The London School of Economics and Political Science
Department of Economics

Essays on Communication, Social Interactions and Information

Diego Ezequiel Battiston

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy, London, June 2018

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 57,292 words.

Statement of conjoint work

I hereby declare that Chapter 1 was jointly co-authored with Jordi Blanes i Vidal and Tom Kirchmaier. Chapter 3 was jointly co-authored with Jordi Blanes i Vidal. I contributed 33% to the work in Chapter 1 and 50% to the work in Chapter 3. I am aware of the London School of Economics and Political Science Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis and have obtained permission from each of the co-author(s) to include the above material(s) in my thesis.

Acknowledgements

This thesis condenses many years of my work as a PhD student at LSE. However, many people contribute to make it possible.

I am deeply indebted to my supervisor Alan Manning for his valuable guidance and thoughtful insights, but also for his encouragement and understanding during these years. I am also extremely grateful to Oriana Bandiera for her generous support, her guidance and for having always wise words of advise.

I own part of my training as an academic to Jordi Blanes i Vidal from whom I learned many non-written rules and insights about how to do research. I will be always grateful to him for having trusted me as a colleague and teammate during these years.

Marcos Vera-Hernandez deserves a special paragraph in the acknowledgments. He has contributed to my research with his valuable academic advise but also by cheering me up and offering me his help whenever I needed it.

I am also grateful to the labour group at LSE, specially Steve Pischke and Guy Michaels for giving me incredible comments and suggestions on how to improve my research. I am also indebted to Tom Kirchmaier who allowed me to be part of his team. Tom's commitment, effort and support has been key for the first chapter of this thesis. Special thanks also to Chief Constable Ian Hopkins QPM, and to Steven Croft, Peter Langmead-Jones, Duncan Stokes, Ian Wiggett and many others at the Greater Manchester Police who contributed to make the project possible. Chapter two has been possible due to the generous help of Matt Nelson and the Minnesota Population Center team, Patricia MacFarlane, Nigel Rogers and Nic Warner who generously devoted time and resources to allow me obtaining and accessing the data for the study.

I am indebted to an incredible group of friends who have been an essential part of my life in London, especially Miguel Espinosa, Nicola Limodio, Maddalena Liverani, Stephan Maurer, Federico Rossi and Francesco Saninno. I feel immensely lucky to have shared all these years with them. Also a big thank you to Andres Barrios, Monica Langella, Maria Molina-Domene and Katalin Szemeredi for the all

the enriching discussions and for the many coffees and dinners.

My eternal gratitude to my parents Roberto and Nancy, my siblings Maria de Lourdes and Emanuel and my brother in law Marcos who always believed in my project. Thanks for staying close to me, no matter the distance. Without their unconditional love and support, I could never have done it. I am also indebted to Roberto, Graciela and my second family from Cipolletti who backed me and encouraged me in this journey from the beginning.

Finally, but most importantly, the biggest thanks to Soledad, my mainstay during all these years. Thanks for being behind every step and for holding me up through all the ups and downs of the PhD. This thesis is dedicated to her.

Abstract

This thesis consists of three papers in the broad field of Applied Economics. I focus on three "soft factors", namely, face-to-face communication, brief social interactions and information updates. I study on how they affect individual and organisational outcomes using different natural experiments. The first chapter provides causal evidence on how the ability to communicate face-to-face (in addition to electronic communication) can increase organisational performance. The study exploits a natural experiment within a large organisation where workers must communicate electronically with their teammates. A computerized system allocates the tasks to workers creating exogenous variation in the co-location of teammates. Workers who share the same room, can also communicate in person. The main findings are that face-to-face communication increases productivity and that this effect significantly varies across tasks, team characteristics and working environments. In the second chapter I construct a novel dataset of immigrants and ships arrived to the US in the early 20th century to study the effects of brief social interactions and their persistence over time. The chapter shows that individuals travelling (during few days) with shipmates that have better connections in the US, have higher quality jobs. Several findings are consistent with the mechanism whereby individuals get information or access to job opportunities from their shipmates. The study highlights the importance of social interactions with unknown individuals during critical life junctures. It also suggests that they are more relevant for individuals with poor access to information or weak social networks. The third chapter shows that executions cause a local and temporary reduction in serious violent crime. The interpretation of this result follows from a theoretical framework connecting information updates with the increasing 'awareness' of individuals about the consequences of crime. Consistently with the predictions of the model, the study finds that effects are stronger when media attention is high and lower in places with high propensity to apply the death penalty.

Contents

List of Tables

List of Figures

Chapter 1

Face-to-Face Communication in Organisations

1.1 Introduction

Workers in teams typically need to communicate effectively with each other, especially when dealing with tasks that are urgent and complex. While a lot of attention has been devoted to understanding the effects of team incentives (Burgess et al. 2010, Bandiera et al. 2013, Friebel et al. 2017) or team composition (Hamilton et al. 2003, Hjort 2014, Lindquist et al. 2017) on performance, the central issue of team communication has been empirically neglected. A necessary step in this direction consists of understanding the causal relation between (access to more) communication and team productivity inside organisations. Unfortunately, even this first step has been impeded by measurement and endogeneity concerns. There are well-known difficulties in gaining access to data on the internal operations of organisations, especially when these are sophisticated enterprises. Yet, without unusually rich data it is not possible to measure communication between teammates. Secondly, the organisational communication infrastructure is typically the result of an efficiency-maximizing decision process, prompting often insurmountable endogeneity concerns.

This paper overcomes these issues by taking advantage of an extremely rich

dataset and a unique natural experiment in a large and complex public sector organisation. In our setting, individuals working in teams are always able to communicate electronically. Some teams, exogenously chosen by a computerised system allocating tasks to workers, can also communicate in person. Therefore, our experiment is best interpreted as identifying the value of communicating face-to-face, in addition to electronically.

Our paper has three objectives. Firstly, we provide the first evidence on a causal link between the ability to communicate face-to-face and team productivity inside organisations. Secondly, we document substantial heterogeneity in the size of this relation. In particular, the ability to communicate face-to-face is more valuable to teams that are demographically homogenous, have experience of working together, face high pressure, and deal with urgent and information-intensive tasks. In contexts where encouraging face-to-face communication is costly, this finding suggests that managers should condition such investments on the nature of the tasks, workers and production environments. Thirdly, we seek to understand and measure the operational costs of communication. In our context, these costs arise from workers being slower to undertake new tasks when they spend time communicating face-to-face on existing tasks. By contrast, we find no displacement of attention away from other tasks that workers are contemporaneously handling.

This Study The setting is the branch in charge of answering 999 calls (the equivalent of the US 911) and allocating officers to incidents in the Greater Manchester Police. An incoming call is answered by a *call handler*, who describes the incident in the internal computer system. When the handler officially creates the incident, its details become available to the *radio operator* responsible for the neighbourhood where the incident occurred. The radio operator then allocates a police officer on the basis of incident characteristics and officer availability. The main metric of performance is the time that it takes for the operator to allocate an officer.[1] Often, delays result from the radio operator's need to gather additional information, which she can do through a variety of channels including communicating with the call handler

[1]We describe this measure in detail in Section 1.2. There, we also list its advantages and potential limitations and explain why the organisation assigned high importance to this measure during our sample period.

in person.

To identify the importance of face-to-face communication we exploit both a natural experiment and highly detailed information throughout the production process. In the Greater Manchester Police, handlers and operators are spread across four rooms, each in a separate part of Manchester. Each room contains the radio operators responsible for the surrounding neighbourhoods as well as a subset of the call handlers, who can take calls from anywhere in Manchester. This arrangement implies that, for some incidents, an operator reads the information inputted in the system by a handler located in the same room. For other incidents, the information will instead have been entered by a handler based in another location. A direct consequence of co-location is that it allows the two teammates, handler and operator, to communicate face-to-face if they wish to do so.[2]

We first exploit the fact that the computerised queuing system matching incoming calls to newly available handlers creates exogenous variation in the co-location of handler and operator. Our baseline finding here is that allocation time is 2% faster when handler and operator work in the same room.[3] This improvement is not at the expense of observable dimensions of the quality of the allocation, such as the seniority of the officer sent. We also show that proximity *within the room* is important - the effect of co-location is twice as high when handler and operator are sitting close together. In fact, allocation time is lower even when *the same pair of workers* are located inside the room closer together. This last finding rules out unobservable characteristics in the match between handler and operator (correlated with co-location) as the explanation for the baseline findings. We provide additional evidence in this respect with a placebo test that exploits an organisational restructure that altered the regular workplaces of handlers and operators.

Having identified the causal effect of co-location on productivity, we proceed to

[2]The alternative to face-to-face communication is further electronic communication. See Section 1.2 for details.

[3]Although not large, this effect compares well with typical annual productivity increases in the public sector (Simpson, 2009). Another comparison is with the effect of introducing team performance pay in the field experiment of Friebel et al. (2017), which they find to be 3%. The effect in our study is twice as large for urgent and information-intensive tasks, among others. At the police force level the baseline effect adds up to approximately 900 hours per month, a substantial magnitude.

establish face-to-face communication as the primary explanatory mechanism. Unsurprisingly, our organisation did not record any information transmitted through informal in-person interactions between co-workers, and therefore we are not able to use these informal messages here. Instead, we provide a set of complementary tests. Firstly, we use several proxies to show that the quality of the handler's *electronic* communication is not higher when a co-located operator will be reading the incident's description. Secondly, we show that operators do not assign higher priority to co-located incidents, at the expense of other contemporaneous incidents. These two findings are counter to the most natural channels (alternative to face-to-face communication) through which co-located teammates could increase productivity.

In addition to results inconsistent with other channels, we also find evidence in favour of the face-to-face communication channel. We do this by examining the behaviour of the handler *after* officially creating the incident. Under the face-to-face communication mechanism, the handler then spends time talking to the operator, which temporarily prevents her from being available to take new calls. Alternative mechanisms, such as better electronic communication by the handler or higher operator effort, do not naturally have that prediction. We show that handlers spend more time 'unavailable' to take new calls following the creation of co-located incidents, and we interpret this as strong evidence that they are communicating with their operators in these incidents.

The second objective of the paper is to uncover conditions under which face-to-face communication is particularly important. We find first that co-location increases productivity more for incidents that are more information-intensive. This is reassuring, in that it is consistent with the notion that having access to an additional communication channel is valuable particularly when more information needs to be transmitted. The effect of co-location is also higher for more intrinsically urgent incidents, as well as during periods when operators face a higher incident workload. These last two findings are consistent with each other, in that they both suggest that operators facing higher time pressure benefit most from being able to gather information through an additional quick, informal channel. Lastly, we investigate the characteristics of the teams associated with a higher effect of co-location on productivity. We provide three separate but mutually consistent results: teams of

the same gender, similar age, and with a longer history of working together benefit more from co-location. Together, the three findings indicate that the ability to communicate face-to-face benefits more teams that are more cohesive, because of either demographic traits or a common, shared, experience.

The third objective of the paper is to identify and highlight the operational costs of face-to-face communication. As mentioned earlier, we do not find that operators distort their attention towards co-located incidents and at the expense of other contemporaneous incidents. Negative spillovers of this type do not therefore seem to be present in our setting. However, we do find that handlers spend more time unavailable to take new calls after creating co-located incidents. This clearly imposes a delay on incoming calls whenever the queue of incoming calls is not empty. In other words, communicating face-to-face has an opportunity cost whenever the organisation has no *slack*. We provide a simple theoretical framework and a set of tests to quantify this cost in our organisation. Empirically, we find that the cost is very small, relative to the benefits of face-to-face communication. As expected, the cost is however higher when the number of on-duty handlers is low relative to the number of incoming calls (i.e. when there is less organisational slack).

Contribution This paper provides, we believe, the first causal evidence on the relation between (face-to-face) communication and team productivity inside organisations. As is common in organisational economics, the study involves a particular setting and production technology.[4] As such, the implications are stronger for high pressure environments such as the healthcare professionals assessing and treating patients in emergency rooms, or the frontline staff and their supervisors in air traffic control, the military, and other time-critical settings.

More generally, the results on the *contingent* value of face-to-face communication have broader applicability. For instance, the results regarding the urgency and information-intensity of tasks indicate the type of production environments where investments encouraging communication are likely to be particularly valuable. Equally significant is the finding that homogeneous teams benefit more from

[4]Other recent papers using data from a single organisation include Bandiera et al. (2010), Bloom et al. (2015) and Chan (2016).

being able to talk to each other. We briefly mention in the conclusion a number of policy prescriptions based on this finding.

Lastly, the insights on the operational (opportunity) costs of communication are of general validity. Of course, increasing communication in the workplace is likely to be associated in many contexts with fixed costs (such as capital, estate or traveling costs) that we do not have the information to analyse here. The costs that we focus on are operational and arise from the fact that every second spent communicating cannot be devoted to other activities. This is a general trade-off, as is the idea that this opportunity cost depends on the alternative use of workers' time and therefore on the amount of slack in the organisation. While our setting is not unusual in the existence of this trade-off, it is unusual in that the highly structured nature of the production process and the granularity of the dataset allow us to estimate it empirically.

Related Literature Despite its importance, field evidence on communication in organisations is scant. Gant et al. (2002) argue that the adoption of innovative HRM practices induces more communication among co-workers. Palacios-Huerta and Prat (2012) use email exchanges to generate a measure of the relative importance of individual managers. Bloom et al. (2014) investigate whether firms adopting technologies such as data intranets altered their spans of control and autonomy levels. None of these papers explore effects on productivity, as we do.

By contrast, a large body of work investigates whether communication affects team performance in laboratory experiments. Early research, typically by psychologists, focused on the shape of the communication networks (Bavelas and Barrett 1951, Leavitt 1951, Guetzkow and Simon 1955). Later on, Weber and Camerer (2003) study how productivity-enhancing languages emerge and are disrupted during mergers. Cooper et al. (1992) and Blume and Ortmann (2007) show that pre-play communication about strategies increases efficiency in weak-link coordination games. An advantage of laboratory experiments is that informal messages between subjects can be observed, something that is much more difficult in real organisations.[5] It is of course unclear how results from the laboratory extrapolate

[5]In our study, we can (partially) observe the electronic messages between teammates, but not

to the field.

The experimental variation in this paper relates to the co-location of teammates. This suggests a link with Catalini (2016), who uses the relocation of departments in a French university to analyse how search costs and monitoring costs vary with physical proximity between academics.[6]. Another related paper is Bloom et al. (2015), who find that working from home increased productivity in a Chinese call centre. The contrast with our finding that co-location increases productivity is likely the result of the many differences between the two settings. A very important one is the complexity of the production process. While the simple individual production of Bloom et al. (2015) can be easily monitored and co-ordinated remotely, we show that, in organisations requiring tight co-ordination between colleagues, working in the same place may have significant advantages, especially when tasks are relatively information-intensive.

Plan We describe the institutional setting in Section 1.2. We introduce the data and the empirical strategy in Section 1.3. We present the main results of the paper in Section 1.4. In Section 1.5, we provide evidence in support of the face-to-face communication mechanism. Section 1.6 explores the heterogeneity of the main results. In Section 1.7, we provide a cost-benefit analysis of the face-to-face communication effect. Section 1.8 concludes.

1.2 Institutional Setting

We exploit a natural experiment in the Operational Communications Branch (OCB) of the Greater Manchester Police (GMP). The OCB is the unit in charge of answering 999 calls from members of the public and managing the allocation of officers to the corresponding incidents. Figures 1.1 and 1.2 provide a simplified visualisation of

the face-to-face messages.

[6]A large body of work examines the relation between geographical proximity, assumed to facilitate face-to-face interactions, and the diffusion and generation of knowledge (Jaffe et al. 1993, Thompson and Fox-Kean 2005) A challenge here is to disentangle geographical distance from other factors, such as knowledge or social distance, correlated with it. In addition, the typical bird's-eye view of these papers does not allow for the isolation of mechanisms explaining why geographical proximity matters.

this production process.

Call Handler Emergency calls requesting the police are allocated to call handlers using a standard computerised queuing system. A result of the system is that any handler can respond to calls from any Manchester location.

The handler questions the caller, assigns an opening code and a grade level, and records any information deemed relevant. The grade level can range from one to three and, very coarsely, determines the official urgency of an incident. The opening code describes, horizontally and at a fairly detailed level, the type of issue that the incident relates to (neighbour dispute, disturbance in licensed premises, etc.). The description of the incident will include information on the individuals involved, their states of mind, the existence of prior history between these individuals and the likelihood of further incidents in the near future.[7]

All the information above is recorded in GMPICS, a specialised IT package used throughout the GMP to create, record and manage incidents.[8] The handler ticks a box in GMPICS to officially create the incident, and then indicates her status as 'not ready' (which allows the handler, among other things, to step away from his desk), or instead 'ready to receive new calls'. Under the 'ready' status, a call can arrive at any point and must immediately be answered by the handler. Once an incident has been created, the handler cannot keep adding details to it.

Radio Operator When an incident is created, it immediately appears on the computer screen of the radio operator overseeing the Manchester subdivision where the incident occurred. The allocation of incidents to radio operators is deterministic, since at any point in time there is only a single operator in charge of a specific subdivision (a corollary of this is that handlers do not decide to which operator

[7]The language used in these descriptions is highly efficient, as it includes a large number of official and unofficial abbreviations for features of incidents that appear repeatedly. For instance, official abbreviations include A/ABAN (apparently abandoned) and NFA (no fixed abode). Unofficial but widely used abbreviations include XXX (very drunk). Despite this, the written descriptions inevitably fail to perfectly communicate the full richness of the information gathered by the call handler.

[8]Our personal conversations with multiple handlers, radio operators and their supervisors indicate that GMPICS is widely regarded as an efficient system. GMPICS was developed in-house and incrementally over more than two decades. OCB staff receive extensive training and accumulate considerable expertise in its use.

they assign an incident). Radio operators are in charge of processing the information inputted by the handler and allocating police officers to incidents, on the basis of incident characteristics and officer availability.

Lacking a direct link with the caller, the radio operator has to rely on the information recorded by the handler in GMPICS. It is, however, often the case that additional information is needed before an officer can be allocated. For instance, written descriptions of incidents are regarded by radio operators as lacking sufficient emotional content, which makes it harder to understand the state of mind of the victim and the impact that the incident has had on it. Similarly, a full characterisation of the physical surroundings where the incident occurred, or of the complex relationships between the people involved are often difficult to communicate in writing. A complete picture of the incident is often necessary to efficiently match incidents with officers, advise the attending officer of important details that she may find at the scene, or even understand the level of priority that the incident merits.[9]

The additional information can be acquired by conducting targeted searches on specific individuals or addresses in the GMP databases, asking the call handler or contacting the initial caller directly. Typically, the allocation of an officer will be delayed until the radio operator can gather this information.

Teamwork In this paper our definition of a team comprises the combination of the call handler and the radio operator. While officially equal in rank, the positions of call handler and radio operator are associated with different status within the OCB. This stems from the fact that the job of radio operator is both more complex and more stressful, as it involves carrying out a variety of tasks in parallel and bearing the ultimate responsibility for the outcomes of incidents. The decision-making authority of radio operators is also wider. For instance, they can overrule the code and grade allocated by the handler (although this is in practice rare). Accordingly, radio

[9]Regarding the optimal matching between incidents and officers, note for instance that some incidents can be responded alternatively by sworn police officers or by PCSOs (police community support officers) and the likelihood that the more extensive legal powers and expertise of police officers may be needed is decision-relevant information. Similarly, incidents involving vulnerable individuals require officers with specialist training, which makes it critical to understand the condition of the caller and other individuals affected. More generally, certain officers are particularly well-suited to dealing with specific types of incidents or individuals.

operators earn a higher salary and have on average more experience in the OCB. Many in fact transferred into radio operations from the call handling desk, a move widely seen in the organisation as a promotion.

Face-to-Face Communication When a radio operator regards the electronic description of an incident insufficient, an efficient and fast way to gather this information is to ask the handler in person. Alternatively, it is often the handlers who decide to complement the written description with additional information delivered face-to-face. When handler and operator are communicating in person, the handler will need to be in 'not ready' status, as she may otherwise be forced to abruptly end the conversation when a new call arrives.

Our conversations with members of the OCB suggest that they attach several advantages to face-to-face communication: firstly, it is a highly efficient channel, in that it allows for rapid, short exchanges that provide immediate feedback to both teammates. Secondly, non-verbal cues can help to communicate fuzzy concepts that in writing would require lengthy descriptions. Thirdly, it is a more natural vehicle for the use of colloquialisms that can succinctly and effectively communicate characteristics of an incident including the physical or mental condition of the individuals involved. For a variety of reasons (including both the potential for misunderstanding and the possibility of future audits of the official GMPICS descriptions) these colloquialisms are less likely to be used in written communication.

Note that face-to-face communication has two features that electronic communication lacks. Firstly, it is oral. Secondly, handler and operator are able to observe each others' faces. Because communication by phone is not a realistic alternative in our setting, we are unable to precisely disentangle which of the two features is responsible for the increase in productivity.[10]

[10]Sending an *electronic message* to the handler is possible in the GMPICS system, although of course the operator then has to wait for the handler's electronic response. This response may not be immediate in the same way that emails are often not immediately answered in standard office environments. Unfortunately, our very rich data does not include information on these potential electronic exchanges. Communicating on the *phone* is theoretically possible but in practice unlikely, as a handler in status 'ready to take new calls' cannot be contacted on the phone without first alerting the handler's supervisor. On the other hand, a handler can easily switch status from 'ready' to 'not ready' if an operator approaches in person with the need to clarify some doubt.

Co-Location In the period between November 2009 and January 2012, OCB staff were spread across four buildings or 'rooms', each in a different part of Manchester: Claytonbrook, Leigh, Tameside and Trafford. Every room accommodated the radio operators overseeing the surrounding subdivisions (Figure 1.3 displays the areas overseen from each of the four locations). As discussed earlier, call handlers were not geographically specialised. However, for historical reasons they were also dispersed across the four locations. This assignment meant that radio operators would sometimes be reading the descriptions of incidents created by same room handlers, while on other occasions the handlers were based in a different part of Manchester.

In January 2012, a major reorganisation of the OCB reassigned all handlers to a single location (Trafford), while radio operators were divided between Claytonbrook and Tameside. This put an end to the natural experiment that we study here.

Measures of Performance As is the case with other public sector organisations (Dewatripont et al., 1999), objectives in the GMP are multifaceted and often vague. The prevention of harm or damage to property, the satisfaction and reassurance of the public, and the application of sufficient but proportionate force are all important objectives that escape precise measurement. Capturing every one of these objectives with explicit measures of performance is therefore an impossible task. Our first measure of performance is the allocation time of an incident: the time elapsed between its creation by the call handler and the allocation of an officer by the radio operator. We also study the effect of distance on response time: the time between creation and the officer reaching an incident's scene.[11]

The two measures that we use are undoubtedly partial. They do not capture, for instance, any notion of whether the 'right' officer was allocated to an incident, or whether the attending officer was in possession of all the relevant information prior to arrival. They also do not indicate whether or not excessive or insufficient

[11]Table 1.2 provides summary statistics for these and other variables. Note that these two measures are strongly correlated, since response time is equal to allocation time plus the officer's travel time. It is worth noting that better information on the part of the radio operator could affect travel time also. Imagine, for instance, a radio operator deciding whether to allocate the closest officer, or an officer who is further away but has a specialised skill. Better information could reveal that the incident does not require the specialised skill, and that the officer with the shorter travel time can be safely allocated.

resources were allocated to resolve an incident.

The two measures are nevertheless very important for the organisation that we study, for two main reasons. The first reason is that the GMP is partly evaluated on the basis of these variables. Specifically, nation-wide numerical targets for maximum allocation and response times were introduced by the UK Home Office in 2008.[12] The second reason is that these measures are regarded as important determinants of the public's satisfaction. UK-wide survey evidence suggests that response time is one of the most important variables predicting citizens' satisfaction with the police forces (Dodd and Simmons, 2002/03).

Table 1.1 provides direct evidence of this in our setting. In the GMP, a subset of callers is regularly questioned about their satisfaction with the treatment they received, after their incident has been closed. We obtained these surveys and linked the response time in our dataset with the answers to the two most important questions (our dataset is described in detail in the next section). Table 1.1 shows that there are very strong correlations between these variables. For instance, in incidents where police response time was below the maximum target prescribed by the Home Office, satisfaction was .14 standard deviations higher. Callers were also more likely to report that their opinion of the police had improved. The effects of response time on satisfaction are not linear, but instead concentrated at the top end of the response time distribution (Figure 1.4).[13]

Overall, there is substantial evidence that the leadership of the GMP internalised the need for minimising allocation and response times. One example can be found in the GMP Incident Response Policy manual April 2011. Allocation and response times are the only tactical performance measures mentioned in the manual. In

[12]For Grade 1 crimes, for instance, these targets were for a maximum of two minutes and fifteen minutes for allocation time and response time, respectively. The equivalent targets for Grade 2 (respectively Grade 3) were 20 and 60 minutes (respectively 120 and 240 minutes). While these targets were nominally scrapped in June 2010, police forces continued to regard them as objectives and to believe that they were being informally evaluated on this basis (Curtis, 2015). Information on response times was also frequently discussed in the reports produced by the HMIC (the central body that in the UK regulates and monitors police forces). For an example, see HMIC (2012).

[13]While we do not claim that these coefficients can be interpreted as causal effects, they suggest at the very least the type of evidence on which the GMP based their decisions. Unfortunately, we are unable to use the victim satisfaction variables as dependent variables in the main analysis of the paper. The number of survey responses is relatively low and it mostly falls outside our baseline sample period.

particular, this indicates that:[14]

> *The OCB will produce daily reports regarding graded response performance. This will include the % of incidents resourced within target and the % attended within target for each division. This will enable ongoing analysis of the accuracy of the resource management of that BCU.*

1.3 Empirical Strategy

In this section we present and discuss the dataset and main variables of the paper. We also first explain the empirical strategy to estimate the effect of co-location on performance, and then justify it with a set of balancing tests. Establishing such a causal effect is not an easy task. In addition to exploiting the idiosyncratic allocation of incidents to handlers, which we outline in this section, we will need to consider the possibility that co-location represents a proxy for unobserved characteristics of the handler, or handler/operator pair. We postpone the discussion of these confounding effects, together with the tests that we use to evaluate them empirically, to Section 1.4.

Dataset Our baseline dataset contains every incident reported through the phone to the GMP between November 2009 and December 2011. We restrict our attention to incidents where the handler allocated the call a grade below or equal to three, therefore transferring responsibility to a radio operator rather than to a divisional commander. For every incident we observe, among others, the allocation and response time, the location of the incident, the grade and (horizontal) opening code, the identity of the call handler and radio operator, and the desk position from which the handler took the call. The dataset was made available to us under a strict confidentiality agreement.

[14] Additional examples include the following. The launching in April 2010 of a website where the public could access up-to-date statistics on response times, separately for each of the twelve divisions (Pilling, 2010). Secondly, the fact that throughout our sample period every report by the GMP to the Manchester City Council Citizenship and Inclusion Overview and Scrutiny Committee provided detailed statistics on response times and, if these were deemed unsatisfactory, a list of reasons for the failure.

Table 1.2 provides basic summary statistics for the main variables in our study. Note first that our sample size is very large, as it includes close to one million incidents. In around one in four observations the handler and operator are in the same room. The performance variables are highly right-skewed. For response, for instance, the median time is 19 minutes, while the average time is more than four times larger.[15]

We find that there is considerable gender and age variation among handlers and operators. Consistently with our earlier discussion of the differences in status, operators are significantly older than handlers. They are also more likely to be female, likely the result of females being more likely to regard the OCB as a long-term career choice.

Intuition of Empirical Strategy The computerised queuing system allocating calls to handlers works as follows. As calls come in, they join the back of a call queue. The system matches the call at the front of the queue with the next handler that becomes available. If the call queue is empty and several handlers start to become available, they form their own queue. The system then matches the handler at the front of the handler queue with the next incoming call. The system creates exogenous variation in the co-location of the handler and operator involved in an incident. We visualise this notion in Figures 1.5A and 1.5B where, for simplicity, we assume that there are only two locations (Trafford and Leigh), rather than four.

Assume that, within a relatively narrow time horizon, two calls (one from Trafford, one from Leigh) reach the queuing system, and that two handlers (one based in Trafford, the other in Leigh) become available. The exact timing at which handlers become available is the result of a large number of factors, including the length of their previous calls, the time at which the calls started, the existence and length of 'not ready' periods etc. Similarly, the exact order at which the calls arrive is the result of many factors, including the times at which the incidents occurred, the delay in dialling 999 and the further delay in opting for a police service and being

[15]The maximum value is more than 15 days, likely the result of some error in the classification of the incident. The fact that the left hand side variables in our regressions are in logarithmic form should dampen the effect of outlying observations. Nevertheless, in Appendix Table 1.A10 we show that our baseline estimates are robust to the exclusion of these outliers.

transferred to the GMP. These factors are arguably orthogonal to the factors determining the order at which handlers become available. It follows that two handlers that are on duty during the same time period should be equally likely to be the one assigned to an incoming call. If, as in Figure 1.5A, the handlers are assigned calls from a subdivision that their room oversees, they will be co-located with the radio operators with whom they have to communicate electronically. For arguably exogenous reasons, they may instead be assigned a call (and have to communicate with an operator) from a different area of Manchester. We capture this variation with the dummy variable *SameRoom*, which is the main independent variable in our study.

We have just argued that, conditional on the exact time period at which a call arrives, on duty handlers should be equally likely to be assigned that call. In practice, some rooms (for instance Trafford) are bigger than others (e.g. Leigh) and therefore contain a larger number of handlers. This implies that the likelihood of $SameRoom = 1$ will be mechanically higher if the call originates in a Trafford neighbourhood, relative to a Leigh neighborhood. Calls originating from Trafford and Leigh may also have different characteristics, which could independently affect their average allocation and response times. Therefore, our claim regarding the exogeneity of the variable *SameRoom* is only conditional on hour (i.e. year X month X day X hour of day) and (handler and operator) room fixed effects.[16]

Estimating Equation Our baseline estimating equation is:

$$y_i = \beta SameRoom_{j(i)k(i)} + \theta_{t(i)} + \lambda_{j(i)} + \mu_{k(i)} + \pi_{g(i)} + \gamma_{h(i)} + \mathbf{X_i} + \epsilon_i \qquad (1.1)$$

where y_i is a measure of OCB performance for incident i. Throughout our paper, allocation and response times are measured in log form, both for ease of interpretation of the coefficients and in the presence of right-skewness to minimise the effect of outlying observations. Consistently with our earlier discussion, we control for $\theta_{t(i)}$ (the fixed effect for the hour t at which the incident arrived) and $\lambda_{j(i)}$ and $\mu_{k(i)}$ (the fixed

[16]In most regressions, we use hour fixed effects to condition on the exact time period at which a call arrives. Our findings are qualitative unchanged if we instead control for the half-hour or quarter-hour period (see, for instance, Appendix Table 1.A7).

effects for the rooms j and k from which the incident was handled and dispatched). Our main independent variable of interest is the dummy $SameRoom_{j(i)k(i)}$, which takes value 1 when rooms j and k coincide.

We also control in our baseline specification for $\pi_{g(i)}$ and $\gamma_{h(i)}$ (the fixed effects for the individual handler g and operator h assigned to the incident) and by other incident characteristics (such as the assigned grade) included in the vector $\mathbf{X_i}$. These latter controls are not essential for identification, but should contribute to the reduction of the standard errors. We cluster these standard errors at the operator room and year/month level. In Appendix Tables A1 and A2 we show that the baseline findings are robust to the inclusion or exclusion of additional controls and to alternative clustering choices.

Balancing Tests Our first set of tests examines the balance of incident (grade, location of the incident scene), worker (gender, age, location of the desk, current workload) and room time-varying (measures of current average workload) variables across the co-location of handler and operator. To perform these tests, we separately regress each variable on $SameRoom$, after controlling for hour and room fixed effects:

$$x_i^s = \beta SameRoom_{j(i)k(i)} + \theta_{t(i)} + \lambda_{j(i)} + \mu_{k(i)} + \epsilon_i \tag{1.2}$$

where the variable x_i^s is a characteristic s of incident i, and the other variables are defined as above. To ease interpretation, non-binary dependent variables are standardised.

The results in Figure 1.6, where we label each row in the left axis by the regression dependent variable, plot the estimated confidence intervals of $SameRoom$. To illustrate the need for our empirical strategy, we report for every variable the estimates of two regressions: with and without the hour and room controls. We find first that $SameRoom$ is (unconditionally) strongly correlated with incident characteristics: the estimates are large and most are statistically significant. The introduction of the hour and room controls, however, greatly decreases both the standard errors and the estimates, which then become extremely small in magnitude. For instance, among the non-binary variables all the estimated coefficients

imply an effect of $SameRoom$ lower than .005 standard deviations of the dependent variable.[17] We also find that, after including the hour and room controls, almost all the incident characteristics are balanced with respect to the variable $SameRoom$.[18] The variables in Figure 1.6 do not include the incident opening code, an important determinant of allocation and response times. The opening code is captured empirically by a large set of dummy variables that are mechanically correlated with each other, which creates a mechanical correlation on the results of balance regressions based on equation (3.3). We therefore switch the dependent and independent variables, and estimate:

$$SameRoom_{j(i)k(i)} = \alpha_i + \theta_{t(i)} + \lambda_{j(i)} + \mu_{k(i)} + \epsilon_i \tag{1.3}$$

where α_i are the fixed effects for the incident opening code. We find that the F-statistic of joint significance of these effects is 1.15 (P-value = .30), suggesting that $SameRoom$ and the opening code dummies are conditionally uncorrelated. Overall, we interpret the results of estimating (1.3) and the regressions of Figure 1.6 as consistent with our assumption that co-location between the handler and operator of an incident is conditionally orthogonal to incident, handler, operator and room time-varying characteristics.

1.4 Baseline Results

In this section we present and interpret the baseline results of the paper. We then use a number of tests to confirm that these estimates can indeed be interpreted as

[17] We report the values of the coefficients and standard errors in Appendix Table 1.A13. Appendix Figure 1.A1 shows that it is the room controls that are critical to the empirical strategy. Failing to control for the hour of the incident does not lead to a stronger correlation between $SameRoom$ and the incident characteristics.

[18] Admittedly, two characteristics are not balanced at the 5% level. Note however that even for these characteristics the differences are extremely small in magnitude. The significant coefficient associated with the Grade 1 regression is both small in magnitude and *negative*, suggesting that same room incidents are slightly more likely to be allocated a low priority by the handler, and should therefore have *higher* allocation and response times. Furthermore, we find in estimating equation (1.3) that the type of incident (horizontally defined) is uncorrelated with $SameRoom$. We also find that all normalized differences of predetermined variables across co-location status (after partialling out hour and room fixed effects) are lower than .01. Imbens and Wooldridge (2009) highlight that unlike t-statistics, normalized differences do not depend mechanically on sample size. They suggest that values below 0.25 indicate good balance.

the causal effect of co-location on performance, rather than the result of co-location being a proxy for unobserved determinants of allocation and response time. We also explore whether the quality of the response is different for co-located incidents. The section concludes with an investigation of potential spillovers onto other (contemporaneous) incidents assigned to the radio operator.

Baseline Estimates Our baseline regressions are variations of equation (3.3). In the first two columns of Table 1.3 we find that allocation and response time are approximately 2% faster on average when handler and operator are located in the same room. At the mean (respectively, median) of the independent variable, this 2% translates into 76 seconds (respectively, 5.4 seconds) saved in terms of allocation time. For response time these savings are of 104 and 20 seconds, evaluated at the mean and median respectively (see Appendix Table 1.A14 for the results when the dependent variables are in levels (minutes)). Aggregated over all the incidents in a month, the savings amount to approximately 900 hours.

We also investigate whether these times are 'on target'. Throughout our sample period, it was an explicit objective of the UK Home Office that allocation and response times should typically be below certain levels.[19] As a result, the GMP recorded information on whether the target maximum time was exceeded for an incident. We use these dummies as dependent variables and find in Columns 3 and 4 that the likelihood of being on target is higher when $SameRoom = 1$. For instance, the coefficient in Column 3 indicates that the likelihood of missing the allocation target decreases by .4 percentage points (around 2% of the mean of .25), when handler and operator are co-located.

Lastly, we find in Column 5 no evidence of co-location affecting the likelihood that incidents classified as crimes are cleared by the GMP.[20]

[19]See Section 1.2 for details about these targets. The fact that the 'on target' dummies are affected by co-location confirms that the results are not disproportionately due to extreme values of the allocation and response time distributions. In terms of understanding whether the reduction in allocation and response times is uniform throughout their distribution, Section 1.6 and Table 1.8 show that it is more urgent incidents (i.e. incidents with shorter expected allocation times) that are more affected.

[20]The absence of a statistically significant effect on the likelihood of clearing the crime may be due to the fact that our sample size is much smaller in this regression, since only around 16% of incidents are crimes. Nevertheless, it is surprising given the findings of Blanes i Vidal and

Estimates by Distance Inside the Room Table 1.3 has established that co-location of handler and operator is associated with higher performance, relative to them working in rooms in separate areas of Manchester. We now investigate whether performance improves as distance decreases *even when handler and operator are already working in the same room.* In addition to providing richer evidence on the functional form of the relation between proximity and teamwork performance, within-room variation allows the introduction of handler/operator pair fixed effects in the regression. We argue in the next subsection that the introduction of these controls strengthens the credibility of our claim regarding the causal interpretation of the estimates.

The assignment of desks to workers was as follows. Inside a room, a fixed desk would be earmarked for the radio operator overseeing a specific subdivision. Handlers, on the other hand, were free to work from any remaining and available desk. To measure the within-room distance between desks, we use yearly-updated floorplans of the four OCB rooms (see Figure 1.7 for an example).[21] We set distance to zero if handler and operator are not in the same room, and add the interaction of distance and the same room variable to our baseline specification.

We provide two types of evidence. In Table 1.4 distance is measured parametrically, in logs. In Figure 1.8 we instead split distance into four categories of approximately equal sample size, and plot the interactions of *SameRoom* with these dummies. The estimates from both specifications indicate that teammates that sit closer together are more productive. In the parametric estimation, a 10% decrease in within-room distance is associated with a 2.6% increase in the effect of *SameRoom* on allocation time. The non-parametric evidence is perhaps more informative. We

Kirchmaier (2017) that a faster response time increases the likelihood of clearing the crime. In that paper, the identification strategy exploits discontinuities in distance across locations next to each other but on different sides of division boundaries. In the current paper, co-location between handler and operator would likely not be a valid instrument for response time. The exclusion restriction is unlikely to be satisfied because co-location could affect clearance likelihood through many channels in addition to faster response times.

[21] The floorplans are unfortunately not to scale, which prevents us from measuring distance in metric units and is likely to introduce measurement error in the within-room distance variable. Instead, desks are depicted in the floorplans in a matrix (x, y) format. Our measure is therefore the euclidean distance between desks inside this matrix. $D = \sqrt{[(y_{RO} - y_H)^2 + (x_{RO} - x_H)^2]}$, where y_{RO} is the position of the radio operator along the row dimension and the other coordinates are defined accordingly. As an example, two adjacent desks in the same row or column are at a distance of one, while the distance between two diagonally-adjacent desks is $\sqrt{2} = 1.4$.

find that incidents assigned to workers separated by a distance lower than 2 (e.g. diagonally adjacent desks at most) are on average allocated and responded 4% faster. The effect decreases monotonically with distance and becomes zero when handler and operator are separated by a distance higher than 4.[22]

The findings in Figure 1.8 indicate that productivity decays very rapidly with within-room distance: being on the other side of the room is equivalent to being on the other side of Manchester. We cannot provide a definitive answer as to why this is the case. Our conversations with GMP staff have however pointed to the fact that some *handlers*' supervisors (labelled to us as 'old-school') discourage the communication between handlers and operators. This is because these supervisors feel mostly responsible for managing the flow of incoming calls and therefore view conversations that occupy the handlers' time (even if they benefit the rapid allocation of officers) as hindering that objective. These attitudes often make handlers unwilling to attract attention by stepping far away from their desks.

Establishing a Causal Interpretation Our preferred interpretation of the findings in Table 1.3 is that: (a) being physically closer allows teammates to communicate face-to-face, and (b) in settings where information is complex and must be processed relatively quickly, this additional communication channel is performance-improving. An alternative interpretation is that call handlers may be better informed or motivated to deal with incidents originating in the geographical area that surrounds their workplace. To understand this potential confounding effect, note in Figures 1.3 and 1.5A that $SameRoom = 1$ when a handler based in a location is allocated an incident from the geographical area surrounding that location. If handlers are more effective at dealing with cases that occur closeby, the findings in Table 1.3 may reflect proximity to the incident scene, rather than to co-location with the co-worker.

A second alternative interpretation is that co-location may be a proxy for some unobserved dimension of similarity between teammates. In an extreme example, imagine that workers communicate through room-specific language, which makes

[22]To interpret this, note that two desks that are three positions apart along both the row and the column dimension are separated by an euclidean distance of 4.2. Two desks separated by three positions along one dimension and two positions along the other are at a distance of 3.6.

electronic communication with individuals outside one's room less efficient. This would be the case if, for instance, there are strong local dialects and the workers in a room are drawn from the neighbourhoods surrounding that room. In that case, co-location would represent a proxy for the ease of electronic communication between teammates, as opposed to providing a performance-improving additional communication channel.

