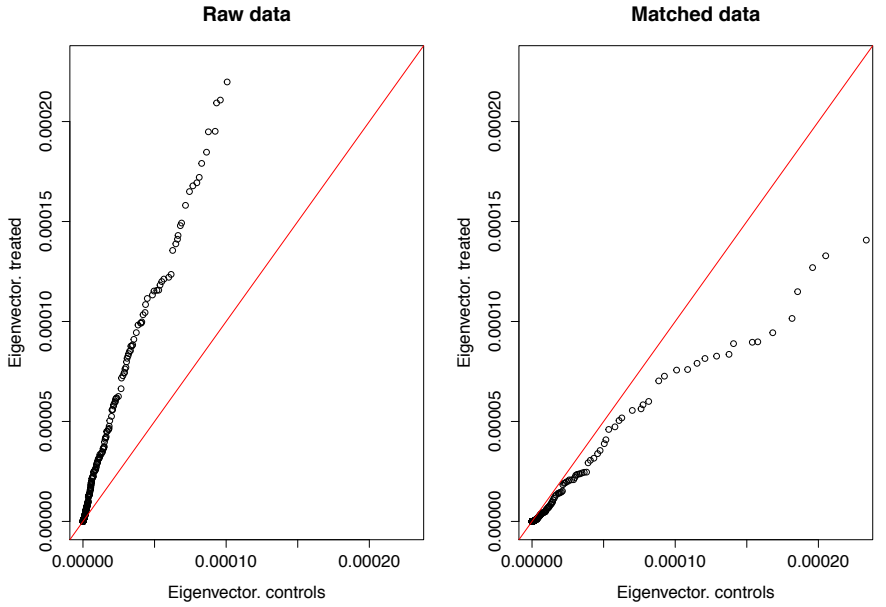


Figure B.10: Star matching – Clustering Coeff.

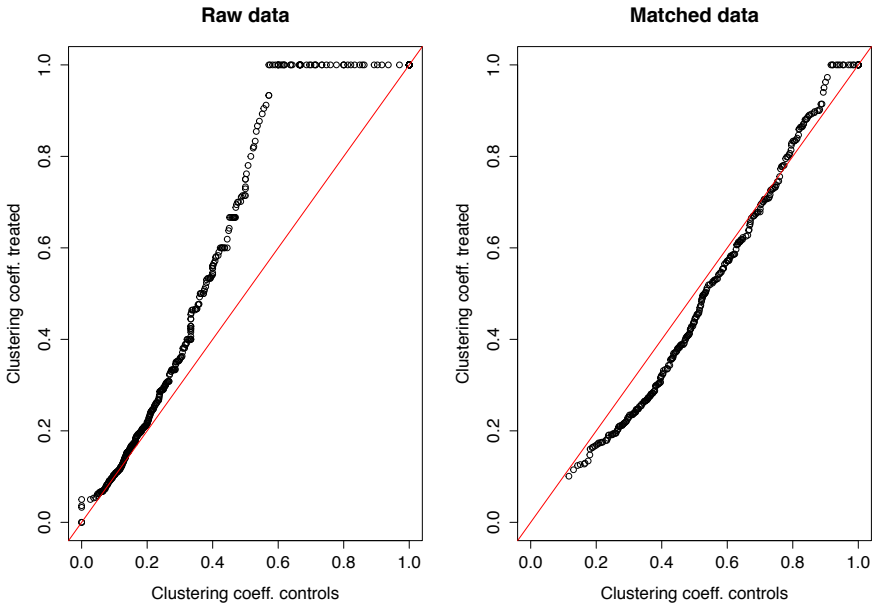


Figure B.11: Star matching – Triangle

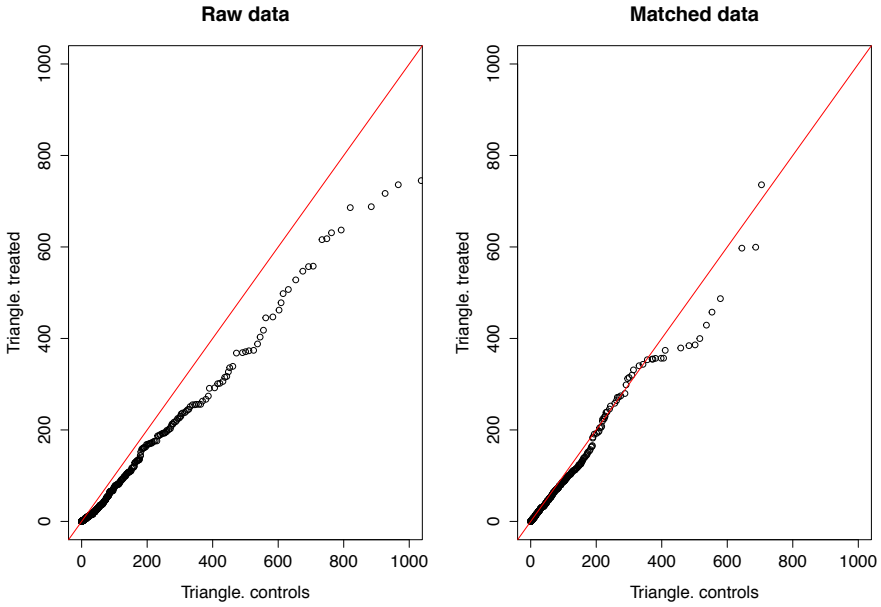


Figure B.12: Coauthor Matching – Cohort

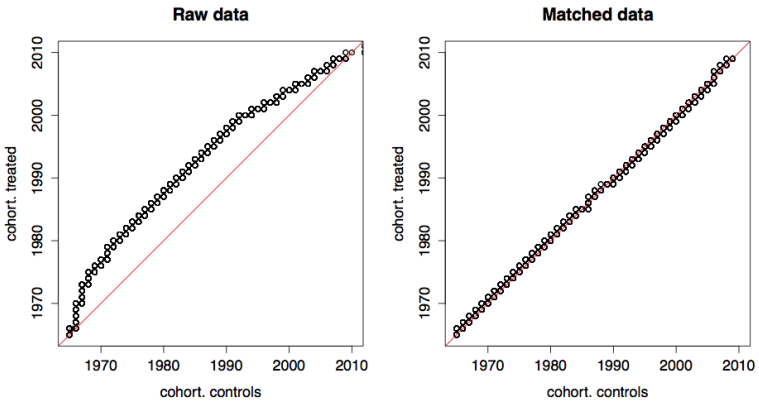


Figure B.13: Coauthor Matching – Productivity

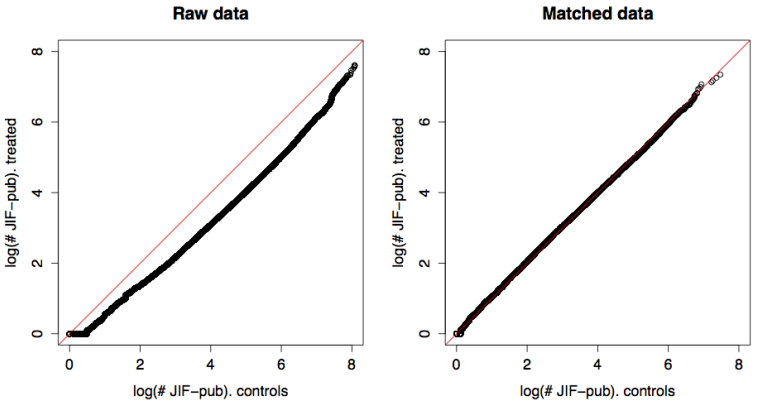


Figure B.14: Coauthor Matching – Connectedness

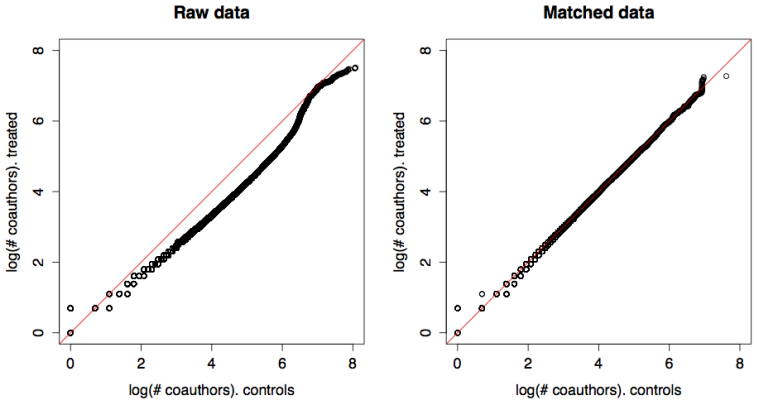


Figure B.15: Coauthorship Matching – Brokerage degree

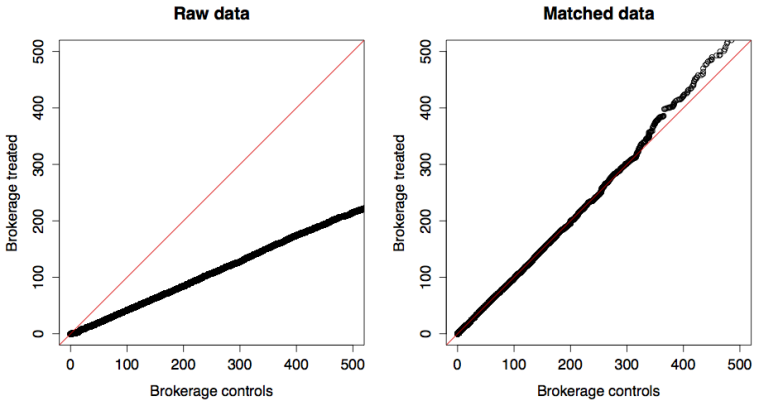


Figure B.16: Coauthorship Matching – Strength of collaboration

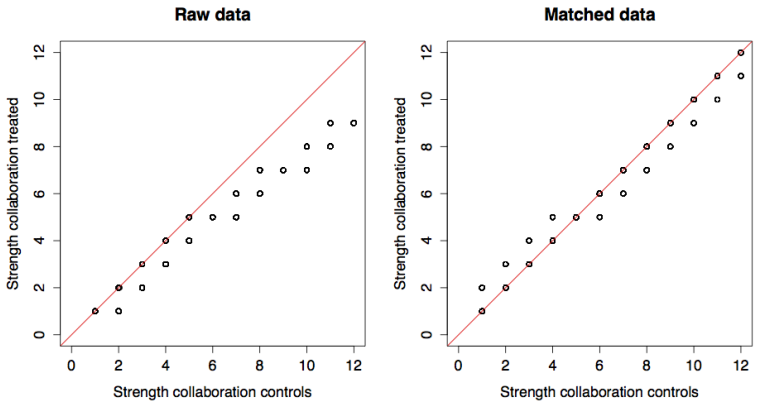


Figure B.17: Coauthorship Matching – Recency of collaboration

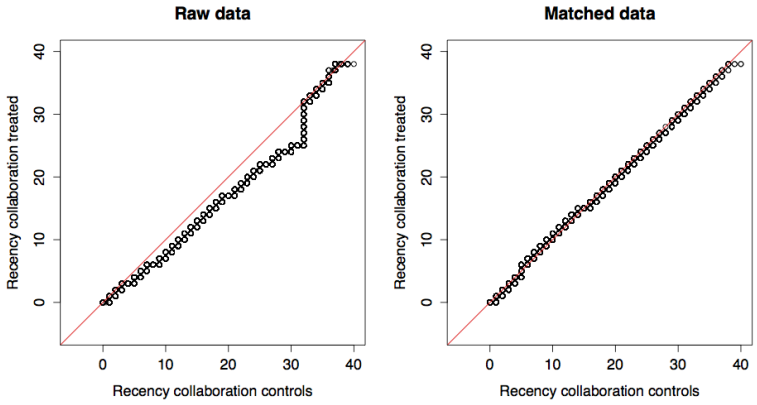
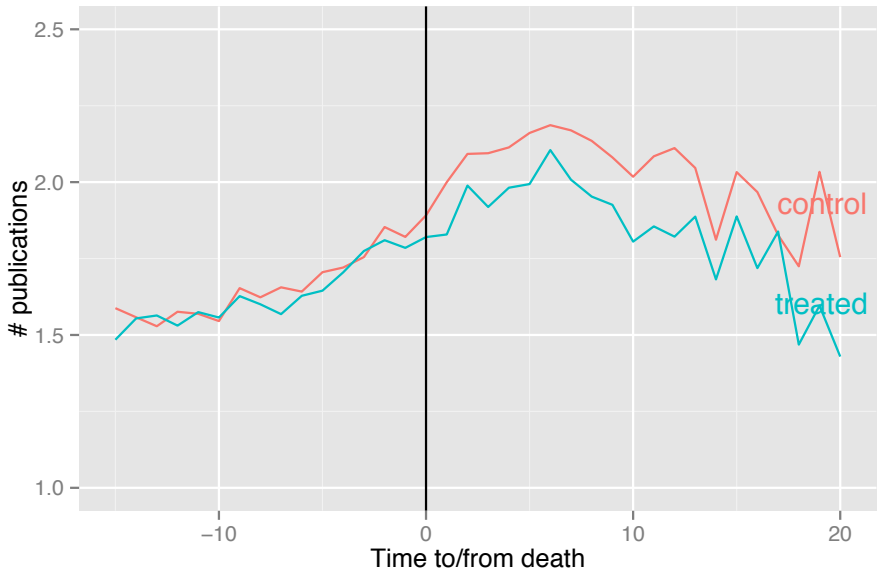


Figure B.18: Publication Trends for Treated and Control Coauthors



Notes: Mean number of publications around the time of the death. The solid red line corresponds to treated group (coauthors of star scientists who suddenly die) and the green line corresponds to the control group (matched coauthors of star scientists).

Figure B.19: Histogram: Brokerage Degree

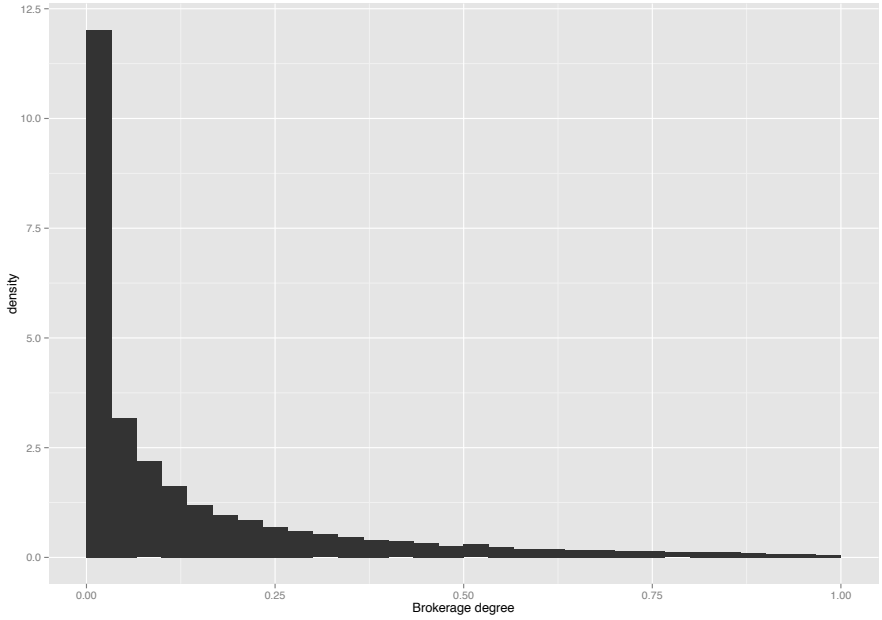


Figure B.20: Histogram: # non-redundant links

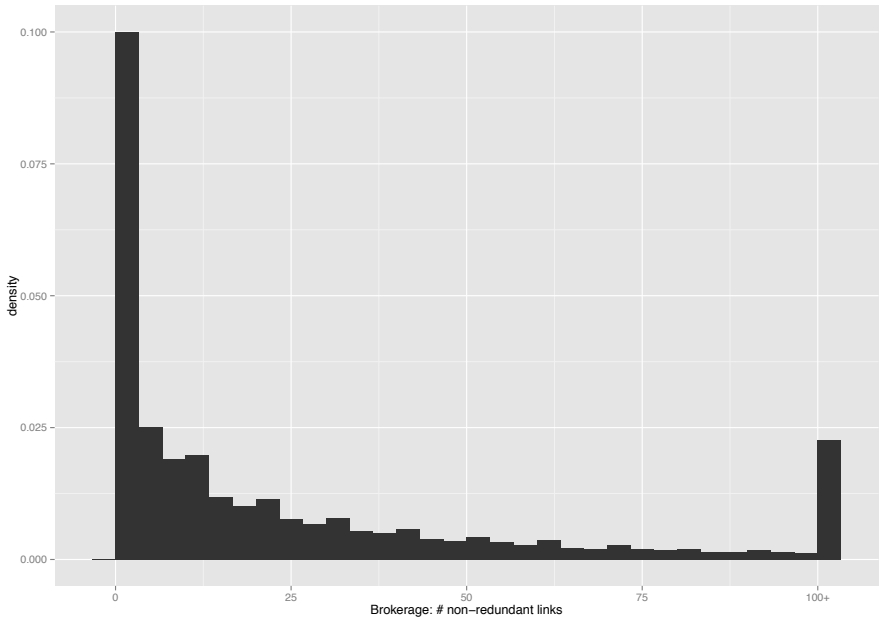


Table D3 Topics of deceased star scientists

Topics	Count
Chemicals and Drugs [D]	865
Diseases [C]	653
Organisms [B]	567
Anatomy [A]	462
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	368
Persons [M]	340
Biological Sciences [G]	245
Health Care [N]	108
Psychiatry and Psychology [F]	76
Geographic Locations [Z]	26
Anthropology, Education, Sociology and Social Phenomena [I]	19
Information Science [L]	8
Physical Sciences [H]	5
Technology and Food and Beverages [J]	3
Humanities [K]	1

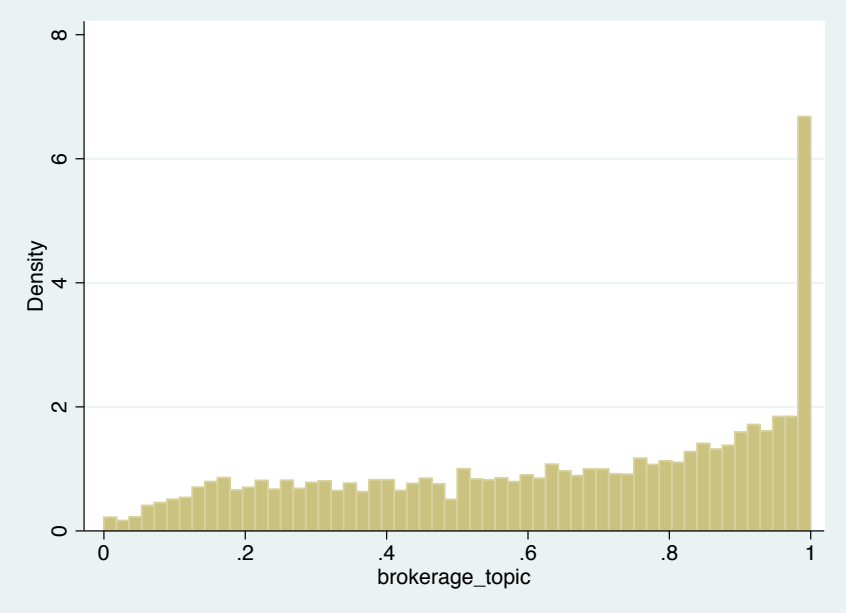
Notes: The speciality of each scientist is defined by the modal topic up to the time of death. In the case of multiple modal topics, all specialities are kept.

Table B.7: Correlation Matrix

Centrality	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Mean brokerage	1.00								
(2) # non–redundant nodes	0.33	1.00							
(3) Degree	-0.46	0.14	1.00						
(4) Embeddedness	-0.13	0.16	0.45	1.00					
(5) Closeness	-0.44	0.11	0.60	0.20	1.00				
(6) Betweenness	-0.14	0.02	0.33	0.09	0.30	1.00			
(7) Eigenvector	-0.04	0.03	0.22	0.59	0.08	0.10	1.00		
(8) Clustering coeff.	0.44	0.03	-0.41	-0.03	-0.51	-0.14	0.02	1.00	
(9) Triangle	-0.05	0.06	0.32	0.89	0.07	0.04	0.70	0.04	1.00

Notes: The centrality measures are based on the coauthorship network three years prior to the death.

Figure B.21: Histogram: brokerage degree in terms of topics



Robustness Checks

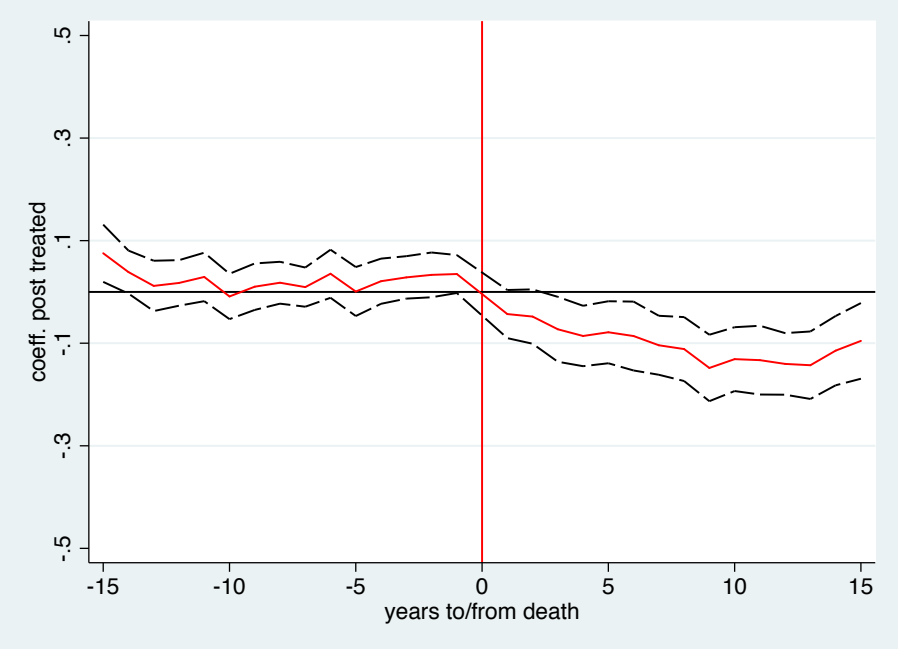


Figure B.22: Publication trends for anticipated deaths

Table B.8: Robustness Checks

Sample: All deaths Dep.Var.: Number of publications per year								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	By age at death			Assumption on career length		Balanced	Placebo	Cohort x
	< 40	< 50	< 60	40 yrs	35 yrs	Panel		Decade
post death	0.011	-0.163	0.009	0.005	-0.038	0.138***	-0.001	0.065
	(0.089)	(0.076)	(0.030)	(0.022)	(0.022)	(0.028)	(0.015)	(0.200)
post death$^*b_{ij,death}$	-0.315	-0.452***	-0.784***	-0.810***	-0.667	-0.721***		-0.890***
	(0.168)	(0.129)	(0.067)	(0.088)	(0.099)	(0.106)		(0.048)
R^2	0.078	0.079	0.071	0.043	0.038	0.060	0.048	0.076
Nb. of obs.	35,972	320,277	1,108,969	2,283,809	2,187,728	388,601	2,339,969	2,339,969
Nb. of deceased	38	153	408	79,918	79,918	10,751	39,776	79,918
Nb. of coauthors	1351	11,600	38,695	697	697	133	692	697

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications (winsorized at the 1% level). The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death.

Table B.9: Robustness Checks, continued

Sample: All deaths Dep.Var.: Number of publications per year							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Fuzzy String Matching		By data source				Anticipated
	Best	Others	Azoulay et al. (2010)		Other sources		deaths only
post death	0.23	0.155***	-0.054**	0.071***	-0.216***	0.071* 0.099***	
	(0.028)	(0.038)	(0.019)	(0.020)	(0.034)	(0.020)	
post death$^*b_{ij,death}$	-0.797***	-1.191***		-0.858***		-0.991***	-0.951***
	(0.069)	(0.160)		(0.117)		(0.097)	(0.078)
R^2	0.052	0.054	0.048	0.049	0.051	0.054	0.064
Nb. of obs.	1,610,110	729,859	954,295	954,295	1,385,674	1,385,674	946,069
Nb. of deceased	419	278	132	132	565	565	194
Nb. of coauthors	53,048	26,935	30,185	30,185	49,817	49,817	30,394

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. The dependent variables are the annual number of publications (winsorized at the 1% level). The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. All columns except column 1 use an OLS specification including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. Column 1 . Column 2 includes all stars and the matched star based on the of table. Column 3 includes all the other matches. Column 4 and 5 are based on the stars named in Azoulay et al. (2010) and column 6 and 7 are all the other used in this paper.

Table B.10: Results by topic

Sample: All deaths Dep.Var.: Number of publications per year								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Chemicals and Drugs [D]	Diseases [C]	Organisms [B]	Anatomy [A]	Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	Persons [M]	Biological Sciences [G]	Health Care [N]
post death	0.086***	0.054	0.086	0.037	0.117	0.090	-0.029	-0.105
	(0.026)	(0.048)	(0.026)	(0.037)	(0.062)	(0.056)	(0.114)	(0.230)
post death$*b_{ij,death}$	-1.057***	-0.933***	-0.727***	-0.745	-1.158***	-1.443***	-0.415	-1.251**
	(0.110)	(0.170)	(0.122)	(0.099)	(0.184)	(0.205)	(0.257)	(0.435)
R^2	0.052	0.053	0.388	0.270	0.151	0.061	0.059	0.097
Nb. of obs.	575,013	293,693	407,118	239,103	218,530	299,022	151,448	59,329
Nb. of deceased	80	57	65	53	45	51	39	16
Nb. of coauthors	4,879	1,355	2,925	1,602	1,357	2,152	1,741	709

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the number of publications. The sample includes the coauthors of star scientists (regardless of their cause of death) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death. Each column represents the speciality of the star scientist defined by the modal topic of publications prior to the year of death (pseudo-death for the matched stars).

Table B.11: Results over time

Sample: All deaths Dep.Var.: Number of publications per year						
	(1)	(2)	(3)	(4)	(5)	(6)
	pre-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004
post death	-0.242	0.270**	0.199*	0.234**	0.013	0.045
	(0.277)	(0.095)	(0.076)	(0.074)	(0.053)	(0.041)
*post death*b*$_{ij,death}$	-0.620	-0.534	-1.260***	-1.112*	-0.462*	-0.585***
	(0.489)	(0.452)	(0.307)	(0.466)	(0.228)	(0.128)
R^2	0.093	0.068	0.066	0.071	0.064	0.075
Nb. of obs.	1,974	11,098	50,371	58,555	139,126	171,931
Nb. of deceased	2	5	11	14	28	33
Nb. of coauthors	48	270	1,348	1,681	4,547	6,323

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications including a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications (winsorized at the 1% level). The sample in each column includes the coauthors of stars scientists who died prior to 1980 (column 1), between 1980-1984 (column 2), between 1985-1989 (column 3), between 1990-1994 (column 4), between 1995-1999 (column 5) and between 2000-2004 (column 6) along with their appropriate control group. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death.

Table B.12: Alternative measure of brokerage

Sample: Sudden deaths						
	(1)	(2)	(3)	(4)	(5)	(6)
	# publ. per year		# JIF-publ. per year		# citations-publ. per year	
post death	-0.136***	0.116**	-0.271***	0.471***	-13.221***	3.982
	(0.023)	(0.040)	(0.065)	(0.115)	(2.242)	(3.260)
post death#non − redundant*		-0.007***		-0.012***		-0.279***
nodes$_{ij,death}$		(0.001)		(0.002)		(0.060)
*post death*degree*$_{j,death}$		-0.0001		-0.002**		-0.043**
		(0.0001)		(0.001)		(0.016)
R^2	0.066	0.071	0.036	0.040	0.031	0.033
Nb. of obs.	503,748	503,748	503,748	503,748	503,748	503,748
Nb. of coauthors	17,272	17,272	17,272	17,272	17,272	17,272

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications includes a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications (winsorized at the 1% level) in columns 1 and 2, the annual number of publication weighted by their journal impact factors in columns 3 and 4 and the number of publications weighted by the number of citations received in columns 5 and 6. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors.

Additional Results

Table B.13: Treated only

Sample: Sudden deaths						
	(1)	(2)	(3)	(4)	(5)	(6)
	# publ. per year		# JIF-publ. per year		# citations-publ. per year	
post death	-0.229***	-0.054	-0.359***	-0.109	-17.582***	-12.240***
	(0.035)	(0.036)	(0.061)	(0.074)	(2.941)	(2.997)
post death$^{*}b_{ij,death}$		-0.912***		-1.296***		-27.728**
		(0.116)		(0.237)		(8.252)
R^2	0.069	0.075	0.038	0.041	0.032	0.033
Nb. of obs.	251,283	251,283	251,283	251,283	251,283	251,283
Nb. of coauthors	8,636	8,636	8,636	8,636	8,636	8,636

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. OLS specifications includes a full set of dyad fixed effect, age and year dummies. Robust standard errors, clustered at the level of the star, are reported in the parentheses. The dependent variables are the annual number of publications in columns 1 and 2, the annual number of publication weighted by their journal impact factors in columns 3 and 4 and the annual number of publications weighted by the citations received in columns 5 and 6. The sample includes the coauthors of star scientists who died suddenly (see table C2 in the appendix for a list of sudden causes of deaths) along with their appropriate control coauthors. The variable post death is a dummy equal to one after the death of a star for all his coauthors. The variable $b_{ij,death}$ is the fraction of non-redundant nodes offered by the star "j" to his coauthor "i" at the time of his death.