In Columns 3 and 4 of Table 1.4 we find evidence that is inconsistent with the two alternative interpretations above. We add a set of handler/operator *pair* fixed effects to the baseline regressions, and estimate the effect of distance within the room on performance. Because handlers and operators do not typically change workplace, the introduction of pair fixed effects effectively absorbs the same room variable.

We find that *the same pair of workers operating from the same room* are more productive when their desks are closer together. The estimated coefficients are in fact almost identical to those in Columns 1 and 2, without the pair fixed effects. These effects absorb any time-invariant characteristics of the match between handler and operator (including the match between the handler and the location of the incident). The robustness to their inclusion therefore confirms that it is the location of the handler relative to the operator that causes the estimated Table 1.3 decreases in allocation and response times.[23]

A second strategy to evaluate the above is to perform a placebo test using the post-2012 information. As we mentioned in Section 1.2, the 2012 reorganisation of the OCB relocated all the call handlers to Trafford, while the radio operators were split between Claytonbrook and Tameside. Therefore, handlers and operators never shared a room after 2012. Using the information on the workplaces of handlers

[23]A potential caveat here is of course that handlers choose daily the desks where they sit, conditional on these desks being unoccupied. Therefore, within-room distance between handler and operator cannot be considered random. This would be problematic to the extent that it is correlated with time-varying characteristics of their match. For instance, it may be that handlers choose to sit next to operators with whom they have worked on more incidents in the past (if these seats are available). While this is a theoretical possibility, we note two things. Firstly, handlers and operators who have worked together on more incidents in the past are empirically *not* more likely to sit closer to each other (Appendix Table 1.A11). Secondly, the effect of within-room distance on allocation time is robust even after controlling for the interaction of the handler/operator pair *and* the year/semester pair (Appendix Table 1.A12). In fact the estimates are very similar, if anything larger. Of course, the introduction of such a large number of fixed effects implies that this regression is highly demanding, as most of the variation in within-room distance is absorbed.

and operators *just before* the reorganisation, we can construct 'placebo same room' variables taking value one when an incident is allocated to a pair of teammates that used to be co-located.[24] In the estimation of (3.3) we now interact the same room variable with dummies for each of the five semesters comprising our baseline period (the last semester of 2009 includes only two months, since the data starts in November). We then use the post-2012 data to estimate (3.3) again, interacting the placebo same room variable with semester dummies. The coefficients are displayed in Figure 1.9.

We find that the same room variable is essentially zero for every semester of the post-2012 period, while it is negative for most of the baseline period. Note in particular the large difference in the estimates between late 2011 and early 2012. This difference suggests that the same pairs of workers that were able to deliver higher performance when jointly assigned to an incident ceased to do so when they stopped being co-located. The evidence in Figure 1.9 reinforces the conclusion that it is indeed distance between co-workers, rather than unobservables correlated with distance, that improves allocation and response times.[25]

Effects on the Type of Officer Sent We now study whether the faster allocation and response times associated with co-located incidents are at the expense of other dimensions of the quality of the response. As we argued in Section 1.2, these are typically difficult to measure empirically. One aspect that we can observe in our dataset is the rank and experience of the officer that was sent to the incident. Officers with the rank of 'response officer' are trained (and accumulate on-the-job experience) specifically to deal with incidents that the police is alerted to. Neighbourhood officers are instead in charge of patrolling but can be called to attend certain types

[24]Following the reorganisation radio operators remained in their previous roles in terms of the subdivisions for which they dispatched officers. Therefore, a post-2012 handler-operator match continues to capture accurately whether the handler is assigned a case from the geographical area around her pre-2012 workplace.

[25]Interestingly, Figure 1.9 also suggests that the effect of co-location on productivity may have been increasing over time. In November 2009 a major reorganisation had taken place that created a Manchester-wide handling system and split the roles of handler and operator. Workers may have taken time to adapt to their new roles, and to fully exploit the sources of higher productivity in the new setting. In particular, the coefficients of Figure 1.9 are consistent with workers learning about the performance-improving potential of co-location over time. We return to this issue in Section 1.6, where we investigate whether individuals that have worked together on more incidents in the past benefit more from co-location.

of incidents, for instance if response officers are temporarily unavailable. If the likelihood of sending an officer with the rank of 'response officer' is lower for co-located incidents, a faster response time might be interpreted as being at the expense of lower 'quality'.

In Column 1 of Table 1.5 we find, however, that this is not the case. In Column 2, we regress the officer's number of years in the force on the same room dummy, and again find no correlation. We conclude that, to the extent that we can measure quality, there is no evidence that co-location is associated with both a faster response and a *worse* response.

Spillovers to Other (Contemporaneous) Incidents We now investigate the existence of potential spillovers from same room incidents into other contemporaneous incidents. Radio operators typically have open (i.e. yet to be allocated) several incidents at the same time. Theoretically same room incidents can generate both positive and negative spillovers. Positive spillovers will occur, for instance, when the time and effort that the operator saves on a same room incident (as a result of being able to gather information more efficiently) is redistributed to other contemporaneous incidents. Negative spillovers are equally plausible. One potential channel would be operators assigning higher priority to incidents that have been created by co-located handlers. If that was the case, the improvement in performance for same room incidents that we document in Tables 3 and 4 would be, at least partially, at the expense of other contemporaneous incidents, as attention is diverted away from them.

To study whether spillovers are in fact present in our setting we first replicate our baseline specification and use as independent variable of interest the percentage of incidents assigned to the operator that, in the period surrounding the index incident, are same room incidents. Positive spillovers should lead to a negative coefficient for this variable because, if same room incidents are easier to deal with, a higher share of those will allow for more time and effort being available for the index incident. Negative spillovers would instead imply that valuable attention or resources are diverted away from the index incident when other incidents are handled in the same room, leading to higher allocation and response times, and a positive coefficient in

this regression.[26]

We find in Table 1.6 no evidence of either positive or negative spillovers. Given the uncertainty about the time horizon on which spillovers might occur, we calculate the independent variable at the 60, 30, and 15 minutes time horizon. We find in every case that a higher share of same room incidents does not translate into different performance for other contemporaneous incidents.

We perform a second exercise by ordering the incidents assigned to each operator according to the time at which they were created. We then create leads and lags for the four incidents that, for a given operator, immediately precede and follow a same room incident.[27] The estimated coefficients in Figure 1.10 are inconsistent with the existence of negative spillovers, since none of the lag and lead coefficients are positive and statistically different from zero. One of the eight coefficients is negative, providing at most weak evidence of some positive spillovers. Overall, we interpret Figure 1.10 as suggesting, consistently with Table 1.6, that the improvement in performance of same room incidents is neither at the expense nor to the benefit of other contemporaneous incidents.

1.5 Mechanism

The findings above have established the existence of a causal relation between co-location and performance. Our preferred explanation is that co-location permits face-to-face interactions which communicate relevant details about incidents. In this section we first discuss alternative mechanisms, and then provide evidence that is consistent with the face-to-face communication mechanism but inconsistent with these alternative mechanisms.

Discussion of the Alternative Mechanisms In addition to the possibility of communication face-to-face, co-location may change other dimensions of the inter-

[26] We use the baseline sample for this exercise, since in principle spillovers could occur both to same room and to non-same room incidents. In Appendix Table 1.A6 we restrict the sample to including only non-same room incidents and find very similar effects.

[27] Because incidents are not dispatched immediately, a same room incident could create spillovers to other incidents that were assigned to the same operator earlier in time.

action between handler and operator. For instance, under co-location handler and operator may be more likely to learn each others' identity. Note importantly that there are only three alternative channels through which these dimensions can affect productivity in our setting. The first alternative channel consists of the handler exerting more effort in the transmission of the GMPICS electronic information under co-location. The second alternative channel is similar: the operator might exert more effort in the interpretation of this information, and the subsequent allocation of an officer, for co-located incidents. A third potential channel would be the preferential allocation of scarce resources, such as police officers, to co-located incidents and in detriment of other incidents. We do not consider this third channel here because the evidence in Section 1.4 showing the lack of negative spillovers is inconsistent with it.[28]

We can think of two plausible reasons why workers may exert more effort under co-location, even in the absence of face-to-face communication.[29] The first reason would be some type of *silent* psychological effect leading to higher priority assigned to incidents that will be read, or were written, by a same room co-worker. The second potential reason would be handler and operator exerting *silent* visual peer pressure on each other, similarly to the visual pressure among supermarket cashiers identified by Mas and Moretti (2009).

We regard this second reason as unlikely, in particular with regards to the handler exerting peer pressure on the operator, as several features of the institutional setting are inconsistent with it. Firstly, while handler and operator are 'teammates', they are not actually 'peers'. As discussed in Section 1.2, operators are both more senior and uniquely responsible for the allocation of the incident, which makes it improbable that they may feel a lot of pressure from handlers. There is in fact little scope for handlers to even be aware of the allocation and response times of the

[28]We anticipate at this point that these three channels are unable to explain the evidence in Table 1.7, which we discuss below.

[29]Note that face-to-face communication could lead to the higher motivation of its receiver (in this case, the radio operator). Storper and Venables (2004) argue persuasively that face-to-face communication can serve as a signal about the importance of a task, thereby stimulating a 'psychological rush' that leads to greater and better efforts. In our context, it is possible that discussing an incident in person may induce the operator to devote more time and effort to it, and this channel is not incompatible with the higher ability to deal with the incident resulting from a richer information set. Similarly, to the extent that the act of communicating face-to-face itself requires effort by the handler, it is by construction correlated with it.

incidents that they created, unless they actively search for them in the GMPICS system. Furthermore, the cognitive and desk-bound activities of the operator are difficult to monitor visually, especially relatively to manual tasks like supermarket item checking. For instance, an operator may appear busy by virtue of looking at her computer screen, while in fact paying little attention to her work. In addition, there are significant physical barriers (computer monitors, desk screens...) between the workers in the rooms of our setting. These barriers make it impossible to observe the behaviour of all but the closest co-workers, unless a handler actively stands up from her desk. While it is possible in theory for a handler to stand up and watch over the operator's shoulder in silence, we think that is an unlikely possibility.

Evidence Inconsistent with the Handler's Effort Mechanism The first alternative mechanism consists of the handler communicating better electronically. We now test whether there is any evidence of the handler being more precise and thorough in the electronic communication of co-located incidents. We have three good measures of this communication. The first one is the handler's creation time: the time elapsed between the handler answering the call and the creation of the incident in the GMPICS system. Remember that this creation time takes place *before* the radio operator is informed of the incident's existence (see Figure 1.2). We expect that a more thorough and precise electronic communication will require more time devoted to writing the description of the incident, and probably also to the elicitation of the information from the caller. In Column 1 of Table 1.7 we however replicate our baseline specification using creation time as dependent variable, and find that it is unaffected by co-location.

As complementary measures of the quality of the electronic communication, we use the number of characters and number of words in the first line of the description of the incident.[30] Unsurprisingly, these two variables are very correlated with each other, even after conditioning on the baseline set of controls (Appendix Table 1.A8).

[30]Unfortunately, due to a combination of technical challenges and the extreme confidentiality of this information, we were not able to obtain the full content of these descriptions. The first line of the incident description consists of a maximum of 210 characters, and serves as a quick summary of the nature of the incident. When operators have more than one incident open at one time, they typically only see the first line of this description, which then plays a role similar to the subject of an email in an inbox.

They are also strongly correlated with the creation time, suggesting that, despite their coarseness, there is valuable information in them. In Columns 2 and 3 of Table 1.7 we find that these variables are not different for co-located incidents.

To conclude, we find no evidence that the electronic information inputted by handlers is better or worse for co-located incidents, relative to other incidents. Therefore, higher effort on the handler's part and the resulting better electronic communication does not appear to be an important mechanism in our setting.[31]

Evidence in favour of the Face-to-Face Communication Mechanism The mechanisms outlined above entail different predictions about the behaviour of the handler after the incident has been created, in particular with respect to the likelihood that the handler is 'not ready' to take a new call. Consider first the alternative mechanism whereby the operator exerts more effort for co-located incidents. Handlers are continually monitored by their supervisors, and are expected to remain at their desks unless there is a reason to leave them. Therefore, any handler exerting visual pressure on an operator would typically be doing so from her desk, an activity that is perfectly compatible with being available to take a new call. Similarly, the notion that operators are psychologically prone to exert more effort for co-located incidents does not require any change in behaviour on the handler's part. In particular, it does not require handlers being more or less willing to take new calls after creating co-located incidents.

Face-to-face communication, on the other hand, is an activity that typically requires the handler's full attention. Being in 'ready' status while talking to an operator risks having to either ignore an incoming call (an offence so serious that it is likely to trigger disciplinary action) or abruptly cut short the discussion of important details. Therefore, a prediction of the face-to-face communication channel is that, following the creation of co-located incidents, handlers will be more likely to be in 'not ready' status. This prediction is not shared by alternative plausible channels.

[31]Table 1.7 also suggests that co-located handlers do not devote less time and effort to the electronic communication, in the expectation of complementing the information face-to-face. One explanation of this lack of substitution may be the fact that an electronic 'paper trail' needs to be established by the handler, so that other staff members can access that information in the future and the handling of the incident is not criticised during later audits.

In Column 4 of Table 1.7 we replicate the baseline specification using the length of the 'Not Ready' interval following an incident as the dependent variable. The *SameRoom* coefficient is 2.5% and statistically significant, suggesting that handlers step away from their desks (or remain on their desks while being unavailable) for longer periods following co-located incidents. In Column 5 of Table 1.7, we repeat this exercise using as dependent variable a dummy for whether the handler signals her immediate availability to take new calls or instead takes some 'not ready' time at all. Again, we find that the likelihood of not being immediately available is higher for co-located incidents. Of course, the organisation did not record informal communication exchanges between co-workers, and therefore we cannot directly observe these exchanges here. In the absence of such direct evidence, we interpret the estimates in Table 1.7 as strong evidence in favour of face-to-face communication being the main mechanism through which co-location improves performance.[32]

1.6 Heterogeneity

In this section we identify characteristics of incidents, teammates and the working environment that are associated with a higher effect of co-location on performance. We regard this exercise as one of the main contributions of the paper. As discussed in the introduction, a better understanding of the specific circumstances in which face-to-face communication has the highest impact can help guide the communication-enhancing investments by managers.[33]

[32]A potential explanation for the effect of co-location on performance that we have not mentioned up to this point is as follows. When the two teammates are within close proximity of each other as the call handler takes a call, the radio operator overhears the exchange with the caller and starts preparing her reaction even before the handler has officially created the incident. This would still represent in-person communication, although of a different kind than the one that we have been discussing throughout. We have however strong reasons to discard this explanation. Firstly, the effects are present even when the two teammates sit relatively far apart, such as at two positions away along the row dimension and three along the column dimension. Secondly, the noise levels in these rooms are incompatible with the ability to overhear or signal across more than the very shortest distances. Thirdly and most importantly, this potential alternative mechanism is unable to explain the evidence in this subsection, whereby the call handler takes longer to be available for the next call following a co-located incident.

[33]We add at this point the standard note of caution that the characteristics of incidents and teams are not randomly allocated, as they may be related to other unobserved characteristics of the same incidents and teams. For instance, it may be that workers of the same gender tend to socialise together during breaks and that it is the unobserved variable 'Socialising during Breaks' that underlies the same gender/same room significant interaction in Table 1.10. This caveat is

Characteristics of Incidents We first examine whether the effects from Table 1.3 are stronger for some types of incidents, relative to others. We focus on two particularly relevant characteristics of incidents: their urgency and the complexity of the information required to understand and describe them. The main hypothesis is that if co-location improves performance because it enables face-to-face communication, we should find a stronger effect for complex incidents where a lot of information must be transmitted. In addition to being intuitive, this hypothesis is consistent with the vast literature arguing that human production is at a lower risk of being substituted by technology for (cognitive) non-routine tasks, relative to routine tasks (Acemoglu and Autor, 2011).

We also study empirically the relation between the urgency of an incident and the effect of co-location on performance. In principle, it is unclear what the sign of this relation should be. On the one hand, the ability to communicate information quickly might be more valuable and therefore used more often when an allocation decision needs to be done faster. On the other hand, in very urgent incidents (e.g. a serious crime in progress) the operator may not want to wait for many nuanced details and will instead allocate an officer as quickly as possible. If that is the case, more urgent incidents will be associated with a lower effect of co-location on allocation time.

Both theoretical concepts, 'urgency' and 'information intensity', have elusive empirical counterparts. The information intensity of incidents is difficult to measure because we unfortunately lack access to complete characterisations of the features of every incident in our dataset. We also lack the full GMIPCS descriptions recorded by handlers, although of course any classification of an incident reliant on the actions taken by its call handler would risk confusing the diligence or ability of the handler with the intrinsic features of the incident.

To overcome the measurement challenges above we use information based on generic incident types to create an indirect measure of information intensity, as follows. We first classify each incident according to its opening code/grade com-

of course present in every study on differential effects by gender and, more generally, in every heterogeneity analysis such as the one here. Our objective here is not to claim causal effects (on the interactions), but instead to understand the type of sub-populations where the effect of co-location is stronger.

bination. We then compute the average creation time (the time elapsed between the handler answering the call and the creation of the incident) for every one of the resulting 144 combinations. The average creation time of an incident's type constitutes our measure of (predicted) information intensity, as it captures how long on average it takes for handlers to extract information from the caller and record it in GMPICS, for that incident type. Although the measure is undoubtedly coarse, our interpretation is that incident types with high average creation time should be those where the amount and complexity of information is typically the largest. We construct our measure of (predicted) urgency in an equivalent way, this time calculating the average allocation time of every incident type (naturally, lower average allocation time is interpreted as higher urgency).

We interact our measures of information intensity and urgency with the same room dummy in the baseline regression. For ease of interpretation, these measures are entered as above-median dummies. The estimates are displayed in Table 1.8. We find first that incident types of high average information intensity are associated with a higher effect of co-location on performance.[34] We also find (weaker) evidence on the urgency of incidents exacerbating the effect of co-location. In particular, the estimate for the interaction with urgency is negative, although statistically significant only in the allocation time regression.[35]

We interpret the estimates from Table 1.8 as indicating that co-location does not increase performance for non-urgent, non-complex incidents. It, however, decreases allocation time (respectively, response time) by 4% (respectively, 2.7%) for incidents that are above-median both in their urgency and their information intensity. The estimate on the interaction with information intensity is, in particular, consistent with the notion that co-location enables an additional communication channel, leading to higher performance for incidents when a lot of communication is necessary.

[34]This finding is robust to measuring information intensity with quintiles (Appendix Figure 1.A2) and in parametric (log) format (Appendix Table 1.A4). It is also robust to building the information intensity prediction exclusively with out-of-sample (i.e. pre-November 2009 and post-January 2012) observations (Appendix Table 1.A3).

[35]Both effects become statistically stronger if information intensity is measured parametrically (Appendix Table 1.A4). However, we find that the effect of co-location does not vary when we use a simpler and coarser measure of the urgency of an incident: its grade. Although the effect is stronger for Grade 1 incidents, relative to Grade 2 and Grade 3, the differences are not statistically significant (see Appendix Table 1.A9).

Characteristics of the Working Environment In our second heterogeneity exercise, we study whether co-location improves performance more when workers have to deal with more incidents. Our main interest is in the workload of the operator, because it is for operators that a high number of incoming incidents in their subdivision can start to accumulate, exerting competing demands on their attention. Our hypothesis is that, if co-location allows operators to quickly resolve any doubt through face-to-face communication, it should be more valuable when the time and effort of the operator are scarce, that is, in periods of higher workload.[36]

Our measure of the operator's workload is the number of incidents created in the subdivision that the operator is overseeing during the hour of the index incident (note that there is a single operator responsible, at any one time, for a subdivision). For ease of interpretation, we enter this measure in the baseline regression as an above-median dummy, both by itself and interacted with the same room variable.

The results are displayed in Table 1.9. We first find that allocation and response times are slower when the operator is busier, as expected. Our main interest is in the estimate of the interaction between the same room variable and the high operator workload dummy, which we find to be negative and statistically significant. The estimated coefficients indicate that co-location reduces allocation time (respectively, response time) by 1.1% (respectively, .8%) during periods of low operator workload, but 2.9% (respectively, 2%) during periods of high workload. This finding lends support to our hypothesis that the benefit of communicating personally with the handler is higher when the operator is more pressured for time and needs to gather information more quickly.[37]

[36]By contrast, our understanding of the institutional environment is that the notion of being 'pressured for time' is less meaningful for handlers. Handlers deal with incidents sequentially and share the responsibility of responding to incoming calls with a large number of colleagues (since every handler can handle incidents from every Manchester area). Together with the fact that handlers are not responsible for the allocation of officers to incidents, this implies that we do not have a strong hypothesis about the relation between our measure below of handler workload and the effect of co-location on performance

[37]Our measure of the handler workload is very coarse, mostly because as discussed earlier, the notion that handlers are busier in some periods relative to others is not clear-cut. We use the (above-median dummy of the) number of incoming calls during the index hour, divided by the number of available handlers. Because this variable is defined at the Manchester-wide level, it is absorbed in the baseline regression by the hour fixed effect. We find in Table 1.9 that the coefficient on the interaction with the same room variable is smaller in magnitude and only weakly statistically significant.

Characteristics of the Workers We now examine whether the effect of co-location on performance is stronger when the teammates share the same age and gender, and have worked together more often in the past. This may be the case for two reasons. Firstly, workers of a similar background (or more familiar with each other) may be more likely to initiate the face-to-face communication exchanges that transmit information regarding an incident. This is because they may be more likely to sit close to each other, or, conditional on the within-room distance, they may be more likely to leave their desk and talk to each other. Secondly, in-person communication may also be more efficient among these types of workers.[38] Alternatively, homogenous teams may be so efficient at communicating electronically that additional in-person communication is more valuable when the team is *not* homogenous.

In Table 1.10 we display estimates of our baseline specification, where we add a same gender dummy, the (log of the) difference in age, and the (log of the) number of past incidents in which handler and operator worked together. We further interact these variables with the same room variable. To isolate the effect of the handler/operator *pair* experience, the specification controls for the individual experiences of handler and operator and their interactions with the same room variable.

Our main finding is that the estimates for the three interactions of interest are statistically significant and of the expected sign. For instance, the effect of co-location is 1.6% higher when handler and operator share the same gender. A 10% increase in the age difference (respectively, number of past interactions) between handler and operator decreases the effect of co-location on performance by 2.5% (respectively, it increases it by 2.1%). These findings are consistent with the notion that face-to-face communication, and therefore co-location, leads to higher performance among co-workers that know and understand each other better.[39] On the

[38]Storper and Venables (2004) discuss how the transmission of uncodifiable information (at which face-to-face communication excels) depends on a 'communication infrastructure' that is specific to a sender-receiver pair. This infrastructure is likely improved through learning by doing, leading to more efficient face-to-face communication as the teammates accumulate experience with each other. It is also likely more efficient among demographically proximate teammates. Alternatively, we could interpret demographic proximity as a proxy for the existence of friendship ties between two co-workers (Bandiera, Barankay and Rasul, 2010). If workers are more willing to and effective at communicating face-to-face with their friends, a similar prediction for the relation between demographic proximity and the effect of co-location on performance would arise.

[39]We find qualitatively similar results when age and past interactions are measured as above-median dummies (see Appendix Table 1.A5). While not the focus of this paper it is interesting

other hand, the non-significant interactions with individual experience suggest that, unless it is specific to the teammate in this particular incident, individual experience does not by itself allow workers to exploit better the potential advantages of co-location.

1.7 The Operational Cost of Face-To-Face Communication

In this section we provide a measure of the *operational* costs of communication.[40] In Section 1.4 we found no evidence of negative spillovers to other incidents being handled contemporaneously by the operator. On the other hand, Section 1.5 has shown that handlers spend 2.5% more time unavailable to take new calls following the creation of co-located incidents. This unavailability imposes a cost on the organisation, as it contributes to incoming calls being answered with a longer delay. We now provide a framework to measure the opportunity cost of the time spent in face-to-face communication, so that it can be compared to its benefit.[41] We then compute this cost in our organisation.

Theoretical Framework We formalise the process by which calls to the police arise, join the call queue and are answered. Assume a population of individuals (of normalised size 1) who can potentially call the police. Every individual can be in one out of three states: dormant (waiting for an incident to happen), in the call queue, or on the phone with the handler. x_i, $i = 1, 2, 3$ denotes the share of individuals in each state. $H < 1$ handlers are on duty to answer calls.

Transitions between states are as follows. Dormant callers join the queue at an

to note that, even when teammates are not co-located, a similar age and a longer experience with each other are still associated with higher performance (although this is not the case for the same gender variable). A potential explanation of these estimates is that, given the complexity of the information that must often be transmitted, even electronic communication is more efficient among these types of teammates.

[40] Building communication channels between workers may entail fixed investments, and we do not have the information to measure the cost of these investments here.

[41] Note that, to the extent that communication takes time and that time cannot be devoted to other activities, the type of cost that we measure here is present in every organisation where communication takes place.

exogenous rate a per unit of time. All callers must spend at least one unit of time there before being assigned a handler. When a call is being answered, it terminates (and the caller re-joins the dormant pool) at a constant rate υ (with $1/\upsilon$ being the average duration of calls). The number of handlers that become available to take new calls per unit of time is therefore υx_3. The total number of calls answered per unit of time is then $min\{\upsilon x_3, x_2\}$, since it is limited both by the number of newly available handlers and by the size of the call queue.

Using this simple framework we can show that the size of the call queue evolves over time depending on the difference between the inflow (the number of dormant individuals who encounter an incident) and the outflow (the number of queued calls answered by handlers):

$$\frac{\Delta x_2}{\Delta t} = a\,(1 - x_2 - x_3) - min\{\upsilon x_3, x_2\} \tag{1.4}$$

Similarly,

$$\frac{\Delta x_3}{\Delta t} = min\{\upsilon x_3, x_2\} - \upsilon x_3 \tag{1.5}$$

If $\upsilon x_3 < x_2$, then it must be that all handlers are busy and $x_3 = H$. Combining equations (1.4) and (1.5) and assuming a steady state, we compute the time in the queue for incoming calls, q^*, as:

$$q^* = \begin{cases} \frac{(1-H)}{\upsilon H} - \frac{1}{a} & \text{if} \quad H < \frac{a}{a+\upsilon+a\upsilon} \\ 1 & \text{if} \quad H \geq \frac{a}{a+\upsilon+a\upsilon} \end{cases} \tag{1.6}$$

This simple framework generates the following predictions. First, incoming calls are answered immediately when there are many handlers (H high), few dormant calls become actual calls (a low) and calls are brief (υ high). Secondly, $\frac{\partial q^*}{\partial(1/\upsilon)} > 0$ so an increase in average call length leads to longer queuing times. Lastly, this effect is lower when the number of handlers is higher, $\frac{\partial^2 q^*}{\partial(1/\upsilon)\partial H} < 0$. We can interpret an increase in H as the increase in organisational slack, as the same amount of incoming work is divided over a higher number of workers. Therefore, this model predicts that an increase in slack both decreases queuing times and reduces the effect of higher

average call duration.

Computing the Opportunity Cost of Face-To-Face Communication Section 1.5 provided evidence of an increase in 'not ready' time following the creation of co-located incidents. This is equivalent in our framework to an increase in the duration of the call, as it mechanically prevents handlers from relieving the pressure in the call queue. We now use information on *all* calls (not just the ones that led to the creation of incidents) to relate call duration, the number of calls and the number of on-duty handlers to the average time spent in the call queue. The resulting coefficients allow us to understand the *opportunity cost* of an additional second spent dealing with a previous call. We estimate:

$$q_i = \alpha + \gamma n_i(\tau) + \delta h_i(\tau) + \beta d_i(\tau) + \epsilon_i \tag{1.7}$$

where q_i is the (log of the) queuing time of incoming call i, n_i and h_i are the (log of) number of calls and on-duty handlers in a time window before i, and d_i is the (log of) average duration of answered calls in the same time window.

Table 1.11 Panel A shows that the estimated elasticity of average call duration on queuing time ranges from .58 to .96. We can compute the effect that an increase in the duration of a single call j has on the queuing time of future calls as follows. First, note that such an increase has an effect on the queuing time of *a single future call* i that can be computed as $\hat{\beta}\frac{exp(q_i)}{TD_i}$, where $exp(q_i)$ is the queuing time of i and TD_i is the total duration of the calls preceding i (which include j). Aggregating over the K calls that follow j, we can write the overall effect of an increase in j's duration as $\hat{\beta}\sum_{i=j+1}^{j+K}\frac{exp(q_i)}{TD_i}$.

The statistic $\hat{\beta}\sum_{i=j+1}^{j+K}\frac{exp(q_i)}{TD_i}$ can be interpreted as the opportunity cost (in terms of additional queuing time of future calls $i = j+1 \dots K$) of increasing the duration of call j by one second. This statistic can be computed directly from our dataset, using the elasticity estimated in Table 1.11 and information on the queuing time of every call, together with the duration of the calls preceding it. Using a time window of 60 minutes to define the calls affected by the increase in the duration of a preceding call, we calculate it as 0.13 seconds. In Table 1.7 we estimated that

co-located incidents increase 'not ready' time by 2.5%. Evaluated at the mean of 'not ready' time (66 seconds), co-located incidents are therefore associated with a cost of $0.13 \times 2.5\% \times 66 = 0.21$ seconds.[42] In our organisation, this is arguably a small cost, when compared with the decreases in allocation and response times of 76 and 104 seconds respectively that we estimated in Section 1.4.[43]

Motivated by our theoretical framework, we expect the opportunity cost of face-to-face communication to be lower when organisational slack, as captured by the relation between on duty handlers and incoming calls, is higher. In Panels B and C we repeat the exercise in Panel A for the subsamples of calls with high and low organizational slack. Consistently with the prediction that increasing a call's duration is less costly when the relative number of handlers is higher, we find a higher elasticity in Panel C (high slack) and a lower in Panel B (low slack). Replicating the analysis above, we calculate costs associated with co-location of 0.15 (respectively 0.31) seconds, for periods of low (respectively high) slack.

Overall, our analysis highlights the importance of measuring the opportunity cost of the time engaged in face-to-face communication, as well as the dependence of this cost on the slack characterising the organisation. In our setting, we find this cost to be much lower than the benefit.

1.8 Conclusion

This paper has provided evidence of a causal relation between co-location and performance, in a teamwork setting characterised by the communication of complex information. A series of additional tests point towards face-to-face communication as the most important mechanism. We have also provided additional evidence on the heterogeneity of the main result and highlighted that face-to-face communication has opportunity costs, as well as benefits. We are not aware of any existing

[42] We repeat this exercise for time windows of 15, 30 and 120 minutes and we estimate the cost in 0.20, 0.18 and 0.22 seconds respectively.

[43] The benefits and costs associated with co-location affect different types of calls. The costs are for the average call, including those which do not lead to incidents and those leading to incidents that are not deemed to merit a response within four hours of the incident creation. The benefits are instead concentrated on the calls that are deemed important enough to be assigned to a radio operator.

study studying these questions, especially one that is comparable in terms of the detail of analysis and the credibility of the estimated effects.

One immediate policy prescription for the specific organisation that we study is in terms of supervisors' awareness of the benefits of communication between co-workers. Discussions between handler and operator following the creation of incidents were not encouraged and were even frowned upon by some supervisors. Because the cost of communication is orders of magnitude smaller than the benefit, one implication is that, in our specific context, there may be too little communication among co-located workers rather than too much. This indicates that a change of norms and culture to encourage more communication could be efficiency-enhancing. More generally, however, the fact that the cost of communication is not zero indicates that the limitations on the information sets of decision-makers highlighted by Hayek (1945) and Arrow (1974) are unlikely to be fully overcome.

Our findings provide direct guidance to managers organising the geographical distribution of activities. Most directly, the evidence casts doubt on the appropriateness of telecommuting policies in settings where workers must communicate complex information to each other. Our results further suggest that telecommuting may be particularly unsuitable (and co-location of teammates particularly valuable) when activities are informationally demanding, workers are homogenous and likely to be busy, and teams are likely to be stable.

There may be additional implications for recruitment policy. A large literature in organisational behaviour is concerned with the advantages and challenges of diversity in the workplace (Shore et al., 2009). In economics, a parallel body of work has studied the differences in productivity between homogeneous and heterogeneous teams (Hamilton et al. 2012, Hjort 2014, Lyons 2016), a question of clear recruitment policy implications. Our results indicate that the relative benefits of homogeneity depend on the geographical configuration of activities. In particular, a more homogeneous organisation is most valuable when workers are likely to be based in the same physical space.

Our results also identify a distinct driver of firm-specific human capital accumulation (Topel, 1991), with implications for staff turnover and team-rotation policies.

Consistently with Hayes et al. (2006) and Jaravel et al. (2016), we find in Section 1.6 that workers accumulate human capital that is specific to a particular co-worker. Importantly, our finding is however that this capital is most valuable (or more rapidly accumulated) among co-located workers. It follows that managers should be wary of the team disruption induced by turnover particularly when the team members work in close proximity.

1.9 Figures of the Chapter

Figure 1.1: Operational Communication Branch

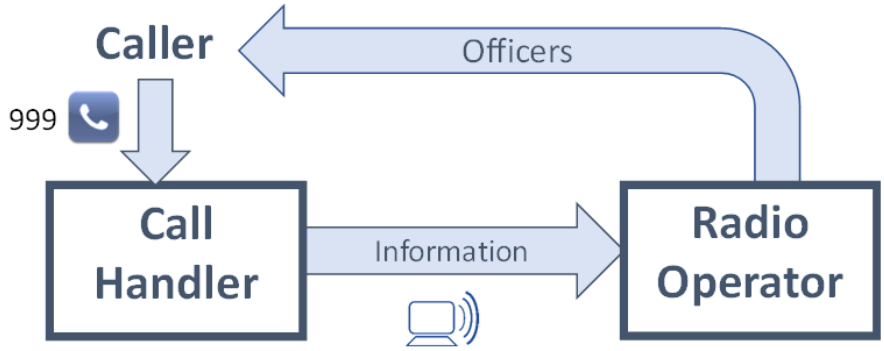


Figure 1.2: Timeline of Actions

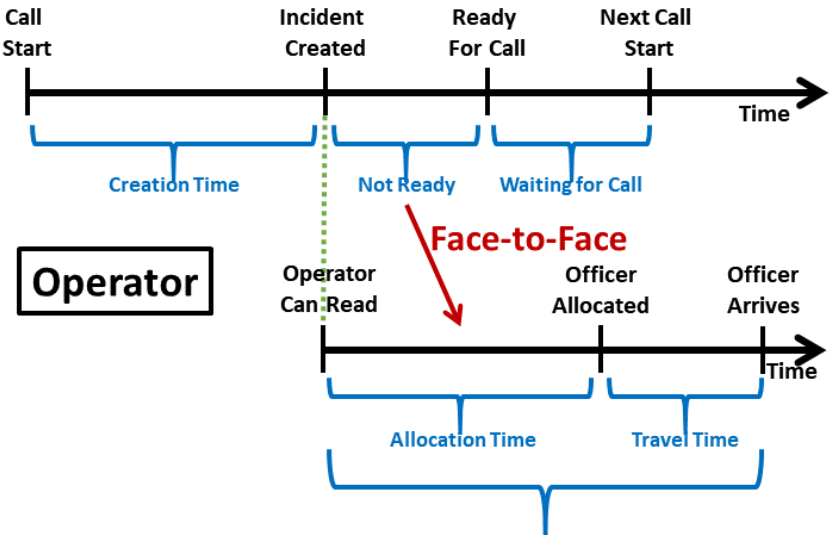


Figure 1.3: Location and Radio Operations Coverage of OCB Rooms

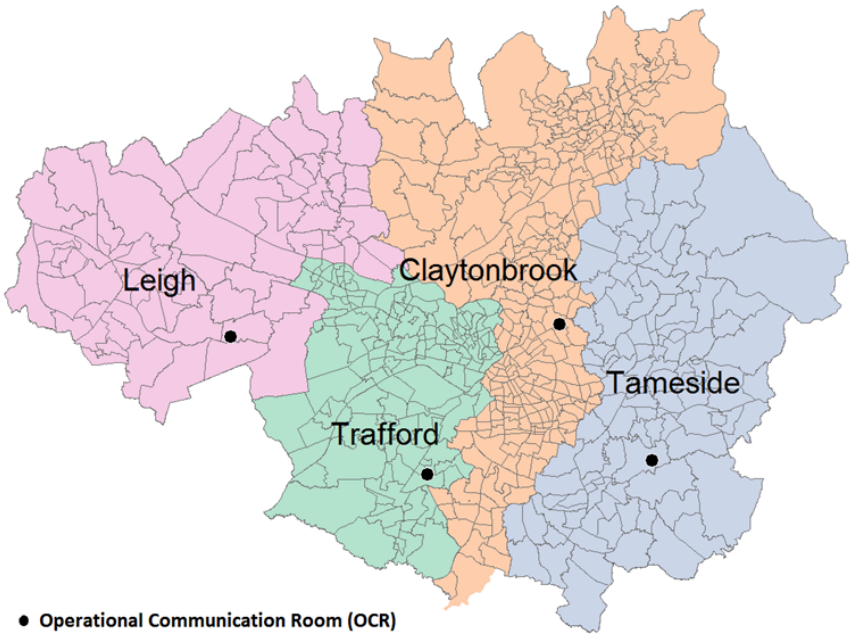
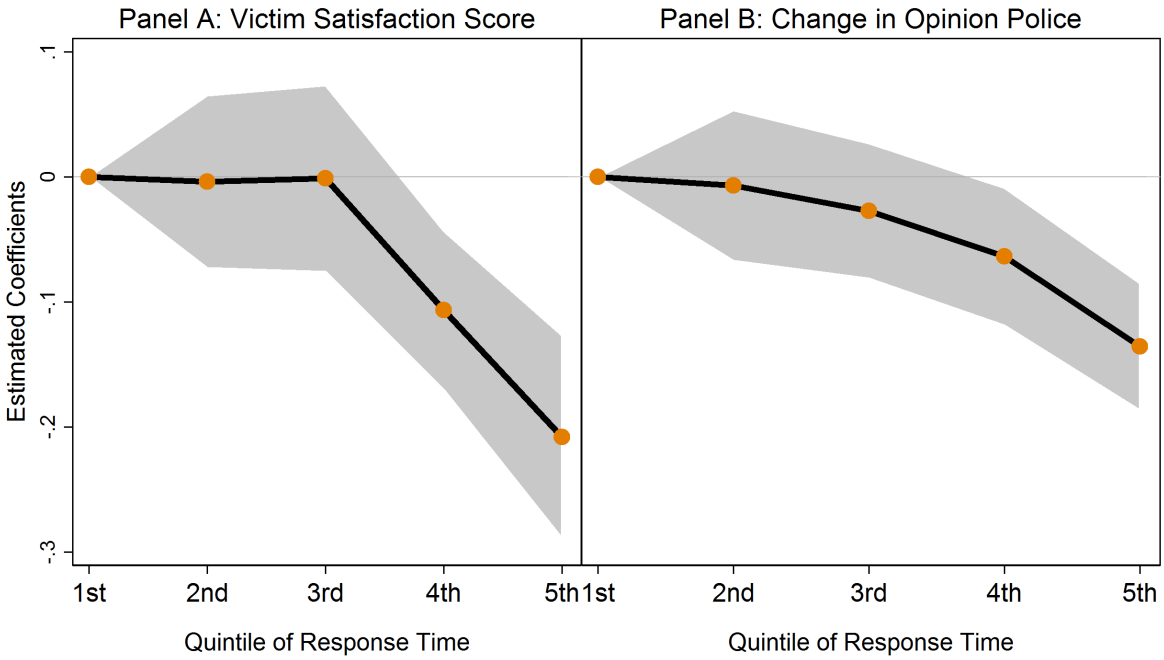


Figure 1.4: Correlation between Response Time and Victim Satisfaction



Each panel displays a different regression. The displayed coefficients are for the Quintiles of Response Time (the first quintile is added to aid visual analysis). 95% confidence intervals are displayed in the shaded grey area. All regressions control for Grade, Call Source, Year X Month X Day, Hour of Day, Division and Opening Code. Standard error are clustered at the Year X Division level.