Figure B.23: Dynamics of the Treatment Effect for Young Scientists

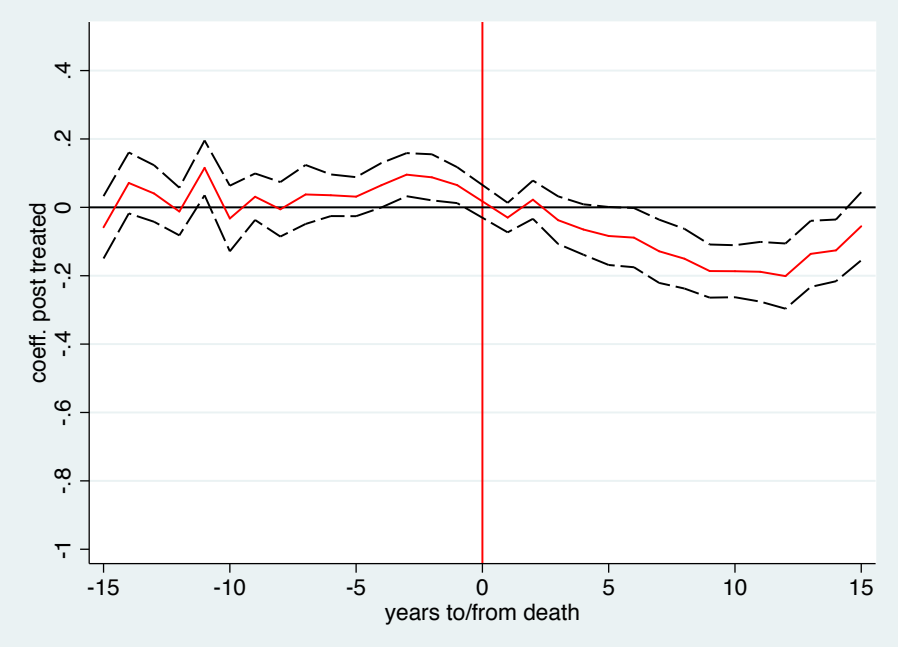


Figure B.24: Dynamics of the Treatment Effect for Experienced Scientists

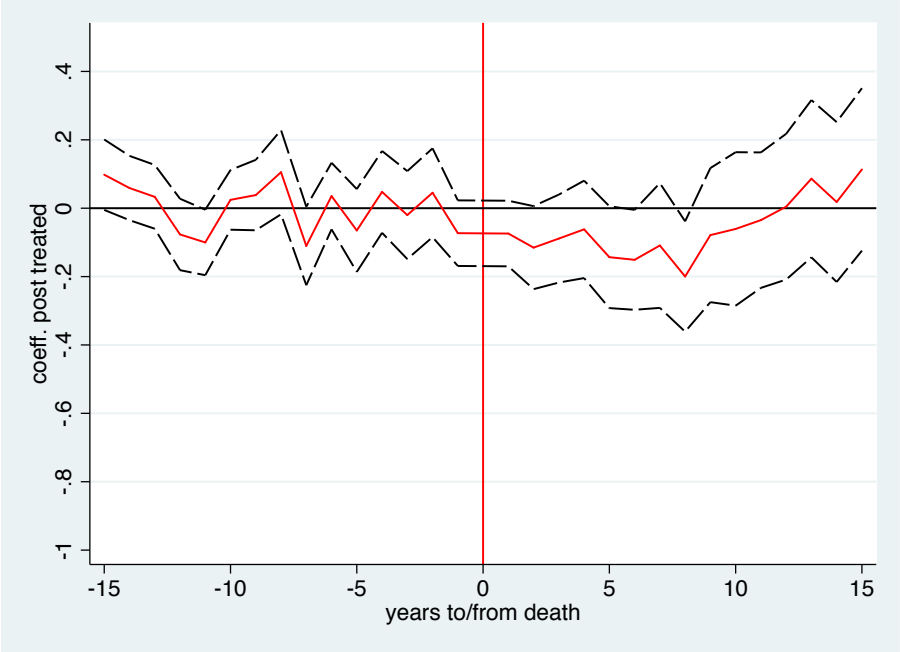


Figure B.25: Dynamics on the Publications as First Author

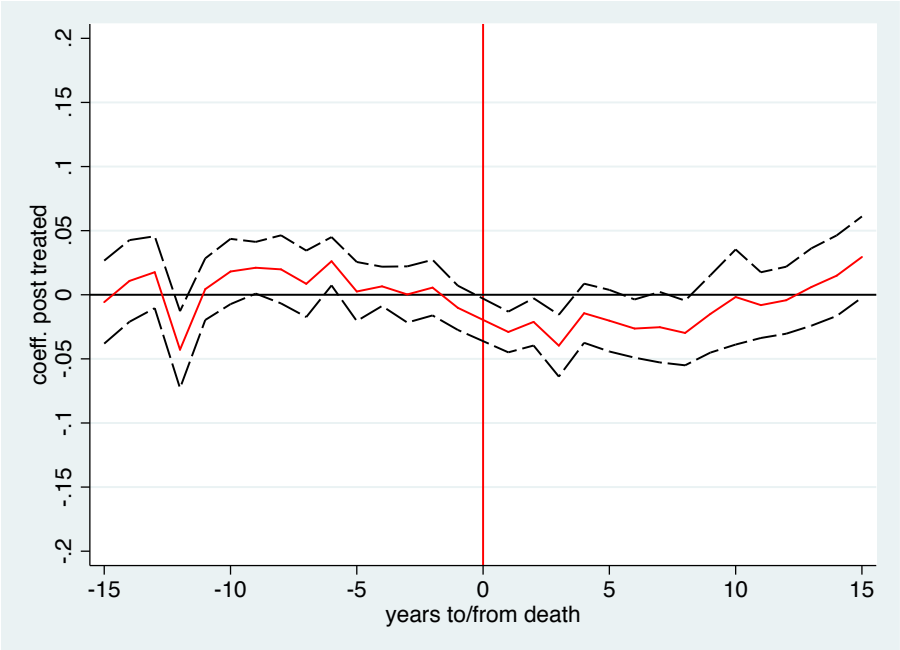
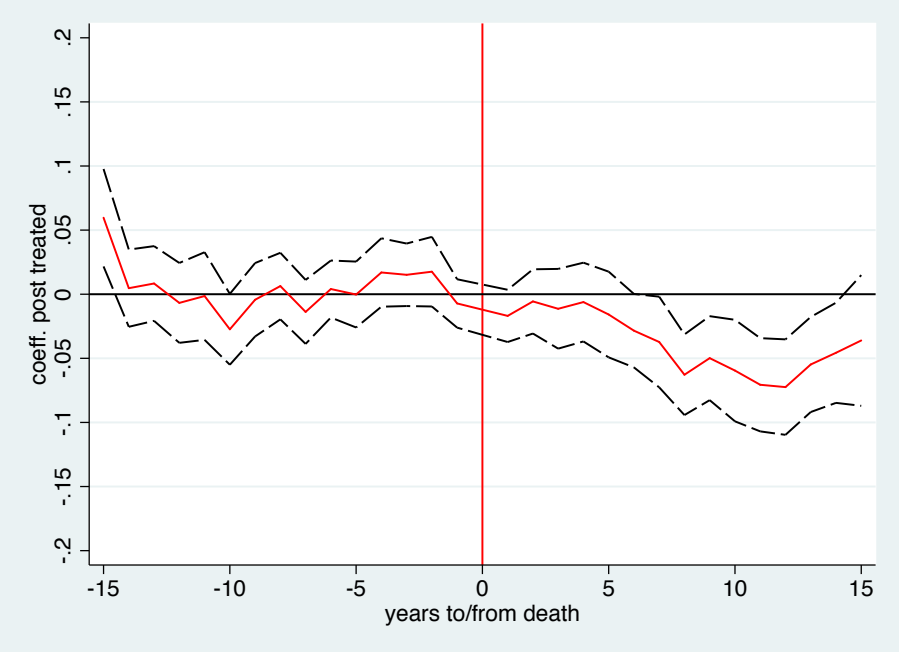


Figure B.26: Dynamics on the Publications as Last Author



Appendix C

Appendix to Chapter 4

Classification

Table C.1: Patent classification codes - Transport

	CLEAN
B60K 1	Arrangement or mounting of electrical propulsion units
B60K 6	Arrangement or mounting of hybrid propulsion systems comprising electric motors and internal combustion
B60L 3	Electric devices on electrically-propelled vehicles for safety purposes: Monitoring operating variables e.g. speed, deceleration, power consumption
B60L 7	Dynamic electric regenerative braking
B60L 11	Electric propulsion with power supplied within the vehicle
B60L 15	Methods, circuits, or devices for controlling the traction-motor speed of electrically-propelled vehicles
B60R 16	Electric or fluid circuits specially adapted for vehicles and not otherwise provided for
B60S 5	Supplying batteries to, or removing batteries form
B60W 10	Conjoint control of vehicles sub-units of different type or different function
B60W 20	Control systems specially adapted for hybrid vehicles
H01M	Fuel cells
	GREY
F02M 39/71	Fuel injection apparatus
F02M 3/02-05	Idling devices for carburettors preventing flow of idling fue
F02M 23	Apparatus for adding secondary air to fuel-air mixture
F02M 25	Engine-pertinent apparatus for adding non-fuel substances or small quantities of secondary fuel to combustion-air, main fuel, or fuel-air mixture
F02D 41	Electric control of supply of combustion mixture or its constituents
F02B 47/06	Methods of operating engines involving adding non-fuel substances or anti-knock agents to combustion air, fuel, or fuel-air mixtures of engines, the substances including non-airborne oxygen
	DIRTY
F02B	Internal-combustion piston engines; combustion engines in genera
F02D	Controlling combustion engines
F02F	Cylinders, pistons, or casings for combustion engines; arragement of sealings in combustion engines
F02M	Supplying combustion engines with combustiles mixtures or constituents thereof
F02N	Starting of combustion engines
F02P	Ignition (other than compression ignition) for internal-combustion engines

Table C.2: Patent classification codes - Electricity Production

	CLEAN
Y02E10	Energy generation through renewable energy sources
Y02E30	Energy generation of nuclear origin
E02B9/08	Tide or wave power plants
F03B13/10-26	Submerged units incorporating electric generators or motors characterized by using wave or tide energy
F03D	Wind motors
F03G4	Devices for producing mechanical power from geothermal energy
F03G6	Devices for producing mechanical power from solar energy
F03G7/05	Ocean thermal energy conversion
F24J2	Use of solar heat, e.g. solar heat collectors
F24J3/08	Production or use of heat, not derived from combustion using geothermal heat
F26B3/28	Drying solid materials or objects by processes involving the application of heat by radiation, e.g. from the sun
	GREY
Y02E50	Technologies for the production of fuel of non-fossil origin
Y02E20/10	Combined combustion
Y02E20/12	Heat utilisation in combustion or incineration of waste
Y02E20/14	Combined heat and power generation
Y02E20/16	Combined cycle power plant, or combined cycle gas turbine
Y02E20/18	Integrated gasification combined cycle
Y02E20/30	Technologies for a more efficient combustion or heat usage
Y02E20/32	Direct C02 mitigation
Y02E20/34	Indirect C02 mitigation, by acting on non C02 directly related matters of the process, more efficient use of fuels
Y02E20/36	Heat recovery other than air pre-heating
	DIRTY
C10G1	Production of liquid hydrocarbon mixtures from oil-shale, oil-sand, or non-melting solid carbonaceous or similar materials, e.g. wood, coal, oil-sand, or the like B03B
C10L1	Fuel
C10J	Production of fuel gases by carburetting air or other gases
E02B	Hydraulic engineering
F01K	Steam engine plans; steam accumulators; engine plants not otherwise provided for; engines using special working fluids or cycles
F02C	Gas-turbine plants; air intakes for jet-propulsion plants; controlling fuel supply in air-breathing jet-propulsion plants
F22	Steam generation
F23	Combustion apparatus; combustion processes
F24J	Production or use of heat not otherwise provided for
F27	Furnaces; kilns; ovens; retorts
F28	Heat exchange in general

Table C.3: Patent classification codes - Radically New Technologies

	3D
H04N 13	Stereoscopic television systems
	IT
G06	Computing; Calculating; Counting
G10L	Speech Analysis or Synthesis; Speech Recognition; Speech or Voice Processing; Speech or Audio Coding or Decoding
G11C	Static Stores
(not G06Q)	Data Processing Systems or Methods; Specially Adapted for Administrative, Commercial, Financial, Managerial, Supervisory or Forecasting purposes; Systems or Methods Specially Adapted for Administrative, Commercial, Financial, Managerial, Supervisory or Forecasting purposes, not otherwise provided for
	Biotechs
C07G	Compounds of unknown constitution
C07K	Peptides
C12M	Apparatus for Enzymology or Microbiology
C12N	Micro-organisms or enzymes; compositions thereof
C12P	Fermentation or Enzyme-using Processes to Synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture
C12Q	Measuring or Testing Processes Involving Enzymes or Micro-Organisms; Compositions or test papers therefor; Processes of preparing such compositions; Condition responsive control in microbiological or enzymological processes
C12R	Processes using micro-organisms
(not A61K)	Preparations for Medical, Dental, or Toilet Purposes
	Nano
B82	Nano-technology
	Robot
B25J 9	Programme-controlled manipulators

Patent Examples

Figure C.1: Patent example US6026921A

US006026921A

United States Patent [19]

Aoyama et al.

[11] **Patent Number: 6,026,921**

[45] **Date of Patent: Feb. 22, 2000**

[54] **HYBRID VEHICLE EMPLOYING PARALLEL HYBRID SYSTEM, USING BOTH INTERNAL COMBUSTION ENGINE AND ELECTRIC MOTOR FOR PROPULSION**

[75] Inventors: **Shunichi Aoyama**, Kanagawa; **Shinichiro Kitada**, Tokyo; **Noboru Hattori**, Kanagawa; **Isaya Matsuo**, Tokyo, all of Japan

[73] Assignee: **Nissan Motor Co., Ltd**, Yokohama, Japan

[21] Appl. No.: **09/271,326**

[22] Filed: **Mar. 18, 1999**

[30] **Foreign Application Priority Data**

Mar. 20, 1998 [JP] Japan 10-072388

[51] **Int. Cl.**[7] **B60K 6/02**

[52] **U.S. Cl.** **180/65.2**; 123/348; 180/65.4; 180/65.8; 701/22

[58] **Field of Search** 701/22, 105, 113; 318/8; 123/347, 348; 180/65.1, 65.2, 65.3, 65.4, 65.8

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,335,429	6/1982	Kawakatsu	701/22 X
5,624,347	4/1997	Yang	180/65.2 X
5,697,466	12/1997	Moroto et al.	701/22 X
5,823,280	10/1998	Lateur et al.	180/65.2
5,823,282	10/1998	Yamaguchi	180/65.2
5,826,671	10/1998	Nakae et al.	180/65.2
5,845,731	12/1998	Buglione et al.	180/65.2
5,865,263	2/1999	Yamaguchi et al.	180/65.2

FOREIGN PATENT DOCUMENTS

5-50865 3/1993 Japan .

Primary Examiner—Michael Mar
Attorney, Agent, or Firm—Foley & Lardner

[57] **ABSTRACT**

A hybrid vehicle employing a parallel hybrid system, using both an internal combustion engine and an electric motor for propulsion, comprises a propulsion mechanism having at least a first electric motor, an internal combustion engine, a clutch, a second electric motor, a transmission, and drive wheels, and constructed so that a crankshaft of the engine and a motor shaft of the first electric motor are connected to an input shaft of the clutch, and that a motor shaft of the second electric motor and an input shaft of the transmission are connected to an output shaft of the clutch to transmit a driving torque from the output shaft of the transmission to the drive wheels. An engine temperature sensor is provided for detecting an engine temperature. An electronic control unit is connected to the propulsion mechanism and the engine temperature sensor, for disengaging the clutch and stopping the engine during a standstill state of the hybrid vehicle, and for propelling the hybrid vehicle by a driving torque generated by the second electric motor when the hybrid vehicle starts, and for starting up the engine by the first electric motor and engaging the clutch after engine startup, during the warm-up of the engine. A variable valve timing control system is electrically connected to the control unit for adjusting an intake-valve closure timing to a predetermined retarded closure timing when the clutch is disengaged and for adjusting the intake-valve closure timing to a predetermined advanced closure timing after the clutch is engaged, after the engine warm-up.

6 Claims, 12 Drawing Sheets

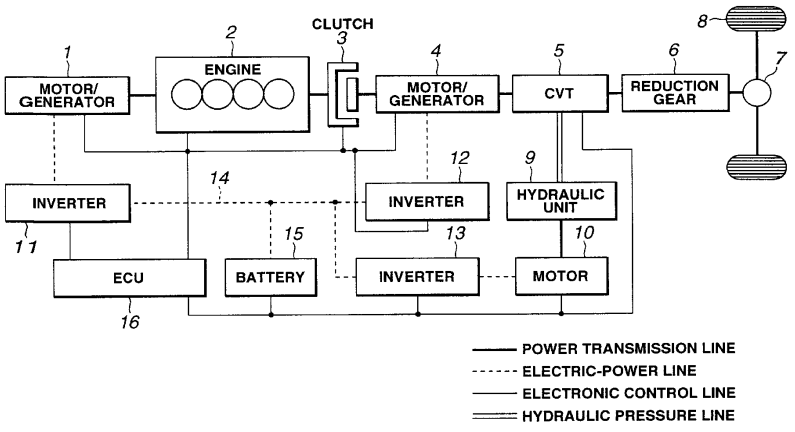


Figure C.2: Patent example US6727670B1

US006727670B1

(12) **United States Patent**
Grabowski et al.

(10) **Patent No.: US 6,727,670 B1**
(45) **Date of Patent: Apr. 27, 2004**

(54) **BATTERY CURRENT LIMITER FOR A HIGH VOLTAGE BATTERY PACK IN A HYBRID ELECTRIC VEHICLE POWERTRAIN**

(75) Inventors: **John Robert Grabowski**, Dearborn, MI (US); **Michael W. Degner**, Novi, MI (US)

(73) Assignee: **Ford Global Technologies, LLC**, Dearborn, MI (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/248,035**

(22) Filed: **Dec. 12, 2002**

(51) **Int. Cl.**[7] **H02P 7/00**

(52) **U.S. Cl.** **318/432**; 318/139; 701/22; 320/121

(58) **Field of Search** 318/432, 139; 701/22; 320/121

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,569,999	A	*	10/1996	Boll et al. 320/136
5,635,805	A		6/1997	Ibaraki et al. 318/139
5,697,466	A		12/1997	Moroto et al. 180/65.2
5,713,814	A		2/1998	Hara et al. 477/5
5,820,172	A		10/1998	Brigham et al. 290/40
5,842,534	A		12/1998	Frank 180/65.2
5,998,952	A	*	12/1999	McLaughlin et al. 318/432
6,026,921	A		2/2000	Aoyama et al. 180/65.2
6,164,400	A		12/2000	Jankovic et al. 180/65.2
6,176,808	B1		1/2001	Brown et al. 477/5
6,215,198	B1		4/2001	Inada et al. 290/40
6,223,106	B1		4/2001	Yano et al. 701/22
6,232,744	B1	*	5/2001	Kawai et al. 320/132
6,232,748	B1		5/2001	Kinoshita 320/132
6,253,140	B1		6/2001	Jain et al. 701/67
6,393,350	B1	*	5/2002	Light et al. 701/54
2003/0088343	A1	*	5/2003	Ochiai et al. 701/22

* cited by examiner

Primary Examiner—Rita Leykin
(74) *Attorney, Agent, or Firm*—Brooks & Kushman; Carlos Hanze

(57) **ABSTRACT**

A battery current limiter and current-limiting method for a battery system and an electric motor in a hybrid automotive vehicle powertrain. The battery current limiter monitors measured battery current and torque commands. A modified current is developed to take a predetermined current margin into account. The modified current reduces battery current in a closed loop fashion simultaneously with a reduction in commanded torque by a feed-forward torque value.

6 Claims, 3 Drawing Sheets

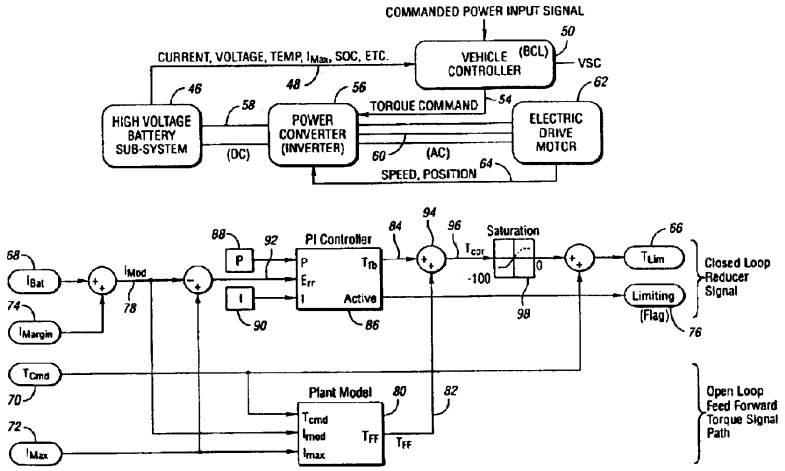


Figure C.3: Patent example US8036340B2

(12) **United States Patent**
Soto Santos

(10) **Patent No.: US 8,036,340 B2**
(45) **Date of Patent: Oct. 11, 2011**

(54) **X-RAY APPARATUS**

(75) Inventor: **Jose-Emilio Soto Santos**, Paris (FR)

(73) Assignee: **General Electric Company**, Schenectady, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 215 days.

(21) Appl. No.: **12/171,314**

(22) Filed: **Jul. 11, 2008**

(65) **Prior Publication Data**

US 2009/0034686 A1 Feb. 5, 2009

(30) **Foreign Application Priority Data**

Jul. 19, 2007 (FR) 07 56591

(51) **Int. Cl.**
H05G 1/34 (2006.01)

(52) **U.S. Cl.** **378/109**; 378/101; 378/102; 378/103; 378/106; 378/110; 378/111; 378/112

(58) **Field of Classification Search** 378/106, 378/101–103, 109–112
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,477,761	A	10/1984	Wolf	
5,283,512	A	2/1994	Stadnick et al.	
6,075,331	A	6/2000	Ando et al.	
6,282,260	B1 *	8/2001	Grodzins	378/87
6,727,670	B1	4/2004	Grabowski et al.	
2003/0107352	A1 *	6/2003	Downer et al.	322/40

FOREIGN PATENT DOCUMENTS

DE	10 2005 052 115	A1	5/2007
EP	0 946 082	A1	9/1999
JP	61-267300		11/1986

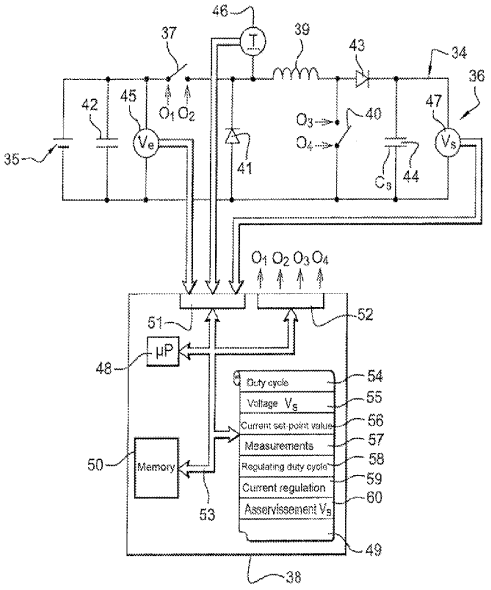
* cited by examiner

Primary Examiner — Hoon Song
Assistant Examiner — Mona M Sanei
(74) *Attorney, Agent, or Firm* — Global Patent Operation; Jonathan E. Thomas

(57) **ABSTRACT**

An X-ray apparatus includes a converter into which there is integrated a control logic circuit configured to regulate the supply voltage of a high-voltage power supply source of the X-ray apparatus. To this end, the intelligent voltage-voltage, converter is placed between the power battery and the capacitor bank. This intelligent converter is capable of determining the optimum voltage to be delivered to the generator for the radiology examination to be undertaken in regulating the current of the power battery at the necessary level of current.

9 Claims, 3 Drawing Sheets



PatentRank Results

Table C.4: Within vs. across-country spillovers

	(1)	(2)	(3)
Dep. var.	PatentRank	PatentRank for "national" citations	PatentRank for "international" citations
Clean invention	0.292***	0.285***	0.361***
	(0.014)	(0.017)	(0.013)
Number of patents	-0.031***	-0.035***	-0.042***
	(0.005)	(0.006)	(0.005)
Family size	0.067***	0.062***	0.073***
	(0.003)	(0.003)	(0.003)
Triadic	0.241***	0.240**	0.331***
	(0.026)	(0.020)	(0.033)
Granted	0.491***	0.435***	0.731***
	(0.021)	(0.016)	(0.028)
Obs.	1,149,988	1,149,988	1,149,988

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variables are the PatentRank index (column 1), the PatentRank index on the pool of national citations (column 2), and the PatentRank index on the pool of international citations (column 3). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.5: Government spending

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All		Transport		Electricity	
Dep. var.	PatentRank index					
Clean invention	0.345***	0.353***	0.153**	0.149**	0.339**	0.347**
	(0.028)	(0.026)	(0.052)	(0.052)	(0.028)	(0.026)
Government spending		0.022***		-0.020		0.021**
		(0.007)		(0.155)		(0.006)
Number of patents	0.012	0.013	-0.040**	-0.040**	0.013	0.014
	(0.008)	(0.008)	(0.015)	(0.015)	(0.008)	(0.008)
Family size	0.060***	0.059***	0.057***	0.057***	0.059***	0.059***
	(0.004)	(0.004)	(0.015)	(0.015)	(0.004)	(0.004)
Triadic	0.285***	0.284***	0.391***	0.394***	0.274***	0.273***
	(0.037)	(0.037)	(0.076)	(0.076)	(0.037)	(0.037)
Granted	0.360***	0.360***	0.534***	0.535***	0.359***	0.358***
	(0.017)	(0.017)	(0.032)	(0.032)	(0.017)	(0.017)
Obs.	497,439	497,439	16,719	16,719	489,531	489,531

Source: International Energy Agency (2013): Energy Technology Research and Development Database (Edition: 2013). Mimas, University of Manchester

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index. The sample includes clean and dirty inventions from the transport sector (columns 3 and 4), electricity sector (columns 5 and 6) and both transport and electricity sectors (columns 1 and 2) . All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.6: University and Firms

	(1)	(2)
Dep. var.	PatentRank index	
Clean invention	0.293***	0.298***
	(0.013)	(0.013)
Number of patents	-0.019***	-0.022***
	(0.004)	(0.004)
Family size	0.063***	0.060***
	(0.003)	(0.003)
Triadic	0.237***	0.229***
	(0.024)	(0.024)
Granted	0.561***	0.552***
	(0.021)	(0.021)
University		0.276***
		(0.014)
Firms		0.206***
		(0.011)
Obs.	826,078	826,078

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors (columns 1 and 2) and the PatentRank index (columns 3 and 4). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.7: Adding inventor and inventor fixed effect

	(1)	(2)	(3)	(4)
Dep. var.	PatentRank index			
Clean invention	0.216***	0.259***	0.274***	0.272***
	(0.004)	(0.006)	(0.017)	(0.027)
Number of patents	-0.028***	-0.023***	-0.002	-0.024***
	(0.002)	(0.003)	(0.007)	(0.008)
Family size	0.027***	0.077***	0.082***	0.085***
	(0.002)	(0.004)	(0.006)	(0.009)
Triadic	0.598***	0.405***	0.250***	0.254***
	(0.009)	(0.015)	(0.042)	(0.056)
Granted	0.721***	0.572***	0.562***	0.574***
	(0.005)	(0.007)	(0.024)	(0.025)
fixed effect	no	inventor	no	applicant
Obs.	697,192	697,192	435,584	435,584

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office, sector, year and month fixed effects.

Table C.8: Intra vs. inter-sectoral spillovers

	(1)	(2)	(3)
Dep. var.	PatentRank	PatentRank intra-sectoral	PatentRank inter-sectoral
Clean invention	0.292***	0.336***	0.248***
	(0.014)	(0.016)	(0.016)
Number of patents	-0.031***	-0.044***	-0.160***
	(0.005)	(0.006)	(0.007)
Family size	0.067***	0.067***	0.068***
	(0.003)	(0.003)	(0.003)
Triadic	0.241***	0.246***	0.259***
	(0.026)	(0.025)	(0.025)
Granted	0.491***	0.456***	0.521***
	(0.021)	(0.021)	(0.017)
Obs.	1,149,988	1,149,988	1,149,988

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variables are PatentRank index (column 1), PatentRank index on citations within their own technological field (based on IPC 3 digit code) (column 2), and the PatentRank index on citations across across technological field (column 3). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.9: Generality and originality as controls

	(1)	(2)	(3)	(4)
Dep. var.	PatentRank index			
Clean invention	0.193***	0.179***	0.193***	0.178***
	(0.007)	(0.007)	(0.007)	(0.007)
Number of patents	-0.010***	0.019***	-0.003	0.016***
	(0.002)	(0.002)	(0.002)	(0.002)
Family size	0.026***	0.022***	0.025***	0.023***
	(0.001)	(0.001)	(0.001)	(0.001)
Triadic	0.130***	0.110***	0.127***	0.111***
	(0.007)	(0.006)	(0.007)	(0.006)
Granted	0.245***	0.203***	0.240***	0.204***
	(0.010)	(0.010)	(0.010)	(0.010)
Generality		0.628***		0.663***
		(0.010)		(0.010)
Originality			0.127***	-0.097***
			(0.006)	(0.008)
Obs.	281,978	281,978	281,978	281,978

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.10: Controlling for age of technological field

	(1)	(2)	(3)	(4)
Dep. var.	PatentRank index			
Clean invention	0.283***	0.267***	0.257***	0.247***
	(0.013)	(0.013)	(0.013)	(0.012)
Number of patents	-0.053***	-0.029***	-0.023***	-0.023***
	(0.003)	(0.003)	(0.003)	(0.003)
Family size	0.065***	0.063***	0.063***	0.063***
	(0.003)	(0.003)	(0.003)	(0.003)
Triadic	0.236***	0.227***	0.210***	0.202***
	(0.025)	(0.025)	(0.025)	(0.025)
Granted	0.487***	0.480***	0.474***	0.470***
	(0.021)	(0.021)	(0.020)	(0.020)
Age of tech field		-0.117***	0.233***	
		(0.006)	(0.014)	
Age of tech fieldˆ2			-0.023***	
			(0.001)	
Age of tech dummies	no	no	no	yes
Obs.	1,149,237	1,149,237	1,149,237	1,149,237

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.11: Clean, Grey and True Dirty

	(1)	(2)	(3)	(4)
Sample	Clean vs. Grey and true Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs. True Dirty
Dep. var.	PatentRank index			
Clean/Grey invention	0.292***	0.121***	0.190***	0.331***
	(0.014)	(0.012)	(0.016)	(0.015)
Number of patents	-0.031***	-0.006	-0.084***	-0.029***
	(0.005)	(0.008)	(0.004)	(0.005)
Family size	0.067***	0.059***	0.065***	0.065***
	(0.003)	(0.006)	(0.004)	(0.003)
Triadic	0.241***	0.278***	0.238***	0.240***
	(0.025)	(0.045)	(0.028)	(0.026)
Granted	0.491***	0.520***	0.508***	0.456***
	(0.021)	(0.022)	(0.022)	(0.019)
Obs.	1,149,988	326,942	978,179	1,006,996

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index. The sample includes clean, grey and truly dirty (column 1), clean and grey (column 2), grey and truly dirty (column 3), and clean and truly dirty (column 4) inventions. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.12: Spillovers from clean and other new technologies

	(1)	(2)	(3)	(4)	(5)
Baseline sector	IT	Biotechs	Nano	Robot	3D
Dep. var.	PatentRank index				
Clean invention	-0.039	0.131***	-0.249***	-0.096*	-0.120***
	(0.028)	(0.023)	(0.040)	(0.043)	(0.018)
Number of patents	-0.031***	-0.029***	0.023***	0.014	0.018*
	(0.005)	(0.006)	(0.008)	(0.078)	(0.008)
Family size	0.017***	0.029***	0.052***	0.053***	0.052***
	(0.003)	(0.004)	(0.006)	(0.006)	(0.006)
Triadic	0.421***	0.435***	0.329***	0.337***	0.333***
	(0.050)	(0.042)	(0.055)	(0.054)	(0.055)
Granted	0.604***	0.413***	0.441***	0.443***	0.448***
	(0.040)	(0.017)	(0.025)	(0.025)	(0.025)
Observations	1,445,552	403,294	180,441	198,602	185,726

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index. The sample includes all clean patents (transport and electricity) and patents from the following technologies: IT (column 1), biotechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Figure C.4: Clean, grey, dirty, and radically new technologies vs. all other technologies – PageRank index

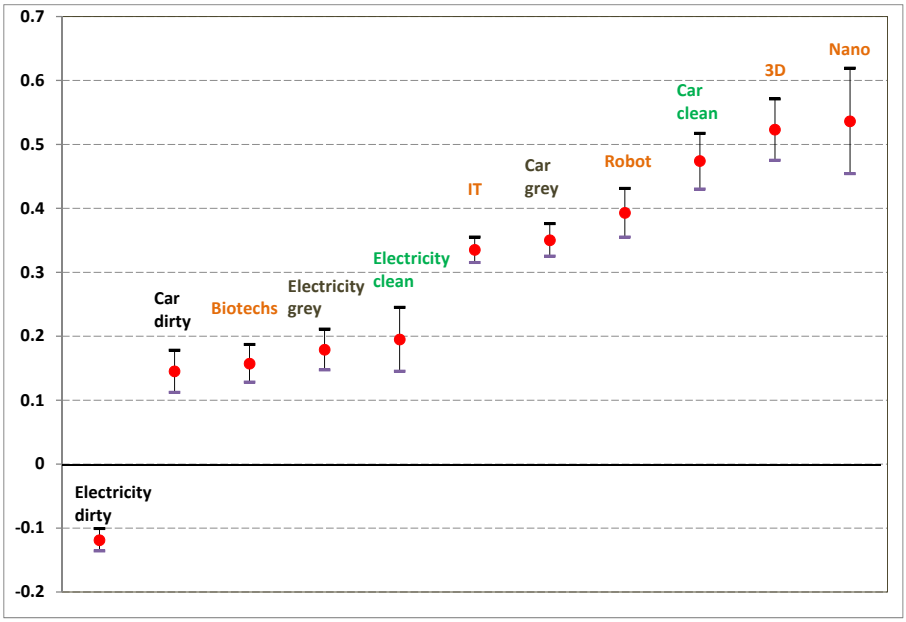


Table C.13: Comparing spillovers from clean and dirty within new technologies

	(1)	(2)	(3)	(4)	(5)
Sector	IT	Biotechs	Nano	Robot	3D
Dep. var.	PatentRank index				
Clean invention	0.129*	0.422***	0.189	0.349	0.290
	(0.053)	(0.067)	(0.100)	(0.325)	(0.461)
Number of patents	-0.037***	-0.074***	0.033	-0.062**	-0.080***
	(0.006)	(0.005)	(0.023)	(0.023)	(0.014)
Family size	0.017***	0.028***	0.070***	0.088***	0.056***
	(0.003)	(0.004)	(0.013)	(0.009)	(0.011)
Triadic	0.401***	0.406***	0.341***	0.261***	0.305**
	(0.049)	(0.041)	(0.070)	(0.057)	(0.060)
Granted	0.624***	0.342***	0.424***	0.443***	0.571***
	(0.044)	(0.019)	(0.075)	(0.042)	(0.044)
Observations	1,270,842	227,100	1,481	22,266	9,359

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank. The sample includes patents from the following technologies: IT (column 1), bioechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.14: Spillovers from clean and CCS technologies

Dep. var.	PatentRank index
Clean invention	0.045
	(0.023)
Number of patents	0.057***
	(0.010)
Family size	0.055***
	(0.005)
Triadic	0.271***
	(0.047)
Granted	0.338***
	(0.019)
Observations	106,700

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. The sample includes clean electricity production inventions and CO2 Capture and Storage technology. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Additional Results

Table C.15: Government spending

	(1)	(2)	(3)	(4)
Sample	Transport		Electricity	
Dep. var.	Citations received			
Clean invention	0.253**	0.253***	0.483***	0.497***
	(0.077)	(0.079)	(0.026)	(0.026)
Government spending		-0.001		0.032***
		(0.033)		(0.007)
Number of patents	-0.070***	-0.070***	-0.006	-0.005
	(0.020)	(0.020)	(0.009)	(0.009)
Family size	0.054***	0.054***	0.066***	0.066***
	(0.012)	(0.012)	(0.004)	(0.004)
Triadic	0.474***	0.474***	0.447***	0.445***
	(0.093)	(0.094)	(0.046)	(0.047)
Granted	0.776***	0.776***	0.696***	0.695***
	(0.055)	(0.055)	(0.026)	(0.026)
Obs.	16,703	16,703	488,896	488,896