Figure 1.5: Natural Experiment

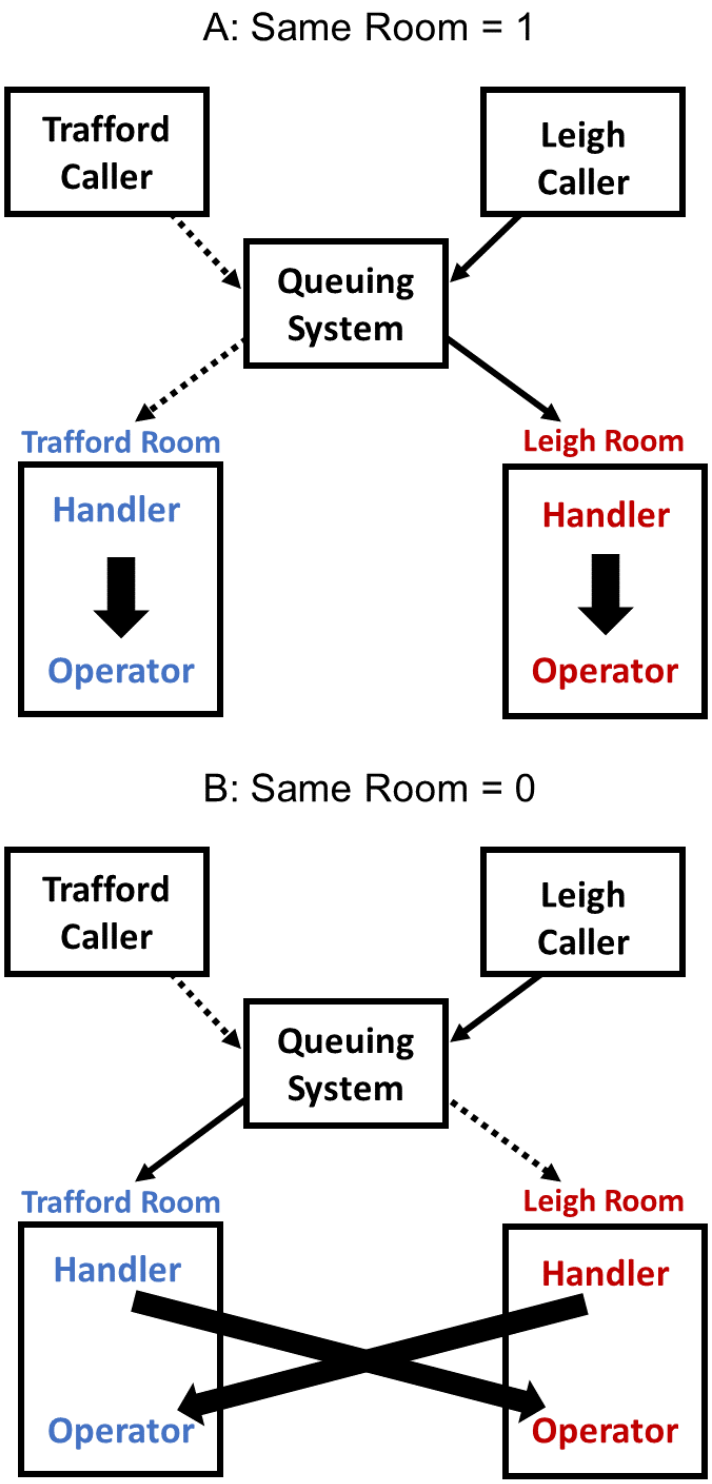
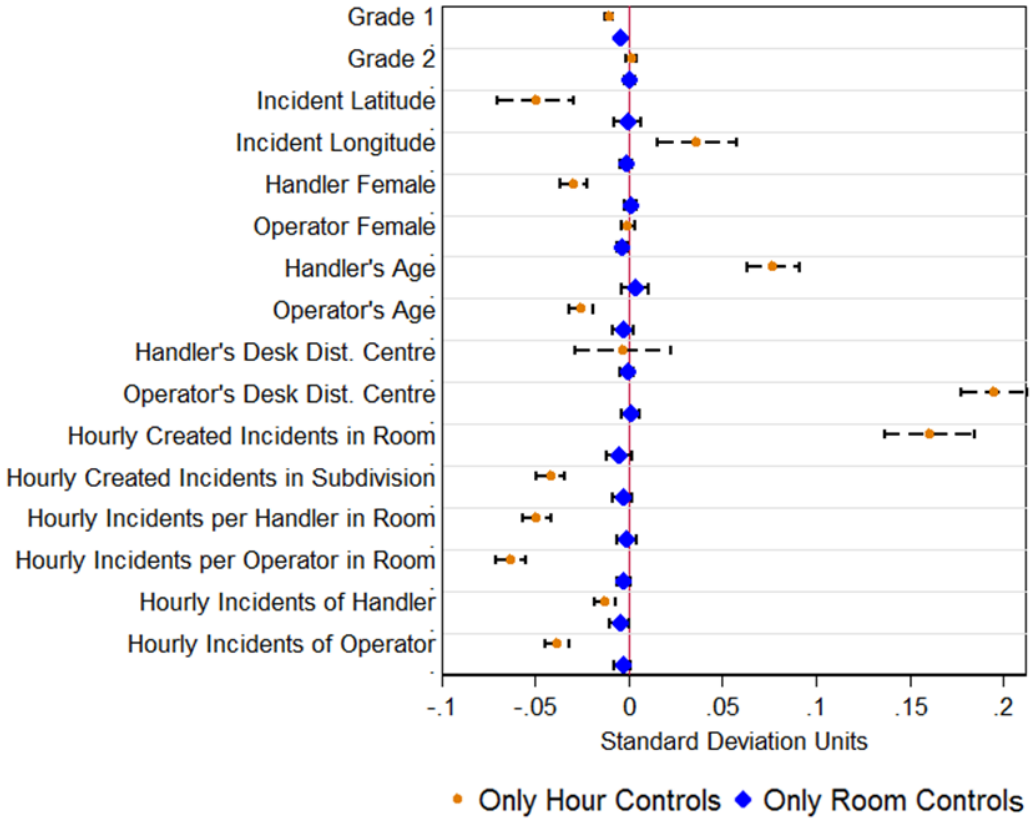


Figure 1.6: Balance of Incident, Worker and Room Characteristics on Same Room



Each row in the figure displays the results of two regressions, where the row variable is the dependent variable and Same Room is the independent variable. The first regression includes no controls and the second regression controls for Year X Month X Day X Hour of Day, Radio Operator Room and Call Handler Room. The displayed 95% confidence intervals are for the coefficient of the Same Room variable. Non-binary dependent variables are standardised. Standard errors are clustered at the Year X Month X Radio Operator Room level. Grade 1, Grade 2, Handler Female and Operator Female are the only dummy variables. Handler's Desk Dist. Centre is the euclidean distance between the handler's desk and the centre of the room. Hourly Incidents per Handler in Room is the number of incidents created during the hour of the index incident, divided by the number of handlers working during that hour. A similar definition applies to Hourly Incidents per Operator in Room. Hourly Incidents of Handler is the number of incidents created by the handler in charge of the index incident, during the hour of creation. Hourly Incidents of Operator is the number of incidents allocated by the operator in charge of the index incident, during the hour of the creation of the incident.

Figure 1.7: Example of OCB Room Floorplan

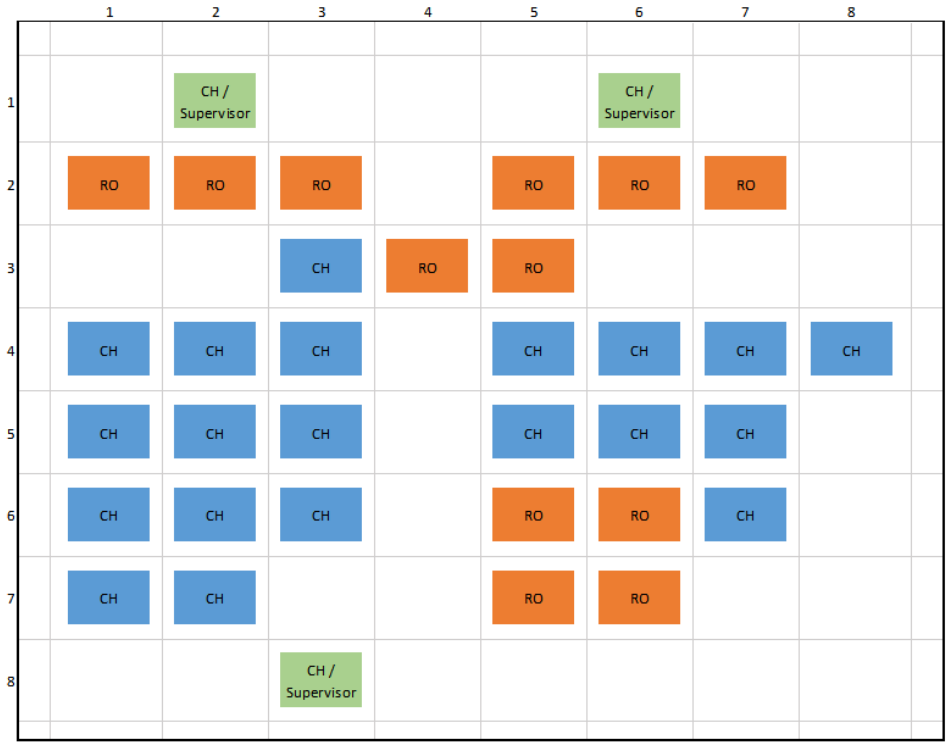
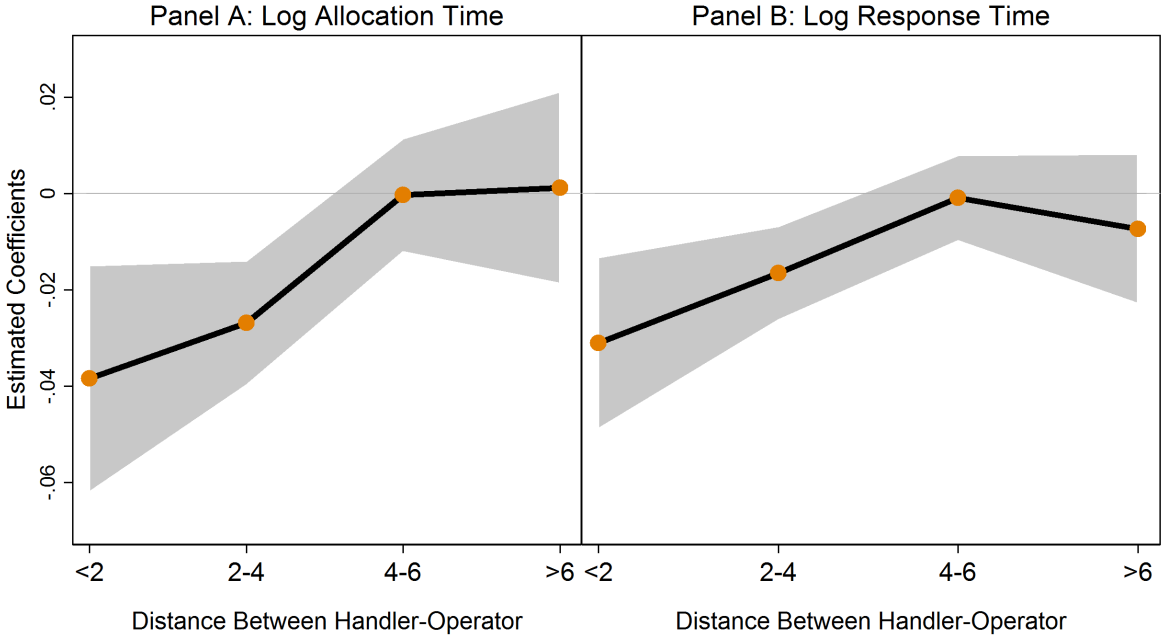
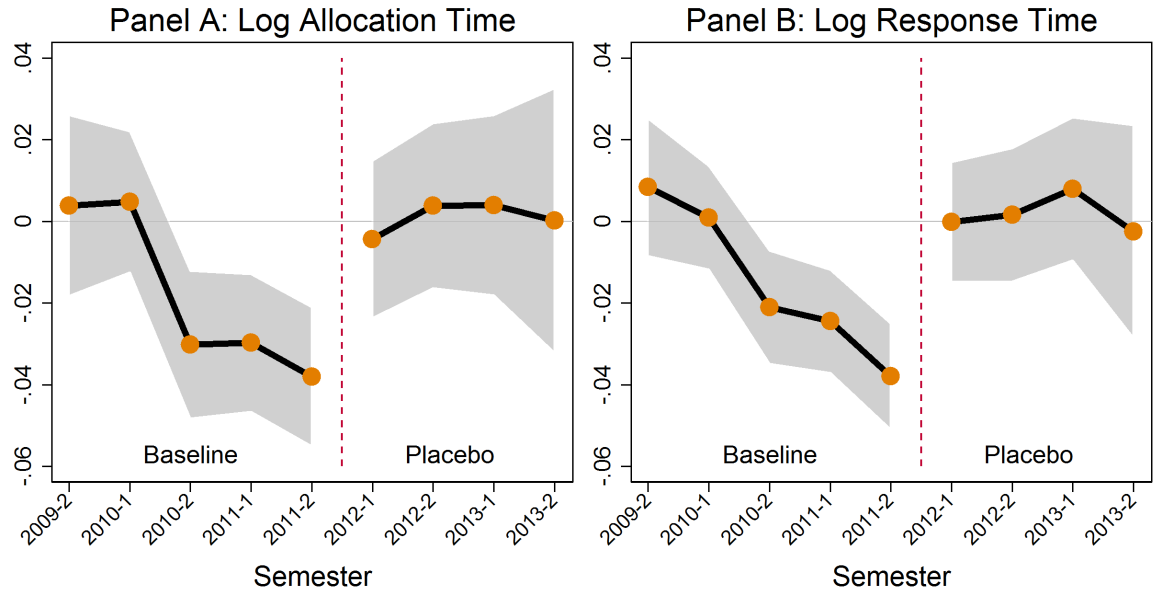


Figure 1.8: Heterogeneity of the Effect of Same Room By Distance Inside Room



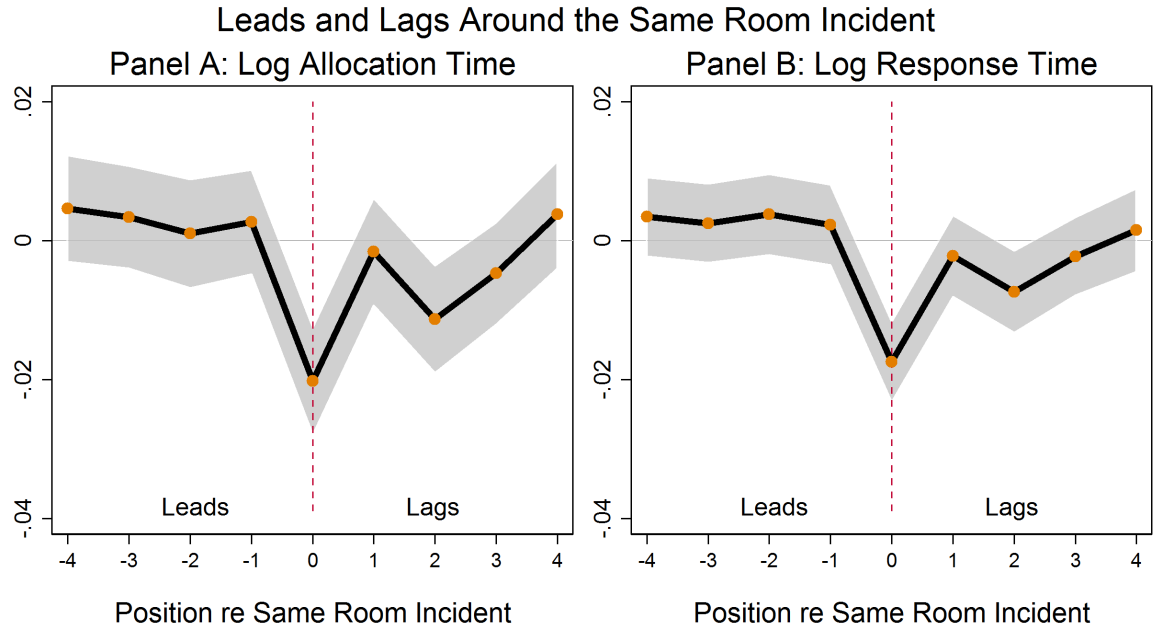
Each panel displays a different regression. The displayed coefficients are for Same Room X Distance Handler/Radio Operator. Distance is the euclidean distance between the desks. 95% confidence intervals are displayed in the shaded grey area. All regressions control for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator and Call Handler. Standard error are clustered at the Year X Month X Radio Operator Room level.

Figure 1.9: Heterogeneity of the Effect of Same Room By Semester, Including Placebo Period



Each panel displays two different regressions. The displayed coeffcients are for the interaction of Same Room (or Placebo Same Room) with the semester indicators. The samples on the left side of each panel are the baseline samples. The samples on the right side of each panel include observations from 2012/13, when all the Call Handlers were based in Trafford, and all the Radio Operators were based in Claytonbrook and Tameside. We regard the 2012/13 period as the placebo period. For this period, the Placebo Same Room variable is based on the Radio Operator and Call Handler locations during the second semester of 2011. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Figure 1.10: Investigating Spillovers from Same Room Incients to Other Incidents



Each panel displays a different regression. The displayed coefficients are for Same Room (position = 0) and four leads and four lags. The leads are the four incidents prior to the Same Room incident assigned to the Radio Operator. The lags are the four incidents following the Same Room incident. All regressions control for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator Identifier and Call Handler Identifier. Standard error are clustered at the Year X Month X Radio Operator Room level.

1.10 Tables of the Chapter

Table 1.1: Correlations Between Allocation/Response Time and Victim Satisfaction Measures

Dep. Variable	**(1) Victim Satisfaction Score**	**(2) Victim Change in Opinion of Police**
Log Allocation Time	-.038***	-.023***
	(.006)	(.004)
Log Response Time	-.055***	-.035***
	(.009)	(.006)
On Target Allocation	.095***	.051***
	(.019)	(.012)
On Target Response	.14***	.082***
	(.028)	(.016)
Observations	9617	7827

This table displays estimates of OLS regressions of measures of caller satisfaction on allocation and response time. Every coefficient is a different regression. The variables in the columns are the dependent variables and the variables in the rows are the independent variables. Victim satisfaction score is the answer by the caller to a survey ranking how satisfied she is with the police dealing with the incident. The score takes values between 1 (Very Dissatisfied) and 8 (Very Satisfied), but has been standardised. Victim change in opinion of police can take values -1, 0 or 1, depending on whether the opinion has worsened, remained the same or improved. All regressions also include indicators for Call Source, Year X Month X Day, Hour of Day, Division, Grade and Opening Code. Standard errors are clustered at the Division X Year level.

Table 1.2: Summary Statistics

	Mean	Median	SD	Min	Max
Allocation Time (min.)	64.124	4.583	276.568	0	21331.78
On Target Allocation	.748	1	.434	0	1
Response Time (min.)	87.484	19.933	311.166	.05	21391.92
On Target Response	.877	1	.328	0	1
Creation Time (min.)	3.889	2.85	4.946	0	219.533
Grade 1	.197	0	.398	0	1
Grade 2	.432	0	.495	0	1
Same Room	.229	0	.42	0	1
Distance inside Room	4.34	4.243	1.782	.5	11.885
Handler Female	.27	0	.444	0	1
Operator Female	.498	0	.5	0	1
Handler's Age	38.406	38	11.471	19	64
Operator's Age	45.15	46	8.243	19	66

This Table reports summary statistics for the baseline sample (N=957137). An observation is an incident. Allocation time is the time between the creation of the incident by the call handler and the allocation of a police officer by the radio operator. Response time is the time between creation of the incident and the police officer arriving at the scene. On target allocation (respectively, response) is a dummy taking value one if the allocation time falls wihin the UK Home Office targets, which are 2, 20 and 120 minutes (respectively 15, 60 and 240 minutes) for Grades 1, 2 and 3. Creation Time is the time between the handler answering the call and the creation of the incident in GMPICS. Grade 1 and Grade 2 are dummies for the grade of the incident. Same Room is a dummy when handler and operator are located in the same room. Distance inside the room is the euclidean distance between the handler and the radio operator desks. This variable is defined in this table only when same room is equal to one (N=219184). Handler female and operator female are dummy variables.

Table 1.3: Baseline Estimates

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**	**(3) On Target Alloc.**	**(4) On Target Response**	**(5) Cleared**
Same Room	-.02***	-.017***	.004***	.002***	-.001
	(.004)	(.003)	(.001)	(.001)	(.003)

This table displays estimates of OLS regressions of five different performance measures on whether the call handler and the radio operator are located in the same room. The sample includes all incidents received by the GMP between November 2009 and December 2011 (N=957137). In Column (1) the performance variable is the log of the allocation time (i.e. the time between the creation of the incident by the call handler and the allocation of a police officer by the radio operator). In Column (2) the performance variable is the log of the response time (i.e. the time between the creation of the incident and the police officer arriving at the scene). In Columns (3) and (4) the dependent variables are dummy variables taking value one if allocation and response times fall within the UK Home Office targets, respectively. The target response times for Grades 1, 2 and 3 are 15, 60 and 240 minutes, respectively. The target allocation times are 2, 20 and 120 minutes. In Column (5) the dependent variable is a dummy taking value one if the crime was cleared. In Column (5) the sample includes only incidents that the police classified as crimes (N=156550). All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.4: Heterogeneity of Same Room by Distance Inside Room

	Individual F.E.		**Pair F.E.**	
Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**	**(3) Log Alloc. Time**	**(4) Log Response Time**
Same Room	-.049***	-.035***	-	-
	(.012)	(.01)	-	-
Same Room	.026***	.018***	.027***	.017**
X Log Distance	(.009)	(.007)	(.01)	(.008)

This table displays estimates of OLS regressions of allocation time and response time on whether the call handler and the radio operator are located in the same room, interacted with the distance between their desks when they are in the same room. The sample includes all incidents received by the GMP between 2009 and 2012 (N=957137). The distance between their desks is calculated as the euclidean distance in the floorplans provided by the GMP. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year and Call Handler Room X Year. Columns (1) and (2) also include Radio Operator and Call Handler Identifiers. Columns (3) and (4) include Radio Operator/Call Handler Pair Identifiers. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.5: Investigating Effects on Type of Officer Sent

Dep. Variable	**(1) Response Rank**	**(2) Log Officer Experience**
Same Room	-.001	.002
	(.001)	(.002)

This table displays estimates of OLS regressions of measures of the type of officer sent on the Same Room dummy. In Column (1) the dependent variable is a dummy for whether the officer sent has the rank of response officer. In Column (2) the dependent variable is the officer's number of years in the GMP. All regressions also include indicators for Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.6: Investigating Spillovers to Other Incidents, by Same Room Incidents

	Spillovers by Same Room Incidents during Period:					
	60 min.		**30 min.**		**15 min.**	
Dependent Variable	**(1) Log Alloc. Time**	**(2) Log Resp. Time**	**(3) Log Alloc. Time**	**(4) Log Resp. Time**	**(5) Log Alloc. Time**	**(6) Log Resp. Time**
% Same Room Incidents Received by Operator	.005 (.005)	.004 (.004)	.006 (.006)	.007 (.004)	.009 (.007)	.007 (.005)

This table investigates potential spillovers from Same Room incidents into other contemporaneous incidents. The dependent variables in the OLS regressions are log of allocation time and log of response time. The independent variable is the percentage of incidents during the index incident time period for which the call handler and the radio operator were located in the same room, excluding the index incident. In Columns (1) and (2) the period comprises of 60 minutes (respectively, 30 minutes for columns (3) and (4) and 15 minutes for columns (5) and (6)). All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. The regressions also include indicators for whether there were no calls received by the Radio Operator during the time period. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.7: Investigating Effects on Other Actions by the Handler

Dep.Var.	**(1) Log Creation Time**	**(2) Log Number of Characters**	**(3) Log Number of Words**	**(4) Log Not Ready**	**(5) Not Ready>0**
Same Room	.00446 (.00326)	-.0004 (.00138)	-.00028 (.0015)	.02513*** (.00928)	.00443** (.00201)

This table displays estimates of OLS regressions of three actions by the handler prior to creating the incident, on whether the call handler and the radio operator are located in the same room. The sample includes all incidents received by the GMP between November 2009 and December 2011. In Column (1) the dependent variable is the log of the creation time (i.e. the time between the handler answering the call and the creation of the incident). In Column (2) the dependent variable is the number of characters in the first line of the description of the incident (maximum number of characters = 210). In Column (3) the dependent variable is the number of words in the first line of the description of the incident. In Column (4) the dependent variable is the log of the not ready time following the creation of the incident. In Column (5) the dependent variable is a dummy for whether the not ready time takes value bigger than zero. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.8: Heterogeneity of Same Room by Incident Characteristics

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	.001	-.001
	(.008)	(.006)
Same Room X Urgent	-.019***	-.007
	(.008)	(.006)
Same Room X Information Intensive	-.021***	-.02***
	(.008)	(.006)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted in measures of the urgency and information intensity of an incident. To compute the information intensity variable we use the sample from 2008 to 2014 and calculate the average creation time (i.e. the time between the handler answering the call and the creation of the incident) of every opening code/grade combinations. We then assign to every opening code/grade incident type its average creation time, and label an incident type as being information intensive if its average creation time is above the median. To compute the urgency variable, we do a similar exercise using allocation time instead of creation time. All regressions also include indicators for Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler, and opening code/grade indicators. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.9: Heterogeneity of Same Room by Worker Workload

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	-.011*	-.008*
	(.006)	(.004)
Same Room X High Operator Workload	-.018**	-.012*
	(.008)	(.006)
Same Room X High Handler Workload	-.006	-.01*
	(.008)	(.006)
High Operator Workload	.128***	.046***
	(.005)	(.004)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted with measures of the workload of the operator and handler. To compute the operator workload measure, we use the number of incidents created in the operator's subdivion during the index hour. To compute the handler workload measure, we use the number of Manchester-wide incidents during the index hour, divided by the number of handlers on duty during that hour. The variables in the regression are dummies taking value one when the workload is above the sample median. We report the uninteracted operator workload measure. The uninteracted handler workload measure is absorbed by the Year X Month X Day X Hour of Day fixed effects. All regressions also include indicators for Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.10: Heterogeneity of Same Room by Handler-Operator Demographic Distance and by Number of Past Interactions

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	-.021	-.031*
	(.023)	(.018)
Same Room X Same Gender	-.016**	-.019***
	(.008)	(.006)
Same Room X Log Difference in Age	.025***	.024***
	(.005)	(.004)
Same Room X Log N Past Interactions	-.021***	-.019***
	(.005)	(.004)
Same Room X Log Handler Experience	-.004	-.003
	(.004)	(.003)
Same Room X Log Operator Experience	.005	.009*
	(.006)	(.005)
Same Gender	-.002	-.003
	(.004)	(.003)
Log Difference in Age	.013***	.01***
	(.003)	(.002)
Log Number Past Interactions	-.073***	-.061***
	(.005)	(.004)
Log Handler Experience	.058***	.045***
	(.009)	(.007)
Log Operator Experience	-.057	-.026
	(.049)	(.036)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted with whether the Radio Operator and the Handler are of the same gender, with the log of their difference in age, and with the number of previous incidents in which they have worked together. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator and Handler. All regressions also control for Handler Experience and Operator Experience and their interactions with Same Room. Standard errors are clustered at the Year X Month X Radio Operator Room level.

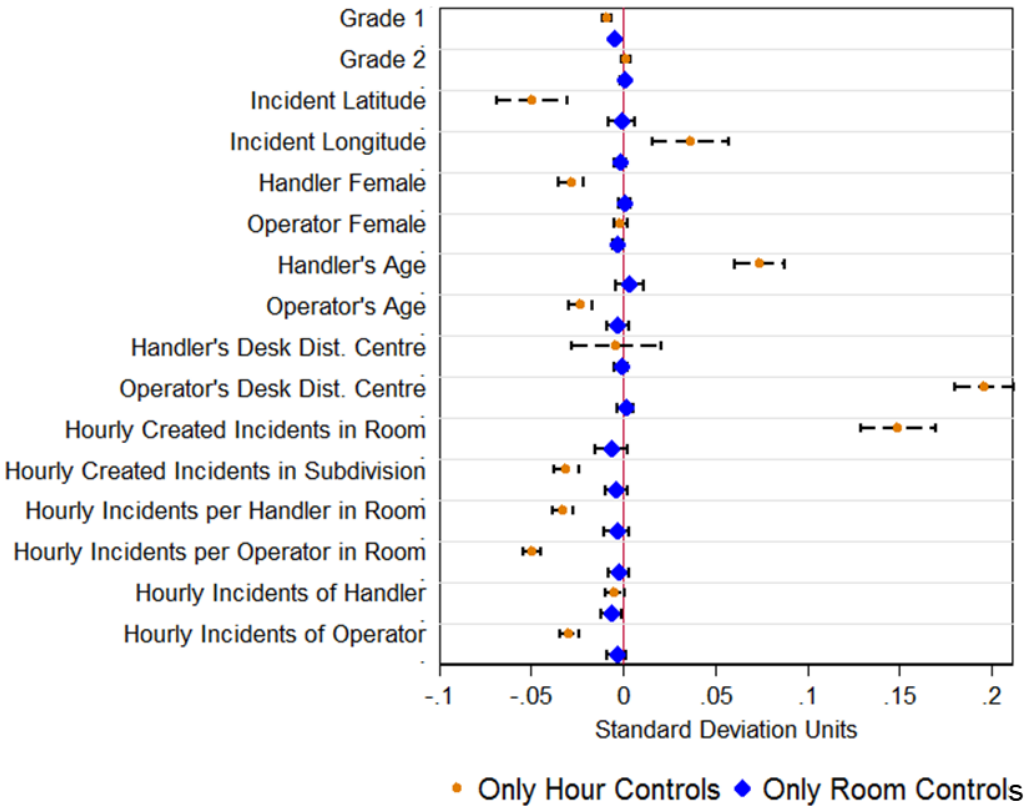
Table 1.11: Opportunity Cost of Higher Call Duration

Dep. Var. = Log Queuing Time	(1) 15 min. Window	(2) 30 min. Window	(3) 60 min. Window
Panel A: All			
Log Calls	.843***	.832***	.734***
	(.005)	(.006)	(.006)
Log Handlers	-.881***	-.872***	-.773***
	(.007)	(.007)	(.008)
Log Avg Call Duration	.582***	.819***	.959***
	(.008)	(.01)	(.011)
Panel B: Low Organisational Slack			
Log Calls	.301***	.35***	.379***
	(.008)	(.009)	(.01)
Log Handlers	-.308***	-.368***	-.418***
	(.01)	(.011)	(.012)
Log Avg Call Duration	.402***	.603***	.776***
	(.009)	(.011)	(.014)
Panel C: High Organisational Slack			
Log Calls	1.827***	1.664***	1.48***
	(.018)	(.02)	(.021)
Log Handlers	-1.653***	-1.514***	-1.326***
	(.018)	(.02)	(.02)
Log Avg Call Duration	.941***	1.184***	1.272***
	(.013)	(.016)	(.018)

This table displays estimates of OLS regressions of queuing time on measures of organisational slack and average call duration in the period preceding the start of the call. We estimate the effects separately at 15, 30 and 60 minutes periods before the call. High organisational slack is defined as periods during which the number of calls per handler was below the median. The sample includes all calls received by the GMP during the second semester of 2011. N=909256 for panel A, N=455023 for panel B and N=454233 for panel C. All regressions include an indicator for whether the call reached the GMP through an emergency line.

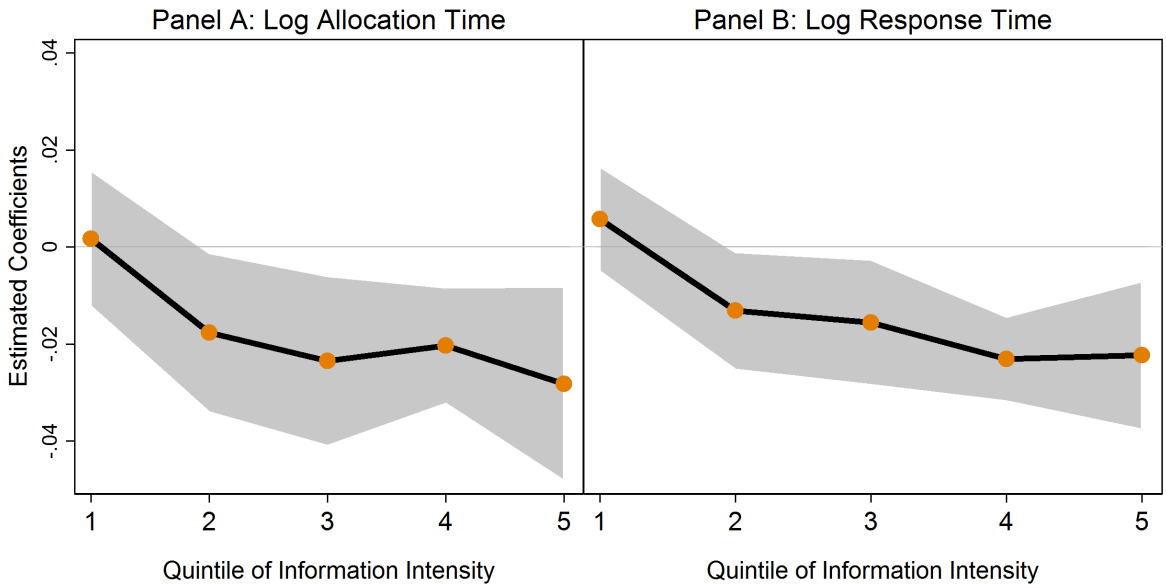
1.11 Appendix A: Additional Figures and Tables of the Chapter

Figure 1.A1: Balance of Incident, Worker and Room Characteristics on Same Room Incidents



Each row in the figure displays the results of two regressions, where the row variable is the dependent variable and Same Room is the independent variable. The first regression includes only Year X Month X Day X Hour of Day controls and the second regression includes only controls for Radio Operator Room and Call Handler Room. The displayed 95% confidence intervals are for the coefficient of the Same Room variable. Non-binary dependent variables are standardised. Standard errors are clustered at the Year X Month X Radio Operator Room level. Grade 1, Grade 2, Handler Female and Operator Female are the only dummy variables. Handler's Desk Dist. Centre is the euclidean distance between the handler's desk and the centre of the room. Hourly Incidents per Handler in Room is the number of incidents created during the hour of the index incident, divided by the number of handlers working during that hour. A similar definition applies to Hourly Incidents per Operator in Room. Hourly Incidents of Handler is the number of incidents created by the handler in charge of the index incident, during the hour of creation. Hourly Incidents of Operator is the number of incidents allocated by the operator in charge of the index incident, during the hour of the creation of the incident.

Figure 1.A2: Heterogeneity of the Effect of Same Room By Information Intensity of Incident



Each panel displays a different regression. The displayed coefficients are for Same Room X Quintile of Information Intensity. To compute the information intensity variable, we use the sample 2008-2014 and regress the log of creation time (i.e. the time between the handler answering the call and the creation of the incident) on the opening code/grade indicators. We then assign to every opening code/grade incident type its predicted creation time, and split the incident types into quintiles of predicted creation time. 95% confidence intervals are displayed in the shaded grey area. All regressions control for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator Identifier and Call Handler Identifier, and Opening Code/Grade Indicators. Standard error are clustered at the Year X Month X Radio Operator Room level.

Table 1.A1: Robustness to Controls

	(1)	(Baseline)	(3)	(4)	(5)
Log Allocation Time	-.023***	-.02***	-.018***	-.019***	-.02***
	(.004)	(.004)	(.004)	(.004)	(.004)
Log Response Time	-.02***	-.017***	-.016***	-.016***	-.017***
	(.003)	(.003)	(.003)	(.003)	(.003)
Hour F.E.	Yes	Yes	Yes	Yes	Yes
Grade/Call Source F.E.	Yes	Yes	Yes	Yes	Yes
Room F.E.	Yes	Yes	No	Yes	Yes
Individual F.E.	No	Yes	No	Yes	Yes
Room/Date F.E.	No	No	Yes	No	No
Individual/Month F.E.	No	No	Yes	No	No
Opening Code/Grade F.E.	No	No	No	Yes	No
Handler Position F.E.	No	No	No	No	Yes

This table displays estimates of OLS regressions of allocation time and response time on whether the call handler and the radio operator re located in the same room. The sample is the basleine sample. Every coefficient is from a different regression. Standard errors clustered at the Year X Month X Operator Room level.

Table 1.A2: Alternative Clustering

Dep. Variable	(1) Log Alloc. Time	(2) Log Response Time
Panel A: Baseline		
Same Room	-.0201***	-.0172***
	(.004)	(.003)
Panel B: By Handler/Operator Pair		
Same Room	-.0201***	-.0172***
	(.0041)	(.0032)
Panel C: By Subdivision		
Same Room	-.0201***	-.0172***
	(.0039)	(.003)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A3: Heterogeneity of Same Room by Incident Characteristics. Prediction with Out of Sample Data

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	-.001	-.002
	(.009)	(.006)
Same Room X Urgent	-.015	-.006
	(.009)	(.007)
Same Room X Information Intensive	-.024***	-.026***
	(.01)	(.007)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted in measures of the urgency and information intensity of an incident. To compute the information intensity variable we use the post-2012 and calculate the average creation time (i.e. the time between the handler answering the call and the creation of the incident) of every opening code/grade combinations. We then assign to every opening code/grade incident type its average time to creation, and label an incident type as being information intensive if its average time to creation is above the median. To compute the urgency variable, we do a similar exercise using the allocation time instead of the handler's time to creation. All regressions also include indicators for Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler, and opening code/grade indicators. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A4: Heterogeneity of Same Room by Incident Characteristics. Interaction with Variables in Logs

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	.081***	.066***
	(.027)	(.02)
Same Room X Non-Urgent	.01***	.006***
(in Logs)	(.003)	(.002)
Same Room X Information Intensive	-.079***	-.064***
(in Logs)	(.021)	(.016)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted in measures of the urgency and information intensity of an incident. To compute the information intensity variable we use the sample from 2008 to 2014 and calculate the average creation time (i.e. the time between the handler answering the call and the creation of the incident) of every opening code/grade combinations. We then assign to every opening code/grade incident type its average time to creation, and use the variable in logs. To compute the urgency variable, we do a similar exercise using the log of the allocation time instead of the log of the handler's time to creation. All regressions also include indicators for Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler, and opening code/grade indicators. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A5: Heterogeneity of Same Room by Demographic Distance (median) by Number of Past Interactions (median)

Dep. Variable	**(1) Log Alloc. Time**	**(2) Log Response Time**
Same Room	-.0107	-.0093
	(.0093)	(.0071)
Same Room X Same Gender	-.0191***	-.0215***
	(.008)	(.0061)
Same Room X Difference in Age High	.0129	.0125**
	(.0079)	(.006)
Same Room X Number Past Interactions High	-.0005*	-.0001
	(.0003)	(.0002)
Same Gender	-.0027	-.0033
	(.0041)	(.003)
Difference in Age High	.0166***	.0077
	(.0063)	(.0048)
Number Past Interactions High	-.0338***	-.0274***
	(.0049)	(.0038)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted with whether the Radio Operator and the Handler are of the same gender, with their difference in age (measured as an above median dummy), and with the number of previous incidents in which they have worked together (measured as an above median dummy). All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator and Handler. All regressions also control for Handler Experience and Operator Experience. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A6: Investigating Spillovers on Non-Same Room Incidents, by Same Room Incidents

	Spillovers by Same Room Incidents during Period:					
	60 min.		**30 min.**		**15 min.**	
Dep. Var	**(1) Log Alloc Time**	**(2) Log Resp Time**	**(3) Log Alloc Time**	**(4) Log Resp Time**	**(5) Log Alloc Time**	**(6) Log Resp Time**
% Same Room Incidents Rece by Operator	.001 (.006)	.001 (.005)	-.001 (.007)	.002 (.005)	-.004 (.008)	-.003 (.006)

This table investigates potential spillovers from Same Room incidents into non-Same Room incidents. The sample includes only incidents where Handler and Operator were in different rooms (N=734767). The dependent variables in the OLS regressions are log of the allocation time and log of the response time. The independent variable is the percentage of incidents during the index incident time period for which the call handler and the radio operator were located in the same room, excluding the index incident. In Columns (1) and (2) the period comprises of 60 minutes (respectively, 30 minutes for columns (3) and (4) and 15 minutes for columns (5) and (6)). All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. The regressions also include indicators for whether there were no calls received by the Radio Operator during the time period. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A7: Robustness to Controlling for the Time Period More Precisely

Dep. Variable	(1) Log Alloc. Time	(2) Log Response Time
Panel A: Baseline (60 minutes)		
Same Room	-.0201***	-.0172***
	(.004)	(.003)
Panel B: 30 minutes		
Same Room	-.0207***	-.0177***
	(.004)	(.003)
Panel C: 15 minutes		
Same Room	-.0198***	-.0179***
	(.0041)	(.0031)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy. All regressions include indicators for Grade, Call Source, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. In Panel A we also include Year X Month X Day X Hour of Day. Panel B substitutes the Hour of Day by the half hour period. Panel C substitutes by the 15 minute period. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A8: Correlation Between Measures of Other Actions by the Handler

Dep. Variable	(1) Log Number of Words	(2) Log Number of Characters	(3) Log Number of Characters
Log Time to Creation	.076***	.076***	
	(.005)	(.005)	
Log Number of Words			.906***
			(0)
Pairwise Correlation	.12	.14	.97

This table displays estimates of the conditional correlation among three actions by the handler during the creation of the incident. The sample includes all incidents received by the GMP between 2008 and 2013 where the dependent and independent variables are available (N=956440). The log of the handler's time to creation is the time between the handler answering the call and the creation of the incident. The number of characters is measured in the first line of the description of the incident (maximum number of characters = 210). The number of words is also measured in the first line of the description of the incident. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level. The unconditional correlation coefficients are also reported.