Source: International Energy Agency (2013): Energy Technology Research and Development Database (Edition: 2013). Mimas, University of Manchester

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received excluding self-citations by inventors. The sample includes clean and dirty inventions from the transport sector (columns 3 and 4), electricity sector (columns 5 and 6) and both transport and electricity sectors (columns 1 and 2) . All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.16: Clean, Grey and true Dirty - Transport

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	Clean vs. Grey and True Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs. True Dirty	Clean vs. Grey and True Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs. True Dirty
Dep. var.	Citations received				PatentRank index			
Clean/Grey invention	0.347***	0.118***	0.304***	0.481***	0.219***	0.090***	0.169***	0.292***
	(0.018)	(0.020)	(0.017)	(0.022)	(0.014)	(0.014)	(0.017)	(0.018)
Number of patents	-0.068***	-0.144***	-0.109***	-0.082***	-0.048***	-0.075***	-0.088***	-0.053***
	(0.009)	(0.010)	(0.009)	(0.009)	(0.006)	(0.006)	(0.006)	(0.005)
Family size	0.070***	0.070***	0.081***	0.065***	0.062***	0.059***	0.074***	0.057***
	(0.008)	(0.011)	(0.010)	(0.007)	(0.007)	(0.010)	(0.007)	(0.006)
Triadic	0.521***	0.483***	0.474***	0.488***	0.279***	0.281***	0.219***	0.284***
	(0.056)	(0.071)	(0.059)	(0.055)	(0.046)	(0.057)	(0.040)	(0.046)
Granted	1.134***	1.122***	1.173***	1.046***	0.620***	0.599***	0.637***	0.588***
	(0.034)	(0.041)	(0.036)	(0.032)	(0.027)	(0.027)	(0.029)	(0.024)
Obs.	419,959	207,524	345,313	287,469	419,959	207,524	345,313	287,469

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors (columns 1 to 4) and the PatentRank index (columns 5 to 8). The sample includes clean, grey and truly dirty (column 1 and 5), clean and grey (column 2 and 6), grey and truly dirty (column 3 and 7), and clean and truly dirty (column 4 and 8) inventions all in the transport sector. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.17: Clean, Grey and true Dirty - Electricity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	Clean vs. Grey and true Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs. True Dirty	Clean vs. Grey and true Dirty	Clean vs. Grey	Grey vs. True Dirty	Clean vs True Dirty
Dep. var.	Citations received				PatentRank index			
Clean/Grey invention	0.488***	0.188***	0.262***	0.499***	0.333***	0.046	0.287***	0.342***
	(0.023)	(0.032)	(0.019)	(0.023)	(0.023)	(0.028)	(0.013)	(0.023)
Number of patents	-0.047***	0.042***	-0.114***	-0.044***	-0.019***	0.062***	-0.073***	-0.01**
	(0.009)	(0.011)	(0.007)	(0.009)	(0.007)	(0.010)	(0.004)	(0.00)
Family size	0.067***	0.070***	0.066***	0.067***	0.060***	0.058***	0.061***	0.061***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.00)
Triadic	0.432***	0.416***	0.396***	0.438***	0.252***	0.228***	0.226***	0.256***
	(0.050)	(0.051)	(0.046)	(0.050)	(0.000)	(0.038)	(0.037)	(0.04)
Granted	0.725***	0.660***	0.738***	0.727***	0.381***	0.331***	0.393***	0.382***
	(0.024)	(0.029)	(0.026)	(0.025)	(0.017)	(0.018)	(0.018)	(0.01)
Observations	748,918	120,752	647,541	733,859	748,918	120,752	647,541	733,859

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors (columns 1 to 4) and the PatentRank index (columns 5 to 8). The sample includes clean, grey and truly dirty (column 1 and 5), clean and grey (column 2 and 6), grey and truly dirty (column 3 and 7), and clean and truly dirty (column 4 and 8) inventions all in the transport sector. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.18: Generality and originality as controls

	(1)	(2)	(3)	(4)
Dep. var.	Citations received			
Clean invention	0.365***	0.332***	0.363***	0.332***
	(0.012)	(0.012)	(0.012)	(0.012)
Number of patents	-0.044***	0.007	-0.025***	0.006
	(0.005)	(0.006)	(0.005)	(0.005)
Family size	0.043***	0.039***	0.041***	0.039***
	(0.002)	(0.002)	(0.002)	(0.002)
Triadic	0.296***	0.264***	0.287***	0.264***
	(0.014)	(0.013)	(0.014)	(0.013)
Granted	0.673***	0.591***	0.659***	0.592***
	(0.023)	(0.021)	(0.022)	(0.021)
Generality		1.149***		1.164***
		(0.019)		(0.019)
Originality			0.371***	-0.036*
			(0.015)	(0.015)
Obs.	281,978	281,978	281,978	281,978

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations by inventors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year and month fixed effects.

Table C.19: Comparing the generality of clean and other new technologies

	(1)	(2)	(3)	(4)	(5)
Sector	IT	Biotechs	Nano	Robot	3D
Dep. var.	Originality measure				
Clean invention	-0.050***	-0.059***	0.009	-0.130***	-0.184***
	(0.004)	(0.004)	(0.018)	(0.004)	(0.006)
Number of patents	-0.070***	-0.033***	-0.049***	-0.051***	-0.051***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Family size	0.003***	0.002***	0.005***	0.005***	0.005***
	(0.0004)	(0.0003)	(0.001)	(0.001)	(0.001)
Triadic	0.010***	0.005	0.014***	0.015***	0.014**
	(0.002)	(0.002)	(0.005)	(0.004)	(0.004)
Granted	0.020***	-0.003	0.029***	0.027***	0.027***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	520,978	155,701	59,651	67,115	62,559

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the originality measure. The sample includes all clean inventions (automobile and electricity production sectors) and inventions from the following technologies: IT (column 1), biotechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by OLS and include patent office-by-year and month fixed effects.

Table C.20: Comparing the generality of clean and other new technologies

	(1)	(2)	(3)	(4)	(5)
Sector	IT	Biotechs	Nano	Robot	3D
Dep. var.	Generality measure				
Clean invention	-0.047***	-0.052***	0.009	-0.126***	-0.204***
	(0.004)	(0.004)	(0.022)	(0.004)	(0.006)
Number of patents	-0.063***	-0.034***	-0.048***	-0.050***	-0.050***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Family size	0.004***	0.003***	0.008***	0.008***	0.008***
	(0.0005)	(0.0003)	(0.001)	(0.001)	(0.001)
Triadic	0.013***	0.016***	0.022***	0.023***	0.020***
	(0.002)	(0.003)	(0.005)	(0.005)	(0.004)
Granted	0.022***	0.022***	0.041***	0.038***	0.039***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Obs.	723,257	207,073	94,437	103,972	98,461

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the generality measure. The sample includes all clean patents (automobile and electricity production sectors) and patents from the following technologies: IT (column 1), biotechs (column 2), nano (column 3), robot (column 4), and 3D (column 5). All columns are estimated by OLS and include patent office-by-year and month fixed effects.

Robustness Checks

Five Years Window

As in section 4.2.4 we look at the number of citations received within a five-year window to at least partially overcome the truncation bias that is due to the fact that we observe citations for only a portion of the life of an invention, with the duration of that portion varying across patent cohorts (see Table C.21). The coefficients obtained for the clean dummy barely change.

Table C.21: Five-year window

	(1)	(2)	(3)	(4)	(5)	(6)
Sector	All	Transport	Electricity	All	Transport	Electricity
Dep. var.	Citations received within 5-year window			PatentRank within 5-year window		
Clean invention	0.382***	0.284***	0.474***	0.210***	0.140***	0.248***
	(0.021)	(0.025)	(0.034)	(0.015)	(0.014)	(0.026)
Number of patents	-0.038***	-0.055***	-0.023*	-0.038***	-0.059***	-0.022**
	(0.008)	(0.001)	(0.010)	(0.005)	(0.006)	(0.007)
Family size	0.075***	0.070***	0.063***	0.069***	0.062***	0.059***
	(0.003)	(0.001)	(0.007)	(0.003)	(0.008)	(0.006)
Triadic	0.508***	0.557***	0.515***	0.306***	0.354***	0.346***
	(0.043)	(0.070)	(0.068)	(0.003)	(0.053)	(0.053)
Granted	1.005***	1.181***	0.756***	0.581***	0.693***	0.473***
	(0.040)	(0.054)	(0.035)	(0.024)	(0.036)	(0.022)
Obs.	1,162,220	419,959	748,918	1,162,220	419,959	748,918

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received within a five-year period after the publication year, corrected for self-citations by inventors (columns 1 to 3) and the PatentRank index on the sample of citations within five years (columns 4 to 6). The sample includes patents which have cited clean or dirty technologies in the automobile sector (columns 2 and 4), electricity sector (columns 3 and 6), and both (colunns 1 and 4). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

Discarding Citations

We discard citations added by patent examiners in Table C.22.[1] By restricting the citation counts to the ones made by the applicant only, we address the concern that patent citations added by examiners might not capture actual knowledge spillovers. The results obtained when all sectors are pooled together barely change but the only noticeable difference is that the clean dummy is no longer significant on the fuel sector when citations added by examiners are excluded. Jaffe and Trajtenberg (1999)) find that patent assigned to the same firm are more likely to cite each other. We therefore correct for self-citations at the level of the applicant (the firm or the individual who filed the patent) rather than at the level of individual inventors in Table C.23. The results don't change qualitatively.

Additional Controls

We add a number of additional controls variables for patent quality in Table C.24. The claims specify the components of the patent invention and hence represent the scope of the invention (Lanjouw and Schankerman (1999)). This information is only available in our patent database for a limited number of patent offices, implying that our sample size is significantly reduced. For this reason we do not include the number of claims in our baseline regressions, but overall the results barely change (coefficient on clean = 0.403***). The number of IPC3 codes is added in order to control for the fact that certain inventions belong to multiple IPC codes. These inventions are likely to be more general and therefore more cited. This effect however does not appear to downplay the clean advantage in terms of spillovers. Finally, we add the number of inventors and still find that clean inventions are

Various subsamples

In Table C.26 we look at different subsamples. We start by restricting the sample to patents that received at least one citation. Given that a large fraction of patents (69%) are never cited, spillovers from clean technologies might be biased if there are disproportionately more dirty patents that are never cited. We also look at highly valuable in-

[1]Note that we restrict the sample to patent offices for which distinction between citation added by patent examiner or applicant is made.

Table C.22: Citations made by *applicants* only

	(1)	(2)	(3)	(4)	(5)	(6)
Sector	All	Transport	Electricity	All	Transport	Electricity
Dep. var.	Citations received excl. citations added by patent examiner			PatentRank index excl. citations added by patent examiner		
Clean invention	0.624***	0.582***	0.659***	0.041***	0.013	0.085***
	(0.018)	(0.025)	(0.026)	(0.009)	(0.011)	(0.009)
Number of patents	-0.010	-0.036***	0.007	-0.016***	-0.016***	-0.018***
	(0.008)	(0.011)	(0.011)	(0.002)	(0.003)	(0.003)
Family size	0.070***	0.070***	0.060***	0.012***	0.017***	0.008***
	(0.003)	(0.007)	(0.004)	(0.001)	(0.002)	(0.001)
Triadic	0.516***	0.537***	0.564***	0.072***	0.049***	0.098***
	(0.042)	(0.064)	(0.066)	(0.008)	(0.009)	(0.011)
Granted	1.144***	1.164***	1.101***	0.102***	0.097***	0.105***
	(0.025)	(0.035)	(0.031)	(0.007)	(0.009)	(0.012)
Obs.	1,162,220	419,950	748,918	1,162,220	419,950	748,918

Notes: Robust standard errors in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, excluding citations made by the patent examiners and corrected for self-citations by inventors (columns 1 to 3) and the PatentRank index on the sample of citations excluding the ones made by the patent examiners (colums 4 to 6). The sample only includes patent offices for which the information on whether citations are added by the applicant or the patent examiner are available. The sample includes patents which have cited clean or dirty technologies in the automobile sector (columns 2 and 4), electricity production sector (columns 3 and 6), in both (colunns 1 and 4). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

Table C.23: Excluding self-citations at applicant level

	(1)	(2)	(3)	(4)	(5)	(6)
Sector	All	Transport	Electricity	All	Transport	Electricity
Dep. var.	Citations received corrected for self-citations at applicant level			PatentRank index corrected for self-citations at applicant level		
Clean invention	0.363***	0.351***	0.350***	0.164***	0.161***	0.155***
	(0.013)	(0.022)	(0.016)	(0.007)	(0.011)	(0.010)
Number of patents	-0.040***	-0.030***	-0.047***	0.0004	0.002	0.001
	(0.005)	(0.007)	(0.007)	(0.003)	(0.004)	(0.004)
Family size	0.044***	0.046***	0.038***	0.033***	0.033***	0.030***
	(0.003)	(0.005)	(0.003)	(0.002)	(0.003)	(0.002)
Triadic	0.218***	0.250***	0.214**	0.114***	0.103***	0.135**
	(0.024)	(0.039)	(0.030)	(0.013)	(0.019)	(0.015)
Granted	0.456***	0.549***	0.388***	0.201***	0.242***	0.168***
	(0.016)	(0.019)	(0.022)	(0.010)	(0.013)	(0.013)
Obs.	421,872	167,711	244,435	421,872	167,711	244,435

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, corrected for self-citations at the applicant level (columns 1 to 3) and the PatentRank index on the sample of citations excluding self-citations at the appliant level (colums 4 to 6). The sample includes patents which have cited clean or dirty technologies in the automobile sector (columns 2 and 4), electricity production sector (columns 3 and 6), in both (colunns 1 and 4). All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

Table C.24: Additional controls

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.	Citations received					
Clean invention	0.404***	0.432***	0.427***	0.432***	0.428***	0.432***
	(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.015)
Number of patents	-0.032***	-0.020***	-0.005***	-0.049***	-0.057***	-0.013
	(0.005)	(0.008)	(0.006)	(0.007)	(0.007)	(0.007)
Family size	0.033***	0.065***	0.061***	0.062***	0.056***	0.013***
	(0.002)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
Triadic	0.239***	0.464***	0.281***	0.401***	0.447***	0.229***
	(0.012)	(0.042)	(0.022)	(0.029)	(0.034)	(0.019)
Granted	0.750***	0.938***	0.922***	0.894***	0.941***	0.855***
	(0.025)	(0.000)	(0.028)	(0.030)	(0.030)	(0.028)
# claims	0.010***					
	(0.0004)					
# IPC 3		0.103***				0.092***
		(0.013)				(0.005)
# inventors			0.321***			0.341***
			(0.014)			(0.167)
# citations made				0.018***		0.017***
				(0.001)		(0.001)
# applicants					0.009***	-0.008***
					(0.0010)	(0.001)
Obs.	175,298	1,161,160	865,607	1,161,160	1,161,160	

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, excluding self-citations by the inventor (columns 1 to 3) and the PatentRank index on the sample of citations excluding self-citations by the inventor (colums 4 to 6). The sample includes clean or dirty technologies in the automobile and electricity production sectors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

ventions by focusing on triadic patents (i.e., patents that have been filed at the USPTO, the EPO and the Japan Patent Office, see above). This can give us some insight into whether the clean advantage is still present for the upper part of the distribution. In addition, we restrict our sample to patents filed at the US patent office and at the European Patent Office. None of these tests modify our main finding (coefficient on clean between 0.319*** and 0.469***).