Table 1.A9: Heterogeneity of Same Room by Incident Grade

Dep. Variable	(1) Log Alloc. Time	(2) Log Response Time
Same Room X Grade 1	-.023***	-.013***
	(.005)	(.004)
Same Room X Grade 2	-.016***	-.016***
	(.006)	(.004)
Same Room X Grade 3	-.014	-.013*
	(.009)	(.007)
P-Value G1 $\neq$ G2	.336	.552
P-Value G1 $\neq$ G3	.412	.955
P-Value G2 $\neq$ G3	.885	.722

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted in the Grade of an incident. All regressions also include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A10: Robustness to Exclusion of Outlying Observations

Dep. Variable	(1) Log Alloc. Time	(2) Log Response Time
Panel A: Excluding .5%		
Same Room	-.0193***	-.0171***
	(.0039)	(.0029)
Panel B: Excluding 1%		
Same Room	-.0196***	-.0164***
	(.0038)	(.0028)
Panel C: Excluding 5%		
Same Room	-.0174***	-.0136***
	(.0036)	(.0026)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy. All regressions include indicators for Grade, Call Source, Radio Operator Room, Call Handler Room, Radio Operator and Call Handler. In Panel A Column (1) (respectively, Column (2)) we drop from the baseline sample the observations with the .5% highest values of allocation time (respectively, response time). In Panels B and C we do the same for the 1% and 5% highest values. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A11: Distance Inside Room and Past Interactions Handler/Operator

Dep. Variable	(1) Log Distance	(2) Log Distance
Log Number Past Interactions	.002	.005
	(.003)	(.005)
Pair Fixed Effects	No	Yes

This table displays estimates of OLS regressions of distance inside room on the number of past incidents on which the handler and the operator worked together. The sample includes all incidents received by the GMP between 2009 and 2012 for which handler and operator were based in the same room (N=209180. The distance between their desks is calculated as the euclidean distance in the floorplans provided by the GMP. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year and Call Handler Room X Year. Column (1) also includes Radio Operator and Call Handler Identifiers. Column (2) also includes Radio Operator/Call Handlers Pair Identifiers. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A12: Heterogeneity of Same Room by Distance Inside Room Controlling for Pair X Semester

Dep. Variable	(1) Log Allocation Time	(2) Log Response Time
Same Room X Log Distance	.032***	.019**
	(.011)	(.009)

This table displays estimates of OLS regressions of allocation time and response time on whether the call handler and the radio operator are located in the same room, interacted with the distance between their desks when they are in the same room. The sample includes all incidents received by the GMP between 2009 and 2012. The distance between their desks is calculated as the euclidean distance in the floorplans provided by the GMP. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year and Call Handler Room X Year, and Radio Operator/Call Handler/Year/Semester Identifiers. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A13: Balance of Incident, Worker and Room Characteristics on Same Room

Controls	(1) None	(2) Hour/Room
Grade 1	-.011***	-.004***
	(.001)	(.001)
Grade 2	.001	.001
	(.001)	(.001)
Incident Latitude	-.05***	0
	(.01)	(.004)
Incident Longitude	.036***	-.002
	(.011)	(.002)
Handler Female	-.029***	.001
	(.004)	(.002)
Operator Female	0	-.003***
	(.002)	(.001)
Handler's Age	.077***	.004
	(.007)	(.004)
Operator's Age	-.025***	-.003
	(.003)	(.003)
Handler's Desk Dist. Centre	-.003	-.001
	(.013)	(.002)
Operator's Desk Dist. Centre	.195***	.001
	(.009)	(.002)
Hourly Created Incidents in Room	.161***	-.005
	(.012)	(.003)
Hourly Created Incidents in Subdivision	-.042***	-.003
	(.004)	(.003)
Hourly Incidents per Handler in Room	-.049***	-.001
	(.004)	(.003)
Hourly Incidents per Operator in Room	-.063***	-.003
	(.004)	(.002)
Hourly Incidents of Handler	-.013***	-.005*
	(.003)	(.003)
Hourly Incidents of Operator	-.038***	-.003
	(.003)	(.002)

This table displays estimates of OLS regressions of allocation and response time on the Same Room dummy, interacted with whether the Radio Operator and the Handler are of the same gender, with the log of their difference in age, and with the number of previous incidents in which they have worked together. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year, Call Handler Room X Year, Radio Operator and Handler. All regressions also control for Handler Experience and Operator Experience and their interactions with Same Room. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Table 1.A14: Baseline Estimates Dependent Variables in Levels

Dep. Variable	**(1) Allocation Time (Minutes)**	**(2) Response Time (Minutes)**
Same Room	-1.275*	-1.921**
	(.685)	(.747)

This table displays estimates of OLS regressions of allocation time and response time on whether the call handler and the radio operator are located in the same room. The regressions are equivalent to those of Columns (1) and (2) in Table 3, with the exception that the dependent variable are in levels. All regressions include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room X Year and Call Handler Room X Year, and Radio Operator/Call Handler/Year/Semester Identifiers. Standard errors are clustered at the Year X Month X Radio Operator Room level.

Chapter 2

The Persistent Effects of Brief Interactions: Evidence from Immigrant Ships.

2.1 Introduction

It has long been shown that social connections play an important role in shaping economic outcomes. (Jackson, 2011; Topa, 2011; Beaman, 2016; Breza, 2016). Evidence to date has focused on connections established over lengthy periods, or among individuals strongly related in their demographic characteristics. However, many social interactions are circumstantial, brief and with previously unknown individuals. These interactions could also have measurable effects, especially for individuals facing critical moments in their lives. For instance, Bandura (1982) argues that "Some fortuitous encounters touch only lightly, others leave more lasting effects, and still others lead people into new life trajectories.". Chance encounters are also at the heart of theories such as those explaining agglomeration economies (Jacobs, 1969; Glaeser, 1999; Sato & Zenou, 2015). The potential value of brief fortuitous interactions has also been recognized by many organisations, which have implemented reforms to encourage these interactions.[1] Despite their potential, brief interactions

[1] The following quote by Scott Birnbaum, Vice President of Samsumg Semiconductors is instructive: *"... our data suggest that creating collisions - chance encounters and unplanned interactions between knowledge workers, both inside and outside the organization-, improves performance."* (Waber, et al., 2014).

have received little empirical attention due to endogeneity and measurement issues.[2]

This paper studies migrants travelling to the US by ship during the first half of the 20th Century. Migrants were placed together in trips lasting no more than a few days. Many faced the need to rapidly learn about potential jobs and final destinations. The dataset follows a large number of individuals who first met while travelling to the US and measures their outcomes many years after arrival. Therefore, this setting provides a unique opportunity to study the value of brief interactions in high-stakes decision contexts.

The dataset links 350,000 male immigrants to their ships of arrival and includes rich geographical information on towns of origin and ports of departure.[3] For every individual, I construct proxies for the quality of his connections upon arrival, exploiting information on the settled immigrants from his town of origin.[4]. More specifically, for each individual, I measure two aspects of his potential connections upon arrival: (a) the average earnings (in the US) of previous migrants from his town of origin and (b) the number of previous migrants from his town of origin. Then, I use these variables to proxy the average quality of an individual's previously unknown shipmates.

The empirical strategy relies on the assumption that, conditional on their towns of origin, individuals departing from the same port and in the same week, were plausibly exogenously assigned to ships. This differential assignment creates variation in the characteristics of the (previously unknown) shipmates of an individual. The identification strategy thus compares individuals (exogenously) allocated to travelling in ships that differ in the quality of previously unknown shipmates. A number of balancing tests supports the notion that, conditional on baseline controls, the assignment of passengers to ships was uncorrelated with the characteristics of their

[2]A body of literature has studied the role of indirect and/or weak (e.g. acquaintances rather than friends) connections. This paper differ from this literature with its focus on the transitory and fortuitous character of the direct interactions between individuals.

[3]Previous studies relying on matched historical data have also used male samples (e.g. Ferrie, 1996; Abramitsky, et al., 2012, 2014). One of the main reasons is that surnames changes were common for females and this makes it difficult to match them across different datasets. In addition to this, female labor force participation is low in this period (Maurer & Potlogea, 2017).

[4]A number of studies have shown the importance of settled immigrants in the assimilation process of new arrived immigrants (Massey et al., 1987, Munshi, 2003; Edin et al., 2003; Lafortune & Tessada, 2012; Beaman, 2012)

previously unknown shipmates. I also provide evidence that the data matching procedure does not induce correlation among shipmates characteristics. In this sense, I perform a number of tests that suggest that, conditional on baseline controls, the probability that a passenger is matched to a census record is uncorrelated with any characteristic of the ship or the individual.

My findings are as follows. Firstly, individuals travelling with higher quality (i.e. better connected) shipmates, end up being employed in higher earnings occupations. This effect is economically significant and persistent in time. For instance, a movement from the lowest to highest quintile in terms of the shipmates' quality is associated with a 4% increase in US labor earnings. This baseline result is robust to: (a) using different measures of occupational earnings, (b) including a large set of additional controls, like, ship-route characteristics, date of arrival and vessel fixed effects, (c) using variation only from individuals boarding at different stops of the same trip and (d) using variation only from repeated trips of the same vessel.

My second set of results suggests that the main mechanism consists of shipmates providing access and/or information about employment opportunities and attractive final destinations. Firstly, I find that the sectors where migrants end up working are affected by the sectors of employment of their shipmates' contacts. Similarly, their final destinations are also affected by the locations of their shipmates' contacts. Secondly, when ships include migrants with different languages, the baseline effects are driven by shipmates speaking the same language. This suggests that some form of verbal communication mediated the effect. Thirdly, the baseline effects are stronger for individuals likely to benefit more from additional connections: (a) individuals travelling by themselves and (b) individuals with poor connections in the US. Overall, my findings provide strong evidence that migrants benefit from their shipmates' information and/or contacts.[5]

Contribution This paper provides, to the best of my knowledge, the first causal evidence on the economic importance of brief social interactions in high-stakes sit-

[5]My dataset is not well suited to disentangling a pure information effect (e.g. shipmates providing information on attractive sectors of employment or final destinations) from a direct access effect (e.g. shipmates providing job referrals or other type of support), and I leave this for future work.

uations. Equally important is the finding that the effects are largely contingent on individual characteristics. In particular, those travelling alone and with fewer connections at destination are more affected than those with a better network at destination. This suggests the existence of a substitution effect between pre-established interpersonal connections and circumstantial contacts.

Findings from this paper have implications beyond its particular historical setting. First, it is possible that there are many situations where individuals face critical decisions that are irreversible or have long term consequences. Examples include, parental choice of school or students choice of college major. Second, results are consistent with studies showing that labor market entry conditions have persistent effects on job assignment and wages (Oreopoulos, et al., 2006; Oyer, 2006; von Wachter & Bender, 2008). In this paper, I show that short-lasting events that take place just before job search started can affect earnings in the long run. Third, this paper contributes to the economic literature on immigrants assimilation process (Borjas, 1995, 2015, Bleakley & Chin, 2009) by providing evidence that information and conditions upon arrival can determine newcomers future economic success.

Finally, this paper also provides a methodological contribution. It is well known that for large datasets, popular record linkage approaches like Fellegi & Sunter (1969) or Feigenbaum (2016) become unfeasible due computational limitations. I develop a Machine Learning approach to link US immigrant and passenger lists that improves the efficiency of previous methods and can serve as a guide to other researchers matching records across large historical datasets.

Related Literature This paper relates to a number of areas of research. First, a large body of literature has shown the effects of networks and social connections in the context of labor markets (Montgomery, 1991; Marmaros & Sacerdote, 2002; Bayer et al., 2008; Ioannides & Loury, 2004; Bentolilla et al. 2010; Dustmann et al., 2015, Bramoullé et al., 2016; Glitz, 2017).[6] Most of this literature has focused on the importance of job referrals and job search methods to access better quality jobs.

[6]There is also a rich theoretical literature in the area of social networks. Recent reviews can be found in Jackson (2009, 2014), Goyal (2015) and Jackson et.al. (2016).

Related to the role of immigrant networks, a number of articles have measured the importance of connections for newly arrived individuals (Munshi, 2003, 2014; Edin et al., 2003; McKenzie & Rapoport, 2007; Beaman, 2012, Battisti et al., 2017). This paper differs from these studies in that I focus on the role of links created while travelling to destination rather than in the role of pre-existing contacts. This also suggests a link with a growing literature documenting how entry conditions to the labor market can have long-run effects on earnings (Brunner & Kuhn, 2009; Genda et al., 2010; Oreopoulos et al., 2006). Also, Kramarz & Skans (2014) find that strong social ties (parents) are an important determinant for the first job of young workers and that social ties become more important when information on potential openings are likely to be scarce.

Theoretical models from different fields have assigned an important role to random social interactions. For instance, in the seminal work of Jacobs (1969) random interactions foster innovation and transmission of ideas and in Glaeser (1999), they influence learning of skills.[7] Despite this theoretical work, there are no empirical studies measuring the importance of random encounters in this field. A notable exception is Fitjar & Rodriguez-Pose (2016) who surveyed 542 Norwegian firms engaged in innovation partnerships. They find that 10% of partnerships emerged from random encounters.

A number of previous studies have analyzed the effects of connections established over long periods (e.g. Sacerdote, 2001; Angrist & Lang, 2004). This paper separates from that literature in that the (initial) exposure to social interaction is short, 10 days on average. On the contrary, peer-effects studies typically focus on connections established over long periods.

This paper also relates to the literature on weak ties. Early research, mainly by sociologists (Granovetter, 1973, 1983) found that a significant number of individuals find their jobs through connections such as "friends of friends". This literature emphasizes the role of weak ties in conveying information not prevalent among relatives or close friends. A recent number of studies have analyzed the "strength of weak ties" hypothesis using recent available data (Yakubovich, 2005). Related to

[7]For a complete review of this literature see Ioannides (2012).

immigrant outcomes, Goel & Lang (2016) study the role of weak ties in job search of recent immigrants to Canada and Giulietti et al. (2014) find that the rural-urban decision is largely affected by weak ties. The type of interactions studied in this paper diverge from the concept of weak ties, usually defined as a subset of acquaintances with lower probability to be socially involved with one another.[8]

Finally, this paper relates to a body of research that study the process of immigrants' assimilation (Chiswick, 1978; Borjas, 1995, 2015; Bleakley & Chin, 2009). A number of determinants have been explored, including the role of language proficiency, age of arrival, macroeconomic conditions or the performance of settled immigrants. Findings from this paper suggest that the first social connections made by immigrants can affect the later economic success of immigrants.

Plan I describe the historical background and institutional setting in Section 2.2. I summarize the construction of the matched census-ships dataset in Section 2.3. The empirical setting and identification strategy is discussed in Section 2.4. Section 2.5, presents the main results of the paper and discuss the economic relevance of them. In Section 2.6, I provide evidence on additional outcomes and heterogeneous effects to establish the social interaction explanation as the preferred interpretation of results. Section 2.7 concludes.

2.2 Historical setting

The period 1850-1924 is often referred to as "The Age of Mass Migration". Official statistics indicate that during this period, more than 30 million individuals arrived in the US (Hatton & Williamson, 1998). This was a period of low administrative barriers to immigration that ended after the imposition of the 1924 Immigration Act which sharply reduced immigrant flows (Goldin, 1994).[9]

[8]Weak ties are defined in different ways in the literature. For instance, Giulietti et al. (2014), define an immigrant's weak ties as those individuals from his same community who are not his relatives. The theoretical model of Sato & Zenou (2015) associate the idea of "random encounters" to weak ties, although they acknowledge the difference with respect to previous studies.

[9]The immigration act of 1892 stated a minimum requirement by banning from entry any person "*unable to take care of himself or herself without becoming a public charge*" (Hutchinson, 1981). In practice this excluded individuals with poor health conditions (including insane) or

The vast majority of immigrants arriving after 1892 entered the US through Ellis Island in New York Harbor.[10] During peak years, Ellis Island registered more than 10,000 arrivals per day. Once arrived, immigrants were inspected and authorized to enter the country. The sub-sections below explain the typical stages of the immigration process. This starts when individuals buy their tickets and finishes with the standardized inspection process at Ellis Island.

Before Departure A typical immigrant would buy his ticket from an agent of the many shipping companies existing at the time.[11] The Passenger Act of 1819 required each vessel arriving from abroad to provide a manifest listing all passengers. Although the information covered by manifests improved over time, after 1904 manifests registered the universe of passengers from any class and nationality (Bandiera, et al., 2016). Given that the cost of any deportation was levied on shipping companies, they faced strong incentives to screen passengers before departing and check that information was accurate. Therefore, individuals were typically required to provide travel documents in advance in order to comply with manifest creation. Additionally, shipping companies carried out their own medical inspection and disinfection before departure.[12] As a result of these requirements, individuals attended the port some days before departing.[13]

The Immigrant Journey Once the medical inspection procedure was completed, passengers were allowed to board the ship for departure. The conditions on the ship

with criminal records as well as those travelling without enough money to support themselves for few days after arrival. By the end of this period, legislation gradually increased the barriers to immigration (Reisler, 1976; Scruggs, 1988). For instance, the 1917 Literacy Act increased the head tax and introduced a literacy test. The 1921 Emergency Immigration Act introduced a system of quotas mainly directed to reduce immigration from eastern and southern Europe. Another exception was the 1882 Chinese Exclusion Act which banned immigration of Chinese workers. The increase in restrictions was mainly driven by the increase of critical perceptions an attitudes towards immigration (Goldin, 1994).

[10] According to official statistics, more than 75% of total arrivals were through Ellis Island and this percentage increased considerably for European immigrants (Ferenczi-Willcox, 1929).

[11] Another common arrangement for travelling was prepaid tickets purchased in advance by relatives residing in the US. These tickets required to follow the same steps and procedures than standard tickets.

[12] Passengers usually received a card certifying the medical inspection and additional information like names, ship and manifest page/line. Passengers were instructed to attach the card to their coats and to show it to inspectors upon arrival.

[13] Some ports had facilities for those passengers waiting for departure. In other cases passengers had to pay for their own accommodation.

were poor for the vast majority, who travelled in steerage class. Rooms usually accommodated large groups and most spaces were shared with other steerage shipmates. Although some individuals traveled with relatives or acquaintances from their home town, a large number of social interactions are likely to have occurred among individuals who had never met before. The duration of the voyage depended on the route and port of departure. By 1910, a trip from Liverpool to New York could take between 6 and 9 days, but departures from Mediterranean ports could take more than two weeks if the route included intermediate stops. Although there was some variation in the duration of the trip, the adoption of the steam engine and other improvements in shipping technology notably reduced the importance of weather conditions (Hopkins, 1910).[14]

Some individuals, specifically those with prepaid tickets and strong connections in the US, had a final destination decided. Indeed, some individuals would have purchased train tickets in advance or relatives would have been waiting in the NY port. However, many passengers travelled with poor information and few contacts on arrival. Lafortune & Tessada (2012) compare the immigrants' answer regarding their intended final destination (if any) with the actual states of residence of recently arrived individuals in the census. They find that only a 45% of answers match with the actual geographical distribution of recent arrivals. Anecdotal evidence suggests that shipmates played an important role in either conveying information on potential destinations and sector of employment or in directly providing job referrals, accommodation and financial support after arrival.[15].

[14]This contrasts with transatlantic voyages during the late 19th century. For instance, there is a well documented evidence that during the Irish famine migration (1840-1850), weather conditions could delay the departure and the arrival of ships by many weeks (Laxton, 1996).

[15]For instance, Taylor (2010) provides an example of how destination within US were sensitive to shipmates' suggestions: *"...His mom gave him all the money she had and told him to go to America. He travelled south on foot until he reached Italy, boarded a ship, and landed in New York. People whom he'd met on the ship told him to go to the city of Buffalo because many Polish people lived there..."*. In a second example, Grossman (2009) illustrates that shipmates were also important in providing jobs and accommodation: *"... He took a boat from Cork to New York City. A priest he had met on the ship got him a room to stay in and his job at New York City's Biltmore Hotel..."*. Anecdotal evidence also document a large number of marriages among partners who met during the trip. Indeed, the "Records of the Board of Trade and of successor and related bodies" from the UK, officially registered 133 marriages *while* travelling to the US.

Arrival at Ellis Island When a ship arrived at New York Harbor, immigration officers requested the certified manifests and steerage passengers were conducted to Ellis Island station.[16] Due to the characteristics of inspection facilities, passengers were divided into groups of (approximately) 30 people following their order in the manifest. Passengers who bought tickets together had close manifest numbers. Therefore, families and close acquaintances were typically inspected as part of the same group and queued at the same desk in the Registry Hall. Immigrants had to pass a quick visual medical screening and then immigration clerks in the Registry Hall checked that the inspection cards and the manifest information matched. Finally, passengers answered a series of questions (with the help of official translators) attempted to detect those with criminal records, extreme political affiliations (e.g. anarchists) or likely to became a public charge.[17] Individuals suspected of not meeting the minimum entry standards were separated for further investigation, a procedure that could take several hours or even days. Despite the strict inspection procedure, official statistics reveal that only 2% of passengers were finally deported (US Bureau of the Census, 1975). After inspection, individuals were discharged to enter the US. At this point, many of them faced the decision of where to seek a new life and/or in which sector to apply for a job. The station had money exchange facilities and many railway agencies from whom they could buy tickets to any destination, including New York City. This paper studies how contacts established during the trip could have influenced decisions at this critical stage.

2.3 Ships-Census Matched Dataset

In this section I summarize the construction of the dataset and main variables used in the study. Some technical details are relegated to Appendix B where I explain in detail the steps involved in the matching process.

[16]First class and cabin passengers were usually inspected on board and discharged to enter the US without going through the main station.

[17]In practice, the criteria for excluding someone for being *likely to became a public charge*, was circumscribed to passengers with several health conditions or those with not enough money to pay for accommodation and food for a few days after arrival.

Data Sources The main dataset in this paper combines information from Passenger Lists and historical Censuses. The Passenger Lists contain the universe of 34,000 ship arriving to the New York port during the period 1909-1924.[18] The set of individual variables available in electronic format are: full name, age, gender, race, marital status and last place of permanent residence. I also observe the date of arrival, port of departure and name of the vessel. I compile additional information on ships' characteristics, ports of departure and European cities from multiple online sources.[19] For most of the analysis, I restrict the sample to ships sailing from non-US ports and located at a distance of 3,000 kilometers or more from the port of New York.[20] Individual census information corresponds to the full count of male immigrants from the Integrated Public Use Microdata Series (IPUMS) for years 1920 and 1930 (Ruggles et al., 2015). Figure 2.1 shows the yearly flow of passengers and the immigrant stock in Census for different sub-samples of the population. As discussed in Bandiera et al. (2016), discrepancies between passenger inflows and Census stock are largely driven by return migration and the large drop in immigration inflows after 1914 is due to the WWI.

Matching Census and Ships Data I match passengers' data with census records using first name(s), surname, year of birth and year of immigration. Passengers are matched to the closest census year after arrival (i.e. arrivals between 1909 and 1919 are matched to the 1920 census and the remaining to the 1930 census). This dataset allows me to observe the characteristics of immigrants once they are settled in the US, but also the details of the voyage to US, including the characteristics of his shipmates.

[18]Information from passenger lists is considered accurate and reliable (Weintraub and Point, 2017). The manifests corresponds to the National Archives and Records Administration microfilms series M237 and T715. Similar data has been used in Bandiera et al. (2016) who discuss in detail the accuracy and coverage of passenger lists during the period.

[19]I obtained information available from a number of websites including *www.jewishgen.org, www.stevemorse.org* and *www.theshiplist.com*. I also used information on passenger lists from the series of Family Archives CDs by Gale Research. Patricia MacFarlane provided generous access to the Immigrant Ships Transcribers Guild (ISTG) database which contains digitized passenger manifests and information on immigration during the period of my study.

[20]This excludes all Caribbean, Mexican and Canadian ports which usually account for voyages of short duration. It also excludes a large number of small vessels transporting workers and supplies from and to the Panama Canal zone. Canadian and Mexican citizens are also excluded from the sample.

The main challenge when matching passenger lists to Census records is the large volume of data.[21] Popular approaches (e.g. Fellegi & Sunter, 1969; Feigenbaum, 2016) can become unfeasible even after following the standard blocking strategy.[22] In Appendix B, I outline a Machine Learning procedure based on Levenshtein Automata that allows me to match records across large datasets. The approach is related to Feigenbaum (2014, 2016) but introduces a number of algorithmic improvements to increase the speed at which the method identifies individuals with similar names and/or surnames.[23] The matched sample consists of 351,289 individuals, 52% of them corresponding to the 1920 census year. The matching rate relative to the Census is around 12%.[24] After excluding individuals sailing from less than 3000 kilometers from New York or missing information on the town of origin or age outside the range 14-65, the sample is reduced to 206,383 individuals.

Geocoding Ports, Routes and Places of Origin I use an algorithm based on the Google Places API to obtain the latitude, longitude and (harmonized) name of departure ports for the universe of ships in the Passenger List data. In total, I identify around 500 different ports, including those located at Caribbean countries, Mexico or Canada. Figure 2.2 displays the ports identified outside the area excluded from the analysis. Using all the ports declared by passengers (regardless of whether the passenger is matched to the Census or not), I reconstruct the whole route of the ship. Appendix C provides more details on the geolocalization procedure.

[21] Matching based on names and surnames requires calculating string similarity measures, which are computationally demanding. Increasing the sample size exponentially increases the number of string comparisons and this usually becomes unfeasible unless further restrictions are imposed.

[22] Blocking restricts the search of potential matches within a smaller set of records, typically individuals with similar years of birth or arrival. Unfortunately, in my setting blocks are so large that the problem remains.

[23] Intuitively, these modifications reduce the number of repeated calculations required to compare among strings. This is (to the best of my knowledge) the first paper in economics implementing this efficient search approach to match historical data (e.g. Radix Tries Search and Block-Specific Dictionaries). A recent literature in Computer Science have studied the problem of matching large string data (e.g. Baeza-Yates & Gonnet, 1996; Schulz & Mihov, 2002). Unfortunately, there is no existing code or software implementation for these methods and most of them remain as theoretical contributions.

[24] The matching rate is comparable to studies tracking immigrants across census years (Ferrie, 1996; Abramitsky et al., 2012, 2014). However, as explained in Appendix B the Machine Learning approach requires a human trained random sample of matched individuals. When creating this sample, I use an strict criteria that resulted in a low number of false positive matches. Cross validation exercises reveal that the matching procedure is highly accurate with a false positive rate below the 0.1%. As discussed in a recent paper by Bailey et al. (2017), false positive matches in linked data are more problematic than false negative matches.

I also geocode information on the "last town of permanent residence" for passengers in the matched sample. The algorithm resembles that used for geocoding ports but it requires some pre-processing steps in order to correct for common typos and abbreviations, towns that disappeared over time and places reported in their original language.[25] The full procedure is described in detail in Appendix C . Overall, I identify around 11,000 different places of origin. Figure 2.3 displays the location of places identified in the matched sample. Appendix Figure 2.A1 shows the relative frequency of the main ports of departure and countries of origin.

Labor Outcomes Since the 1920 and 1930 censuses did not record information on individual income, I follow previous studies (Abramitsky et al. 2012, 2014; Maurer & Potlogea, 2017) and use the *Occupational Earnings Score* which assigns each individual the percentile rank of his occupation in terms of median earnings in 1950. Naturally, this measure is invariant to wage differences within occupations but it captures whether an individual is employed in a job that pays relatively more. As a robustness check, I use two additional measures. The first one is the *Duncan Socioeconomic Index*, which assign a (subjective) prestige rating to each occupation based on earnings, education and the 1947 National Opinion Research Center Survey (NORC). The second additional measure is the *Nam-Power-Boyd Index* (Nam & Boyd, 2004) which measures the percentage of the labor force employed in occupations with combined levels of education and earnings below the incumbent occupation.[26] Finally, in order to aid the interpretation of the results, I construct a measure of occupational earnings by assigning to each individual the median earnings of his occupation in 1940. Information on sectors of employment and occupations is created and harmonized by IPUMS based on unstructured text questionnaires answers.[27]

[25]The algorithm generates the following information: latitude and longitude of the place, name identified by the Google Places Api and the south-west/north-east coordinates of the smallest rectangle containing the place. A 20% of the records have missing information on the place of origin and a 15% of the observations are geocoded with a precision above the locality level (e.g. province).

[26]All these variables are created by the Minnesota Population Center and are comparable across individuals and census years (Ruggles et al. 2015).

[27]Although these variables are not directly comparable with more recent industry or occupation classifications (e.g. SIC or NAICS for industries or SOC for occupations), the disaggregation is comparable to 3-digits level and consistent accross census years.

Summary Statistics Table 2.1 presents some summary statistics of the data. Panel A reports aggregated information on the number of individuals, ships and places of origin for different sub-samples and data sources. The first column (full sample) includes individuals from any origin and age group. The matching rate, defined as the number of matched individuals with respect to the individuals observed in the Censuses, is 12.4%. Matched individuals are observed in approximately 34,000 different ships, departing from 422 ports and proceeding from 10,900 different places of origin.[28] After restricting the sample to individuals in the age group 14-65 with non-missing information on the place of origin and to ships departing from ports at a minimum distance of 3000 km. from New York, approximately 206,000 individuals from 15,000 ships, 170 ports and 8,200 places of origin remain in the sample.

Panel B reports basic statistics on individual and ship characteristics. Ships in the regression sample travelled an average distance of 6,500 kilometers (whole route). This distance would take about 10 days at 15 nautical knots, the average speed for steamers in that period. In the full passenger list data, an average ship transported 173 male passengers in the age group 14-65 (excluding those boarding at less than 3000 km from New York). Ship size is consistent with the findings in Bandiera et al. (2013) for the same period.[29]The average number of passengers per ship observed in the matched sample was about 20. Ships were very diverse in terms of places of origin: an average ship transported individuals from 15 different towns of origin (in the matched sample). A large proportion of passengers were single and travelled without any relative. At destination, most immigrants settled in urban places and 21% were observed living in New York in the next Census after their arrival.

[28]Table 2.1 indicates that 15% of places of origin are geographical units above the locality level (e.g. province). As a robustness check, in Appendix B I re-estimate the main results excluding these geographical units

[29]Bandiera et al. (2013) find that for the period 1892-1924, the average number of passengers per ship was approximately 500. However, after 1911, the average number of passengers drops below 200 per ship. After accounting for the gender, age and port restrictions in my sample, the average number of passengers is in the same range.

2.4 Empirical Setting

In this section, I explain the empirical strategy to estimate the effects of brief social interactions, and then justify it with a set of balancing tests. Establishing this causal effect is not an easy task. In addition to considering the exogenous allocation of individuals across ships, I need to consider the possibility that shipmates' characteristics can affect earnings through channels that do not require social interaction. I postpone the discussion of these confounding effects to Section 2.6, were I provide additional evidence on the social interaction mechanism.

Defining Brief Social Interactions The first step in the analysis requires defining the set of individuals who met for the first time during the voyage. For every individual, I identify this set by *excluding* any shipmate such that 1) shares the same town of origin or 2) has a similar surname, defined as a *Jaro-Winkler* distance below 0.1.[30,31] Along the paper, I will refer to them as the set of *unrelated shipmates*. In Section 2.5, I perform a set of exercises to rule out the chance that effects are driven by a weak definition of unrelated shipmates.

Connections on Arrival An important variable that I use below is the quality of potential contacts that immigrants had in the US. This is a key variable in the empirical strategy as I will proxy the quality of shipmates based on this dimension. Following a number of influential papers (e.g. Wegge, 1998; Munshi, 2003; McKenzie & Rapoport, 2007, 2010) I define the set of potential contacts at destination, as those individuals who emigrated in the past from the same place of origin. There are two additional reasons to use the community of origin as the relevant unit to define the social network at destination. First, there is a strong consensus among historians on the importance of settled immigrants in triggering chain migration and supporting new arrivals from the same community (Daniels, 2002). Second, during this period

[30]The Jaro-Winkler distance (Winkler, 1990) measures the similarity between two words based on the number and position of common characters.

[31]In addition to these conditions, I use the smallest rectangular area containing the place of origin to exclude any shipmate with area overlapping above 50%. This additional condition assures that no shipmate is considered "unrelated" due to a poor geocoding information (e.g. a shipmate with the same province of origin but without information on the exact town of origin). In Section 2.5, I show that the main results are robust to more strict conditions (e.g. excluding close towns)

the outcomes of newcomers are strongly correlated with the characteristics of settled immigrants from the same community.

To measure the quality of contacts on destination, I focus on two variables:[32]

1) *The average earnings score of settled immigrants from the same town of origin.*

2) *The number of individuals from the same town who emigrated to the US in the past.*[33]

The first variable proxies the economic status of potential contacts, based on the notion that wealthier connections can provide information or referrals on better jobs. The second variable proxies the size of the network at destination.[34]

Formally, I define $x_{c(k),t(k)}$ as the earning score for an individual k from town $c(k)$ and who travelled in period $t(k)$. This notation emphasizes the fact that each individual in the data is associated to a unique town of origin and emigration period. The average earnings of potential connections on land for individual j is defined as $X_{c(j),t(j)} = \sum_{r(k)=1}^{t-1} x_{c(k),r(k)}/N_{c(j),t(j)}$ with $N_{c(j),t(j)}$ being the number of individuals from town $c(j)$ who emigrated before period $t(j)$ and are observed in the census.[35] The number of potential contacts upon arrival for individual j, defined as $Z_{c(j),t(j)}$, can be measured as the size of emigration flows from town $c(j)$ to the US *before* period $t(j)$. Note that $Z_{c(j),t(j)}$ is measured using the whole passenger list but $X_{c(j),t(j)}$ and $N_{c(j),t(j)}$ are calculated using the matched sample only. This underlines the complementarity of the two measures. Table 2.1 Panel B, reports summary statistics about these variables. Earnings of potential contacts are measured in the scale of 0 to 100 and the average in the sample is 49.7. The average number of potential contacts of an individual is 9,300.

[32] As a robustness check, in Section 2.5, I re-estimate the main results using alternative definitions of connections on arrival.

[33] The earnings of settled immigrants are calculated only for towns observed in the matched sample as I have no information on earnings of non-matched individuals. The number of emigrants from each town is calculated using the full flow of passengers observed in the passenger lists since 1900. For a given immigrant, either variable is calculated using only individuals who travelled at least one month before him.

[34] Previous studies have measured the migrant network size in different ways. For instance, Munshi (2003) measures it as the share of immigrants from the home community while Beaman (2012) uses the number of individuals from the same country living in a given city.

[35] Note that earnings scores of individuals arrived in different years are usually observed in the same census year.

Figure 2.4 illustrates the relevance of previous definitions. Each panel of the figure displays the coefficients of the following regressions between individual outcomes and the quintiles of his potential contacts' characteristics, conditional on ship and predetermined individual characteristics:

$$Y_i = \sum_{q=1}^{5} \beta_q ContactsChar_i^q + \sigma_{s(i)} + \alpha I_i + \epsilon_i \tag{2.1}$$

where Y_i is an outcome of individual i (measured at the next Census after arrival), $ContactsChar_i^q$ is a dummy for the quintile q of some characteristic of the potential contacts of the individual (e.g. the number of individual's contacts $Z_{c(i),t(i)}$). Each regression controls for ship fixed effects $\sigma_{s(i)}$ and a set of predetermined individual characteristics I_i. Panel A shows the correlation between individual earnings and the average earnings (and number) of settled immigrants from the same town of origin. Panels B to D shows that the location of individuals and the sector of occupation are strongly correlated with those of previous emigrants from the same place. Thus, even if newcomers never interact with settled immigrants, we can think that at the moment of the trip, the previous definitions are predetermined predictors of immigrants' economic success.

Identification Strategy In order to identify the effects of brief social interactions, I rely on the assumption that, conditional on their towns of origin, individuals departing from the same port and in the same week, were plausibly exogenously assigned to ships. The plausibility of this assumption is empirically validated later in this section. The intuition behind the identification strategy can be illustrated with the following example: Assume that an individual with residence in Benevento (Italy) has decided to emigrate from the port of Naples (the closest to his town). Naturally, individuals departing in different years or seasons, may face different conditions at departure or arrival. Consequently, shipmates' characteristics can be correlated with unobserved determinants of the individual's earnings at destination. Consider, however, all the ships departing from Naples within a relatively narrow time horizon (e.g. a week). The identification strategy relies on the assumption that the individual assignment is uncorrelated with the characteristics of the unrelated

shipmates boarding the same ship.[36]

A number of historical facts support this assumption. First, the selection among passengers of different income took place mainly within ships, as every vessel had different classes and service upgrades. For instance, wealthy individuals usually travelled in first or cabin classes. Second, during a short window of time, the fares for lower class categories (e.g. third class or steerage) were remarkably similar across shipping lines for a given route.[37] The vast majority of immigrants travelled in steerage class. Third, delays due to paperwork or unexpected changes announced by the shipping company were common. Finally, passengers bought their tickets days or weeks in advance, without being able to anticipate the characteristics of their potential shipmates. Naturally, the exogeneity claim must be validated in the data, and in this section I discuss a number of empirical exercises that support this assumption.

A potential concern is that some vessel characteristics (for instance, their external look or capacity) can influence the individual decision, creating some endogenous sorting of passengers. In Section 2.5, I show that results are robust to the inclusion of a large set of ship characteristics and even of vessel fixed effects. Moreover, as shown below in this section, ship characteristics are strongly balanced with respect to the average shipmates' quality.

The exogenous allocation across ships, creates quasi-experimental variation in the pool of (unrelated) shipmates of each passenger. This implies that similar individuals can be exposed to a pool of shipmates with different quality of connections on land. An advantage of this strategy follows from the fact that the characteristics of contacts upon arrival are predetermined variables at the moment of the trip, thus not affected by any shock occurring after departure.

[36] In Section 2.5, I explore two alternative identification strategies based on the variation created by repeated voyages of the same vessel and by individuals boarding at different ports during the same trip.

[37] For instance, Hopkins (1910) reports that in 1909, all the steamers covering the Mediterranean service of the Cunard Line, North German Lloyd, White Star Line and Italian Royal Mail Lines had a basic minimum fare of $65 for third class (steerage). Indeed, when including all routes and services, more than 80% of steamers had a basic minimum fare between $55 and $65. This basic fare excluded any additional service or railway transportation.

Estimating Equation The baseline estimating equation is:

$$Y_i = \beta_1 \bar{X}_i + \beta_2 \bar{Z}_i + \theta_{p(i)} \times \lambda_{w(i)} + \delta_{c(i)} \times \pi_{t(i)} + \epsilon_i \tag{2.2}$$

where Y_i is a labor market outcome for immigrant i in the US. Consistently with the earlier discussion, I control for the interaction between $\theta_{p(i)}$ (a fixed effect for the port of departure) and $\lambda_{w(i)}$ (the fixed effect for the week of arrival).[38]

The main variables of interest, $\bar{X}_i$ and $\bar{Z}_i$, measure the quality of the connections of i's shipmates. The first variable is the average earnings score of the potential connections on land among i's shipmates. The second measure, is the average number of potential contacts among i's shipmates. As discussed in Section 2.3, potential connections on land for individual j are defined as the set of emigrants from the same town of origin. Formally, if $u(i,s)$ is the subset of passengers travelling in ship s and unrelated to i, I define $\bar{X}_i = \sum_{j \in u(s,i)} X_{c(j),t(j)}/n_{u(s,i)}$ with $n_{u(s,i)}$ being the number of unrelated shipmates for individual i. Similarly, I define $\bar{Z}_i = \sum_{j \in u(s,i)} Z_{c(j),t(j)}/n_{u(s,i)}$.[39] As defined before in this Section, for a given individual j, $X_{c(j),t(j)}$ is the average earnings in the US among individuals from town $c(j)$ who emigrated before period $t(j)$ and $Z_{c(j),t(j)}$ is the total emigration flow from town $c(j)$ to the US before period $t(j)$.

The baseline specification also controls for the interaction between $\delta_{c(i)}$ (a fixed effect for the town of origin of immigrant i) and $\pi_{t(i)}$ (a fixed effect for the semester of arrival). The inclusion of this interaction serves two purposes. First, it controls for unobserved time-variant characteristics that could result in individuals from specific

[38] Note that I do not observe the week of departure, however, conditional on the port of departure, this is similar to control for the week of departure. Moreover, the route of the ship accounts for almost all the variation in voyage duration. In Section 2.5, I present evidence that results are robust to the inclusion of the route fixed effects.

[39] Some technical aspects involved in the calculation are worth mentioning: (a) Note that both variables are averaged across unrelated shipmates, thus unaffected by their number; (b) As discussed in Section 2.2, most social interactions are likely to be among passengers boarding at the same port. For this reason I only calculate the average characteristics among this set of unrelated shipmates. In Section 2.5, I modify this definition and use the characteristics of shipmates from different ports; (c) I only use the characteristics of shipmates in the matched sample. As discussed by Ammermueller & Pischke (2009) and Sojourner (2013), failing to account for the full set of relevant peers, can introduce some attenuation bias in the results. Of course, the identification strategy assumes that the probability that shipmates' are matched is not systematically correlated with unobserved characteristics of the individual, after conditioning for the baseline controls. I address this concern later in this Section.

towns boarding certain ships with higher probability. This would be the case, for instance, if agencies sold tickets for different ships with varying intensity across regions of the country. Second, given that potential connections on land are defined at the town of origin level, it absorbs any characteristic of individual's own contacts. As discussed in Caeyers & Fafchamps (2017), this strategy eliminates any negative exclusion bias (Guryan et al., 2009) introduced by the fact that i's connections are excluded in the calculation of $\bar{X}_i$ and $\bar{Z}_i$.[40] All regressions cluster standard errors at the week of arrival level. In Appendix Table 2.A2, I show that baseline estimates are robust to alternative clustering choices.

Balancing Tests and Evidence of Exogenous Sorting This subsection discusses a number of tests supporting the identifying assumption outlined before. This is critical to establish a causal interpretation of the effects of shipmates' characteristics on future labor outcomes.

The first test consists of studying the correlation between the predetermined variables of an individual and those of his unrelated shipmates. The exogeneity claim requires that this correlation must be zero after conditioning on the interaction between the port of departure and the week of arrival. Therefore, for every individual in the matched sample, I calculate the average characteristics of his unrelated shipmates. In order to avoid the negative mechanical bias of leave-one-out correlations, I follow Bayer et al. (2008) and sample one individual per ship when performing these calculations. Column 1 of Table 2.2 reports the unconditional correlations and Column 2 conditions on Port of Departure X Week of Arrival.[41] Results indicate that the unconditional correlations are high and significant but all of them become low and insignificant (at 5% level) after controlling for Port of Departure X Week of Arrival.[42]

[40] I define $\pi_{t(i)}$ at semester level due to the relatively small size of most towns of origin. For instance, I observe very few week-port cells with more than one individual from the same town boarding different ships. In Section 2.5, I show that results are robust to controlling for the interaction between town of origin and the month of arrival.

[41] A number of predetermined characteristics in the test vary at the town of origin level, for this reason, I do not control for the town of origin fixed effect, but on a larger geographical level (e.g. provinces in the case of italy). Note however, that this imposes a more demanding condition for balance.

[42] Significance levels are bootstrapped by repeating 500 times the procedure of sampling one individual per ship.

The second set of tests is given by standard balance regressions. This consists of OLS regressions of a number of predetermined passenger and ship characteristics on the two main variables of interest, $\bar{X}_i$ and $\bar{Z}_i$. The results in Figure 2.5, where I label each row in the left axis by the dependent variable, plot the estimated 95% confidence intervals of the regression. Panel A plots the confidence intervals for the average earnings of unrelated shipmates' contacts on land. Similarly, Panel B corresponds to the average number of shipmates' potential connections on land. To illustrate the importance of the baseline controls, I report the estimates with and without the Port of Departure X Week controls.[43] To ease interpretation, all variables in the regressions are standardized.

I find that shipmates' characteristics are (unconditionally) correlated with individual and ship characteristics: the estimates are statistically significant for most dependent variables. The introduction of the baseline controls, however, greatly decreases the estimates which become extremely small in magnitude. For any left hand side variable, the coefficients imply that one standard deviation in either the number or the earnings of unrelated shipmates' contacts on land, has an effect lower than 0.05 standard deviations. Indeed, after controlling for Port X Week, only two of the 32 displayed coefficients are statistically different from zero at the 5% level.[44]

Overall, I interpret the results of this subsection as supporting the exogeneity of the variation of shipmates' characteristics among unrelated individuals departing from the same port during a given week. Consequently with these findings, In Section 2.5 I provide additional support for the identification assumption, by showing that the results are robust to the inclusion of a large set of additional controls.

Census-Ships Data Matching and Non-Random Sampling A potential concern in the study is that the matching process creates a non-random sample of the ships. A number of additional findings suggest that, conditional on baseline controls,

[43]Following the discussion in footnote 41, regressions include fixed effects for large administrative units. Additionally, in order to eliminate any potential downward exclusion bias (Guryan et al., 2009), I control for the earnings and number of passenger's own potential connections. Appendix Figure 2.A3 displays similar balancing tests using the same controls and sample used in the baseline specification (variables defined at town of origin level are then excluded)

[44]Since the right hand side variables can be correlated with each other, Appendix Figure 2.A2 displays the F-statistics of the joint significant test of each regression.

matching is not systematically correlated with individual or ship characteristics.

First, note that the dependent variable in the last row of Figure 2.5 is the (standardized) share of matched passengers within the ship. Conditional on the Week X Port controls, the correlation is extremely low in magnitude: One standard deviation increase in $\bar{X}_i$ or $\bar{Z}_i$, changes the matching rate in less than 0.02 standard deviations. Figure 2.6 further explores this idea and estimates the balance equation for quintiles of the shipmates' contacts characteristics.

Second, I estimate the correlation between the ship matching rate and a set of individual predetermined characteristics conditional on similar controls than those in the balance regressions. Figure 2.7 plots this regression. Estimated coefficients are insignificant for 12 out of 13 variables and low in magnitude in every case. Along with the balance tests, this evidence suggests that conditional on baseline controls, the matching algorithm does not correlate with individual outcomes. This is not surprising as surname characteristics are the main determinants of the matching rate, and within the Week X Port cell, they are not systematically different.

Finally, I use the full Passenger List data to study whether the probability of being matched correlates with ships characteristics. I regress a dummy variable indicating if the passenger was matched to Census on the full set of Ship fixed effects. Table 2.3 reports the F-statistic for the joint significance test of Ship fixed effects. Column (1) shows that without further controls, Ship fixed effects have significant predictive power on the matching rate. However, as shown in Column (2), after including the Week X Port fixed controls, Ship fixed effects are jointly insignificant.[45]

These findings also highlight an advantage of the empirical strategy: Even if matching is non-random for the whole sample (e.g. because some nationalities are easier to match), narrowing the variation to the Week X Port of Departure level eliminates any significant difference in matching rates across ships or individuals.

[45] A different concern is related to the partial observability of the relevant network structure. Under (conditional) exogenous sorting of individuals across ships, this would result in coefficients attenuated to some extent as discussed in Ammermueller & Pishcke (2009) & Sojourner (2013). In Appendix D, I discuss how the baseline results vary according to the matching rate and the implications for potential attenuation bias. Additionally, I discuss a number of simulations suggesting that the attenuation bias is relatively low in this setting.

2.5 Baseline Results

This section describes and interprets the baseline results of the paper. I also show that the effects of travelling with better connected shipmates persisted for years after the arrival. I then discuss a number of robustness tests aimed to provide additional support for the identification assumption. Finally, I discuss the robustness of results to alternative specifications and clustering of standard errors.

Baseline Estimates Table 3.1 reports estimates of Equation (2.2) for different measures of earnings and job quality. Column (1) indicates that both dimensions of shipmates' contacts quality have a positive and significant effect on individual earnings score. Exposure to shipmates with connections employed in jobs one percentile higher in the earnings distribution, increases individual earning score in 0.14 points. Similarly, every thousand additional (average) connections among shipmates increases earnings score by 0.05. Columns (2) to (3) reports the results for the alternative measures of job quality discussed in Section 2.3. Estimates indicate effects of a similar magnitude.[46] Although these variables are correlated with the earning score, they measure different aspects of job quality. Understanding the size of effects based on Earnings Score is not straightforward as the earning distribution is typically left-skewed. In order to ease the interpretation of my findings, I also report the estimates of Equation (2.2) when the dependent variable is the logarithm of the earnings derived from the 1940 Census.[47] Findings in Column (4) mean that an upward shift of 10 percentiles along the income distribution of shipmates' connections, increases individual earnings by 2,7%. Every thousand additional average connections among unrelated shipmates, increases earnings by 0.7%.[48]

Equation (2.2) can hide some non-linear relationship between individual earnings

[46]The Duncan Socioeconomic Index, reflects the social perception of the "prestige" associated to an occupation. The Nam-Power-Boyd index captures differences in the education-earning composition of different occupations. Both variables have the same scale than the earnings score (0 to 100).

[47]The construction of this variable is described in Section 2.3.

[48]Appendix Table 2.A1 reports the results for two additional variables based on the 1950 Census. The dependent variable in Column (2) replicates the last column in Table 3.1 but using 1950 Census. Column (3) assign each individual the median earnings of the percentile associated to his occupation according to the earnings distribution in 1950. Results are robust to these alternative earnings measures.

and shipmates' connections quality. A potential concern is that results are driven by few ships with outlier characteristics. Figure 2.8 displays non-parametric evidence that the effects are increasing in the quintiles of the variables of interest. In the case of shipmates' connections earnings, effects are monotonically increasing and statistically significant for quintiles 3 to 5. Travelling in a ship in the highest quintile, increases individual earnings score in 1.8 points with respect to the lowest quintile (an effect of 4% according to the regression with log-earnings in panel B). In the case of the number of connections, the effects are weakly increasing but only significant for the highest quintile. Travelling in a ship among the highest quintile of this variable, increases individual earnings score by 1 point with respect to the lowest quintile (an increase of 2% based on the regression with log-earnings displayed in panel B). It is useful to compare these figures with the estimated correlations between earnings and the characteristics of individual's own connections in the US (Panel A of Figure 2.4). Although the later is not necessarily causal, it is a useful benchmark for interpreting the magnitude of the effects. Not surprisingly, the effects of shipmates' connections on earnings are lower than the correlation with respect to the own contacts' characteristics. For instance, relative to the lowest quintile, the effect of travelling with shipmates in the highest quintile of contacts' earnings is three to four times lower than the effects of having connections in the highest quintile of earnings.

Appendix Table 2.A3 explores the interaction between the two measures of quality of shipmates' connections. The estimated coefficients correspond to an OLS regression (analogous to Equation (2.2)) where the explanatory variables are the interactions between two sets of dummies indicating whether the number of shipmates' connections or their average earnings are above/below the median of its distribution. Both measures of connections' quality are relevant. Starting from a situation where shipmates have low-quality connections in terms of both earnings and number, an increase in either dimension has a positive impact on earnings. Table 2.A3 also suggests that the earnings of shipmates' connections is relatively more important than the number of shipmates' connections.

The baseline effects display some heterogeneity at geographical level. Appendix Figure 2.A4 plots the estimates of Equation (2.2) where the shipmates contacts' earnings variable is interacted with dummies for the country of origin of the indi-

vidual. The map shows the relative size of the effects for Europe. Among countries with more emigrants in the data, effects are stronger for Ireland, Poland and Greece. Naturally, other factors correlated with the country of origin can drive the heterogeneous effect. For instance, the estimated effect for Italians is significant but slightly below the median for Europe. This could be partially explained by the fact that Italians from distant regions typically spoke different languages. Unsurprisingly, the potential benefits of social interactions might depend on the ability to communicate with those well connected shipmates.

Persistence of the Effects Due to the low number of arrivals between 1914 and 1919, most immigrants in the data are observed many years after arrival (7.5 years on average). This suggests that effects of social interactions with unrelated shipmates is highly persistent. Figure 2.9 explores this idea in more detail and displays the estimates of the baseline equation where the right hand side variables are interacted with dummies for each year since arrival. Although this disaggregation can confound other characteristics correlated with the time since arrival, the figure suggest that effects are not only driven by recent migration. Moreover, estimated effects are statistically significant even 10 years after arrival.[49]

Additional Controls In this subsection I show that results are robust to the inclusion of a large number of additional controls. This evidence is important to rule out some potential threats to the validity of the identification strategy. Table 2.5 summarizes all these findings. Columns (2) and (3) show that estimates are robust to the inclusion of a set of individual characteristics (age, race, marital status, language, and an indicator for the individual travelling with some relative) and a set of characteristics of the ship and the route (e.g. ship capacity, number of passengers, distance travelled, number of stops, share of male passengers, etc.).

[49]There are two main confounders for this heterogeneous effect. First, earlier arrivals are older when observed in the Census, and additionally, given the high rate of return migration in this period, likely positively selected. Second, immigrant cohorts can differ in terms of skills and other unobserved determinants of earnings. Whereas the later can't be controlled for, I alleviate the first concern by controlling for the interaction between the right hand side variables and the age of the individual. An additional source of heterogeneity over time is the 1921 Immigration Act, which mainly affected immigration from eastern and southern European countries. Appendix Table 2.A4 shows the effects of shipmates' contacts characteristics interacted with dummies of pre/post 1921 Immigration Act. Results suggest that baseline findings are mainly driven by arrivals before 1921.

Robustness to these controls is consistent with the assumption that, conditional on baseline controls, the pool of shipmates is not correlated with individual or ship characteristics. In a more general way, I want to rule out that individuals select into ships due to unobservable characteristics of the ship. This would be the case if for instance, more educated individuals (which potentially correlates with their connections quality) select into ships with higher capacity or higher speed. Such situation would confound the effect of better connected shipmates with individual's different characteristics. Column (6) shows that effects are similar after controlling for vessel fixed effects and this finding is inconsistent with such interpretation.

Note that the baseline specification (Equation (2.2)) absorbs any shock at the Town of Origin X Semester level. Although this is an already narrow time-space grid, some concerns may arise regarding the relevant time horizon in which local shocks can affect passengers' predetermined characteristics.[50] Column (4) extends the baseline specification to a shorter window of time by controlling for the interaction between fixed effects of the town of origin and the month-year of arrival. Since most towns are relatively small, there are fewer cells with multiple individuals from the same town boarding different ships within the same month. Despite of the lower number of observations, results remain statistically significant with coefficients of similar magnitudes. Column (5) narrows the time horizon to the week level but uses a larger spatial aggregation grid (administrative units above the locality level, e.g. provinces in the case of Italy). In this case, results are similar for the earnings of shipmates' contacts and non-significant for the number of connections on land, although standard errors are also larger due to the introduction of a large number of fixed effects.

As discussed in Section 2.4, it is possible that ships departing from the same port during the same week, followed a different route. Although the vessel fixed effect controls for most of this variation, some vessels could have covered different routes over time. Column (7) shows that baseline results are robust to the inclusion of fixed effects for each route identified in the data. This rule out that results are driven by some correlation among shipmates' characteristics created by individuals

[50]For instance, it could be the case that a local shock greatly changes the quality of individual's own connections within a semester.

selecting across ships based on the travelled route.[51]

Finally, Columns (8) and (9) aim to control for a narrow set of individual characteristics and labor market conditions upon arrival. Column (8) includes fixed effects for the NYSIIS phonetic coding of surnames (Atack and Bateman, 1992) which accounts for approximately 8000 groups of surnames.[52] Column (9) includes fixed effects for the date of arrival. Despite of a lower number of observations, estimates are robust to the inclusion of the additional controls. These findings have a number of implications. First, surnames embeds some important unobserved characteristics of individuals. Thus, findings are consistent with the claim that conditional on baseline controls, passengers do not select into ships according to individual characteristics that correlate with earnings. Second, surname is the most critical variable when matching between Passenger Lists and Censuses. Some surnames are more difficult to match either because they are too frequent, or because they are more likely to be misspelled when transcribed. Therefore, results in Column (8) are inconsistent with a non-random matching across ships driving the results. Lastly, results in Column (9) rule out that some correlation between shipmates' characteristics and daily conditions upon arrival explains my findings. This would be the case if for instance, the arrival of passengers from certain towns triggered some events like a higher demand for train tickets to some destinations or a lower availability of temporary accommodation in New York City.

Narrowing the Definition of Unrelated Shipmates One potential concern when establishing a causal interpretation of Equation (2.2) is the possibility that shipmates from different places of origin are already connected before travelling. Although this is an unlikely event for the vast majority of passengers, I restrict in two ways the pool of shipmates assumed to be unrelated. First, I use the fact that travelling together (or buying the ticket from the same agent) typically implied nearby manifest line numbers. In Table 2.6, I report the estimates of the baseline equation but for every individual, I restrict the set of his unrelated shipmates by

[51] This is not surprising given that the ports concentrating most of the departures in this period are usually covered by few routes, and in many cases by a unique route.

[52] Including surname fixed effects is problematic for two reasons. First, the large variety of different surnames would absorb most of the variation at individual level. Second, a non-negligible part of the variation in surnames can be due to transcription errors or typos.

imposing a minimum distance in their ID numbers (which follows the same order than manifest line numbers). The first two rows of the table exclude any shipmate with a difference in ID numbers lower than 10 and 15 respectively. The second way in which I restrict this set is by excluding passengers with towns of residence located at less than 100 kilometers from each other. The last row of Table 2.6 displays the baseline results after imposing both sets of restrictions (Minimum ID number difference and minimum distance). Point estimates are somewhat lower for the earnings of shipmates' contacts (but they remain statistically significant at 1%) and they are similar for the number of shipmates' connections. Note that either restriction can introduce some attenuation bias if true unrelated shipmates are excluded.[53] Moreover, due to language constraints and social preferences, interaction with unrelated individuals can be more likely to occur among those from closer towns.

Alternative Definition of Connections on Arrival As discussed in Section 2.4, defining potential contacts in the US at the town of origin level is in line with a number of previous studies. However, in the setting of this paper, it is possible to think that narrower definitions of connections are also relevant (for instance, relatives who emigrated in the past). In Appendix Table 2.A5, I re-estimate the baseline specification using two alternative definitions of potential connections upon arrival to the US. First, in Column (2) I consider individuals with similar surname (based on the NYSIIS coding) who previously emigrated from the same province or large administrative unit. Second, I consider past emigrants from the same town of origin who share a similar surname (Column (3)). For small places, the second definition captures to a large extent, relatives who emigrated in the past. In order to ease the comparison across definitions, I standardize all the right hand side variables. Column (1) corresponds to the baseline definition.[54] Alternative definitions result in estimated effects of similar magnitude, and in both cases, higher than the baseline

[53] See footnote [45]

[54] Narrowing the definitions for potential contacts significantly reduce the number of observations and statistical power since, for instance, very few individuals from same town and with the same surname migrate in the same semester. For this reason, the specification in Table 2.A5 includes fixed effects for the group at which contacts are defined (e.g. Town of Origin X Surname) but interacted with census year instead of semester fixed effects.

results. Higher estimates can be due to a number of reasons. First, unique surnames are not included in the pool of unrelated individuals when computing earnings of shipmates contacts. Second, given a narrower definition, within ship variation in shipmates' characteristics is also larger. Finally, connections with settled emigrants of similar surname can be the main source of information and support upon arrival, or just better predictors of economic success for immigrants.

Alternative Identification Strategies I explore two different sources of variation in the characteristics of shipmates. The first strategy exploits the fact that many vessels travel from the same port repeatedly during the year. Therefore, I only compare passengers travelling in the same vessel within the same semester.. Column (1) in Appendix Table 2.A6 estimates the following equation:

$$Y_i = \beta_1 \bar{X}_i + \beta_2 \bar{Z}_i + \psi_{v(i)} \times \theta_{p(i)} \times \lambda_{y(i)} + \delta_{c(i)} \times \pi_{t(i)} + \eta_{r(i)} \times \pi_{t(i)} + \epsilon_i \quad (2.3)$$

where $\psi_{v(i)}$ is a vessel fixed effect, $\eta_{r(i)}$ is a route fixed effect and the rest of variables are defined identically to Equation (2.2). Estimates for this specification are displayed in Column (1). Point estimates are highly significant for the case of earnings of shipmates' contacts and the magnitude is approximately 40% lower than the baseline effects. These results provide additional evidence that baseline effects are not driven by passengers sorting across vessels.

The second alternative variation follows from the fact that some ships stopped at different ports before arriving in New York. In this case, I exploit the variation in shipmates' characteristics created by passengers from different ports. A potential concern of this specification is that some ports can be very distant from each other reducing the potential interaction between these shipmates. Moreover, in many cases, shipmates boarding at different ports spoke different languages.[55] Additionally, the sample size is largely reduced because either there were no intermediate stops or because only few individuals boarded at a different port. Indeed, I exclude any ship where more than 90% of the passengers boarded in the same port. Column

[55]This was true not only for ships stopping at different countries. For instance, italians boarding at different ports typically spoke different languages/dialects and fluent communication among them was very unlikely.

(2) estimates the following equation:[56]

$$Y_i = \beta_1 \bar{X}_i^{sp} + \beta_2 \bar{Z}_i^{sp} + \alpha_1 \bar{X}_i^{dp} + \alpha_2 \bar{Z}_i^{dp} + \psi_{v(i)} + \delta_{c(i)} \times \pi_{t(i)} + \eta_{r(i)} \times \pi_{t(i)} + \epsilon_i \tag{2.4}$$

where $\bar{X}_i^{sp}$, $\bar{Z}_i^{sp}$, $\bar{X}_i^{sp}$ and $\bar{Z}_i^{sp}$ are defined similarly to Equation (2.2) but I distinguish between the characteristics of passengers boarding in the port (superscript sp) and that of those boarding at a different port (superscript dp). Estimates from this equation are displayed in Column (2). Point estimates are higher for the characteristics of same-port shipmates and only significant for the earnings of shipmates' contacts. Finally, in Column (3) I only consider individuals who boarded the ship at the first departing port and use the variation created by shipmates boarding at subsequent ports. Remarkably, point estimates for the earnings of shipmates' contacts are similar to those in Column (2).[57]

2.6 Mechanism: Establishing a Social Interaction Interpretation

The findings in the previous Section, show a causal link between the short run exposure to a pool of better connected individuals and future performance in the labor market. However, this reduced form result is compatible with a number of mechanisms that do not necessary require social interaction among unrelated shipmates. In this Section, I provide evidence supporting the social interaction interpretation as the most plausible one.

I start by showing, that the effect is stronger for passengers with fewer connections and that results are driven, to a larger extent, by shipmates speaking the same language (a natural mediator of social interaction). Then, I show that the sector of employment and place of residence of shipmates' contacts have predictive power on

[56]Note that I don't include the interaction between port of departure and the time dimension in order to exploit the variation across ports of the same route. Instead, I control for the interaction between the route and the semester of arrival. This is a less demanding specification compared to the baseline, but unfortunately, statistical power is too low to include Route X Week fixed effects.

[57]This also illustrates that my identification strategy is robust to exclusion bias (see Angrist, 2014).

the occupational and residential outcomes of the individual. Finally, as a reassuring exercise, I show that conditional on arriving in the same week and from the same port, the correlation in labor and residential outcomes is stronger among shipmates.

Heterogenous Effects As described in Section 2.3, this was a period of high-stakes decisions for most immigrants. Consequently, the effects of brief social interaction are expected to be higher for individuals with poor connections and no access to relevant information. Table 2.7 displays estimations of the baseline regression where each measure of shipmates' connections is interacted with dummies indicating how well connected is the individual himself. Column (1) explores the quality of connections on board, that is, whether the passenger is travelling alone or with relatives.[58] Individuals travelling alone are more benefited by travelling with higher quality shipmates. Column (2), shows that individuals with poor connections on land[59] are more affected by their shipmates' contacts characteristics. Finally, Column (3) shows that effects are stronger for individuals travelling alone *and* with poor connections on land.

A subset of ships in the sample contains shipmates who spoke different mother tongue (using Census definition). As verbal communication is an essential component of social interaction, I expect that the characteristics of shipmates who speak the same language are more relevant.[60] Table 2.8 displays the estimates of the baseline equation but separating among the characteristics of unrelated shipmates with similar and different language. Although the average effects are lower compared to baseline results, the coefficients associated to shipmates of similar language are always higher compared to those of different language.[61] In the next subsection, I

[58] In order to avoid confounding the effect with the surname prevalence, I include surname NYSIIS code fixed effects. This explains why, consistent with Table 3.1, the average effect is higher compared to the baseline.

[59] An individual is defined as low connected when the median earnings and the median number of past emigrants from his town are below the median.

[60] For instance, Bertrand et al. (2000) use common language to measure links within neighborhoods.

[61] A large number of ships are dropped from the sample because all matched passengers spoke the same language, thus, a number of reasons can explain the lower average effects. First, the remaining ships are larger than the average, with social interactions more difficult to detect or subjected to higher attenuation bias. Second, departures of "multilingual" ships are more concentrated after 1921, where social interactions were less important as shown in Appendix Table 2.A4. Finally, it could be the case that remaining ships covered routes and ports where individuals were less prone to social interaction or less benefited from it.

find evidence that shipmates that spoke the same language are also more relevant in explaining the sector of employment and place of residence of immigrants.

Sector of Employment and Residence Place of Shipmates' Connections According to the social interaction hypothesis, shipmates are important in providing information on potential destinations within US. They can also affect labor decisions either by granting access to their networks on arrival or by directly providing job referrals. Consequently, I expect that a number of immigrants migrated toward places where shipmates' contacts concentrates. Similarly, a number of immigrants should have got jobs in sectors where shipmates' contacts were employed with more intensity. I explore this idea with a number of tests.

First, I run three OLS regressions where the dependent variables are dummies indicating whether the individual is employed in primary activities, manufactures, or services.[62] The main explanatory variables are the share of shipmates' contacts employed in primary activities and the share employed in manufactures (services is the ommited category). Regressions also include the set of fixed effects in Equation (2.2). Table 2.9 displays the results of this exercise. Notably, results reveal that individuals travelling with shipmates' contacts employed more intensively in some sector, are also more likely to be employed in that sector.

Second, given that New York City was the most popular destination for immigrants, I study to what extent this desicion depended on the place of residence of shipmates' contacts. Figure 2.10 plots the OLS regression of a dummy variable indicating whether the immigrant remained in New York on the share of shipmates' contacts living in New York (I estimate this non-parametrically for the quintiles of the explanatory variable and controlling for the same set of fixed effects in baseline Equation (2.2)). The estimated effect is monotonically increasing and significant for the two highest quintiles.

Finally, similar conclusions are obtained using with a more granular definition of sector of employment and place of residence. I show this by estimating the following

[62]Based on IPUMS detailed industry classification.

OLS regression(s):

$$Y_{ij} = \beta S_{ij} + \ \gamma_i \ + \phi_{j(i)} \times \theta_{p(i)} \times \lambda_{w(i)} + \phi_{j(i)} \times \delta_{c(i)} \times \pi_{t(i)} + \epsilon_i \tag{2.5}$$

where Y_{ij} is an indicator variable that takes one if individual i is employed in sector j (or lives in place j) and zero otherwise, γ_i is an individual fixed effect[63] and $\phi_{j(i)}$ is a sector of empoyment (or place of residence) fixed effect. Consistently with the main identification strategy, $\phi_{j(i)}$ is interacted with the fixed effects in the baseline Equation (2.2). The main variable of interest is S_{ij}, the share of shipmates' contacts employed in sector j (or living in place j).

Table 2.10 displays the estimates of Equation (2.5). In Panel (A), the sector of employment is defined alternatively at one and two digits based on the IPUMS detailed classification. In either case, coefficients are highly significant. An increase of 10 percentage points in the share of shipmates contacts employed in sector j, increases by 0.8% the probability of working in that sector. In Panel B, I use two definitions for the place of residence. Column (1) shows the result for the state of residence and Column (2) for the city of residence.[64] Coefficients have a magnitude comparable to those in Panel A. Finally, Panel (C) displays the estimates of Equation (2.5) for the Sector of Occupation and the State of Residence with S_{ij} measured separately for shipmates of same and different language. Similar to previous findings, estimates are significantly higher for the characteristics of same-language shipmates.

Correlation in Labor and Residential Outcomes among Shipmates Baseline estimates exploit the variation in predetermined characteristics of shipmates. In this subsection, I follow a different approach by directly looking at labor and residential choices of unrelated shipmates. This exercise is complementary to the previous analysis in two ways. First, it is not affected by measurement issues related to the definition of networks characteristics (e.g. baseline estimates require to measure the earnings of settled immigrants). Second, it can account for social interaction effects, in situations where connections on land are less important for

[63]Note that each individual enters multiple times in this specification

[64]In the later case I exclude individuals with missing information on the city of residence or living in rural locations.

immigrant decisions.

I extend the empirical approach that Bayer et al. (2008) use to identify social interaction effects among neighbors. In this case, I compare the correlation in outcomes among individuals arrived during the same week, conditional on departing from the same port. As already shown in Section 2.4, predetermined characteristics of shipmates are uncorrelated once we control for the Week X Port interaction. Thus, it is plausible to assume that unobservable determinants of labor and residential outcomes are also uncorrelated. Under this assumption, (conditional) correlation in shipmates' outcomes can be interpreted as the causal effect of travelling together. Naturally, the main limitation of this test is that it does not rule out the presence of common shocks after departure.

More specific, I estimate the following equation using the combination of all possible (non-repeated) pairs of individuals arrived during the same week:

$$Y_{ih} = \beta SameShip + \gamma_i + \gamma_h + \theta_{p(i)} \times \theta_{p(h)} \times \lambda_{w(ih)} + \delta_{d(i)} \times \delta_{d(h)} + \epsilon_{ih} \qquad (2.6)$$

where Y_{ih} is a measure of similarity between the outcomes of (unrelated) individuals i and h. Variables γ_i, γ_h are passenger fixed effects. In order to compare individuals departing from the same port and week, the regression controls for the interaction between $\theta_{p(i)}, \theta_{p(h)}$ (port of departure fixed effects) and $\lambda_{w(ih)}$ (week fixed effect). As suggested above, common shocks experienced during the voyage or upon arrival can create some correlation in individual outcomes even in the absence of social interaction. To alleviate this problem, I control for $\delta_{d(i)} \times \delta_{d(h)}$, the interaction between the fixed effects for the dates of arrival of each passenger in the pair. Although this does not eliminate ship-specific shocks, it controls for any shock affecting passengers arrived during the same day. For instance, some types of jobs could have been advertised only during weekends.

Table 2.11 displays the estimates of Equation (2.6) for different outcomes. The dependent variable in Column (1) takes one if the pair of individuals works in the same sector and has the same occupation.[65] Travelling in the same ship, has an

[65]Occupations and sectors are defined at the most detailed level available in IPUMS created variables.

effect of 0.15 percentage points which corresponds to a 10% increase in the mean of the dependent variable. In Column (2), the dependent variable measures whether the pair works in the same sector within the same state. In this case, the effects are in the magnitude of 26% over the mean of the dependent variable. Columns (3) an (4) suggest that travelling in the same ship creates some geographical agglomeration. Column (3) shows that travelling in the same ship, is associated with a 3% reduction in the distance between the US residence place of (unrelated) individuals. Column (4) shows that the probability of living in the same city is 0.2 percentage points higher for unrelated shipmates.

Finally, I estimate the effects of *SameShip* interacted with a number of pair-specific characteristics. Consistent with previous findings in this section, Appendix Table 2.A8 shows that the effects are only driven by pairs of individuals who spoke the same language. Appendix Table 2.A9 explores the idea that pairs of individuals with strong connections upon arrival, should be less affected by brief social interactions. Results are consistent with this interpretation.

2.7 Conclusion

Although the role of chance encounters with previously unknown people has been long recognized by academics, and more recently by companies and managers, empirical evidence on this subject is largely absent. This paper provides causal evidence that brief social interaction with unknown people has economic relevance, provided they occur during critical life junctures. In particular, I study the effects of interactions among immigrants who met for the first time while travelling to the US during the period 1909-1924. Using a dataset of matched immigrants-ship with detailed geographical information, I have shown that conditional on their town of origin, individuals travelling with (previously unrelated) better connected shipmates, ended up being employed in better quality jobs. I identify this effect using the variation within the same port and week of departure and controlling for the town of origin. A number of tests show that this variation is plausibly exogenous and thus, results are credibly driven by differences in shipmates' characteristics.

A second set of estimations, provides suggestive evidence that the underlying mechanism is related to shipmates providing access and/or information on potential job opportunities or places of destination within the country. At the same time, heterogeneous baseline results highlight that random social encounters are more important for individuals with lower access to pre-established networks (i.e. contacts with immigrants from the same community and had settled in the US).

This paper prompts a number of implications beyond the particular setting of the study. First, my results indicate that the benefits of brief social interactions are larger for uninformed individuals or individuals with lower access to stronger forms of networks, like friends or relatives. Second, my results highlight the influence that interactions with unknown people can have in situations where individuals have to make critical decisions and information is scarce. This extends to a large number of settings, for instance, parental choice among schools or students choice of college major. A closely related implication is that economic outcomes among recent waves of refugees to Europe can be affected by the characteristics of those who they interact in the days surrounding the voyage (which include boat-mates but also border agents, NGO workers, etc.). More generally, results from this paper illustrates that brief episodes can have long-lasting effects on future earnings.

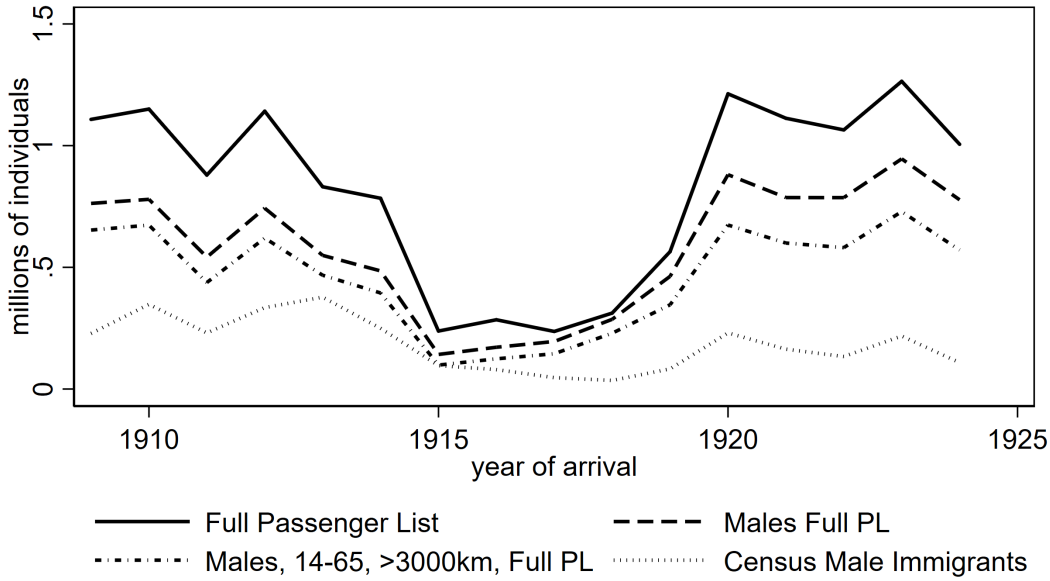
In recent years, a growing volume of individual level data has become available for researchers. In many cases, information is dispersed across multiple sources and merging across them relies on noisy string variables. Examples vary from historical full count census to recent automatic web generated data. This paper illustrates that incorporating tools from Computer Science can be highly valuable for applied researchers.

Finally, this paper leaves a set of open questions for future research. The extent to which brief social interactions matter in less critical situations can't be answered in the context of this study. Similarly, the setting is not suitable to disentangling between the pure information effect of brief interactions from the direct effect of providing access to better connections or financial support. In this sense, it would be relevant to study settings where individuals meet for a brief period and they never meet again. Finally, despite of recent trends in management practices, the

productivity effects of chance encounters within organizations remains largely unexplored.

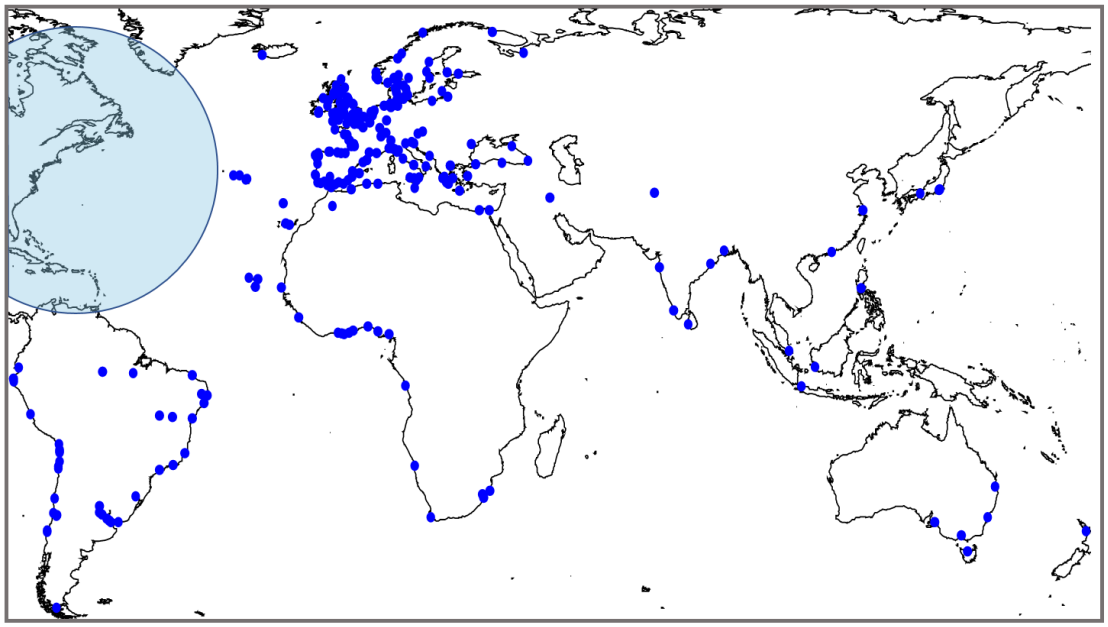
2.8 Figures of the Chapter

Figure 2.1: Passenger List and Census Data



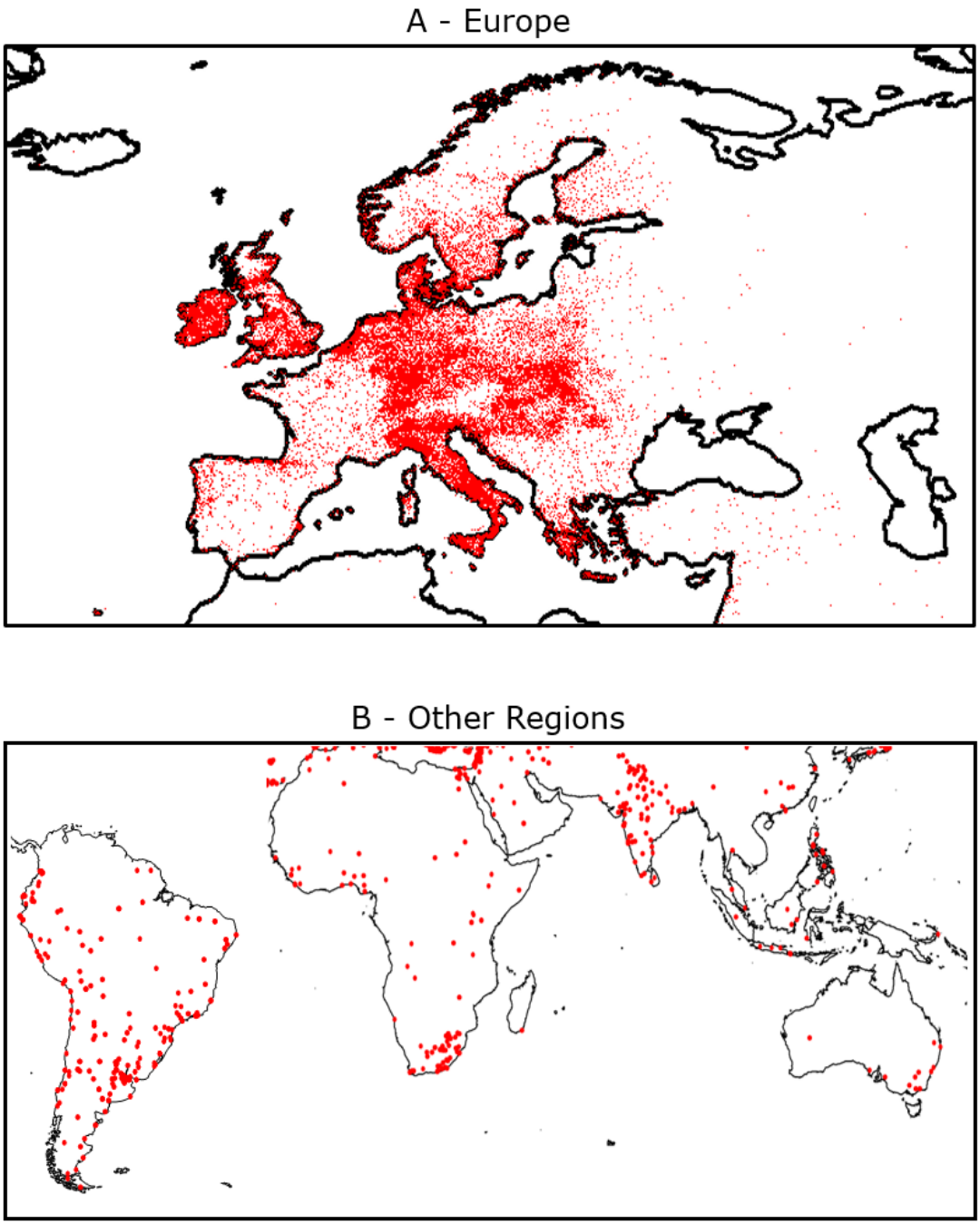
The graph displays the number of passengers arrived in the period 1909-1924 in the passenger list data and the number of foreign born individuals in the 1920 and 1930 census. For passenger lists, The samples displayed correspond to the full number of individuals departed from any port, the subsample of male individuals and the subsample of males 14-65 years old who departed from ports more than 3000km from New York port. The census figure correspond to male immigrants with country of birth other than Mexico and Canada.