Table C.25: Additional controls

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.	PatentRank index					
Clean invention	0.239***	0.295***	0.294***	0.294***	0.291***	0.298***
	(0.009)	(0.014)	(0.013)	(0.014)	(0.014)	(0.013)
Number of patents	-0.0002	-0.005	-0.022***	-0.027***	-0.031***	0.003***
	(0.002)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
Family size	0.021***	0.061***	0.023***	0.059***	0.058***	0.023***
	(0.001)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
Triadic	0.120***	0.244***	0.138***	0.193***	0.236***	0.095***
	(0.005)	(0.030)	(0.017)	(0.022)	(0.026)	(0.014)
Granted	0.336***	0.484***	0.517***	0.462***	0.488***	0.475***
	(0.016)	(0.021)	(0.018)	(0.019)	(0.021)	(0.017)
# claims	0.004***					
	(0.0002)					
# IPC 3		0.077***				0.062***
		(0.007)				(0.003)
# inventors			0.216***			0.238***
			(0.009)			(0.010)
# citations made				0.014***		0.012***
				(0.001)		(0.001)
# applicants					0.006***	-0.008
					(0.002)	(0.001)
Obs.	175,298	1,161,160	865,607	1,161,160	1,161,160	

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index on the sample of citations received, excluding self-citations by the inventor. The sample includes clean or dirty technologies in the automobile and electricity production sectors. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects

Table C.26: Different subsamples

	(1)	(2)	(3)	(4)
Sample	No zero	triadic	US patent office	EU patent office
Dep. var.	Citations received			
Clean invention	0.321***	0.387***	0.429***	0.491***
	(0.012)	(0.019)	(0.019)	(0.050)
Number of patents	-0.045***	-0.041***	-0.054***	-0.010
	(0.005)	(0.008)	(0.009)	(0.019)
Family size	0.056***	0.021***	0.049***	0.048***
	(0.002)	(0.002)	(0.003)	(0.011)
Triadic	0.365***		0.134***	0.447***
	(0.022)		(0.021)	(0.048)
Granted	0.625***	0.663***	0.957***	0.641***
	(0.025)	(0.045)	(0.069)	(0.045)
Obs.	514,865	45,129	134,664	10,248

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the total number of citations received, excluding self-citations by the inventor The sample includes (i) patents that receive at least one citation in column 1; (ii) triadic patents (filed at EPO, USPTO and JPO) in column 2; (iii) patents first filed in the US patent office only in column 3; (iv) patents first filed in the European patent office only in column 4. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

Table C.27: Different subsamples

	(1)	(2)	(3)	(4)
Sample	nozero	triadic	US patent office	EU patent office
Dep. var.	PatentRank index			
Clean invention	0.173***	0.212***	0.254***	0.340***
	(0.008)	(0.009)	(0.014)	(0.032)
Number of patents	-0.013***	-0.003	-0.016***	-0.007
	(0.002)	(0.003)	(0.004)	(0.008)
Family size	0.039***	0.007***	0.031***	0.038***
	(0.002)	(0.001)	(0.002)	(0.003)
Triadic	0.164***		0.070***	0.234***
	(0.013)		(0.010)	(0.023)
Granted	0.249***	0.294***	0.573***	0.337***
	(0.016)	(0.024)	(0.040)	(0.026)
Observations	514,865	45,129	134,664	10,248

Notes: Robust standard errors, p-values in parentheses (* p<0.05, ** p<0.01, *** p<0.001). The dependent variable is the PatentRank index on the sample of citations received, corrected for self-citations by inventors. The sample includes (i) patents that receive at least one citation in column 1; (ii) triadic patents (filed at EPO, USPTO and JPO) in column 2; (iii) patents first filed in the US patent office only in column 3; (iv) patents first filed in the European patent office only in column 4. All columns are estimated by Poisson pseudo-maximum likelihood and include patent office-by-year-by-sector fixed effects, and month fixed effects.

Appendix D

Note on Coauthored Work

The research presented in chapter 2 of this thesis, "Nation–Building Through Compulsory Schooling During the Age of Mass Migration", is co-authored with Oriana Bandiera, Martina Viarengo and Imran Rasul. Each author contributed equally to this project.

The research presented in chapter 4 of this thesis, "Knowledge Spillovers from Clean and Dirty Technologies: Evidence from Patent Citations", is co-authored with Antoine Dechezleprêtre and Ralf Martin. Each author contributed equally to this project.

I thank my coauthors for the fruitful cooperation.

Bibliography

Abramitzky, R., B. L. and Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(467-506).

Abramitzky, R., Boustan, L., and Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102:1832–56.

Abramitzky, R., Boustan, L., and Eriksson, K. (2014). Cultural assimilation during the age of mass migration. *mimeo, Stanford University*, 15:2015.

Acemoglu, D., Aghion, P., Bursztyn, L., and Hemous, D. (2012). The environment and directed technical change. *The American Economic Review*, 102(1):131.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly Journal of Economics*, 117:1231–94.

Acemoglu, D., S., J., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91:1369–401.

Aghion, P., Akcigit, U., Cagé, J., and Kerr, W. R. (2016). Taxation, corruption, and growth. *European Economic Review*.

Aghion, P., Dechezleprêtre, A., Hemous, D., Martin, R., and Van Reenen, J. (2012a).

Carbon taxes, path dependency and directed technical change: evidence from the auto industry. Technical report, National Bureau of Economic Research.

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

Aghion, P. and Howitt, P. (1992b). A model of growth through creative destruction. *Econometrica*, 60 (2):323–351.

Aghion, P. and Howitt, P. (1996). Research and development in the growth process. *Journal of Economic Growth*, 1(1):49–73.

Aghion, P. and Howitt, P. (1998). *Endogenous growth theory*. MIT press.

Aghion, P., Persson, T., and Rouzet, D. (2012b). Education and military rivalry. Technical report, NBER WP 18049.

Agrawal, A., McHale, J., and Oettl, A. (2013). Collaboration, stars, and the changing organization of science: Evidence from evolutionary biology. Technical report, National Bureau of Economic Research.

Aizenman, J. and Kletzer, K. (2008). Networking, citation of academic research, and premature death. *NBER working paper*.

Alcacer, J. and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4):774–779.

Alcacer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in us patents: An overview and analysis. *Research Policy*, 38(2):415–427.

Alesina, A. and Reich, B. (2015). Nation building. *NBER WP 18839*.

Allison, J. R., Lemley, M. A., Moore, K. A., and Trunkey, R. D. (2003). Valuable patents. *Geo. Lj*, 92:435.

Allison, P. D. and Waterman, R. P. (2002). Fixed–effects negative binomial regression models. *Sociological methodology*, 32(1):247–265.

Ambec, S. and Barla, P. (2006). Can environmental regulations be good for business? an assessment of the porter hypothesis. *Energy studies review*, 14(2):1.

Ambec, S., Cohen, M. A., Elgie, S., and Lanoie, P. (2013). The porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness? *Review of Environmental Economics and Policy*, 7(1):2–22.

Anderson, R. (1995). *Education and the Scottish People 1750-1918*. Oxford: Clarendon Press.

Angrist, J. (2002). How do sex ratios affect marriage and labor markets? evidence from america's second generation. *Quarterly Journal of Economics*, 117:997–1038.

Arlington, M. (1991). English-only laws and direct legislation: The battle in the states over language minority rights. *Journal of Law and Politics*, pages 325–52.

Arrow, K. J., Bernheim, B. D., Feldstein, M. S., McFadden, D. L., Poterba, J. M., and Solow, R. M. (2011). 100 years of the american economic review: The top 20 articles. *The American Economic Review*, 101(1):1–8.

Audretsch, D. and Feldman, M. (2004). Knowledge spillovers and the geography of innovation. *Handbook of regional and urban economics*, 4:2713–2739.

Aumann, R. and Myerson, R. (1988). Endogenous formation of links between players and coalitions: an application of the shapley value. *The Shapley Value*, pages 175–191.

Azoulay, P., Fons-Rosen, C., and Zivin, J. S. G. (2014). Does science advance one funeral at a time? *NBER working paper*.

Azoulay, P., Graff Zivin, J., and Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics*, 25:549–589.

Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G., and Detsky, A. S. (2007). The meaning of author order in medical research. *Journal of Investigative Medicine*, 55(4):174–180.

Bala, V. and Goyal, S. (1999). A non-cooperative theory of network formation. Technical report, Tinbergen Institute.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: the key player. *Econometrica*, 74(5):1403–1417.

Ballester, C., Zenou, Y., and Calvó-Armengol, A. (2010). Delinquent networks. *Journal of the European Economic Association*, 8(1):34–61.

Bandiera, O., Rasul, I., and Viarengo, M. (2013). The making of modern america: Migratory flows in the age of mass migration.

Bandle, O., Baunmuller, K., Jahr, E., Karker, A., Naumann, H.-P., Telemann, U., Elmevik, I., and Widmark, G. (2005). *The Nordic LLanguage. An International Handbook of the History of the North Germanic Languages*, volume 2. Berlin: De Gruyter.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013a). The diffusion of microfinance. *Science*, 341(6144):1236498.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2014). Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research.

Banerjee, A. V., Duflo, E., Glennerster, R., and Kinnan, C. (2013b). The miracle of microfinance? evidence from a randomized evaluation.

Banks, A. and Wilson, K. (2011). Cross-national time-series data archive. databanks international, jerusalem: Israel.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barro, R. and McCleary, R. (2005). Which countries have state religions? *Quarterly Jou*, 120:1331–70.

Becker, S. O. and Hvide, H. K. (2013). Do entrepreneurs matters? *The Warwick Economics Research Paper Series (TWERPS) 1002, University of Warwick, Department of Economics*.

Benavot, A. and Riddle, P. (1998). The expansion of primary education, 1870-1940: Trends and issues. *Sociology of Education*, 61:191–210.

Bennedsen, M., Nielsen, K., Perez-Gonzales, F., and Wolfenzon, D. (2007). Inside the family firm. the role of families in succession decisions and performance. *Quarterly Journal*, 122:647–691.

Besley, T. and Persson (2010). State capacity, conflict, and development. *Econometrica*, 78:1–34.

Bisin, A. and T., V. (2000). "beyond the melting pot": Cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics*, 115:955–88.

Biskup, M. (1983). Preußen und polen: Grundlinien und reflexionen. *Jahrbücher für Geschichte Osteuropas. Neue Folge*, 31:1–27.

Bjørner, T. B. and Mackenhauer, J. (2013). Spillover from private energy research. *Resource and Energy Economics*.

Black, S. and Sokoloff, K. (2006). *Long-Term Trends in Schooling: The Rise and Decline (?) of Public Education in the United States*, volume 1. Handbook of the Economics of Education.

Bloch, F., Jackson, M. O., and Tebaldi, P. (2016). Centrality measures in networks. *Available at SSRN 2749124*.

Borgatti, S. P. (2003). The key player problem. In *Dynamic social network modeling and analysis: Workshop summary and papers*, page 241. National Academies Press.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71.

Borgatti, S. P. and Jones, C. (1996). A measure of past collaboration1.

Borjas, G. J. and Doran, K. B. (2012a). Cognitive mobility: Labor market responses to supply shocks in the space of ideas. Technical report, National Bureau of Economic Research.

Borjas, G. J. and Doran, K. B. (2012b). The collapse of the soviet union and the productivity of american mathematicians*. *The Quarterly Journal of Economics*, 127(3):1143–1203.

Borjas, G. J. and Doran, K. B. (2013). Which peers matter? the relative impacts of collaborators, colleagues, and competitors.

Bramoullé, Y., Kranton, R., and D'amours, M. (2014). Strategic interaction and networks. *The American Economic Review*, 104(3):898–930.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative science quarterly*, pages 518–539.

Braun, F., Schmidt-Ehmcke, J., and Zloczysti, P. (2010). Innovative activity in wind and solar technology: Empirical evidence on knowledge spillovers using patent data. *DIW Berlin Discussion Papers*, (993).

Braun, S. and Kvasnicka, M. (2013). Men, women, and the ballot: Gender imbalances and suffrage extensions in the united states. *Explorations in Economic History*, 50:405–26.

Breschi, S. and Catalini, C. (2010). Tracing the links between science and technology: An exploratory analysis of scientists and inventors networks. *Research Policy*, 39(1):14–26.

Breschi, S. and Lissoni, F. (2005). Knowledge networks from patent data. In *Handbook of quantitative science and technology research*, pages 613–643. Springer.

Breschi, S., Lissoni, F., and Montobbio, F. (2007). The scientific productivity of academic inventors: new evidence from italian data. *Econ. Innov. New Techn.*, 16(2):101–118.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108.

Brumberg, S. (1997). Going to america, going to school - the immigrant-public school encounter in turn-of-the-century new york city. *American Jewish Archives*, pages 86–135.

Burt, R. S. (1980). Models of network structure. *Annual review of sociology*, pages 79–141.

Burt, R. S. (1993). The social structure of competition. *Explorations in economic sociology*, 65:103.

Burt, R. S. (1997). The contingent value of social capital. *Administrative science quarterly*, pages 339–365.

Burt, R. S. (2000). The network structure of social capital. *Research in organizational behavior*, 22:345–423.

Burt, R. S. (2002). The social capital of structural holes. *The new economic sociology: Developments in an emerging field*, pages 148–190.

Burt, R. S. (2004). Structural holes and good ideas. *American journal of sociology*, 110(2):349–399.

Burt, R. S. (2008). Information and structural holes: comment on reagans and zuckerman. *Industrial and Corporate Change*, 17(5):953–969.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press.

Buskens, V. and Van de Rijt, A. (2008). Dynamics of networks if everyone strives for structural holes1. *American Journal of Sociology*, 114(2):371–407.

Caballero, R. and Jaffe, A. (1993). How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. In *NBER Macroeconomics Annual*, volume 8, pages 15–86. MIT press.

Calvo-Armengol, A. and Jackson, M. O. (2004). The effects of social networks on employment and inequality. *American economic review*, pages 426–454.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Calvó-Armengol, A. and Zenou, Y. (2004). Social networks and crime decisions: The role of social structure in faciliating delinquent behaviour. *International Economic Review*, 45(3):939–958.

Cantoni, D., Chen, Y., Yang, D., and Yuchtman, N. (2014). Curriculum and ideology. *NBER WP 20112*.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19:22–64.

Carillo, M. R., Papagni, E., and Capitanio, F. (2008). Effects of social interactions on scientists' productivity. *International Journal of Manpower*, 29(3):263–279.

Carter, L. (2009). Evening schools and child labor in the united states, 1870-1910. *Doctoral Disseration, Vanderbilt University.*

Case, D. O. and Higgins, G. M. (2000). How can we investigate citation behavior? a study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7):635–645.

Caselli, F. and Coleman, W. (2001). The us structural transformation and regional convergence: A reinterpretation. *Journal of Political Economy*, 109:584–616.

Cassi, L. and Plunket, A. (2010). The determinants of co-inventor tie formation: proximity and network dynamics.

Cassi, L. and Plunket, A. (2012). Research collaboration in co-inventor networks: combining closure, bridging and proximities.

Chaney, T. (2011). The network structure of international trade. Technical report, National Bureau of Economic Research.

Chay, K. and Munshi, K. (2013). Black networks after emancipation: Evidence from reconstruction and the great migration. *mimeo, Brown*.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Clay, K., Lingwall, J., and Stephens, M. (2012). Compulsory attendance laws and nineteenth century schooling. *NBER WP 18477*.

Clots-Figueras, I. and Masella, P. (2013). Education, language and identity. *Economic Journal*, 123:F332–57.

Cockburn, I., Kortum, S., and Stern, S. (2003). Are all patent examiners equal? examiners, patent characteristics, and litigation outcomes. *Patents in the knowledge-based economy*, 35.

Cohen, G. (1996). *Education and Middle-Class Society in Imperial Austria 1848–1918*. West Lafayette: Purdue University Press.

Cole, J. and Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the" science citation index". *The American Sociologist*, pages 23–29.

Coleman, J., Katz, E., and Mentzel, H. (1966). Medical innovation: Diffusion of a medical drug among doctors. *Indianapolis: Bobbs-Merrill*.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American journal of sociology*, pages S95–S120.

Colle-Michel, M. (2007). Le cours d'histoire dans l'enseignement secondaire. pour une education à la citoyenneté, institut d'histoire ouvrière. *économique et Sociale: Les Analyses du IHOES*.

Collins, P. and Wyatt, S. (1988). Citations in patents to the basic research literature. *Research Policy*, 17(2):65–74.

Collins, W. (1997). When the tide turned: Immigration and the delay of the great black migration. *Journal Economic History*, 57:607–32.

Collins, W. and Margo, R. (2006). Historical perspectives on racial differences in schooling in the united states. *Handbook of the Economics of Education*, 1:107–54.

Commission, E. (2010). Organisation of the education system in luxemburg, eurybase: The information database on education systems in europe. http://eacea.ec.europa.eu/education/eurydice/documents/eurybase/eurybasefullreports/LUEN.pdf.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *The American Economic Review*, pages 35–69.

Cook, K. S. and Emerson, R. M. (1978). Power, equity and commitment in exchange networks. *American sociological review*, pages 721–739.

Cook, K. S., Emerson, R. M., Gillmore, M. R., and Yamagishi, T. (1983). The distribution of power in exchange networks: Theory and experimental results. *American journal of sociology*, pages 275–305.

Coupé, T., Ginsburgh, V., and Noury, A. (2010). Are leading papers of better quality? evidence from a natural experiment. *Oxford Economic Papers*, 62(1):1–11.

Cowen, R. and Kazamias, A. (2009). *International Handbook of Comparative Education*. Dordrecht: Springer.

Criscuolo, P. (2006). The'home advantage'effect and patent families. a comparison of oecd triadic patents, the uspto and the epo. *Scientometrics*, 66(1):23–41.

Cubberley, E. (1920). *The History of Education. Educational Progress Considered as a Phase of the Development and Spread of Western Civilization*. Boston: Houghton Mifflin Company.

Cubberley, E. (1947). *Public Education in the United States: A Study and Interpretation of American Educational History*. Boston: The Riverside Press Cambridge.