Figure 2.2: Ports of Departure



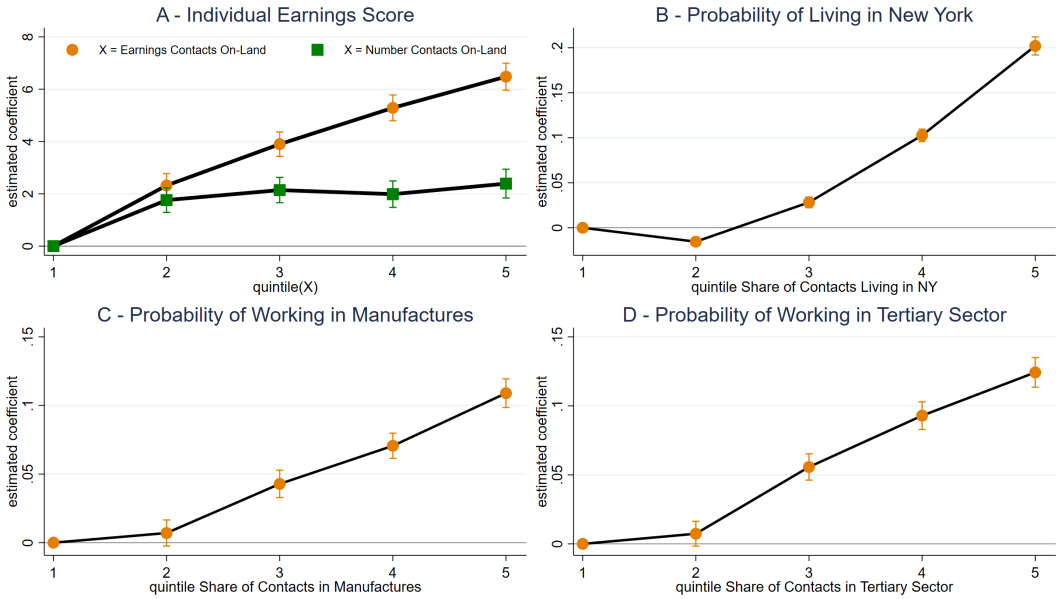
The figure shows all the geolocalized ports of departure for years of arrival 1909-1924. Ports located at less than 10km are displayed as a single port. The shaded area indicates the ports located at a distance below 3000km from New York City.

Figure 2.3: Places of Origin in Matched Sample



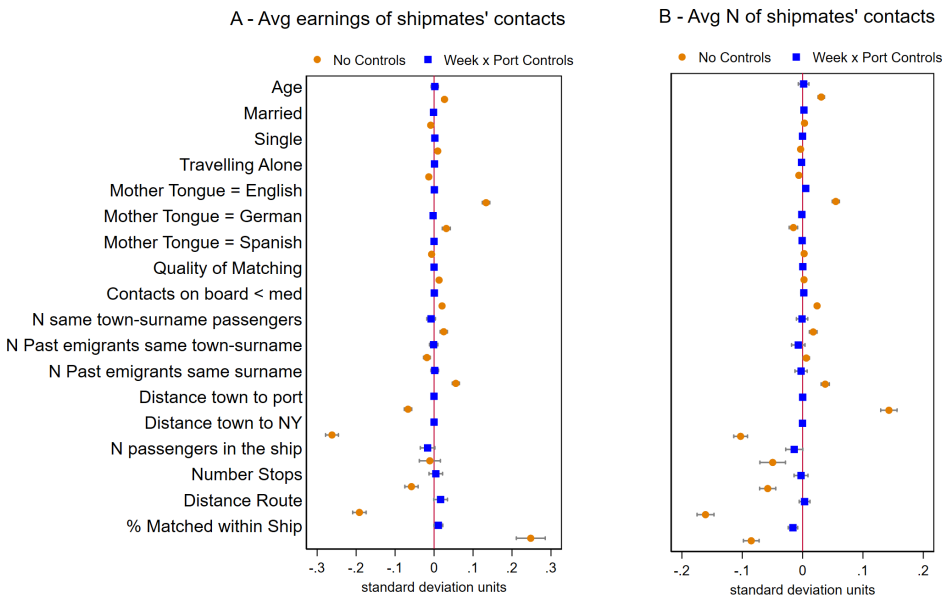
The figure shows all the geolocalized places of origin declared by male passengers in the matched sample for years of arrival 1909-1924. For places identified with precision above the locality level, the map reports the centroid of the administrative unit.

Figure 2.4: Individual Outcomes and Settled Immigrants from Same Town



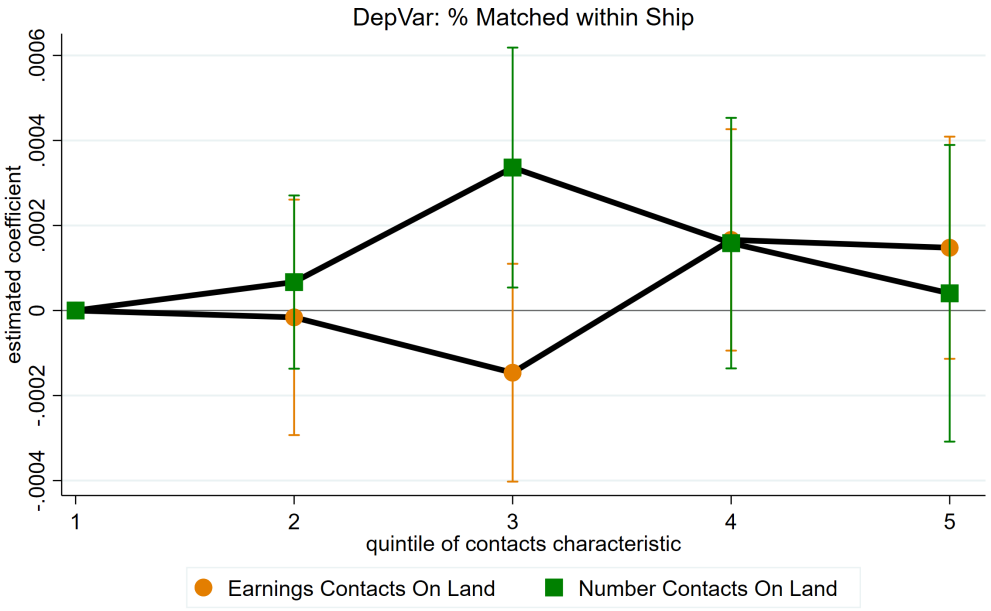
Panel A displays the coefficients of OLS regressions of the individual earnings score on the quintiles of the average earnings and the average number of immigrants from the same town of origin. Panel B, displays the coefficients of a regression of a dummy indicating whether the individual lives in New York city as a function of the share of immigrants from the same town of origin settled in New York city. Panels C and D, show the coefficients of a regression for dummies of sector of occupation on the share of immigrants from the same town of origin working in that sector. All regressions control for Ship fixed effects and individual characteristics. Standard errors clustered at the week of arrial level.

Figure 2.5: Balance of Predetermined Characteristics



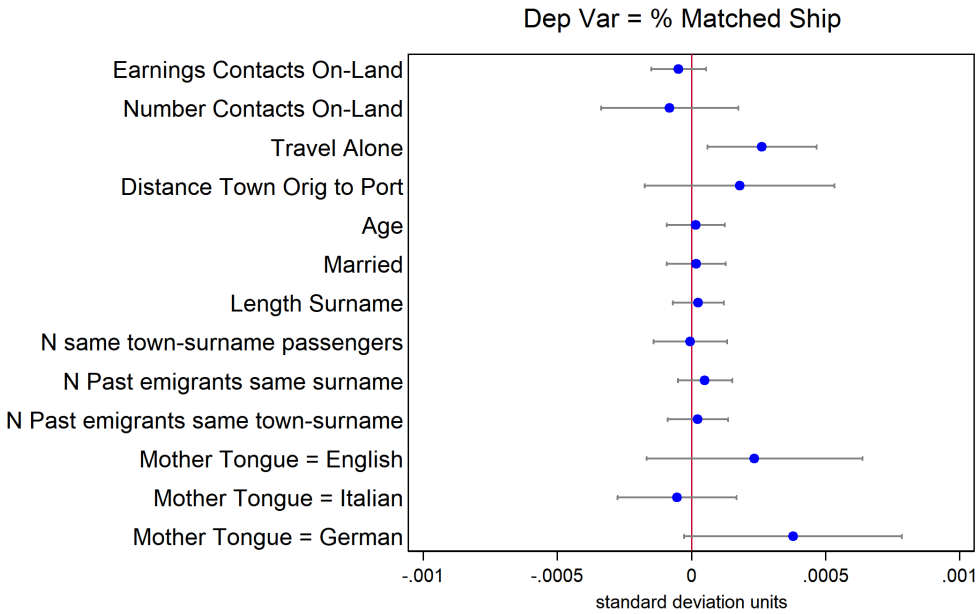
Each panel displays the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and on the average number of (unrelated) shipmates' contacts. Each row is a different regression. All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. Regressions with squared (blue) markers control for fixed effects of the interaction between week of arrival and port of departure and also include fixed effects for large administrative region interacted with port. Standard errors clustered at week of arrival.

Figure 2.6: Data Matching and Quality of Own Contacts



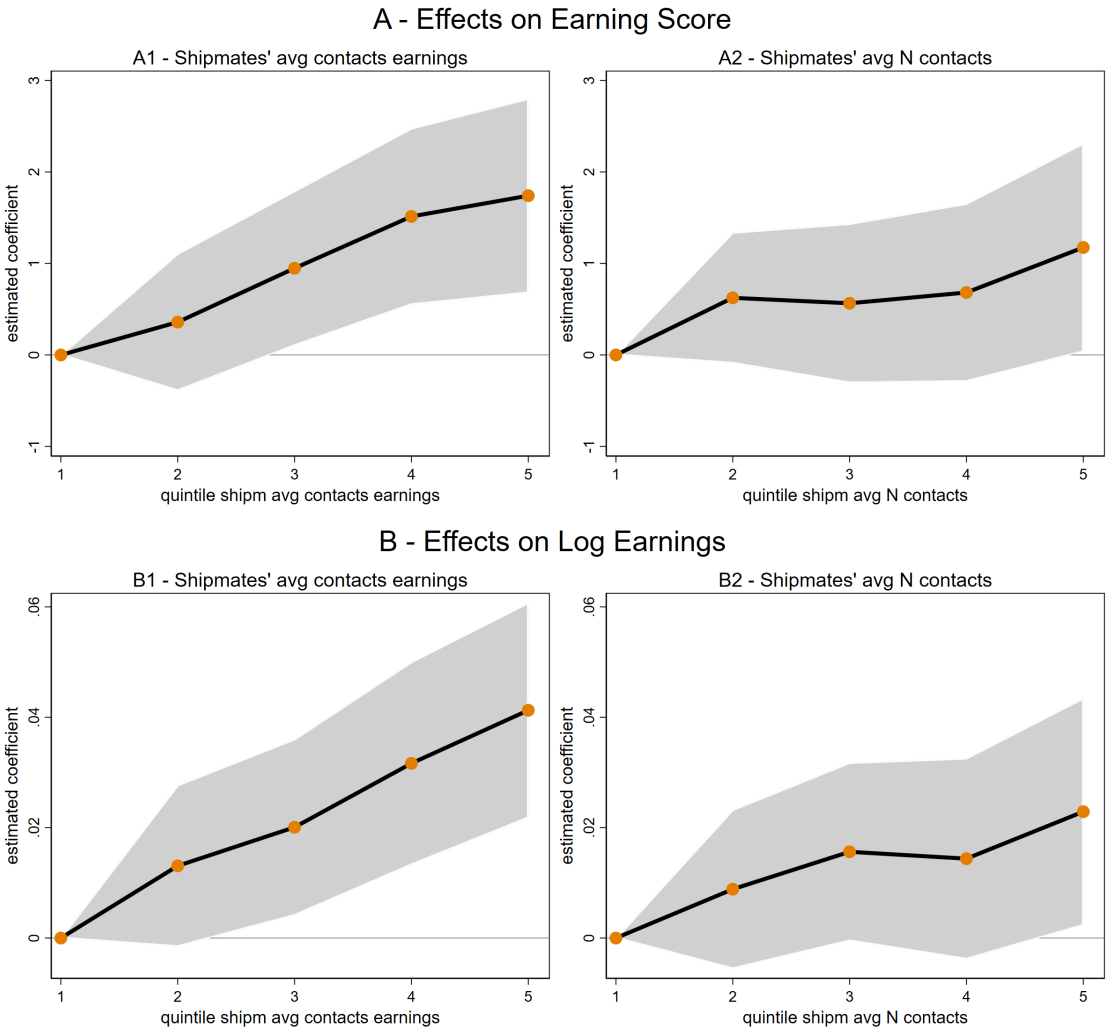
The figure displays the coefficients of an OLS regression where the dependent variable is the % of passengers matched with census within the ship. Circle markers correspond to the quintiles of earnings of individual's contacts on land. Squared markers correspond to the quintiles of the number of individual's contacts on land. Contacts on land are defined as previous emigrants from the same town of origin. Standard errors clustered at week of arrival.

Figure 2.7: Data Matching and Individual Characteristics



The figure displays the point estimates and confidence intervals of an OLS regression of the percentage of passengers matched within the ship on a set of individual characteristics of the passenger. The regression controls for Port of Departure X Week of Arrival and for the administrative unit of origin. Standard errors clustered at the week of arrival level.

Figure 2.8: Effect of Shipmates' Connections on Earnings

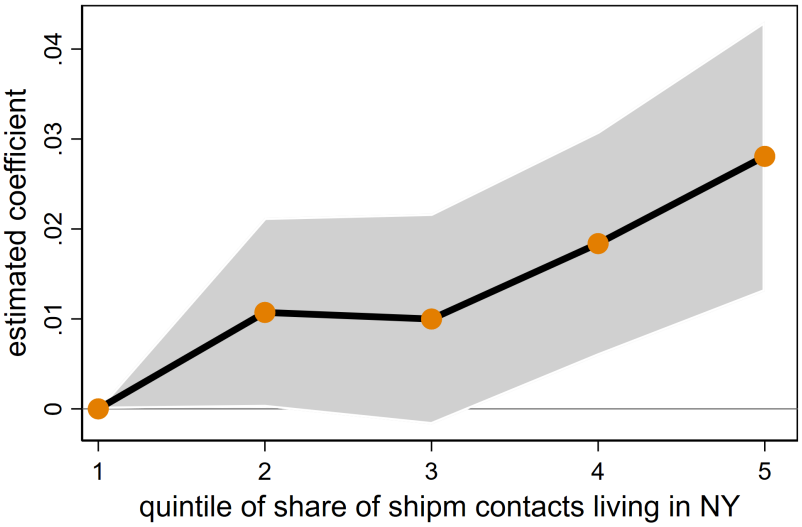


Each panel displays the estimated coefficients of a regression of earnings on the quintiles of shipmates' characteristics. In Panel A, the dependent variable is the occupational earning score and in Panel B the log occupational earnings (1940 based). Sub-panels A1 and B1 shows the effect for the quintiles of the average earnings of (unrelated) shipmates' contacts. Sub-panels A2 and B2 shows the effect of the quintiles of the average number of (unrelated) shipmates' contacts. Shipmates' contacts are defined as individuals from the same town of origin who emigrated in the past. Regressions control for fixed effects of the interaction between the week of arrival and port of departure and the interaction between the town of origin and the semester of arrival. Standard errors clustered at week of arrival.

Figure 2.9: Effect on Earnings by Time Since Arrival

The figure displays the coefficients of an OLS regression of the individual earnings score on the average characteristics of (unrelated) shipmates contacts interacted with dummies for the number of years since arrival to the country. Regressions control for the Week x Port of Departure, Place of origin X Census Year linear trends and the interaction between the age of arrival and the characteristics of shipmates contacts. Robust standard errors clustered at the week of arrival level.

Figure 2.10: Probability of Staying in NY as a Function of Shipmates' Contacts Residing in NY



The figure displays the OLS estimation of the probability of residing in New York city as a function of the quintiles of the share of (unrelated) shipmates'contacts living in New York city. Regressions also controls for quintiles of individuals contacts living in New York city, the number of contacts of the individual and of his (unrelated) shipmates. Regressions include the baseline controls mentioned in the text. Robust standard errors clustered at the week of arrival level.

2.9 Tables of the Chapter

Table 2.1: Descriptive Statistics

Panel A	**Full Sample**	**Reg Sample**[1]		
N Male Individuals Full Passenger List	9,297,026	4,716,934		
N Male Immigrants Census 1920-1930	2,836,404	2,469,503		
N Matched Individuals	351,289	206,383		
N of Ships Matched Sample	34,091	14,910		
N of Vessels Matched Sample	5,138	1,152		
N Ports Matched Sample	422	166		
N Routes Matched Sample	865	454		
N Places of Origin Matched Sample[2]	10,909	8,250		
Panel B	**Avg**	**Std**	**Min**	**Max**
Min Linear Distance Travelled (thousands of km) [3]	6.5	1.2	3	31
Estimated Days Full Voyage at 15 Knots Speed	9.7	1.9	4.6	46.5
Distance Town to Port of Departure	526.6	913.1	0	19214
Passengers per Ship in Passenger List [4]	173	303.2	1	3749
Passengers per Ship in Matched Sample [4]	20.1	23.2	1	262
Past Emigration from Same Place (thousands)	9.3	22.7	0	168
Earnings of Past Emigrant from Same Place	49.7	11.6	.6	100
Avg N of Potential Contacts of Shipm (thousands) [5]	6.2	9.6	0	168
Avg Earnings of Potential Contacts of Shipmates [5]	49.8	6.4	3.1	100
N of Different Places of Origin in the Ship	14.9	17	1	178
Age at arrival	23	10.4	0	68
Married at arrival	.29	.45	0	1
Share Travelling Alone [6]	.74	.44	0	1
Share Living in Urban Places at Destination	.82	.38	0	1
Share Individuals Staying in New York City	.21	.4	0	1
Average N of Ships in Week X Port	2.8	1.8	0	15

[1] The regression sample includes individuals 13-65 years old. For the case of Passenger List information, it only includes ships departing from ports more than 3000km away from New York port and without missing information on the place of origin. [2] Places of origin with at least two matched individuals during one semester in the regression sample. [3] The Minimum Linear Distance of the voyage is estimated as the sum of the straigth distance between subsequential ports identified within the route, sorted by their proximity to New York port. [4] Only individuals in the regression sample. [5] Potential contacts are defined as past emigration from the same town or place of origin. [6] Individuals travelling alone are defined as those without any other passenger in the ship with same place of origin and surname.

Table 2.2: Correlation of Characteristics within Ship

	(1) Unconditional Correlation	**(2) Conditional on Week x Port**
Age	0.079***	-0.007
Married	0.088***	0.003
Single	0.094***	0.003
Travelling alone	0.075***	0.001
Mother tongue = English	0.642***	0.004
Mother tongue = German	0.488***	0.002
Mother tongue = Spanish	0.462***	-0.011
Quality of machting	0.110***	0.008
N same town passengers	0.079***	0.031*
N same town-surname passengers	0.085***	0.006
N Past emigrants same town-surname	0.008	-0.004
N Past emigrants same surname	0.114***	-0.006
N Past emigrants same town	-0.014***	-0.015
Avg earnings of land contacts	0.176***	-0.010
Distance town to port	0.165***	-0.004
Distance town to NY	0.701***	0.000

The table displays unbiased estimates of the correlation individual and average shipmates' characteristics , excluding those residing in the same place or with similar surname. Column (2) controls for Week of arrival X Port of Departure and Adm Region X Port. Sample of 15-65 males not residing in the US before departure. Bootstrapped significance levels.

Table 2.3: Probability of Matching Passenger List - Census

	Dep Var = Passenger Matched with Census	
	(1) No Controls	(2) Week X Port
F-Stat Joint Significance of Ship FE	5.92	0.60
p-value	0.00	1.00
N Individuals	5008017	4996193

The table reports the joint significance F-statistic for the Ship Fixed Effects, in a regression where the dependent variable is a dummy for whether the passenger is matched in the census. The sample is the full passenger list for non-american citizens in the age group 14-65.

Table 2.4: Effect of Shipmates' Connections on Earnings and Job Quality

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Average Contacts Earnings	0.14***	0.08***	0.11***	0.27***
	(0.03)	(0.02)	(0.03)	(0.06)
Number of Contacts	0.05**	0.04**	0.05***	0.07**
	(0.02)	(0.02)	(0.02)	(0.04)
Mean DepVar	50.89	23.24	44.36	881.88
N individuals	97395	97818	97395	96484
R2	.338	.359	.368	.384
F excl	12.2	9.9	11.4	12.9

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. † Coefficients multiplied by 100 in this column. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 2.5: Additional Controls

	Depvar = Earnings Score								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Avg. Earnings	0.14***	0.14***	0.14***	0.09**	0.11**	0.16***	0.15***	0.18***	0.21***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.05)
N of contacts	0.05**	0.04**	0.04**	0.07**	0.02	0.04*	0.04*	0.07**	0.07**
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)
N individuals	97395	95115	95115	67765	74775	95096	95069	78127	77952
R2	.338	.342	.342	.41	.394	.346	.348	.467	.488
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual Controls	-	✓	✓	✓	✓	✓	✓	✓	✓
Ship-Trip Controls	-	-	✓	✓	✓	✓	✓	✓	✓
Town Origin X Month Arriv	-	-	-	✓	-	-	-	-	-
Week X Port X Admin Reg	-	-	-	-	✓	-	-	-	-
Vessel FE	-	-	-	-	-	✓	✓	✓	✓
Route FE	-	-	-	-	-	-	✓	✓	✓
Surname Nysiis FE	-	-	-	-	-	-	-	✓	✓
Date of Arrival FE	-	-	-	-	-	-	-	-	✓

Each column displays estimates of OLS regressions of the earnings score on the the average earnings of (unrelated) shipmates potential contacts and their average number of contacts (in thousands). The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. Potential contacts are defined as past emigrants from the same town of origin. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. Baseline controls include fixed effects for Week of Arrival X Port of Departure and Place of Origin X Semester. Individual controls are age, marital status, indicators for individuals travelling with relatives, white ethnicity and english native tongue. Ship-Route controls are the number of passengers in the ship, max capacity of the ship, number of stops, total distance, number of trips made by the ship, average number of stops of the ship, days since the last trip observed in the sample, share of married passengers, share of male passengers and number of US resident passengers. Surnames fixed effects are based on the NYSIIS codification system. Column (4), include interactions between the week of arrival, port of departure and the administrative region of the place of origin. Robust standard errors clustered at week of arrival level.

Table 2.6: Investigating Potential Contacts Before Travelling

	(1) Shipmates contacts avg earnings	(2) Shipmates avg Number of contacts
Baseline	0.14***	0.05**
	(0.03)	(0.02)
$\|ID_i - ID_j\| > 10$	0.12***	0.04**
	(0.03)	(0.02)
$\|ID_i - ID_j\| > 15$	0.12***	0.04**
	(0.03)	(0.02)
$Dist(Town_i, Town_j) > 100km$	0.09***	0.04***
	(0.03)	(0.02)

Each row in the table reports the coefficients of a different OLS regression of the earnings score on the shipmates contacts characteristics. Each column variable is a different explanatory variable of the same regression. The second and third row exclude any shipmate j with ID number difference below 10 and 15 respectively. The last row excludes any shipmate j with ID number difference below 15 and with town of origin located at less than 100km of individual's town of origin. All regressions control for the characteristics of individual own contacts. All regressions include the baseline controls described in the text and fixed effects fo rthe interaction between port of departure and week, and the interaction between port of departure, administrative area of residence and year-semester. Robust standard errors clustered at week of arrival level.

Table 2.7: Estimated Effects by Individual's Connections On-Board and On-Land

	Depvar = Earnings Score		
	(1)	(2)	(3)
Definition of Low Connections:	No Contacts On Board (Same Town or Surname)	Quality of Potential Contacts on Land	No Contacts On Board + Quality of Land Contacts
Shipmates Contacts Earnings x Low Connections	0.24*** (0.06)	0.19*** (0.06)	0.34*** (0.07)
Shipmates Contacts Earnings x High Connections	0.15*** (0.04)	0.11*** (0.03)	0.12*** (0.03)
Shipmates N of Contacts x Low Connections	0.07* (0.04)	0.10** (0.05)	0.13** (0.05)
Shipmates N of Contacts x High Connections	0.07** 0.03	0.05** 0.02	0.05** 0.02

Each column shows the coefficients of a different OLS regression of the earnings score on the average earnings of contacts and on the number of contacts of (unrelated) shipmates' interacted with a dummy variable indicating the quality of connections of the individual. In Column (1), an individual is defined as low connected if he is travelling without any person of same surname from the same place of origin. In Column (2) an individual is defined as low connected if the number of persons from the same place in the ship is below the median and if the average earnings of past emigrants from same place is below the median. Column 3 defines an individual as low connected if the number of emigrants and the average earnings of past emigrants from the same palce of origin is below the median and if there is no other passenger from the same place of origin in the ship. Surname similarity is defined based on nysiis phonetic coding. All regressions include fixed effects of Week X Port of Departure, Place of Origin X Semester of Arrival, indicators for the route and a dummy variable indicating if the individual is high or low connected according to the definition in the column. Column (1) includes fixed effects for each nysiis surname category. Standard errors clustered at the week of arrival level.

Table 2.8: Effects by Language of Shipmates

	(1) Earnings Score	(2) Log Earns†
Average Earnings of **Similar Language** Shipmates' Contacts	0.05**	0.14***
	(0.02)	(0.04)
Average Earnings of **Different Language** Shipmates' Contacts	0.03	0.06
	(0.03)	(0.06)
Average Number of **Similar Language** Shipmates' Contacts	0.01	0.03
	(0.01)	(0.02)
Average Number of **Different Language** Shipmates' Contacts	-0.01	-0.02
	(0.01)	(0.02)

Each column of the Table displays estimates of an OLS regression of a measure of individual earnings on the average characteristics of (unrelated) shipmates contacts. The main explanatory variables are calculated separately for shipmates who spoke similar and different mother tongue. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Column (2) the dependent variable is the (log) median earnings of the occupation in 1940.The sample includes all (male 14-65) matched passengers in the period 1909-1924 with at least one shipmate speaking a different mother tongue. Mother tongue definition is constructed based on IPUMS categories. † Coefficients multiplied by 100 in this column. Regressions also control for baseline controls as defined in the text. The number of observations in the regressions is 62,890. Standard errors clustered at the week of arrival level.

Table 2.9: Effects on Sector of Employment

Shipmates Characteristics	(1) Primary Sector	(2) Manufactures Sector	(3) Services Sector
Share of Contacts in Primary Sector	0.08**	-0.010	-0.07*
	(0.03)	(0.04)	(0.04)
Share of Contacts in Manufactures	0.01	0.07*	-0.08**
	(0.03)	(0.04)	(0.04)
Mean DepVar	0.3	0.4	0.4
N individuals	83459	83459	83459

This table displays estimates of OLS regressions of the sector of employment of the individual on the share of (unrelated) shipmates contacts employed in each sector. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable a dummy indicating whether the individual is employed in agriculture and other primary activites. In Column (2) the dependent variable a dummy indicating whether the individual is employed in the manufacturing sector. In Column (3) the dependent variable a dummy indicating whether the individual is employed in services or public sector. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions control for the share of individual contacts in each sector, the number of contacts of the individual, the average number of contacts of his shipmates, the average earnings of the shipmates contacts, the average earnings of his own contacts and indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 2.10: Shipmates Effects on Sectors of Occupation and Place of Residence

Panel A: Sector of Occupation of Individual

	(1) Sector 1 digit	(2) Sector 2 digits
Share of Shipmates Contacts Working in the Same Sector	0.079***	0.076***
	(0.014)	(0.008)

Panel B: Place of Residence of Individual

	(1) State of Residence	(2) City of Residence
Share of Shipmates Contacts Living in Destination Place	0.084***	0.073***
	(0.009)	(0.010)

Panel C: By Language of Shipmates

Share of Contacts Working/Living in the Same Sector/State:	(1) Sector of Occup (1d)	(2) State of Residence
Shipmates of **Similar Language**	0.078***	0.086***
	(0.012)	(0.007)
Shipmates of **Different Language**	0.015	0.016**
	(0.012)	(0.008)

Panel A displays the coefficients of an OLS regression of $Y_{ij}(t)$, a dummy that takes one if individual i works in sector j, on X_{ij}^{SM}, the share of (unrelated) shipmates contacts working in sector j. Regressions include individual fixed effects, fixed effects of the interaction between sector of occupation, week and port of departure and fixed effects of the interaction between sector of occupation, administrative region of origin and year of arrival. Regressions also control for the share of individual contacts working in sector j. In Column (1) sector of occupation is defined at 1 digit and in Column (2) at 2 digits, in both cases based on the 3 digits classification created by IPUMS. Panel B displays the coefficients of an OLS regression of $Y_{ic}(t)$, a dummy that takes one if individual i lives in place c, on X_{ic}^{SM}, the share of (unrelated) shipmates contacts residing in place c. Regressions include individual fixed effects, fixed effects of the interaction between the place of residence, week and port of departure and fixed effects of the interaction between place of residence, administrative region of origin and year of arrival. Regressions also control for the share of individual contacts living in place c. In Column (1) the place of residence is defined as the state of residence. In Column (2) the place of residence is based on 85 cities with the highest share of individuals from the sample, excluding those residing in non-classified cities or small rural areas. In Panel C, the share of shipmates contacts working in different sectors or living in different states are calculated separately for shipmates with similar and different mother tongue. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. The number of observations are 712,440 for Panel A Column(1), 5,303,720 for Panel A Column(2), 7,563,689 for Panel B Column(1), 8,971,750 for Panel B Column(2), 464,445 for Panel C Column (1) and 5,110,210 for Panel C Column (2). Standard errors clustered at the week of arrival level.

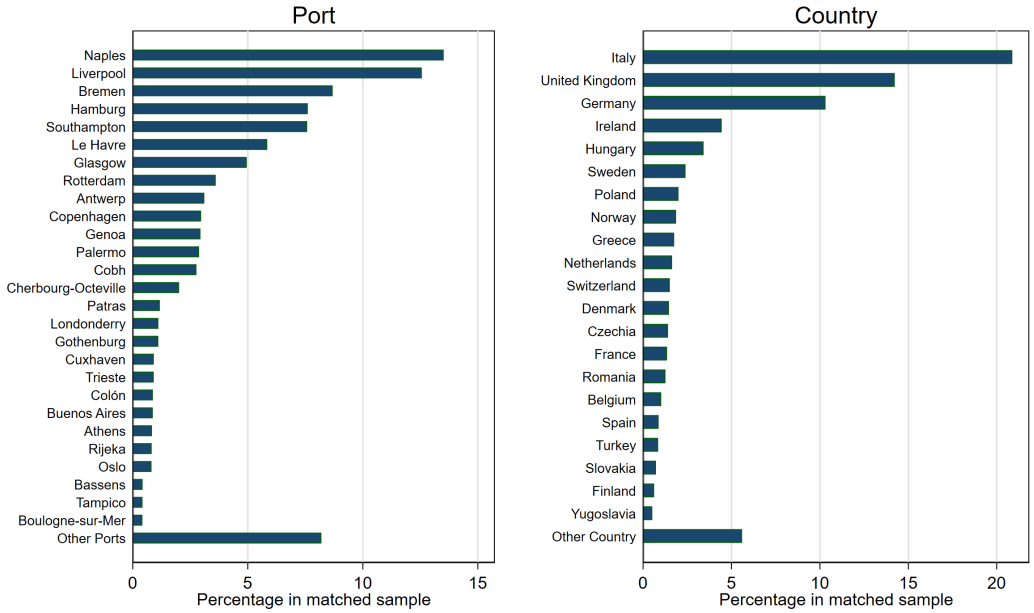
Table 2.11: Correlation in Labor and Spatial Outcomes of Shipmates

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
Same Ship	0.15***	0.09***	-3.15***	0.20**
	(0.04)	(0.03)	(0.60)	(0.09)
% Effect Over the Mean	9.4	11.09	-3.15	2.18
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship. The sample consists of all matched male passengers arrived during the period 1909-1921 grouped into non repeated pairs of individuals who arrived during the same week. The sample only include pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state . In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair, fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) and fixed effect of Date Arrival(i) X Date Arrival(j) where i and j index individuals in the pair. Standard Errors clustered at week of arrival level.

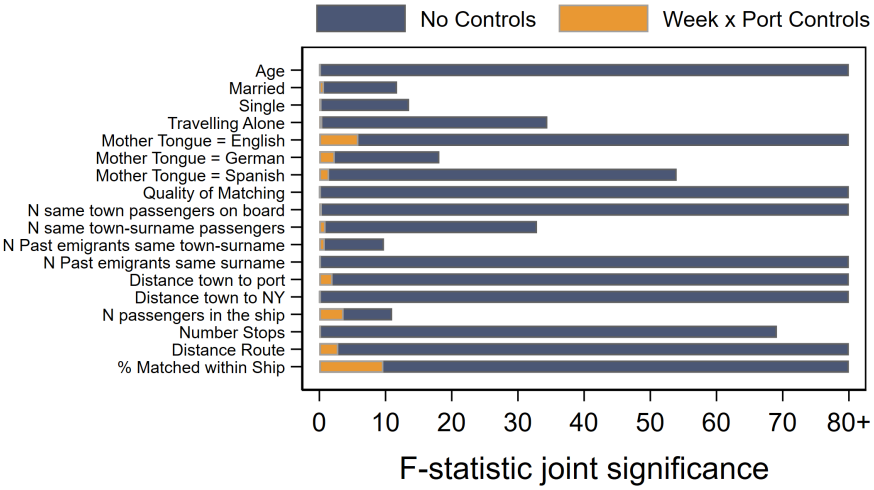
2.10 Appendix A: Additional Tables and Figures of the Chapter

Figure 2.A1: Main Ports of Departure and Countries of Origin



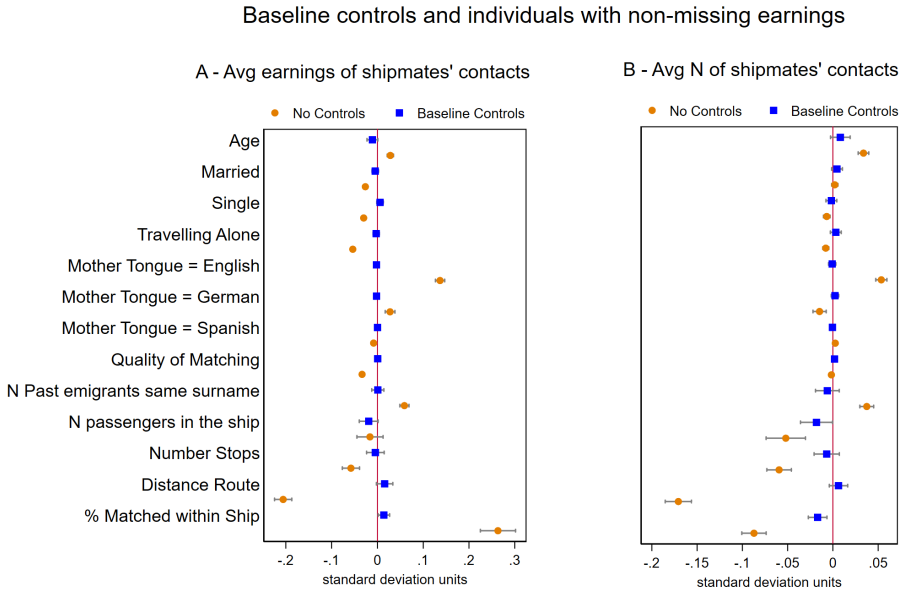
Data include all matched passengers during in 1909-1924 who are not US citizens. Data exclude arrivals from ports located at less than 3000km from the port of New York.

Figure 2.A2: Balance Regressions, Joint Significance



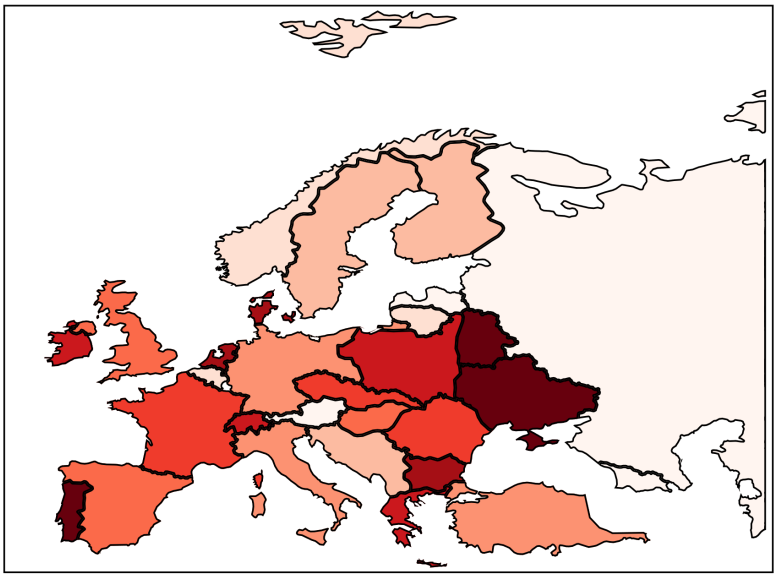
Each panel displays the joint F statistic of the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and on the average number of (unrelated) shipmates' contacts. Each row is a different regression All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. The regression with additional controls include fixed effects for the interaction between week of arrival, port of departure and fixed effects for large administrative region interacted with port. Standard errors clustered at week of arrival level.

Figure 2.A3: Balance of Predetermined Characteristics



Each panel displays the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and the average number of (unrelated) shipmates' contacts. Each row is a different regression. All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. In panel B, regressions control for fixed effects of the interaction between week of arrival and port of departure and also include fixed effects for the interaction between town of origin and semester of arrival. Standard errors clustered at week of arrival.

Figure 2.A4: Heterogenous Effects by Country of Origin of Passengers (Europe)



The figure displays the OLS estimation of the effects of shimates' contacts earnings interacted with dummies for countries of origin of the individual. Regressions control for dummies of country of origin, week of arrival and port of departure interactions fixed effect and town of origin fixed effects interacted with the year of arrival. The plotted coefficients are normalized in the scale zero to one. Effects are significant at 10% for the following list of countries/regions: Belarus, Czechoslovakia, Denmark, Germany, Greece, Hungary, Ireland, Italy, Netherlands Poland, Portugal, Switzerland, UK, Ukraine.

Table 2.A1: Alternative Measures of Earnings Based on 1950 Census

Shipmates Characteristics	(1) Log Earnings Occupation 1940	(2) Log Earnings Occupation 1950	(3) Log Earnings Percentile 1950
Average Contacts Earnings	0.27***	0.25***	0.28***
	(0.06)	(0.06)	(0.08)
Number of Contacts	0.07**	0.07*	0.08
	(0.04)	(0.04)	(0.05)
Mean DepVar	881.88	2219.54	2235.79
N individuals	96484	97395	97395

This table displays estimates of OLS regressions of different measures of earnings on the average characteristics of shipmates contacts. All coefficients are multiplied by 100. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the (log) median earnings of the occupation in 1940. Column (2) is similar to Column (1) but using 1950 1% census sample. In column (3) the dependent variable is the (log) median earnings of the percentile ranking of the occupation in 1950. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 2.A2: Alternative Clustering of Standard Errors

	Clustering Level				
	(1) Week of Arrival	(2) Month of Arrival	(3) Ship x Port Depart	(4) Region of Origin	(5) Multicluster Week-Ship
Avg. Earnings	0.14***	0.14***	0.14***	0.14***	0.14***
	(0.031)	(0.037)	(0.028)	(0.032)	(0.031)
N of contacts	0.05**	0.05**	0.05**	0.05**	0.05**
	(0.020)	(0.020)	(0.019)	(0.019)	(0.020)
N individuals	97391	97391	97391	97391	97391

Each column shows the coefficients of a different regression for alternative levels of clustering in standard errors. The dependent variable is the earnings score. The reported explanatory variables are the average earnings of (unrelated) shipmates potential contacts and their average number of contacts (in thousands). Potential contacts are defined as past emigrants from the same town of origin. All regressions control for the characteristics of individual own contacts. Baseline controls include interaction for week of arrival and port of departure and the interaction between administrative region of origin and port of departure.

Table 2.A3: Effects by Interactions Between Variables of Interest

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Contacts' Earnings High X	1.73***	1.23***	1.74***	3.91***
N Contacts High	(0.48)	(0.36)	(0.43)	(0.90)
Contacts' Earnings High X	1.42***	0.93***	1.26***	2.51***
N Contacts Low	(0.47)	(0.35)	(0.43)	(0.89)
Contacts' Earnings Low X	0.70	0.64**	0.70*	1.63**
N Contacts High	(0.43)	(0.30)	(0.38)	(0.79)
N individuals	97395	97818	97395	96484
R2	.338	.358	.367	.383

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. Each column shows the coefficients for the interaction between two set of dummies. The first set of dummies indicates whether the shipmates connections earnings are above or below the median of its distribution and the second set of dummies indicates whether the shipmates number of connections is above or below the median of its distribution. The omitted category is shipmates below the median of contacts earnings and contacts number. † Coefficients multiplied by 100 in this column. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 2.A4: Effects Before and After the 1921 Emergency Quota Act

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Avg Contacts Earnings x Pre-Quota	0.15***	0.09***	0.13***	0.30***
	(0.03)	(0.02)	(0.03)	(0.07)
Avg Contacts Earnings x Post-Quota	0.07	0.03	0.03	0.12
	(0.07)	(0.05)	(0.06)	(0.12)
Number of Contacts x Pre-Quota	0.05**	0.04**	0.05**	0.08*
	(0.02)	(0.02)	(0.02)	(0.04)
Number of Contacts x Post-Quota	0.04	0.04	0.05	0.04
	(0.04)	(0.03)	(0.03)	(0.07)
Mean DepVar				
N individuals	97391	97814	97391	96480

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the reported explanatory variables are the average earnings score of the potential contacts of (unrelated) shipmates and the average number of them, in both cases interacted with a dummy indicating whether the individual emigrated before or after the introduction of the 1921 Emergency Quota Act. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. † Coefficients multiplied by 100 in this column. Standard errors are clustered at the Week of Arrival level.

Table 2.A5: Alternative Definitions of Contacts on Land

Shipmates Characteristics	(1) Baseline Definition (Same Town)	(2) Same Admin Region and Similar Surname	(3) Same Town and Similar Surname
Avg Contacts Earnings	1.25*** (0.14)	2.77*** (0.32)	2.78*** (0.47)
Number of Contacts	0.38** (0.15)	0.72** (0.35)	0.87* (0.47)
N observations	130684	35552	17297

Each column shows the coefficients of a different regression of the individual earning score on the average earnings and number of potential contacts of (unrelated) shipmates'. Explanatory variables are standardized with zero mean and standard deviation one in every regression. Each column corresponds to a different definition of potential contacts residing in the US. In Column (1), potential contacts are defined as past emigrants from the same town of origin. In Column (2), potential contacts are defined as past emigration from same administrative area of origin and with similar surname. In Column (3), potential contacts are defined as past emigrants from the same town of origin and with similar surname. Surname similarity is based on nysiis phonetic coding. All regressions control for fixed effects for Week of Arrival X Port of Departure and fixed effects for the group at which potential contacts are defined interacted with census year. Standard errors clustered at the week of arrival level.

Table 2.A6: Alternative Identification Strategies

	(1)	(2) Different Stops of Same Ship	(3) Different Stops of Same Ship
	Repeated Trips of Same Vessel	Any Port in the Route	Only First Port of Departure
Boarding at Same Port:			
Average Earnings of Shipmates' Contacts	0.08***	0.11***	-
	(0.02)	(0.04)	-
Average Number of Shipmates' Contacts	0.03*	0.02	-
	(0.02)	(0.02)	-
Boarding at Different Port:			
Average Earnings of Shipmates' Contacts	-	0.07**	0.07*
	-	(0.03)	(0.04)
Average Number of Shipmates' Contacts	-	-0.01	0.01
	-	(0.02)	(0.02)
N observations	93305	48463	23791
Vessel × Port × Year Arriv	✓	-	-
Place of Origin × Semester	✓	✓	✓
Route × Semester	✓	✓	✓
Vessel FE	-	✓	✓

This table displays estimates of OLS regressions of the occupational earnings score on the average characteristics of shipmates contacts. Each column is a different regression. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. All regressions control for the characteristics of own contacts, the number of passengers and the days elapsed since the previous trip of the vessel. Column (1) only includes vessels with at least two trips during the year. The characteristics of the shipmates in rows are the average earnings of contacts in land and the average number of contacts in land, calculated separately for shipmates boarding the ship at the same port and at different ports of the same route. Columns (2) and (3) exclude any ship with more than 90% of total passage boarding in the first port. Column (3) only includes passengers boarding in the first port of the route. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. Standard errors are clustered at the Week of Arrival level.

Table 2.A7: Subsample of Places of Origin Geolocalized with High Precision

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Average Contacts Earnings	0.25***	0.12***	0.20***	0.46***
	(0.05)	(0.04)	(0.04)	(0.14)
Number of Contacts	0.04	0.05**	0.05**	0.08
	(0.03)	(0.02)	(0.03)	(0.08)
Mean DepVar	52.04	24.09	45.53	714.03
N individuals	70925	71257	70925	70295

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. The sample is restricted to those individuals for whom the town of origin is geocoded with locality or sublocality precision level. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. † Coefficients multiplied by 100 in this column. Standard errors are clustered at the Week of Arrival level.

Table 2.A8: Correlation in Outcomes by Spoken Language

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
SameShip x Same Lang	0.20***	0.19***	-6.63***	0.52***
	(0.05)	(0.04)	(0.77)	(0.10)
SameShip x Diff Lang	0.03	-0.05	2.13***	-0.36***
	(0.05)	(0.04)	(0.70)	(0.12)
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship, interacted with a dummy indicating if the pair speaks the same mother tongue (based on census categories). The sample consists of all matched male passengers arrived during the period 1909-1924, grouped into non repeated pairs of individuals who arrived during the same week. The sample only include pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state. In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair, fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) where i and j index individuals in the pair and fixed effects for Date Arrival (i) X Date Arrival (j). Standard Errors clustered at week of arrival level.

Table 2.A9: Correlation in Outcomes by Contacts On-Land

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
SameShip X	0.12***	0.06**	-2.34***	0.16*
HighCont-HighCont	(0.04)	(0.03)	(0.59)	(0.10)
SameShip X	0.16***	0.13***	-3.82***	0.24**
HighCont-LowCont	(0.06)	(0.04)	(0.79)	(0.11)
SameShip X	0.29***	0.21***	-7.09***	0.43***
LowCont-LowCont	(0.07)	(0.06)	(1.02)	(0.13)
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship, interacted with a set of dummies indicating if both individuals have a high number of contacts on land, only one individual has high contacts on land or both have high number of potential contacts on land. Contacts on land are defined as the number of past emigrants from the same town of origin. The sample consists of all matched male passengers arrived during the period 1909-1921, grouped into non repeated pairs of individuals who arrived during the same week. The sample only includes pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state. In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair and fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) where i and j index individuals in the pair. Standard Errors clustered at week of arrival level.

2.11 Appendix B: Matching Passenger Lists and Census using Machine Learning

In this section I provide further details on the matching procedure used to merge Passenger Lists with Census Data. I start by describing the potential problem faced by researchers dealing with large historical records. Then, I explain the steps involved in the matching algorithm and the techniques used to increase its speed.

The Dimensionality Problem An important challenge when matching across large datasets follows from the need of relying on fuzzy and noisy variables like names and surnames. Economists have used a number of approaches to address this problem, for instance, Fellegi & Sunter (1969), Christien & Churches (2005), Goeken (2011) and more recent Feigenbaum, (2016). However, in many cases, these approaches become unfeasible when data is large.[66] Not surprising, many studies relying on historical data have tried to overcome this problem by either using small random sub-samples or by imposing restrictive assumptions during the matching process.

Although recent advances in computer science have improved the search and matching techniques (see for instance, Schulz & Mihov, 2002), they remain unfamiliar and probably inaccessible to most applied Economists. The lack of easy implementations and the high entry costs to this literature has contributed to their low adoption. In this Appendix, I address the problem of matching across large historical datasets by improving on existing Machine Learning approaches (Feigenmabum, 2016). I introduce some simple modifications, popular among Computer Scientists, which significantly increase the speed and reduce the computational requirements of the matching process.

Two problems contribute to make matching unfeasible. First, the number of calculations required to compare records increases exponentially with the sample size. Intuitively, if there are N individuals in each dataset, the matching process

[66] I tried replicating the approaches described by Christien & Churches (2005) and Feigenbaum (2016) using a 20% random sample of the data. Both procedures resulted unfeasible for a desktop PC with intel-i7 processor and 24GB ram.

involves comparing the name similarity of each pair of individuals which result in N^N calculations. Second, measuring similarity between string variables, involves computationally intensive algorithms. For instance, the most extended measure to compare two strings is the Levenshtein Distance (Levenshtein, 1966). It is defined as the minimum number of character insertions, deletions or substitutions required to transform the first string into the second one. Some statistical packages include commands to calculate Levenshtein distances but they are typically slow due to the complexity of the algorithm (usually based on Wagner & Fischer, 1974).

Blocking In some cases, researchers alleviate the first problem by narrowing the subset of potential matches *before* comparing names. In my setting, this *blocking* strategy, consists in defining for every individual i in the passenger list, a set of Census individuals such that: 1) They arrived in the US during the same year than i and 2) The distance in reported year of birth with respect to i is below 2. Then, for each passenger, I search for census individuals with similar names and surnames, *only* within the relevant block. In some cases, blocking solves the dimensionality problem and matching performs reasonably well.

Unfortunately, in many cases like in my setting, blocks are too large and the number of pair comparisons remain unfeasible. Some restrictive assumptions (like blocking on phonetic coding, or on the first two characters of the surname) are not recommended, particularly when dealing with non-English surnames, as they significantly reduce the accuracy of the matching.[67]

Matching Procedure The whole procedure follows a number of steps described below. Some steps are similar to those in Feigenbaum (2016), but some modifications are introduced to increase the feasibility and accuracy of the method. For efficiency reasons, the direction of the match is performed from the Passenger List to the Census data.

1. **Preliminar Cleaning:** I start by using a dictionary of US places (states,

[67] An important advantage of the algorithm used in this paper is that the Levenshtein distance, although computationally more intensive than the Jaro-Winkler distance, captures to a larger extent different sources of string differences, (e.g. not only typos but also phonetic transcriptions, etc.)

cities and acronyms), to detect passengers that are either US citizens, or have residence in the US. These individuals are excluded from the matching. Then, I use a dictionary of names acronyms and abbreviations (e.g. Jno. for John) and replace them in Passenger Lists and Census.[68].

2. **Unmatchable Cases:** I drop multiple observations with same name, surname, year of arrival and year of birth. These individuals cannot be distinguished from each other in the Census data, and therefore matching them is not possible.

3. **Set of Candidates:** For every passenger arriving during year y_a with year of birth y_b, find a set of "potential matches" in the census with year of immigration y_a and year of birth $y_b \pm 2$ and with a Levenshtein distance in given name and surname below a threshold d.[69] This is the key step in the procedure and usually unfeasible if performed without any additional restriction. I explain later in this Section, two modifications that allow to identify candidates with similar names and surnames significantly faster compared to existing algorithms available in some statistical packages. Lastly, I drop any passenger for whom the set of candidates includes multiple census individuals that match exactly in name, surname and year of birth.

4. **Human Trained Sample:** The previous step defines a set of potential matches for each passenger. I randomly sample 2000 sets, and for each one, I decide whether there exists a candidate who is a "true match" for the reference passenger. As noted by Feigenbaum (2016), human criteria to detect true matches is highly reliable and accurate compared to automatized heuristic procedures. In a recent paper Bailey et al.(2017) find that supervised procedures, based on human trained samples, result in higher matching quality compared to other methods like Ferrie (1996). The training step is performed using all information available to the researcher, this includes the distance in names, surnames and year of birth but also additional information on the whole set of candidates and even the whole sample. Note that it is possible that no can-

[68] This dictionary is constructed based on information from genealogy sites

[69] Census data can be affected by rounding bias in the year of birth. For this reason, I also include the closest round year of birth.

didate is declared a true match. This would happen in two situations. First, if no candidate looks similar enough to the reference passenger (e.g. surnames are too different to be considered a typo or phonetic translation). Second, because more than one passenger looks similar to the reference passenger. When deciding whether a candidate is a true match, I also consider the number of candidates in the block, how similar is the second best candidate, how popular is the name or surname, and any type of information that can be relevant. In this step, the researcher sets the level of accuracy of the match as the following steps are aimed to "imitate" the heuristic behavior of the researcher.[70]

5. **Prediction of True Matches:** Based on the human trained sample, I use a Machine Learning approach to predict the true matches for the whole sample. Feigenbaum (2016) proposes a double-threshold probit procedure and Goeken et al. (2011) describe a Support Vector Machine approach[71]. In my case, I use a Random Forest Classifier (Breiman, 2001) due to its well known out-of-sample prediction properties. Additionally, the inclusion of a large set of variables describing the whole set of candidates combined with the ability of the method to detect highly non-linear patterns, notably reduces the number of multiple predictions (i.e. two candidates are matched with the same passenger).[72] Indeed, cross-validation exercises reveals that the method results in a negligible number of false positives matches.[73] Bailey et al. (2017) shows that the bias introduced by false positive links are more harmful than the biased resulting in smaller matched samples and suggest that the quality of inference can be improved by increasing the precision of match (at the cost of reducing the number of matches). Table 2.B1 at the end of this section describes the main variables used as inputs in the Random Forest Classifier.

6. **Refining Predictions:** The fact that each Census candidate can belong to

[70]Other linking approaches that use human trained samples are Goeken et al. (2011) and Cristien & Churches (2005).

[71]This is similar to the procedure used by IPUMS to create census linked samples.

[72]For the few cases where multiple matches are predicted, I only consider the highest probability match. Alternatively, the difference in the matching probability between the best and the second best matching can be considered, but I find no significant differences in my case.

[73]The Scikit Phyton package includes an straightforward implementation of the Random Forest Classifier

the set of potential candidates of multiple passengers implies that for a small number of cases, the same Census individual is matched with two different passengers. In those cases, I use the matching probability of the Random Forest model to assign as a true match the pair with highest probability. Then I run the Random Forest Classifier again excluding from the set of Census candidates those already matched to a passenger.[74]

As mentioned above, Step 3 is unfeasible even after blocking on year of birth and year of arrival. Some improvement in the algorithm that searchs among similar names and surnames is required to make any progress. The modifications I propose are the following: **1)** Reduce the number of comparisons by using indexed dictionaries of names and surnames specific for every block. **2)** Use a Levehnstein automata approach for searching among "similar names". A Levehnstein automata is a function that identify all the words within a list that are below a certain string distance. The automata significantly reduces the speed of calculations by transforming the dictionaries of names and surnames into a data structure called "radix trie" which decomposes words into a tree of common suffixes. Intuitively, the speed gain comes from the fact that when two words are detected to be above a certain string distance, every word sharing the same "branch" of the second word, will be at least at the same distance, and many searches are skipped.

Indexed Dictionaries A simple way of increasing search speed is by eliminating repeated calculations. This is achieved by creating a set of dictionaries for names and surnames, specific to every year of immigration and year of birth block. For instance, the target dictionary of surnames for a passenger arrived in 1911 with year of birth 1891, will contain the set of (non-repeated) surnames in Census data corresponding to all individuals arrived in 1911 with years of birth 1889 to 1893. Each surname is associated to a numerical id number. Similarly, names and surnames in the Passenger List are stored in dictionaries specific to the year of immigration and year of birth. Denote $W_S^P(y_b, y_a)$ to the dictionary of surnames constructed with individuals in the passenger list arrived in year y_a and born in year y_b. In a similar

[74] All the results in the paper are robust to dropping individuals who were originally matched to multiple passengers.

way, denote $W_S^C(y_b, y_a)$ to the dictionary of surnames based on census individuals arrived in year y_a and born in year $y_b \pm 2$. Instead of comparing among individuals, dictionary search is reduced to find for every entry in $W_S^P(y_b, y_a)$, a set of entries in $W_S^C(y_b, y_a)$ below a maximum Levenshtein distance defined by the researcher.

Levenshtein Automata and Radix Tries Search across dictionaries is more efficient than comparing individuals records, however, calculating string distance measures is slow. If dictionaries are too large, search remains unfeasible. I start by decomposing each dictionary into a Radix Trie, a structure that store words as a combination of suffixes (nodes) and paths connecting them.[75] Figure 2.B1 below, shows an example of it for a dictionary of 8 surnames. Note that each word is associated to a parent branch and child nodes can emerge after a word terminates.

After transforming dictionaries into Radix Tries, I program a Levenshtein Automata that searches within the Trie and that for each entry in $W_S^P(y_b, y_a)$, retrieves a set of "similar" surnames from $W_S^P(y_b, y_a)$ (similarly for given names dictionaries). This Levenshtein Automata is thousands of times faster than any sequential word comparison. The reason is the lower number of required computations. Intuitively, as words are organized into branches, once the Automata detects a word not satisfying the similarity criteria, it stops searching into subsequent nodes. Remaining words in the branch, won't satisfy the criteria as well.[76]

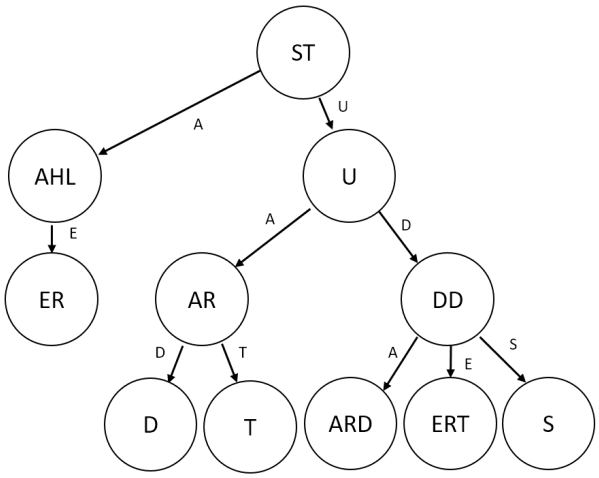
In order to further increase the speed, I add two additional elements. First, search is adaptative: for short words I start with a lower tolerance (maximum distance of 2) and only increase this threshold if few similar words are found. For longer words, the Automata starts to search with a tolerance of 3. The reason is that setting a high tolerance bound for short words is inefficient as it would retrieve most of the target words of similar length. Second, I store results as numerical matrices, where each cell contains the id number that indexes the word and the first column correspond to the Passenger List dictionary entries.[77]

[75] Radix tries are a common way in Computer Science to storage large volumes of string data. Beyond the search speed increase, they are also useful to storage information in a sequential way.

[76] See author's website for a simple example of an implementation of a Levenshtein Automata based on Radix Tries based on Python.

[77] I further restrict the number of "similar words" to the closest 300 entries identified by the Automata. This number is non-biding for the vast majority of names, but restricts the matrix

Figure 2.B1: Radix Trie



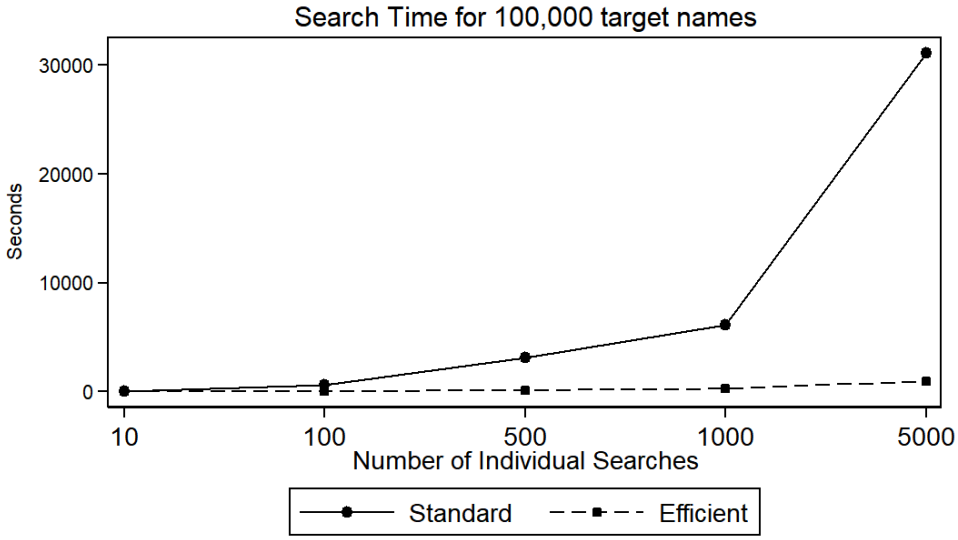
Dictionary: {*stahl, stahler, stud, stuart, stuard, studdard, studdert, studs*}

The final step to find the "set of potential candidates", is as follows: for every individual in the Passenger List, find the Census individuals with given names *and* surnames identified in the numerical matrices mentioned above. Since this step entirely relies on numerical variables, the process is fast even for large volumes of data. Figure 2.B2 below illustrates the efficiency gain of the improved algorithm. The figure compares the time required to find potential candidates for different number of individual records using a target database with 100,000 individuals.[78] The standard method uses the stata command *strdist* to calculate Levenshtein distances and sequentially searches for candidates with names and surnames at a maximum distance of 3. The efficient algorithm incorporates Radix Tries Search and Dictionaries as explained in the text. The difference is significant, for instance, the standard algorithm takes more than 8 hours to perform 5,000 candidates searches while the improved algorithm does the same job in 16 minutes.

dimension for a small number of short names that match with any word of similar length. The criteria to sort entries is based on the Jaro-Winkler distance (a variation of the Levenshtein distance that accounts for the length of the string and the relative position of the unmatched characters, (Lynch and Winkler, 1994)). This is convenient because it has a denser scale compared with Levenshtein distance. Furthermore, Feigenbaum (2016) uses a Jaro-Winkler treshold of 0.2 to restrict the pool of potential matches

[78]This size corresponds to the average size of a Year of Immigration X Year of Birth block, although the number of searches is substantially lower than the one performed to construct the dataset.The calculations were performed with an i7-7th generation Intel processor and 24 GB of ram memory.

Figure 2.B2: Comparing Search Algorithms



The Figure displays the search time in seconds of a standard search procedure and an improved version incorporating Radix Tries Search and Dictionaries as explained in the text. The improved algortihm used in this figure is a simplified version of the algorithm used to create the main dataset in the paper as it does not incorporate further improvements like efficient memory allocation (using numerical codes for strings storage) or alternative search methods for composed names. The target dictionary contains 100,000 individuals and the horizontal axis corresponds to different number of searches.

Table 2.B1: Variables used for Random Forest Matching

Pair specific variables	Jaro Winkler Distance in first names
	Jaro Winkler Distance in surnames
	Jaro Winkler Distance of names and surnames combined
	Any match in the first name (relevant when multiple first names)
	First names match in Soudex code
	Surnames match in Soudex code
	Difference in age
	Round year of birth in Census
	Round year of birth in Passenger List
	Exact first name-surname match
	Exact first name-surname-yearbirth match
	First letter of first name matchs
	First letter of surname matchs
	Last letter of first name matches
	Last letter of surname matches
	Midle name initial matches (when multiple names)
	First name case Census(e.g. multiple names, middle initial, etc.)
	First name case Passenger List (e.g. multiple names, middle initial, etc.)
Block and aggregated variables	Number of potential candidates (and square)
	N of first name matches within block of candidates
	N of surname matches within block of candidates
	Average first name (Jaro Winkler) distance to all candidates in block
	Average surname (Jaro Winkler) distance to all candidates in block
	Jaro Winkler distance in first name to next candidate in block
	Jaro Winkler distance in surname to next candidate in block
	N of exact name-surname matches within block of candidates
	Frequency of surname in Census
	Frequency of first name in Census
	Frequency of surname in Passenger List
	Frequency of first name in Passenger List
	Frequency of first name-surname combination
	N of individuals in census year of birth cell

Note: The table does not list interactions between the variables included in the model.

2.12 Appendix C: Geocoding geographical information

This section describes the algorithm used to geocode the geographical units used in the main analysis.

Places of Origin The data contains information on the "last town of permanent residence". I first identify those individuals reported as US residents and exclude them from the matching process. For the matched sample, I pre-process the data by correcting for common typos and abbreviations in city or country names (e.g. Liverpool abbreviated as lpool). Then, I run a geocoding algorithm that uses the Google Places Api to identify the following information: Latitude and Longitude of the place, Name identified by Google Places and the Shouth-West/North-East coordinates of the smallest rectangle that contains the place. This rectangle is used in the main analysis to further restrict the set of shipmates assumed to be unrelated before the voyage.

The algorithm runs in several steps. It first starts by running an automatized search of the place of origin reported in the Passenger List (after cleaning). I only keep the cases where Google Places retrieves a unique place and it refer to a locality (city, village, etc.). For the remaining cases, I use a dictionary of country abbreviations and acronyms to split the sample by country of origin. Then, I search with Google Places using biasing parameters corresponding to the country. In a second step, I set the language parameter consistently with the country[79]. Finally, I manually search for the remaining cases where more than one observation is observed in the data. In many cases, the manual process consists in homogenizing names spelled with typos and re-running the Google Places search. In other cases, it consists in checking genealogy sites, and simple Google search for towns' name changes or translations.

In a number of cases (around 18% of the sample), the exact town can't be identified either because the individual report a broader administrative unit (e.g.

[79]This is useful for some eastern European cities, transcribed in their native language

the Italian province or region instead of the town), or only a larger administrative unit transcription is recognized by the algorithm, or the exact town does not exist anymore.[80] These cases are codified under the larger administrative region and the corresponding rectangle accounts for this. Finally, a number of observations can only be associated to disappeared historical regions (e.g. Kingdom of Galicia in the actual border between Poland and Ukraine). For these cases, I manually assign a coded name and the rectangle that covers the area of the historical region.

For towns identified with high precision, I use a reverse geocoding algorithm to find the larger administrative region containing it. Broadly, this corresponds to the Google Place Api *administrative_area_level_3* category. Of course, due to changes in political divisions during the 20th century, measurement error can be significant for this codification.

Ports and Routes Following similar steps than those used to geocode the town of origin, I obtain the latitude and longitude of every port of departure in the whole sample (including not matched observations). When two ports belong to the same city or they are located at less than 10 kilometers, I group them into the same unit (e.g. Liverpool and Birkenhead). Some observations include not only the port of departure but a whole list of ports covered during the voyage. In those cases, I only consider the first port reported by the individual as the departure port. Notably, the procedure geolocalizes the port of departure for more than 99% of the passenger records in the period 1909-1924.

Using all the ports of departure in the ship identified at the passenger level, I reconstruct the whole route of the vessel and calculate the total distance of the trip. I assume that stops are sorted by their distance to New York port and the travel distance is calculated as the sum of the minimum linear distance connecting the stops. In the case of ships that stop at Caribbean ports, when constructing Route Fixed Effects, I group them into the same category. There are three reasons for this grouping. First, the distance between Caribbean ports is small and total distance, other trip characteristics, and Caribbean ports' conditions are quite similar if we

[80]The advantage of using towns of origin instead of regions or provinces, is that fewer towns changed their names during the 20th Century.

ignore this variation. Second, routes are identified based on the port of embarkation of all the passengers within the ship. Given that relatively few passengers board the ship in these small ports, differences in the estimated route can be due to measurement error. Finally, the main analysis do not use individuals departing from these ports.

2.13 Appendix D: Baseline Effects by Ship's Matching Rate and Potential Attenuation Bias

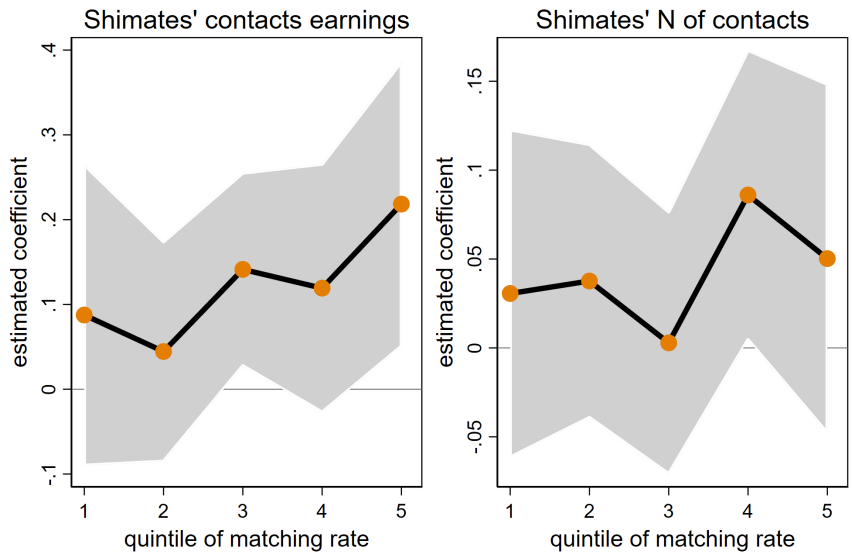
Figure 2.D1 reports the baseline effects by quintiles of the ship's matching rate.[81] Effects are weakly increasing in the matching rate for both measures of shipmates' connections. This suggests that some attenuation bias could be expected due to the partial observability of the set of unrelated shipmates'. However, the fact that effects are not uniquely driven by the highest quintile also indicates that attenuation bias is not extremely large. This is not surprising given that many passengers within the ship shared either the same town or the same region of origin. Thus, since matching is orthogonal to individual characteristics (conditional on baseline controls), the sampling variation is lower relative to a case where shipmates' characteristics vary at individual level (i.e. within towns of origin). For instance, in the extreme case where individuals are matched proportionally to the share of their towns of origin within ship, there is no attenuation bias if the matched sample is large enough to include at least one individual per town of origin in the ship.

In order to explore this idea more explicitly, I perform a series of exercises based on simulated data. Using a distribution of ships and passengers that replicates the one observed in the full Passenger List, I generate the individual earnings as $Y_{i(c)} = \alpha + \bar{X}_{i(-c)} + X_{i(c)} + \epsilon_i$, where $X_{i(c)}$ is a simulated town of origin-specific component and $\bar{X}_{i(-c)}$ is the average town of origin component across all the unrelated shipmates' contacts (in other words, it replicates the construction of the average earnings of unrelated shipmates' contacts as used in the previous sections.) The term ϵ is an idiosyncratic individual component. The variance of $X_{i(c)}$ and ϵ_i are calibrated based on the distribution of their analogues observed in the matched sample data. Then, I create a random sample of passengers for each ship and recalculate the variable $\bar{X}_{i(-c)}$ using only the sampled passengers. Finally, I calculate the attenuation bias for different sampling percentages using OLS estimations of the earnings equation.

The main difficulty when estimating the distribution of towns of origin within the ship (before sampling), is that this variable is harmonized only for the matched

[81] To avoid some confounding effects, the quintiles are calculated conditional on the Port of Departure and the Year of Arrival.

Figure 2.D1: Estimated Effects on Individual's Earnings Score by Ship's Matching Rate



The figure displays the OLS estimation of the effects of shimates' contacts earnings interacted with dummies for the quintiles of the percentage of individuals matched within the ship on individual's earnings score. Quintiles are calculated conditional on Port of Departure. Regressions control for quintiles of the ship matching rate and other controls included in the baseline specification. Regressions exclude ships with less than 10 individuals matched. Robust standard errors clustered at the week of arrival level.

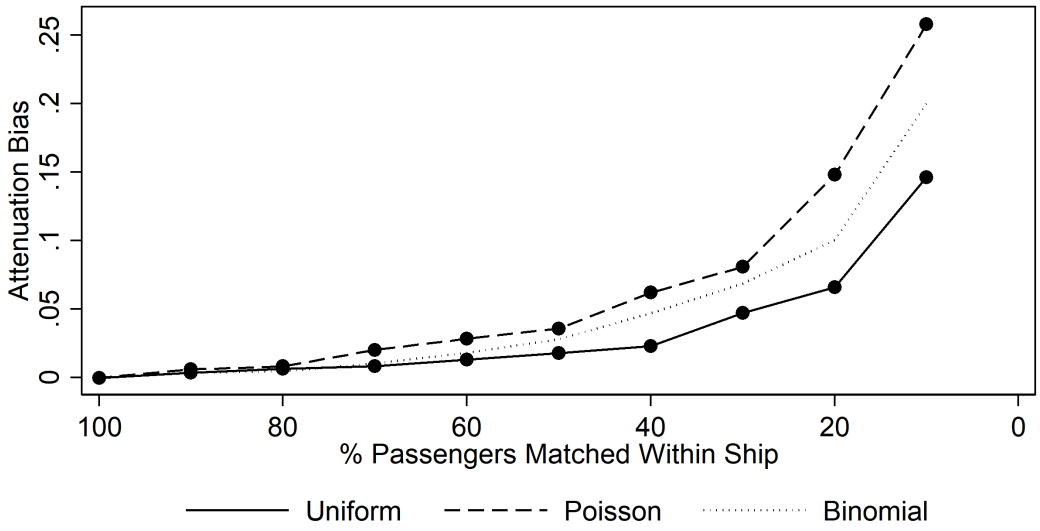
sample. The distribution of towns plays an important role in the attenuation bias as all the shipmates' characteristics ultimately depend on their town of origin. Hence, I simulate the distribution of passengers among towns of origin using three alternative assumptions: Uniform, Poisson and Binomial distributions. The parameters of each distribution is calibrated to replicate the number of average towns per ship in the matched sample data. [82]

Figure 2.D2 shows the results of the simulations discussed above. The exercise reveals that for all distributional assumptions, the attenuation bias is relatively low. For instance, even for matching rates of 10%, the attenuation bias varies from 15% to 25%. The low attenuation bias is mainly driven by the fact that the number of different towns within the ship is not extremely large. Although these simulations rely on a number of arbitrary assumptions (e.g. homogenous

[82]More specific, I start by simulating the distribution of towns with a low value for the distributional parameter. The distributional parameter is expressed as a percentage of the size of the ship (e.g. a uniform distribution with parameter of 0.5 implies that, on average, there is a town every two passengers within the ship). Then, I create a random sub-sample of passengers for each ship where the sampling rate is equal to the average matching rate in the data (approximately 12%). Finally, I calculate the average number of different towns per ship in the simulated random sub-sample. If this value is below the average number of different towns per ship in the matched data, I increase the distributional parameter and repeat the process until these values match.

matching rate across ships), the findings from this section suggest that baseline results shouldn't be seriously downward biased due the partial observability of the pool of shipmates.

Figure 2.D2: Simulated Attenuation Bias



The graph displays the estimated attenuation bias using a simulated dataset with a distribution of ships and passengers similar to the one observed in the full Passenger Lists in the period 1909-1924. The distribution of passengers across different towns of origins is simulated assuming alternatively a uniform, poisson or exponential distribution. The parameter of each distribution is set set proportionally to the ship's size and calibrated to match the average number of towns in the matched sample. The attenuation bias is based on the OLS estimation of a linear m odel where the dependant variable is the individual income and the explanatory variable is the average income of shipmates' town of origin. Simulated income is calibrated to have a similar variance than the observed in the data. The horizontal axe measures the simulated share of passengers matched within the ship.

Chapter 3

The Local Effect of Executions on Serious Crime

3.1 Introduction

The death penalty is arguably the most controversial criminal justice policy in the US. Its divisiveness in the US political debate is demonstrated by the differences across states in its application, by the back-and-forth Supreme Court decisions on its constitutionality, and by the number of pressure groups devoted to advocate against (and sometimes for) it. Proponents claim that it generates respect for law and order and that it constitutes justifiable retribution for heinous crimes. On the opposite side it is argued that it cheapens the value of life and that its uneven use contributes to perpetuate social injustices.

The core issue from an academic perspective is whether the death penalty reduces serious crime, most intuitively through a deterrence mechanism.[1] Unfortunately empirical evidence studying the link between the death penalty and serious crime remains highly flawed, most importantly because of its inability to identify causal effects (Donohue and Wolfers 2005, 2009, NRC 2012, Charles and Durlauf 2013, Nagin 2013). Studies using state-year panel datasets are for instance hampered by

[1]While deterrence is the most commonly discussed mechanism, NRC (2012) note that there are others, such as the social censure that is signaled by such an extreme form of punishment. On the opposite side, it has been argued that the death penalty has a brutalisation effect that leads to an increase in crime (Bailey 1998, Shepherd 2005)

the likely correlation between unobserved determinants of crime (including other features of the sanctions regime) and a state-year use of the death penalty. The instrumental variables sometimes used to account for this endogeneity have not been credible (Donohue and Wolfers, 2009).

In this paper we use a county-date panel dataset to provide causal evidence on the local effect of executions on crime. For every execution we identify both its exact date and the county where the crime motivating the execution was originally committed (the 'original-crime county'). We then investigate the existence of a local effect, that is an effect in the original-crime county and during the days surrounding the execution. Our focus is on serious but not necessarily capital-offense violent crime, as we believe that highly salient penalties could have effects that spill over into criminal behaviour not specifically covered by such penalties.[2]

We start the paper with a simple behavioural model, which we use to understand the rationale for a short-term effect of executions on crime. In our model, the comparison between the cost and benefit of crime is moderated by a psychological variable which we label 'awareness' of the cost of crime (Loewenstein, 1996). Highly salient criminal punishments raise awareness, but can be forgotten as time passes (Mullainathan, 2002). The first prediction of the model is that executions reduce crime, although this effect is only temporary. The second prediction is that the additional effect of an execution decreases in the number of recent executions.

We test the first prediction by regressing serious crime on a local execution dummy, which takes value one for the original-crime county and a short time window around the execution date. The high granularity of the dataset permits the introduction of a rich set of controls (i.e. date and county-month fixed effects), which enhances the credibility of the estimates. Executions are scheduled long in advance, and typically take place in prisons located far away from the original-crime county. Because of this, reverse causality is implausible in our setting. Furthermore,

[2]Our measure of serious crime combines homicides, rapes and assaults with weapons. However, we also show the baseline results separately for each category of crime, including homicides. There are several ways to understand these potential spillovers. In a purely Bayesian framework with partially uninformed criminals and correlated sanctions, learning about an execution could lead any violent offender to update on the likely punishment for his behaviour. In the behavioural model that we outline in Section 3.2, highly salient punishments could raise general awareness about the cost of crime even for non-capital offense criminals.

omitted variables such as the sanctions regime remain arguably constant during very short windows of time. The identifying assumption, which we regard as plausible, is that executions do not occur in dates of abnormal levels of crime (in the original-crime county, relative to other counties).

The main finding of the paper is that serious crime is .1 units lower (around 20% of the sample mean) during the local execution window. This finding is robust to using either a one-day or a three-day window, and it is qualitatively similar when studying separately homicides, rapes and assaults with weapon. The baseline finding is also robust to controlling for state and date interactions, to using non-linear models such as the Poisson model, and to alternative sample choices. We complement the baseline specification with an event analysis to study the dynamics of the crime reduction around the execution date. The examination of the estimated leads and lags suggests that serious crime remains largely unchanged in the days leading up to an execution, decreases in the day before and for two additional days, and then returns to its pre-existing level.

We test the second prediction of the model by interacting the local execution variable with the number of past executions in the original-crime county and during the previous five years. The estimated interaction is positive and statistically significant, confirming that the crime reduction effect of an additional execution is lower in counties that are highly prone to the application of the death penalty.

A limitation of the literature studying the effects of the death penalty is the inability to directly measure its impact on criminals' perceptions about the execution risk. While we do not have access to perception measures in this paper, our empirical analysis takes seriously the notion that criminals must be made aware of the existence of an execution, if such an event is going to affect their behavior. This notion provokes the focus on the original-crime county, as it is residents of this county that should be more likely to receive information (or to be affected by information) about an execution. We also test this notion directly by interacting the local execution variable with a measure of the media attention associated with that execution. We find that the baseline effect is stronger for executions that are associated with a lot of media interest, thus providing evidence on the mechanism

through which execution events affect criminal behaviour.

While awareness about an execution is predicted to increase most in the location of the original crime, nearby locations may also be partially affected. We study this in the last heterogeneity exercise of the paper, where we allow the effect of an execution to differ non-parametrically depending on the distance between a county and the original-crime county. The plotted estimates are consistent with the notion that publicity about an execution is highest in the county most associated with it, and then dissipates as one moves away from such epicentre.

Despite the voluminous literature on the death penalty, our findings arguably represent the first causal evidence that executions have an effect on crime. The focus on short-term effects implies however that we are reluctant to draw unwise policy conclusions of the type 'each execution prevents x units of crime' (Ehrlich 1975, Dezhbakhsh et al. 2003). The amount of crime prevented could certainly be larger than what we identify here. It may be, for instance, that some criminals do not respond to each additional execution but would do so if capital punishment disappeared completely from the statute book. This would be consistent with our first heterogeneity exercise, which indicates the presence of a non-linear relation between executions and crime. On the other hand, the total amount of crime prevented could be lower than our estimates suggest. The focus on the short-term implies that we cannot comprehensively study whether the effects that we identify lead to a permanent reduction, or instead to a temporal displacement of criminal activity.[3] As Chalfin and McCrary (2017) argue, both possibilities would constitute evidence of responsiveness. However, from a policy perspective distinguishing between them is obviously important.

Our objective in this paper is relatively modest. Prior to arguing that the death penalty prevents enough crime to pass any type of cost-benefit analysis, we believe it is important to credibly show that executions affect crime at all. We therefore interpret our findings as providing a necessary first step in the academic evaluation of the crime effects of the death penalty. The importance and controversy of this

[3]We fail to find any evidence of displacement to the week after the execution. However, the lack of statistical power implies that we need to be cautious about this negative finding. The same lack of statistical power makes us reluctant to examine effects beyond a single week.

question implies that we would not want it to be the last step.

Related Literature The most influential study on the effect of the death penalty is probably Ehrlich (1975), who claimed that each execution saved on average eight lives. This study was comprehensively criticised by a National Research Council report (1978). A later group of studies used state-year panel datasets and often claimed to find very large deterrence effects (Dezbakhsh et al. 2003, Mocan and Gittings 2003, Zimmerman 2003, Ekelund et al. 2006, Kovandzic et al. 2009). These papers have in turn been subject to a wide array of persuasive criticisms (National Research Council, 2012). In addition to the endogeneity concerns mentioned above, several issues have been raised. Firstly, the fact that the number of executions per state-year is highly skewed implies that estimates are disproportionately caused by a small number of outlying observations (Berk, 2005). More generally, within-state year-to-year variation in executions is low relative to variation in crime, and this makes it difficult to disentangle the effect of executions from other factors affecting the crime rate (Donohue and Wolfers, 2005). Secondly, insufficient attention has been devoted to understanding conceptually how criminals should alter their perceptions about the sanctions regime in response to executions. In addition, there has been no consensus or clear criteria as to what the best empirical measure for execution risk is. This is particularly worrying as estimates have been highly sensitive to the measure chosen. Lastly, existing literature has typically assumed a common effect, while there are reasons to expect effects to vary widely across states.