Daniels, R. (1990). *Coming to America: A History of Immigration and Ethnicity in American Life*. New York: Harper Collins.

Daubanes, J., Grimaud, A., and Rougé, L. (2013). Green paradox and directed technical change: The effect of subsidies to clean r&d.

De Maeyer, J. (2005). *Religion, Children's Literature and Modernity in Western Europe 1750-2000*. Leuven: Leuven University Press.

Deng, Y. (2008). The value of knowledge spillovers in the us semiconductor industry. *International Journal of Industrial Organization*, 26(4):1044–1058.

Dernis, H. and Khan, M. (2004). Triadic patent families methodology. Technical report, OECD Publishing.

Dillingham, W. (1911). *Report of the United States Immigration Commission (1907–1910)*. US Government Printing Office.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2012). The intergenerational transmission of risk and trust attitudes. *Review of Economic Studies*, 79:645–77.

Donnermair, C. (2010). *Die Staatliche Übernahme des Primarschulwesens im 19. Jahrhundert: Maßnahmen und Intentionen. Vergleich Frankreich–Österreich*. PhD thesis, Doctoral thesis, Univ. of Vienna.

Ductor, L., Fafchamps, M., Goyal, S., and van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, 96(5):936–948.

Dutta, B., Ghosal, S., and Ray, D. (2005). Farsighted network formation. *Journal of Economic Theory*, 122(2):143–164.

Easley, D. and Kleinberg, J. (2010). Networks, crowds, and markets. *Cambridge University Press*, 1(2.1):2–1.

EFA (2000). County report armenia. http://www.unesco.org/education/wef/countryreports/armenia/contents.ht

Einhorn, E. (2005). *Modern Welfare States: Scandinavian Politics and Policy in the Global Age*. Westport: Praeger Publishers.

Eisenberg, M. (1989). The passage of compulsory attendance laws, 1870-1915. *Journal of the Midwest History of Education Society*, 17:184–98.

Engerman, S. (2005). The evolution of suffrage institutions in the new world. *Journal of Economic History*, 65:891–921.

Fadlon, I. and Nielsen, T. (2015). Household responses to severe health shocks and the design of social insurance. Working Paper.

Fafchamps, M., Leij, M. J., and Goyal, S. (2006). Scientific networks and co-authorship.

Fafchamps, M., Leij, M. J., and Goyal, S. (2010). Matching and network effects. *Journal of the European Economic Association*, 8(1):203–231.

Fagernas, S. (2014). Papers, please! the effect of birth registration on child labor and education in early 20th century usa. *Explorations in Economic History*, 52:63–92.

Ferenczi, I. and Willcox, W. (1929). *International Migrations*. Cambridge, MA: NBER.

Fernandez, R. (2007). Women, work, and culture. *Journal of the European Economic Association*, 5:305–32.

Ferrie, J. and Kuziembo, I. (2015). *The Role of Immigrant Children in their Parents' Assimilation in the US, 1850-2010*. in L.P. Boustan, C.Frydman and R.A.Margo (eds.) Human Capital in History: The American Record, forthcoming.

Field, A. (1979). Economic and demographic determinants of educational commitment: Massachusetts,1855. *Journal of Economic History*, 39:439–59.

Fischer, C. and Newell, R. (2008). Environmental and technology policies for climate mitigation. *Journal of environmental economics and management*, 55(2):142–162.

Fischer, C., Newell, R. G., and Preonas, L. (2014). Environmental and technology policy options in the electricity sector: Interactions and outcomes.

Fishback, P. (2000). Workers' compensation project data. http://eacea.ec.europa.eu/education/eurydice/documents/eurybase/eurybase_full_reports/LU_EN.pdf

Fleming, L., Mingo, S., and Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly*, 52(3):443–475.

Flora, P., Alber, J., Eichenberg, R., Kohl, J., Kraus, F., and Pfenning, W. (1983). *State, Economy and Society in Western Europe 1815-1975. A Data Handbook in Two Volumes*. Frankfurt am Main: Campus Verlag.

Forster, S. (2008). *L'école et ses réformes*. Lausanne: Presses Polytechniques et Universitaires Romandes.

Fort, M. (2006). Educational reforms across europe: A toolbox for empirical research, mimeo. http://www2.dse.unibo.it/fort/files/papers/fort_reforms.pdf.

Forti, E., Franzoni, C., and Sobrero, M. (2013). Bridges or isolates? investigating the social networks of academic inventors. *Research Policy*.

Fouka, V. (2014). Backlash: The unintended effects of language prohibition in us schools after world war i. *mimeo UPF*.

Fracassi, C. and Tate, G. (2012). External networking and internal firm governance. *The Journal of Finance*, 67(1):153–194.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. (2010). Network games. *The review of economic studies*, 77(1):218–244.

Galor, O. and Moav, O. (2006). Das human-kapital: A theory of the demise of the class structure. *Review of Economic Studies*, 73:85–117.

Galor, O., Moav, O., and Vollrath, D. (2009). Inequality in landownership, human capital promoting institutions and the great divergence. *Review of Economic Studies*, 76:143–79.

Garnier, M., Hage, J., and Fuller, B. (1989). The strong state, social class, and controlled school expansion in france, 1881-1975. *American Journal of Sociology*, 64:279–306.

Gathmann, C., Jorges, H., and Reinhold, S. (2012). Compulsory schooling reforms, education and mortality in twentieth century europe. *IZA DP 6403*.

Geddes, R., Lueck, D., and Tennyson, S. (2012). Human capital accumulation and the expansion of women's economic rights. *Journal of Law and Economics*, 55:839–67.

Gerlagh, R., Kverndokk, S., and Rosendahl, K. E. (2009). Optimal timing of climate change policy: Interaction between carbon taxes and innovation externalities. *Environmental and resource Economics*, 43(3):369–390.

Gkolia, C. and Brundrett, M. (2008). *Educational Leadership Development in Greece*. in M.Brundrett and M.Crawford (eds.),*Developing School Leaders. An International Perspective*, Oxon: Routledge.

Glaeser, E., Ponzetto, G., and Shleifer, A. (2007). Why does democracy need education? *Journal of Economic Growth*, 12:77–99.

Glaeser, E., Sacerdote, B., and Scheinkman, J. (2002). The social multiplier. harvard institute of economic research. *Discussion paper number 1968*.

Glaeser, E. L. and Sacerdote, B. (1996). Why is there more crime in cities? Technical report, National Bureau of Economic Research.

Glaeser, E. L., Sacerdote, B. I., and Scheinkman, J. A. (2003). The social multiplier. *Journal of the European Economic Association*, 1(2-3):345–353.

Glenn, C. (2002). *The Myth of the Common School*. Oakland, CA: ICS Press.

Go, S. and Lindert, P. (2010). The uneven rise of american public schools to 1850. *Journal of Economic History*, 70:1–26.

Goldin, C. (1994). *The Political Economy of Immigration Restriction: The United States, 1890-1921*. in C.Goldin and G.Libecap (eds.), *The Regulated Economy: An Historical Analysis of Government and the Economy*, Chicago: University of Chicago Press.

Goldin, C. (1999a). A brief history of education in the united states. *NBER HWP 119*.

Goldin, C. (1999b). Egalitarianism and the returns to education during the great transformation of american education. *Journal of Political Economy*, 107:S65–94.

Goldin, C. and Katz, L. (2003). Mass secondary schooling and the state: The role of state compulsory schooling in the high school movement. *NBER WP 10075*.

Goldin, C. and Katz, L. (2008). *The Race Between Education and Technology*. Cambridge, MA: Belknap Press for Harvard University Press.

Goldin, C.and Katz, L. (2001). The legacy of us educational leadership: notes on distribution and economic growth in the 20th century. *American Economic Review*, 91:18–23.

Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, pages 112–149.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best response dynamics.

Goodchild, L., Jonsen, R., Limerick, P., and Longanecker, D. (2014). *Higher Education in the American West: Regional History and State Contexts*. Palgrave McMillan.

Goulder, L. H. and Schneider, S. H. (1999). Induced technological change and the attractiveness of co2 abatement policies. *Resource and Energy Economics*, 21(3):211–253.

Goyal, S., Van Der Leij, M. J., and Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy*, 114(2):403–412.

Goyal, S. and Vega-Redondo, F. (2007). Structural holes in social networks. *Journal of Economic Theory*, 137(1):460–492.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American journal of sociology*, pages 481–510.

Granovetter, M. (1995). *Getting a job: A study of contacts and careers*. University of Chicago Press.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, pages 1360–1380.

Greaker, M. and Heggedal, T.-R. (2012). A comment on the environment and directed technical change. *Oslo Centre for Research on Environmentally Friendly Energy Working Paper*, 13.

Greene, W. H. (2007). Fixed and random effects models for count data. *Working paper, Department of Economics, Stern School of Business, New York University, New York*.

Grewal, R., Lilien, G. L., and Mallapragada, G. (2006). Location, location, location: How network embeddedness affects project success in open source systems. *Management Science*, 52(7):1043–1056.

Griliches, Z. (1981). Market value, r&d, and patents. *Economics letters*, 7(2):183–187.

Griliches, Z. (1992). The search for r&d spillovers. Technical report, National Bureau of Economic Research.

Griliches, Z., Hall, B. H., and Pakes, A. (1991). R&d, patents, and market value revisited: is there a second (technological opportunity) factor? *Economics of Innovation and new technology*, 1(3):183–201.

Grossman, G. and Helpman, E. (1991). Innovation and growth in the world economy.

Grupp, H. (1998). *Foundations of the economics of innovation: theory, measurement, and practice*. Edward Elgar.

Grupp, H., Münt, G., and Schmoch, U. (1996). Assessing different types of patent data for describing high-technology export performance. *Innovation, patents and technological strategies*, pages 271–287.

Guellec, D. and Van Pottelsberghe de la Potterie, B. (2004). From R&D to productivity growth: Do the institutional settings and the source of funds of R&D matter? *Oxford Bulletin of Economics and Statistics*, 66(3):353–378.

Guimerà, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20:23–48.

Hahn, Y., Islam, A., Patacchini, E., and Zenou, Y. (2015). Network structure and education outcomes: Evidence from a field experiment in bangladesh.

Haines, M. (2010). *1890 County Yearbook*. Inter-university Consortium for Political and Social Research. Historical, Demographic, Economic, and Social Data: The United States, 1790-2002 (ICPSR02896-v3), Ann Arbor, MI: ICPSR.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. Technical report, National Bureau of Economic Research.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, pages 16–38.

Hall, B. H., Mairesse, J., and Turner, L. (2007). Identifying age, cohort, and period effects in scientific research productivity: Discussion and illustration using simulated and actual data on french physicists. *Econ. Innov. New Techn.*, 16(2):159–177.

Han, A. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35:303–16.

Han, X., Small, D. S., Foster, D. P., Patel, V., et al. (2011). The effect of winning an oscar award on survival: correcting for healthy performer survivor bias with a rank preserving structural accelerated failure time model. *The Annals of Applied Statistics*, 5(2A):746–772.

Harary, F. (1969). Graph theory.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8):1343–1363.

Hart, R. (2004). Growth, environment and innovation - a model with production vintages and environmentally oriented research. *Journal of Environmental Economics and Management*, 48(3):1078–1098.

Hart, R. (2007). Can environmental regulations boost growth? In *Sustainable Resource Use and Economic Dynamics*, pages 53–70. Springer.

Hart, R. (2008). The timing and balance of policies for CO_2 abatement when technological change is endogenous. *Journal of Environmental Economics and Management*, 55:194–212.

Hausman, J. A., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship.

Haynie, D. L. (2001). Delinquent peers revisited: Does network structure matter? 1. *American journal of sociology*, 106(4):1013–1057.

Hegde, D. and Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, 105(3):287–289.

Hellerstein, J. K., Kutzbach, M. J., and Neumark, D. (2015). Labor market networks and recovery from mass layoffs before, during, and after the great recession. Technical report, National Bureau of Economic Research.

Henderson, R., Jaffe, A. B., and Trajtenberg, M. (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and Statistics*, 80(1):119–127.

Higham, J. (1988). *Strangers in the Land: Patterns of American Nativism, 1860-1925*. (Second Edition), New Brunswick: Rutgers University Press.

Hirschman, C. and Daniel Perez, A. (2010). *Immigration and Nativism in the United States and Europe: Demography and Globalization versus the Nation-State*. in J.Alber and N.Gilbert (eds.) *United in Diversity? Comparing Social Models in Europe and America*, New York: OUP.

Hobsbawm, E. (1990). *Nations and Nationalism Since 1780: Programme, Myth, Reality*. Cambridge, UK: Cambridge University Press.

Hollis, A. (2001). Co-authorship and the output of academic economists. *Labour Economics*, 8(4):503–530.

Horner, W., Dobert, H., Von Knopp, B., and Mitter, W. (2007). *The Education Systems of Europe*. Dordrecht: Springer.

Hove, O. (1967). The system of education in norway. *Paedagogica Europea*, 3:192–228.

Hutt, E. (2012). Formalism over function: Compulsory schooling, courts, and the rise of educational formalism in america, 1870-1930. *Teachers College Record, Columbia Univ.*, 114:1–27.

Isen, A. (2015). Dying to know: Are workers paid their marginal product? *Working Paper*.

Jackson, M. and Wolinsky, A. (1996). A strategic model of economic and social network. *J Econ Theory*, 71:44–74.

Jackson, M. O. et al. (2008). *Social and economic networks*, volume 3. Princeton university press Princeton.

Jackson, M. O. and Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *The American economic review*, pages 890–915.

Jackson, M. O., Rogers, B. W., and Zenou, Y. (2015). The economic consequences of social network structure. *Available at SSRN*.

Jaffe, A. B., Newell, R. G., and Stavins, R. (2005). A tale of two market failures: Technology and environmental policy. *Ecological Economics*, 54(2):164–174.

Jaffe, A. B. and Trajtenberg, M. (1996). Flows of knowledge from universities and federal labs: Modeling the flowof patent citations over time and across institutional and geographic boundaries. NBER Working Papers 5712, National Bureau of Economic Research.

Jaffe, A. B. and Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology*, 8(1-2):105–136.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, 90(2):215–218.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598.

Jaravel, X., Petkova, N., and Bell, A. (2015). Team-specific capital and innovation. *Available at SSRN 2669060*.

Jeynes, W. (2008). *The Widespread Growth of the Common School and Higher Education in American Educational History: School, Society and the Common Good*. Thousand Oaks, CA: SAGE.

Jones, B., Reedy, E., and Weinberg, B. A. (2014). Age and scientific genius. Technical report, National Bureau of Economic Research.

Jones, B. F. (2009). The burden of knowledge and the death of the renaissance man: is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Jones, B. F. (2010). Age and great invention. *The Review of Economics and Statistics*, 92(1):1–14.

Jones, B. F. and Olken, B. (2005). Do leaders matter? national leadership and growth since world war ii. *Quarterly Journal of Economics*, 120:835–864.

Jones, C. I. (1995). R & d-based models of economic growth. *Journal of political Economy*, pages 759–784.

Jones, C. I. (2005). Growth and ideas. *Handbook of Economic Growth*, 1:1063–1111.

Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., Sung, W.-K., and Lee, J.-H. (2009). On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97.

Karsten, S. and Majoor, D. (1994). *Education in East Central Europe: Educational Changes after the Fall of Communism*. Waxmann Verlag GmbH: Germany.

Katz, L. F., Kling, J. R., and Liebman, J. B. (2001). Moving to opportunity in boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, 116(607-654).

Keller, W. (2004). International technology diffusion. *Journal of Economic Literature*, 42(3):752–782.

Keyssar, A. (2000). *The Right to Vote*. New York: Basic Books.

Kirk, D. (1946). *Europe's Population in the Interwar Years*. Princeton, NJ: Princeton Univ. Press.

Kverndokk, S. and Rosendahl, K. E. (2007). Climate policies and learning by doing: Impacts and timing of technology subsidies. *Resource and Energy Economics*, 29(1):58–82.

Kverndokk, S., Rosendahl, K. E., and Rutherford, T. F. (2004). Climate policies and induced technological change: Which to choose, the carrot or the stick? *Environmental and Resource Economics*, 27(1):21–41.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. (1998). Law and finance. *Journal of Political Economy*, 106:1113–55.

Laband, D. and Tollison, R. (2006). Alphabetized coauthorship. *Applied Economics*, 38(14):1649–1653.

Lai, R., D'Amour, A., Yu, A. Y., Sun, Y., and Fleming, L. (2011). Disambiguation and co-authorship networks of the u.s. patent inventor database (1975 - 2010). *http://hdl.handle.net/1902.1/15705, Harvard Dataverse, V5*.

Landes, W. and Solomon, L. (1972). Compulsory schooling legislation: An economic analysis of law and social change in the nineteenth century. *Journal of Economic History*, 32:53–91.

Lanjouw, J. and Mody, A. (1996). Innovation and the international diffusion of environmentally responsive technology. *Research Policy*, 25(4):549–571.

Lanjouw, J. O. and Schankerman, M. (1999). The quality of ideas: measuring innovation with multiple indicators. Technical Report 7345, National Bureau of Economic Research.

Lanjouw, J. O. and Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465.

Laurance, W. F. (2006). Second thoughts on who goes where in author lists. *Nature*, 442(7098):26–26.

Leddon, L. (2010). *Compulsory Attendance: An Analysis of Litigation*. PhD thesis, Graduate School, University of Alabama.

Lee, S. and Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social studies of science*, 35(5):673–702.

Lemley, M. A. and Sampat, B. (2012). Examiner characteristics and patent office outcomes. *Review of Economics and Statistics*, 94(3):817–827.

Leopold, A. (1973). Games scientists play. *BioScience*, 23(10):590–594.

Levin, S. G. and Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review*, pages 114–132.

Library, P. (2013). The provision about the common labor school of rsfsr, published 16 october 1918. http://www.prlib.ru/en-us/history/pages/item.aspx?itemid=693.

Lindert, P. (2004). *Growing Public Social Spending and Economic Growth Since the Eighteenth Century*. (Vol. 2): Cambridge: Cambridge University Press.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, 10(2):145–162.

Lleras-Muney, A. (2002). Were compulsory attendance and child labor laws effective? an analysis from 1915 to 1939. *Journal of Law and Economics*, 45:401–35.

Lleras-Muney, A. and Shertzer, A. (2015). Did the americanization movement succeed? an evaluation of the effect of english-only and compulsory school laws on immigrant's education. *American Economic Journal: Economic Policy*, 7:258–90.

Long, J. and McGinnis, R. (1982). On adjusting productivity measures for multiple authorship. *Scientometrics*, 4(5):379–387.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*.

Lott, J. J. (1999). Public schooling, indoctrination and totalitarianism. *Journal of Political Economy*, 107:S127–57.

Lukach, R. and Lukach, M. (2007). Ranking uspto patent documents by importance using random surfer method (pagerank). *Available at SSRN 996595*.

Luttmer, E. and Singhal, M. (2011). Culture, context and the taste for redistribution. *American Economic Journal: Economic Policy*, 3:157–79.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Margo, R. and Finegan, T. (1996). Compulsory schooling legislation and school attendance in turn-of-the-century america. *Economics Letters*, 53:103–10.

Marsden, P. V. (1982). Brokerage behavior in restricted exchange networks. *Social structure and network analysis*, 7(4):341–410.

Maslov, S. and Redner, S. (2008). Promise and pitfalls of extending google's pagerank algorithm to citation networks. *The Journal of Neuroscience*, 28(44):11103–11105.

Maurseth, P. and Verspagen, B. (2002). Knowledge spillovers in europe: a patent citations analysis. *The Scandinavian journal of economics*, 104(4):531–545.

Mehra, A., Kilduff, M., and Brass, D. J. (2001). The social networks of high and low self-monitors: Implications for workplace performance. *Administrative science quarterly*, 46(1):121–146.

Melton, J. (1988). *Absolutism and the Eighteenth-century Origins of Compulsory Schooling in Prussia and Austria*. Cambridge: Cambridge University Press.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American sociological review*, pages 635–659.

Meyer, J., Tyack, D., Nagel, J., and Gordon, A. (1979). Public education as nation-building in america: Enrollments and bureaucratization in the american states, 1870-1930. *American Journal of Sociology*, 85:591–613.

Michalopoulos, S. and Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica*, 81:113–52.

Michel, J. and Bettels, B. (2001). Patent citation analysis. a closer look at the basic input data from patent search reports. *Scientometrics*, 51(1):185–201.

Milligan, K., Moretti, E., and Oreopoulos, P. (2004). Does education improve citizenship? evidence from the united states and the united kingdom. *Journal of Public Economics*, 88:1667–95.

Mitchell, B. (2007a). *International Historical Statistics: Europe, 1750-2005*. London: Palgrave Macmillan.

Mitchell, B. (2007b). *International Historical Statistics: The Americas, 1750-2005*. London: Palgrave Macmillan.

Moehling, C. (1999). State child labor laws and the decline of child labor. *Explorations in Economic History*, 36:72–106.

Mohr, R. D. (2002). Technical change, external economies, and the porter hypothesis. *Journal of Environmental economics and management*, 43(1):158–168.

Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press.

Mokyr, J. (2005). The intellectual origins of modern economic growth. *The Journal of Economic History*, 65(02):285–351.

Morrisson, C. and Murtin, F. (2009). The century of education. *Journal of Human Capital*, 3:1–42.

Muller, C. (2007). *Histoire de la Structure, de la Forme et de la Culture Scolaires de l'enseignement Obligatoire à Genève au XXe Siècle (1872-1969)*. PhD thesis, University of Geneva.

Naidu, S. (2012). Suffrage, schooling, and sorting in the post-bellum us south. *NBER WP 18129*.

Narin, F. and Breitzman, A. (1995). Inventive productivity. *Research Policy*, 24(4):507–519.

Neuhoff, K. (2005). Large-scale deployment of renewables for electricity generation. *Oxford Review of Economic Policy*, 21(1):88–110.

Newman, M. E. (2001a). Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131.

Newman, M. E. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.

Newman, M. E. (2001c). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Nguyen, B. and Nielsen, K. (2010). The value of independent directors: Evidence from sudden deaths. *Journal of Financial Economics*, 89:550–567.

Noailly, J. and Shestalova, V. (2013). Knowledge spillovers from renewable energy technologies: Lessons from patent citations. Technical Report 22, Centre for International Environmental Studies, The Graduate Institute.

O Buachalla, S. (1988). *Education Policy in Twentieth Century Ireland*. Dublin: Wolfhound Press.

Oelkers, J. (2009). Schillers schulen, lecture at the municipal museum of ludwigsburg, november 6, 2009. http://paed-services.unizh.ch/user_downloads/1012/Ludwigsburg.pdf.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science*, 58(6):1122–1140.

Otto, V. M., Löschel, A., and Dellink, R. (2007). Energy biased technical change: a cge analysis. *Resource and Energy Economics*, 29(2):137–158.

Otto, V. M., Löschel, A., and Reilly, J. (2008). Directed technical change and differentiation of climate policy. *Energy Economics*, 30(6):2855–2878.

Packalen, M. and Bhattacharya, J. (2012). Words in patents: Research inputs and the value of innovativeness in invention. Technical report, National Bureau of Economic Research.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Patacchini, E. and Zenou, Y. (2012). Juvenile delinquency and conformism. *Journal of Law, Economics, and Organization*, 28(1):1–31.

Patnam, M. (2011). Corporate networks and peer effects in firm policies. Technical report, Working Paper.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics*, 87(2):308–322.

Perry-Smith, J. E. and Shalley, C. E. (2003). The social side of creativity: A static and dynamic social network perspective. *Academy of management review*, 28(1):89–106.

Persson, T. and Tabellini, G. (2000). *Political Economics: Explaining Economic Policy*. Cambridge, MA: MIT Press.

Popp, D. and Newell, R. (2012). Where does energy R&D come from? examining crowding out from energy R&D. *Energy Economics*, 34(4):980–991.

Popp, D., Newell, R., and Jaffe, A. (2009). Energy, the environment, and technological change. NBER Working Papers 14832, National Bureau of Economic Research, Inc.

Porter, M. E. and Van der Linde, C. (1995). Toward a new conception of the environment-competitiveness relationship. *The journal of economic perspectives*, 9(4):97–118.

Provasnik, S. (2006). Judicial activism and the origins of parental choice: The court. *History of Education Quarterly*, 46:311–47.

Rablen, M. D. and Oswald, A. J. (2008). Mortality and immortality: The nobel prize as an experiment into the effect of status upon longevity. *Journal of Health Economics*, 27(6):1462–1471.

Ramirez, F. and Boli, J. (1987). The political construction of mass schooling: European origins and worldwide institutionalization. *Sociology of Education*, 60:2–17.

Reagans, R. and McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly*, 48(2):240–267.

Reagans, R. E. and Zuckerman, E. W. (2008). Why knowledge does not equal power: the network redundancy trade-off. *Industrial and Corporate Change*, 17(5):903–944.

Redelmeier, D. A. and Singh, S. M. (2001). Longevity of screenwriters who win an academy award: longitudinal study. *BMJ: British Medical Journal*, 323(7327):1491.

Ricci, F. (2007a). Channels of transmission of environmental policy to economic growth: A survey of the theory. *Ecological Economics*, 60(4):688–699.

Ricci, F. (2007b). Environmental policy and growth when inputs are differentiated in pollution intensity. *Environmental and Resource Economics*, 38(3):285–310.

Riggs, K. R., Reitman, Z. J., Mielenz, T. J., and Goodman, P. C. (2012). Relationship between time of first publication and subsequent publication success among non-phd physician-scientists. *Journal of graduate medical education*, 4(2):196–201.

Ritter, G. (1986). Entstehung und entwicklung des sozialstaats in vergleichender perspektive. *Historische Zeitschrift*, 243:1–90.

Rivera, M. T., Soderstrom, S. B., and Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *annual Review of Sociology*, 36:91–115.

Rocha, R., Ferraz, C., and Soares, R. (2015). Human capital persistence and development. *mimeo, PUC-Rio.*

Rodan, S. and Galunic, D. C. (2004). More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, 25:541–556.

Rogers, E. (2003). Diffusion networks. *Networks in the knowledge economy*, pages 130–179.

Romer, P. M. (1990a). Endogenous technological change. *Journal of Political Economy*.

Romer, P. M. (1990b). Endogenous technological change. *Journal of political Economy*, pages S71–S102.

Rosen, S. (1981). The economics of superstars. *The American economic review*, pages 845–858.

Rosenblat, T. S. and Mobius, M. M. (2004). Getting closer or drifting apart? *The Quarterly Journal of Economics*, pages 971–1009.

Ruppenthal, J. (1920). English and other languages under american statutes. *American Law Review*, 54:39–90.

Rust, V. (1990). The policy formation process and educational reform in norway. *Comparative Education*, 26:13–25.

Salimova, K. and Dodde, N. (2000). *International Handbook on History of Education*. Moscow: Orbita-M.

Sampat, B. N., Mowery, D., Nelson, R. R., Rai, A., Schankerman, M., and Stern, S. (2005). Determinants of patent quality: an empirical analysis, working paper.

Scherer, F. M. (1965). Firm size, market structure, opportunity, and the output of patented inventions. *The American Economic Review*, 55(5):1097–1125.

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics*, 26(1):193–211.

Schmookler, J. (1966). *Invention and economic growth*, volume 26. Harvard University Press Cambridge, MA.

Schneider, R. (1982). Die bildungsentwicklung in den westeuropischen staaten 1870-1975. *Zeitschrift Soziologie*, 11:207–26.

Schneider, S. H. and Goulder, L. H. (1997). Achieving low-cost emissions targets. *Nature*, 389(6646):13–14.

Schriewer, J. (1985). Weltlich, unentgeltlich, obligatorisch.' konstitutionsprozesse nationaler erziehungssysteme im 19. jahrhundert. *Francia Forschungen zur westeuropischen Geschichte*, 13:663–74.

SCImago. (2007). Sjr – scimago journal & country rank. Retrieved August 12, 2015, from http://www.scimagojr.com.

Sefa, E. and Lushnje, A. (2012). Albanian women during monarchy (1928-1939). *International Journal of Humanities and Social Science*, 2:299–305.

Segerstrom, P. S. (1998). Endogenous growth without scale effects. *American Economic Review*, pages 1290–1310.

Shaffer, M. J. (2011). *Entrepreneurial Innovation: Patent Rank and Marketing Science*. PhD thesis, Washington State University.

Silva, S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Silverman, A. B. (2003). Duty to disclose prior art to the united states patent and trademark office. *JOM Journal of the Minerals, Metals and Materials Society*, 55:64–64.

Simola, H. (2002). From exclusion to self-selection: Examination of behavior in finnish primary and comprehensive schooling from the 1860s to the 1990s. *History of Education*, 31:207–26.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management science*, 51(5):756–770.

Slaje, J. (2009). *Bildungssysteme in Polen und Österreich im Vergleich Unter Besonderer Berücksichtigung des Bologna-Prozesses*. PhD thesis, Diploma Thesis, University of Vienna.

Small, H. and Griffith, B. C. (1974). The structure of scientific literatures i: Identifying and graphing specialties. *Science studies*, pages 17–40.

Smulders, S. and De Nooij, M. (2003). The impact of energy conservation on technology and economic growth. *Resource and Energy Economics*, 25(1):59–79.

Smulders, S. and Withagen, C. (2012). Green growth–lessons from growth theory. *World Bank Policy Research Working Paper*, (6230).

Sorenson, O. and Fleming, L. (2004). Science and the diffusion of knowledge. *Research Policy*, 33(10):1615–1634.

Sorenson, O. and Singh, J. (2007). Science, social networks and spillovers. *Industry and Innovation*, 14(2):219–238.

Soysal, Y. and Strang, D. (1989). Construction of the first mass education systems in nineteenth-century europe. *Sociology of Education*, 62:277–88.

Steinhilber, A. and Sokolowski, C. (1966). *State Law on Compulsory Attendance. U.S. Department of Health, Education, and Welfare*. Circular No. 793. Washington, D.C.: GPO.

Stephens, M. and Yang, D. (2014). Compulsory education and the benefits of schooling. *American Economic Review*, 104:1777–92.

Stolze, W. (1911). Friedrich wilhelm i. und die volksschule. *Historische Zeitschrift*, 107:81–92.

Stuart, T. E. and Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1):21–38.

Terza, J., Basu, A., and Rathouz, P. (2008a). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27:531–43.

Terza, J., Bradford, W., and Dismuke, C. (2008b). The use of linear instrumental variables methods in health services research and health economics: A cautionary note. *Health Services Research*, 43:1102–20.

MINISTRO DOS NEGOCIOS DO REINO (1835). *Regulamento Geral da Instrucção Primaria*. Lisbon: Government Publication.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, pages 450–460.

Torgler, B. and Piatti, M. (2011). A century of american economic review.

Torvik, V. I. and Smalheiser, N. R. (2009). Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3):11.

Torvik, V. I., Weeber, M., Swanson, D. R., and Smalheiser, N. R. (2005). A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology*, 56(2):140–158.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, pages 172–187.

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, 5(1):19–50.

Tyack, D. (1976). Ways of seeing: An essay on the history of compulsory schooling. *Harvard Education Review*, 46:355–89.

UNESCO (2007). International bureau of education–world data on education 6th edition. http://www.ibe.unesco.org/en/services/online-materials/world-data-on-education/sixth-edition-2006-07.html.

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review*, pages 674–698.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly*, pages 35–67.

Uzzi, B. and Spiro, J. (2005). Collaboration and creativity: The small world problem1. *American journal of sociology*, 111(2):447–504.

Van Dalen, H. P. and Klamer, A. (2005). Is science a case of wasteful competition? *Kyklos*, 58(3):395–414.

Van Pottelsberghe, B., Denis, H., and Guellec, D. (2001). Using patent counts for cross-country comparisons of technology output. Technical report, ULB–Universite Libre de Bruxelles.

Veefkind, V., Hurtado-Albir, J., Angelucci, S., Karachalios, K., and Thumm, N. (2012). A new epo classification scheme for climate change mitigation technologies. *World Patent Information*, 34(2):106–111.

Verdolini, E. and Galeotti, M. (2011). At home and abroad: An empirical analysis of innovation and diffusion in energy technologies. *Journal of Environmental Economics and Management*, 61(2):119–134.

Waldinger, F. (2010). Quality matters: The expulsion of professors and the consequences for phd student outcomes in nazi germany. *Journal of Political Economy*, 118(4):787–831.

Waldinger, F. (2012). Peer effects in science: Evidence from the dismissal of scientists in nazi germany. *The Review of Economic Studies*, 79(2):838–861.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Watson, J. (2012). *The double helix*. Hachette UK.

Weber, E. (1976). *Peasants into Frenchmen: The Modernization of Rural France, 1870-1914*. Stanford, CA: Stanford University Press.

Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, pages 331–360.

Wellman, B. and Berkowitz, S. D. (1988). *Social structures: A network approach*, volume 2. CUP Archive.

West, M. and Woessmann, L. (2010). 'every catholic child in a catholic school': Historical resistance to state schooling, contemporary private competition and student achievement across countries. *Economic Journal*, 120:F229–55.

Wielemans, W. (1991). Comprehensive education in belgium: A broken lever? *European Journal of Education*, 26:167–78.

Wines, E. (1851). The biblioteca sacra and american biblical repository. Volume 8.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Mass.: MIT Press.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Zapf, W. and Flora, P. (1973). Differences in paths of development: An analysis for ten countries. *Building States and Nations*, 1:161–211.

Zucker, L. G. and Darby, M. R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences*, 93(23):12709–12716.