Our empirical design alleviates or circumvents the limitations above. Our simple theoretical model is explicit about the assumptions underlying a potential short-term reaction of criminals to news about an execution. Secondly, the use of disaggregated data implies that the independent variable has to be binary, thereby reducing the scope for arbitrary choices in its formation. The use of a binary independent variable also reduces concerns about outlying observations.[4] Thirdly, the focus on the days surrounding executions and the original-crime counties makes inference easier. Awareness about an execution is likely to be higher on these counties and dates,

[4]The number of executions in our sample is 493. Naturally, they represent a very small proportion of county-date observations.

so we maximise the statistical power of the study when focusing on them. Lastly, we conduct a set of heterogeneity exercises, thereby allowing for the effects to differ across executions and counties.

Our paper is most related to a small set of studies that estimate short-term effects of executions (Phillips 1982, Grogger 1990, Hjalmarsson 2009). The focus on the short-term represents definite progress in terms of identification. However, these studies are also single-jurisdiction time-series analyses, and as a result lack a control group that can capture time effects. The restriction to a single jurisdiction both limits the number of executions examined and therefore the power of the study, and forces the researchers to select states where the frequent use of the death penalty implies that the effect of an additional execution should be lower.[5]

3.2 Conceptual Framework

In this section we outline a simple framework to rationalise why executions could have a (transitory) effect on crime. At the core of the model is the notion that psychological forces that can change significantly over short time periods can affect behaviour. The effect of these forces is distinct from the Bayesian updating on the expected consequences following conviction that is at the heart of the cost benefit analysis of crime (Becker, 1968). Our model instead builds upon ideas in psychology, criminology and behavioural economics.[6]

We assume that criminal acts yield instant utility u and the expected cost of

[5]In addition to the panel studies discussed above, there is a body of work using time series econometric methods to link capital punishment and crime (Stolzenberg and D'Alessio 2004, Land et al. 2009, Cochran et al. 1994, Bailey 1998). The major limitation of these studies is that they are not well placed to identify causal effects (Charles and Durlauf, 2013).

[6]For instance, the drift theory of crime states that criminals often develop the ability to temporally neutralize the internal cost of not complying with social norms (Matza, 1964). In a related explanation, Gottfredson and Hirschi (1990) argue that criminal acts are often non-controlled, impulsive, opportunistic and short-sighted. They then relate poor parental inputs and crime through the development of a low self-control capacity. Building upon the notion that moral codes are not objective or universal, Gibbs (1989) claims that public punishment actions communicate the signal that 'society condemns some acts'. In economics, Loewestein (1996, 2000) assumes that tastes and attention can be affected in the short run by emotions and drives. These visceral factors typically change fast and are predictably correlated with external circumstances. Laibson (2001) proposes a model where trivial variations in situational cues can elicit temporary but powerful changes in marginal utility. Card and Dahl (2011) empirically show that unexpected emotional cues can trigger violent actions by changing the reference point of a 'gain-loss' utility function.

crime is c. Note that c is time-invariant and therefore not affected by informational updating on the consequences of crime.[7] Instead we follow Loewenstein (1996) in assuming that the comparison between benefits and costs of actions is distorted by a psychological state s_t that we label 'awareness' of criminal consequences. Specifically, the individual commits a crime whenever $u > s_t c$. While we refer to s_t as 'awareness', it can be alternatively interpreted as the visceral factors of Loewenstein (1996), the perceived social norms of Matza (1964) or the level of self-control in Thaler and Shefrin (1981). The important assumption for our purposes is that s_t is affected by information on events such as executions, especially when these relate to individuals or locations that feel proximate to the decision-maker.

Define $x_j = 1$ as the occurrence of an event (such as a proximate execution) in period j. $x_j = 0$ if the event did not occur. Under full persistence, $s_t = 1$ if at least one event occurred in the past, and zero otherwise. We instead posit that the state variable awareness $s_t = s(X_t, W_t)$ is a function of the history of events $X_t = (x_1, ..., x_t)$ and of a corresponding system of weights $W_t = (w_{1t}, w_{2t}, ..., w_{tt})$. We assume a random recalling process where events increase awareness but their effect is forgotten over time (Mullainathan, 2002). In particular, awareness is a linear combination of past events, $s_t = \sum_{j=0}^{t} w_{jt} x_j / \sum_{j=0}^{t} x_j$, where the weight w_{jt} can be interpreted as the persistence of event x_j in s_t. This implies that at time t event x_j is remembered with probability w_{jt}, which we assume is given by:

$$w_{jt} = m + pR_{j,(t-1)} \tag{3.1}$$

where the term m is a baseline probability of recalling an event and $R_{j,(t-1)} = 1$ if the event j was remembered in the previous period. We assume that $\alpha \equiv m/(1-p)p < 1$.

The expected recalling probabilities are obtained by backward solving (3.1) for $E(w_{jt}|x_j)$ using the facts that $E(R_{j,(t-1)}|x_j) = w_{jt-1}$ and that $w_{jj} = 1$ (i.e. the individual always remembers what just happened). Therefore,

[7] If that was the case, executions could affect these perceived consequences and therefore have medium or even long-term effects. This is of course unless criminals form their beliefs using signals from a short span of time.

$$E(w_{jt}|x_j) = p\alpha + (1-\alpha)p^{t-j}$$

This equation implies that recalling probabilities decay exponentially after an event if no new events occur. Denote $\widetilde{x}_t$ as the number of accumulated events before period t. The expected level of awareness is given by the following expression:

$$E(s_t|X_t) = p\alpha + \frac{(1-\alpha)}{\widetilde{x}_t}\sum_{j=0}^{t} x_j p^{t-j}$$

From this expression we can easily derive our first result:

> **Result 1:** *In a setting with imperfect (random) recall probabilities, a new event reduces crime temporarily.*

To prove this result, we calculate $\Delta_t = E(s_t|X_{t-1}, x_t = 1) - E(s_t|X_{t-1}, x_t = 0)$ for a given sequence of events X_{t-1}. We find that:

$$\Delta_t = \frac{1-\alpha}{\widetilde{x}_t(\widetilde{x}_t+1)}\sum_{j=0}^{t}(1-p^{t-j})x_j \tag{3.2}$$

The assumption that $\alpha < 1$ implies that $\Delta_t > 0$. Because the individual commits a crime whenever $u < s_t c$, this implies that crime is (weakly) reduced when there is a new event. The fact that $E(w_{jt}|x_j)$ decreases over time implies that this negative effect fades over time.

> **Result 2:** *In a setting with imperfect (random) recall probabilities, the effect of a new event decreases with the number of accumulated past events.*

It is straightforward to show in equation (3.2) that $\frac{\partial^2 \Delta_t}{\partial x_j \partial \widetilde{x}_t} < 0$. Intuitively, the marginal effect of a new execution is lower in counties with lots of executions.

3.3 Data

We use several data sources to construct an unbalanced date and county panel dataset. The main variables are the level of crime and the presence of an execution

caused by a crime in that county. Table 3.1 displays a set of descriptive statistics. Figure 3.1 displays the set of counties in the sample, and highlights the counties with at least one associated execution during the period that that county appears in the panel.

The information on criminal activity is extracted from the National Incident-Based Reporting System (NIBRS). The NIBRS is a voluntary program where participating law enforcement agencies report detailed information to the FBI, on a monthly basis. Our sample contains every county and month reported to the NIBRS during the 1997-2015 period. Note however that the number of counties covered has been increasing over time. Around 30% of the US population was covered in 2013, up from close to zero in 1997.[8] The number of county-date observations in the main estimating sample is 8,462,202.

We use information at the incident level, taking into account that an incident can be associated with multiple offences (e.g. an armed assault being inputted as an assault and an illegal carrying of weapons). Our dependent variable 'serious crimes' is the sum of homicides (murders and non-negligent manslaughters), rapes (forcible rapes and sex assaults) and aggravated assaults accompanied by the use of a fire weapon.[9] In Panel B of Table 3.1 we find that the average of serious crimes per county-day is .14. This low number results both the relative rareness of these crimes and the fact that many counties covered by the NIBRS are quite small. Rapes and assault with weapons are similarly prevalent, with homicides being less common.

The data on executions is extracted from the website *https:deathpenaltyinfo.org*. For every execution we identify the county where the capital offense was originally committed (i.e. the 'original-crime county') and use this to compute our main independent variables of interest. In total, we find 493 executions for which the original-crime county is covered by the NIBRS during that particular month (see Panel A of Table 3.1). These executions occur in 143 different counties, with an average of 3.45 executions per county. Figure 3.A1 in the Appendix shows the number of executions per year for the period and counties in our sample. Figure

[8]Our empirical strategy, which controls for the interaction of county, month and year, is designed to account for within-county changes in coverage over time.

[9]Aggravated assaults are those regarded as unlawful attacks for the purpose of inflicting severe or aggravated bodily injuries.

3.A2 in the Appendix displays the distribution of executions and serious crimes by day of the week.

Lastly, we use 'death penalty' search results from Google Trends to measure the media attention received by each execution. This data is available only from 2004 onwards. Google Trends does not report the total number of searches, but instead an index of the relative intensity of searches. We compute our measure in the following way. We first compute g_{st}, which is the index for state s (relative to all states in the US) in date t. We then use US-aggregate information to calculate z_{mt}, which is the index for date t in the month m to which t belongs.

Our measure of death-penalty related media attention in a state-date combination is $M_{st} = g_{st} \times z_{mt}$. The measure captures the relative interest in death-penalty related topics within a state and date. Note that by construction this measure is normalised by the amount of media attention present within a particular month. It is therefore orthogonal to the large increase in the use and coverage of the internet over our sample period.

3.4 Event Study Analysis

In this section we study the evolution of serious crime in the days preceding and following an execution, in the county where (many years earlier) the crime leading to that execution was originally committed. Our main independent variable of interest is the dummy $Executionj_{it}$, which takes value one (for a county i date t combination) on the day j relative to the date of an execution motivated by a crime in that county. The estimating equation is:

$$crime_{it} = \sum_{j=-6}^{+6} \beta_j Executionj_{it} + \gamma_t + (\alpha_i \times \pi_{m(t)} \times \lambda_{y(t)}) + \epsilon_{it}$$

where $crime_{it}$ is the number of crimes, γ_t represents a set of date indicators and $(\alpha_i \times \pi_{m(t)} \times \lambda_{y(t)})$ is a set of interactions between county, month and year indicators. The date indicators absorb any US-wide shocks to crime occurring on a particular

date.[10] Including the interacted county, year and month indicators controls for any county-specific shocks within the relatively narrow time window of a month. Standard errors are clustered at the state and year level, a choice that we regard as conservative.

The identifying assumption implicit in the estimation of (3.3) is that executions are not scheduled to take place on days of idiosyncratically high or low crime in the original-crime county, relative to other counties on the same date and to that same county on that same month. We regard this assumption as plausible. Executions are scheduled well in advance and they take place in a single maximum-security prison per state. The prison is typically not located in the county where the original crime occurred.[11]

Figure 3.2 displays the estimated effects $\hat{\beta}_{-6} \dots \hat{\beta}_{+6}$, using the number of serious crimes (homicides, rapes and assaults with weapon) as dependent variable. We find that crime remains largely stable prior to the eve of an execution (the dip at $j = -4$ is not statistically significant). A statistically significant decrease on the eve of the execution persists approximately for two additional days, and then crime returns to its pre-existing level.[12] Figure 3.A3 in the Appendix shows that the estimates are essentially unchanged when controlling for the interaction of date and state indicators.

We interpret the evidence in Figure 3.2 as indicating that executions have a negative effect on crime. This negative effect is consistent with news about the execution affecting would-be criminals' awareness of the potential consequences of crime. Figure 3.2 further indicates that this effect is very short lived, lasting around three days. In light of this evidence and in order to increase statistical power, our baseline regressions use dummy variables capturing time windows around the execu-

[10] In one of the robustness tests, we find that the results are essentially unchanged when interacting the date indicators with state indicators.

[11] The Death Penalty Information Center maintains a list of upcoming executions (https://deathpenaltyinfo.org/upcoming-executions). Executions are scheduled up to five years in advance. For a list of the prisons where executions take place in each state, see https://en.wikipedia.org/wiki/Execution_chamber.

[12] The finding that serious crime starts to decrease already the day before the execution is consistent with Hong and Kleck (2018) finding that newspaper and television stories start to report on executions the day before they take place.

tion date.[13] We estimate our main results using a one-day time window (comprising exclusively of the execution date), as well as a three-day window (which additionally includes the day before and the day after the execution). We show below that the results do not depend on the choice of window.

3.5 Main Results

In this section we display the main results of the paper. We estimate variations of the equation:

$$crime_{it} = \beta ExecutionWindow_{it} + \gamma_t + (\alpha_i \times \pi_{m(t)} \times \lambda_{y(t)}) + \epsilon_{it}$$

where $ExecutionWindow_{it}$ is a dummy variable taking value one in a time window (either one or three-day) around an execution, in the county where the original crime was committed. The other variables are defined as above. We present first the baseline results and then a set of additional tests exploring their robustness.

Baseline Estimates Table 3.2 displays the results of estimating (3.3), separately for the serious crime variable and for each of its three components. We find first that the number of serious crimes is approximately .10 units lower in the window around the execution date. Note that this is a very large effect: the mean of the variable serious crimes in the sample is only slightly higher, at .14. A better reference point is perhaps the mean in the subset of counties associated with at least one execution throughout our sample period. As Panel B of Table 3.1 shows this is .48. Evaluated against this benchmark, the estimate represents a decrease of approximately 20% in serious crime.

We find qualitatively similar results when evaluating separately the effects on homicides, rapes and assaults with weapons. Evaluated against the mean of the dependent variable, the effects are strongest for homicides. The .012 coefficient

[13]The baseline regressions implicitly assume that dates close to the execution date but outside the chosen time window are not associated with higher or lower levels of crime (relative to that same county in that same month). In light of the evidence in Figure 3.2, this assumption appears to be justified.

of the one-day window represents approximately half of the mean of the homicide rate in the counties associated with at least one execution. The relative rarity of homicides implies however that the estimate is not statistically significant for the three-day window. For rapes and assaults with weapons the estimated effects represent between 15% and 25% of the means of the respective dependent variables in the counties associated with at least one execution.

Robustness In Table 3.3 we evaluate the robustness of the main findings to modifying the set of controls, the estimating equation, the clustering strategy and the estimating sample. An extensive robustness exercise is particularly important in our context, given the well-documented finding that death-penalty estimates can be extremely fragile to specification choices (Donohue and Wolfers, 2009).

We first add interactions between date and state indicators to the baseline regression (3.3). In doing this, the identification assumption is that executions are not scheduled in days of idiosyncratically high or low crime in the original-crime county, relative to other counties *in the same state*. Because executions do not typically take place in the original-crime county, we regard this assumption as particularly plausible. The estimates are largely unchanged.

Relative to non-linear methods with a large number of fixed effects, OLS has the advantage of being more robust and easy to interpret (Angrist and Pischke, 2009). In the fifth row of Table 3.3 we estimate however a Poisson model, which may be more appropriate given the count nature of the dependent variable (Hjalmarsson, 2009). The coefficients are strongly statistically significant and similar in magnitude to those in the baseline regression. The estimated coefficients indicate that executions are associated with a 16%-20% decrease in serious crime, in the original-crime county. In the fourth row, we find that the coefficients are strongly significant, albeit much smaller in magnitude, when the number of serious crimes (plus one) is entered in logs.[14]

In the baseline regression, the standard errors allow for correlation within the same state and year. In the fourth row of Table 3.3, we allow the date to be an

[14]The log model is likely misspecified in our context, given the high proportion of zeros in the serious crime variable.

additional dimension of error correlation (Cameron, Gelbach and Miller, 2011). The two-way clustered standard errors are essentially identical to the baseline one-way standard errors.

In the last three rows of Table 3.3, we examine the robustness of the estimates to changes in the estimation sample. We first exclude counties from the state of Texas, which can be regarded as an outlier in its enthusiastic application of the death penalty.

Secondly, we limit the sample to including only counties from states where at least one execution took place during the sample period. Note that the non-execution states only contribute in the baseline regression to the estimation of the date fixed effects. It could be argued, however, that excluding these non-execution states generates a better counterfactual to the original-crime counties during the execution window.

In the last row of Table 3.3 we account for the fact that the coverage of the NIBRS has increased significantly throughout the sample period. Our baseline empirical strategy (where we control for the interaction of county, year and month indicators) is designed to absorb secular trends and even relatively short-term variations in crime. Nevertheless, we repeat the estimations with a (balanced) panel of counties that are present throughout the period 2002-2015.

Overall, we find that the baseline estimates are not sensitive to these reasonable alterations of the estimation sample.

3.6 Heterogeneity

The theoretical framework in Section 3.2 provides a number of testable additional predictions. Firstly, it suggests that the awareness impact of an additional execution should be higher in counties where executions are relatively rare. Secondly, the model is based upon the notion that it is news about execution that increases awareness of the negative consequences of crime. This suggests that executions that are widely covered by the media should impact criminal decision-making more strongly than those that receive less attention. It further indicates that, while the strongest

effects should be concentrated on the original-crime county, neighbouring counties may also be partially affected, as there may be informational spillovers towards these counties.

Death Penalty Propensity We first study whether the increase in awareness decreases in the number of past executions of the county. Our measure is the number of past executions in the five year window previous to (the month before) an execution, in the original-crime county that the execution is associated with. Note that by defining the variable in this way, we ensure that it does not mechanically increase over the sample period. We interact this measure with the main independent variable of interest, the time window around an execution. An additional complicating factor is the strong positive correlation between the number of executions in a county and its population, a correlation that one would expect. The complication arises because the population of a county is also related to its level of crime. Therefore, in this heterogeneity exercise we also control for the interaction with the county's (log of) population.

Panel A of Table 3.4 displays the results. The positive coefficient of the interaction with the county's death penalty propensity indicates that the effect of executions on serious crime is lower in counties with a higher number of past executions, as predicted by our conceptual framework.This finding provides a potential reconciliation of our paper with Hjalmarsson (2009), who finds no effect of execution on homicides in a sample from the state of Texas. Given that Texas carries out a disproportionately high number of executions every year, our findings predict that the effect of each additional execution should be relatively low.

Media Attention In our second heterogeneity exercise, we differentiate between executions receiving different levels of media attention. Our initial measure of media attention is based on Google Trends search results and explained in detail in Section 3.3 above. We use this measure to split the main independent variable into two different variables, depending on whether the execution windows coincide with (within state and date) above median media attention regarding death penalty is-

sues.[15] Because our measure of media attention varies within a county/month and also within a date, we need to explicitly control for it in the regression.

We find in Panel B that it is only executions associated with above median media attention that are associated with decreases in crime. For these execution windows, the effect is in fact much larger than the baseline effect. For instance, for the one day window the effect is -.175. The above median and the below median dummies are statistically different from each other at the 7% level.

Neighbouring Counties Lastly, we examine in Panel C whether counties that neighbour the original-crime county also experience a decrease in crime during the execution window. To study this, we use an additional variable: a dummy taking value one during an execution window and for counties sharing a border with the original-crime county. Naturally, the regression maintains the main independent variable of interest.

The resulting estimate for the one-day window indicates that the number of serious crimes is approximately .025 units lower on the day of an execution, in the counties neighbouring the original-crime county. Note that this effect is economically meaningful and statistically significant. Reassuringly, it is also much smaller than the original-crime county estimate, which remains largely unchanged.[16]

Panel C of Table 3.4 suggests that the awareness regarding an execution (and its impact on crime) may ripple away from its epicentre at the original-crime county. To study the functional form of this relation, we calculate the distance between (the centre of) a county and (the centre of) the original-crime county. We then create a set of distance dummies and combine these with the original time windows around executions. The corresponding estimates from introducing these in equation (3.3), for the one-day window, are displayed in Figure 3.3. At zero kilometres we display the baseline effect for the original-crime county.

We find that surrounding counties with a (centre to centre) distance of less than

[15]Because the information from Google Trends is only available from 2004 onwards, these regressions exclude all years before that. This reduces the number of executions to 311. Out of these, 189 are classified as coinciding with high media attention.

[16]It is one fourth of the original-crime estimate, although as we can see in Panel B of Table 3.1 the average number of serious crimes is also smaller, at .21 rather than .48.

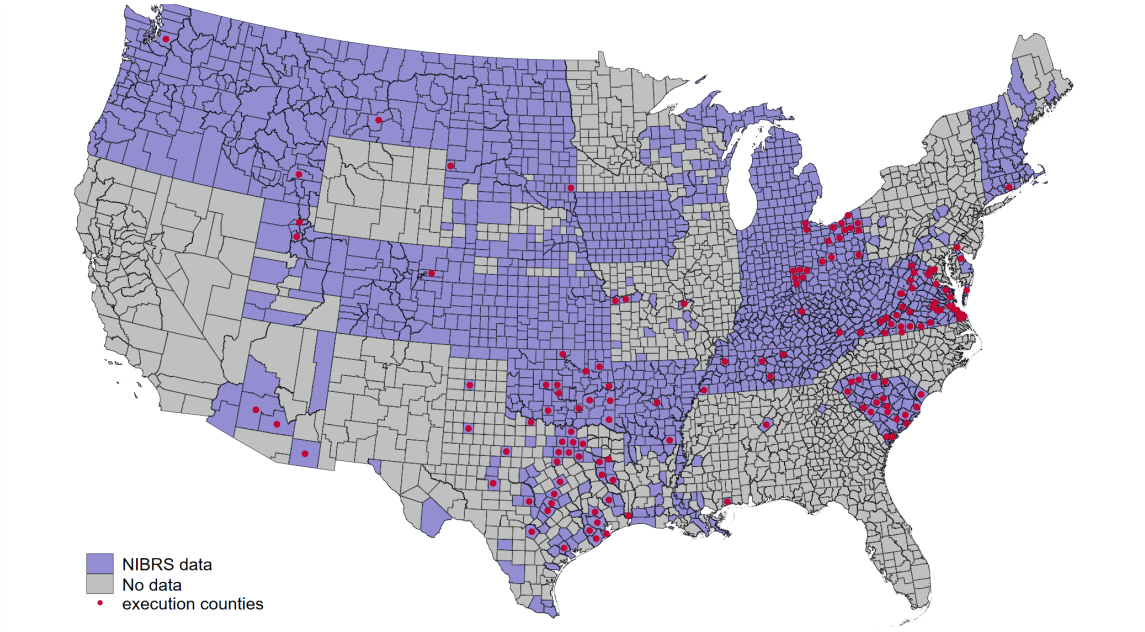
50 kilometres experience a .04 decrease in the number of serious crimes, during the execution day. The effect is .02 decrease for counties between 50 and 100 kilometres away, smaller at higher distances, and statistically insignificant beyond 2,000km. Overall, these results are consistent with the notion that awareness about an execution (and its effect on behaviour) is strongest for would-be criminals that are closer to the location of the original crime.

3.7 Conclusion

Although the death penalty is a controversial criminal justice policy in the US, credible causal evidence about its effect is scarce. In this study, we focus in the county-day level variation as a plausibly source of identification. We show that executions cause a local and temporary reduction in serious violent crime (homicides, rapes and assaults with weapon). We interpret this result using the simple behavioural model explained in Section 3.2. The empirical findings of Sections 3.4, 3.5 and 3.6 are consistent with the predictions of the theoretical framework. Namely, that the effect is decreasing in the number of recent executions in the county, and higher for executions associated with a lot of media attention. We interpret our findings as providing a necessary first step in the academic evaluation of the crime effects of the death penalty. Our focus on short-term effects implies however that we are reluctant to draw unwise policy conclusions regarding the effectiveness of death penalty to prevent crime. Indeed, the magnitude of our results are much lower that previous studies claiming significant deterrence effects this policy. However, as discussed in Section 3.1, we find reasons to think that compared to a counterfactual scenario with no executions, the total amount of crime prevented could certainly be larger or smaller than what we identify.

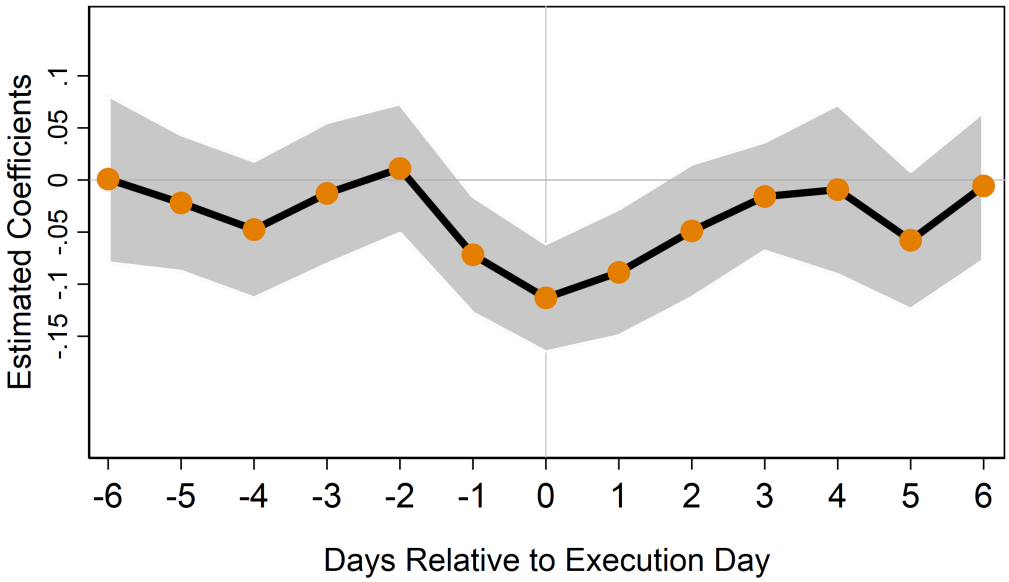
3.8 Figures of the Chapter

Figure 3.1: Sample of Crimes and Executions



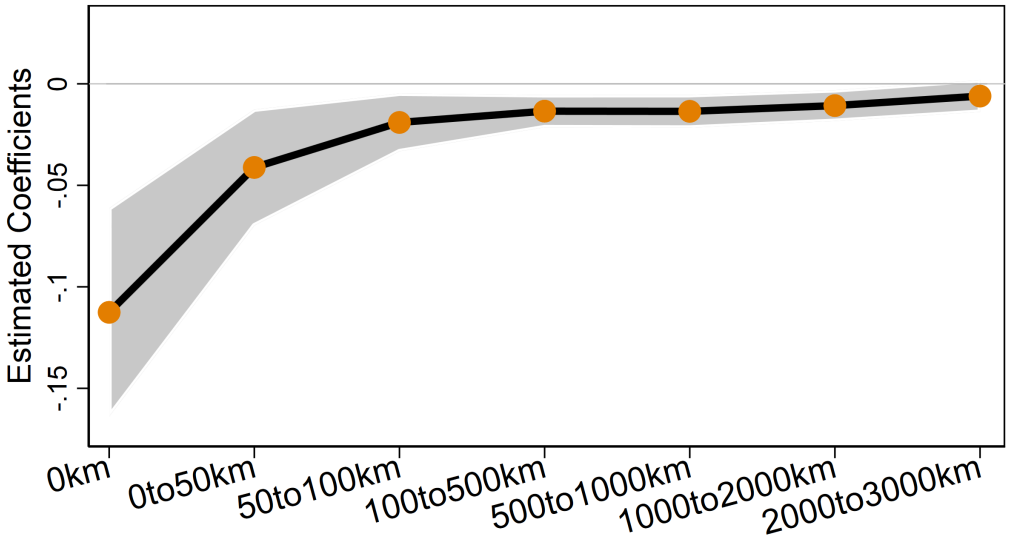
The graph displays information on the counties with crime information (NIBRS) in the sample. The graph also displays markers for the counties with at least one execution during the sample period.

Figure 3.2: Event Study: Evolution of Serious Crime in Days Around Execution Day



This figure displays the coefficients for the interactions between the days relative to the execution date and the county where the capital crime relating to that execution was originally committed. The dependent variable in the regression is the number of serious crimes (homicides, rapes, assaults with weapon). The regression includes Date and County X Month X Year indicators. The standard errors are clustered at the state X year level. The number of observations is 8,462,173 forevery regression.

Figure 3.3: Effects by Distance to Original-Crime County



This figure displays the coefficients for the combinations of the one-day time window around an execution and the distance to the centre of the county where the capital crime relating to that execution was originally committed. At zero we have the original-crime county. The dependent variable in the regression is the number of serious crimes (homicides, rapes, assaults with weapon). The regression includes Date and County X Month X Year indicators. The standard errors are clustered at the state X year level. The number of observations is 8,462,173 for every regression.

3.9 Tables of the Chapter

Table 3.1: Descriptive Statistics

Panel A: Executions

	N in Sample	**N of Counties**	**Only Execution or Neighboring Counties**			
			Avg N	Std	Min	Max
Executions	493	143	3.45	7.86	1	70
Executions in Neighboring County	1,496	350	4.27	8.04	1	70

Panel B: Crimes

	All Counties in Sample			**Execution Counties**		**Neighboring Counties**	
	Total	Avg	Std	Avg	Std	Avg	Std
Serious Crimes	1,173,252	0.14	0.66	0.48	1.27	0.21	0.69
Homicides	50,257	0.01	0.09	0.02	0.17	0.01	0.11
Rapes	586,928	0.07	0.35	0.21	0.64	0.10	0.43
Assaults Weapons	540,346	0.06	0.43	0.25	0.84	0.10	0.41

The table displays descriptive statistics for executions and crimes for the sample period. The sample consists of 8,462,202 day-county observations during the period 1997-2015.

Table 3.2: Baseline Estimates

Dependent Variable	(1) Serious Crime	(2) Homicides	(3) Rapes	(4) Assaults Weapons
1 Day Execution Window	-.101***	-.012**	-.053***	-.036*
	(.026)	(.006)	(.021)	(.019)
3 Days Execution Window	-.085***	-.006	-.045***	-.033***
	(.016)	(.004)	(.012)	(.011)
Mean Dependent Variable	.139	.006	.069	.064
In Execution Counties	.481	.024	.209	.248

This table displays estimates of OLS regressions of crime on the combinations of a time window surrounding an execution and the county where the crime motivating that execution was originally committed. The 1 Day Execution Window comprises of the execution date. The 3 Days Execution Window includes also the day before and the day after the execution. Every coefficient results from a different regression. All regressions control for Date indicators and County X Month X Year indicators. The standard errors are clustered at the State X Year level. The dependent variable in column (1) is the sum of the dependent variables in columns (2), (3) and (4).

Table 3.3: Robustness

Dep. Variable = Serious Crimes	**(1) 1 Day Window**	**(2) 3 Days Window**
Baseline	-.101***	-.085***
	(.026)	(.016)
Adding State X Date Indicators	-.094***	-.078***
	(.026)	(.016)
Poisson Regression	-.202***	-.157***
	(.064)	(.038)
Dep. Variable in Logs	-.041***	-.032***
	(.012)	(.007)
Multi-Way Clustering	-.101***	-.078***
	(.026)	(.016)
Sample Excludes Texas	-.12***	-.094***
	(.045)	(.026)
Sample Includes only States with Executions	-.094***	-.078***
	(.026)	(.016)
Balanced Panel 2002-2015	-.095***	-.076***
	(.028)	(.019)

This table displays estimates of OLS regressions of the number of serious crimes on the interaction between a time window surrounding an execution and the county where the crime motivating that execution was originally committed. Every Panel/Column combination displays a different regression. In Column (1) the time window comprises of the execution date. In Column (2) the time window includes also the day before and the day after the execution. In Panel A the independent variable is interacted with the propensity to carry out the death penalty in that county, measured as the (log of the) accumulated number of executions in the county. The regression also controls for the interactions with the (log of the) number of days since the last execution, and the (log of the) county's population. In Panel B an additional independent variable is added to the regression, capturing the interaction between a time window surrounding an execution and the counties neighbouring the county where the crime motivating the execution was originally committed. In Panel C, the main independent variable is interacted with a weekly indicator of media attention, measured using the Google Trends result at the state level for death penalty topics. All regressions control for Date indicators and County X Month X Year indicators. The standard errors are clustered at the State X Year level.

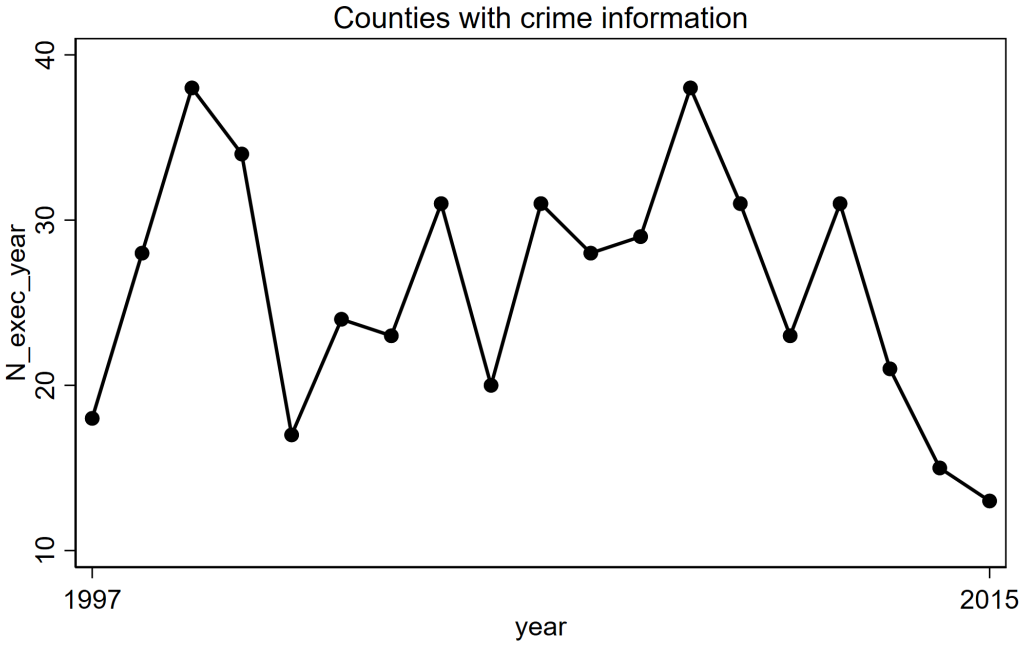
Table 3.4: Heterogeneity

Dep. Variable = Serious Crimes	(1) 1 Day Window	(2) 3 Days Window
Panel A: Death Penalty Propensity		
Execution Window	.362	.405***
	(.288)	(.171)
Execution Window X Death Penalty Propensity	.075*	.045***
	(.039)	(.019)
Execution Window X County Population	-.043*	-.043***
	(.026)	(.015)
Panel B: Media Attention		
Execution Window X High Media Attention	-.175***	-.125***
	(.048)	(.034)
Execution Window X Low Media Attention	.008	-.062
	(.074)	(.048)
High Media Attention	-.001	-.001
	(.001)	(.001)
Panel C: Effect on Neighbouring Counties		
Execution Window	-.102***	-.085***
	(.026)	(.016)
Execution Window (Neighbouring County)	-.024**	-.027***
	(.012)	(.008)

This table displays estimates of OLS regressions of the number of serious crimes on the interaction between a time window surrounding an execution and the county where the crime motivating that execution was originally committed. Every Panel/Column combination displays a different regression. In Column (1) the time window comprises of the execution date. In Column (2) the time window includes also the day before and the day after the execution. In Panel A the independent variable is interacted with the propensity to carry out the death penalty in that county, measured as the (log of the) accumulated number of executions in the county. The regression also controls for the interactions with the (log of the) number of days since the last execution, and the (log of the) county's population. In Panel B an additional independent variable is added to the regression, capturing the interaction between a time window surrounding an execution and the counties neighbouring the county where the crime motivating the execution was originally committed. In Panel C, the main independent variable is interacted with a weekly indicator of media attention, measured using the Google Trends result at the state level for death penalty topics. All regressions control for Date indicators and County X Month X Year indicators. The standard errors are clustered at the State X Year level.

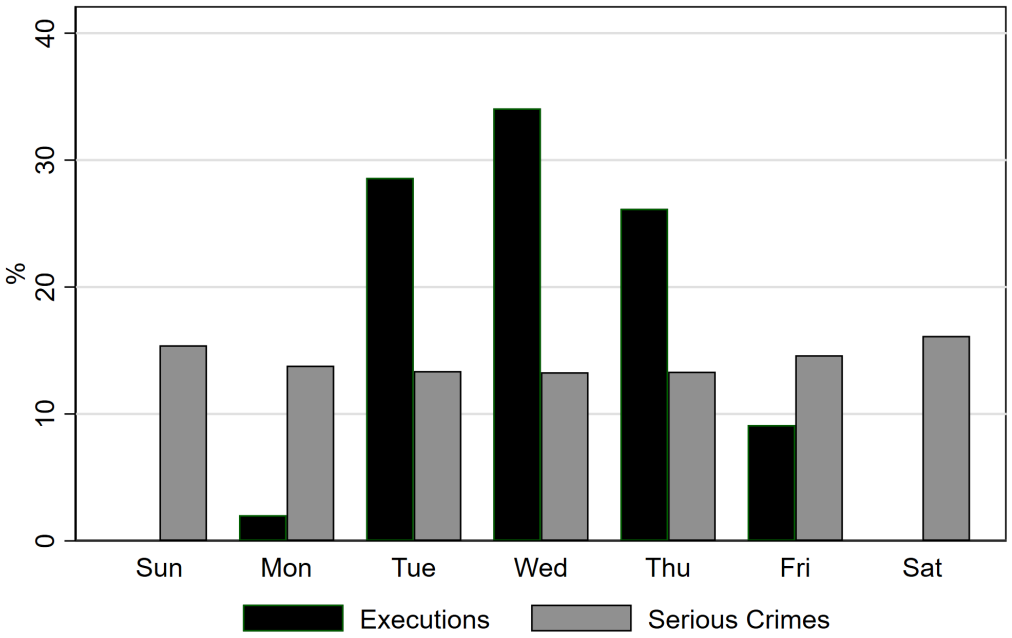
3.10 Appendix A: Additional Tables and Figures of the Chapter

Figure 3.A1: Executions by Year



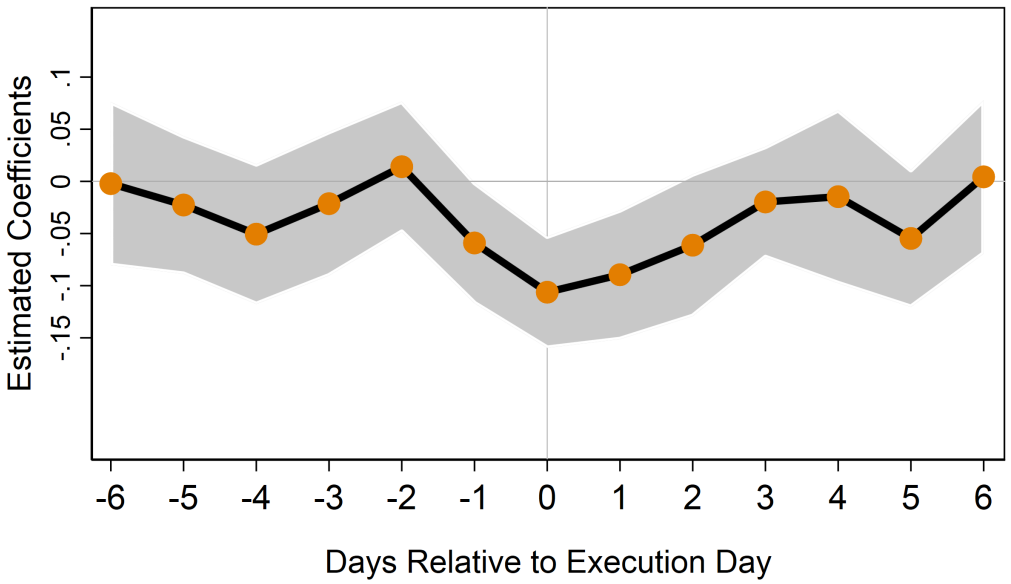
The graph displays information on the number of executions for counties with crime information from NIBRS in the sample.

Figure 3.A2: Distribution of Executions and Crime by Day of the Week



The figure displays the distribution of executions and serious crimes observed in the sample by day of the week.

Figure 3.A3: Event Study: Evolution of Serious Crime in Days Around Execution Day. Controlling for State X Day Indicators



This figure displays the coefficients for the interactions between the days relative to the execution date and the county where the capital crime relating to that execution was originally committed. The dependent variable in the regression is the number of serious crimes (homicides, rapes, assaults with weapon). The regression includes Date X State and County X Month X Year indicators. The standard errors are clustered at the state X year level. The number of observations is 8,462,173 forevery regression.

Conclusion

This thesis used three different natural experiments to study, in three specific settings, the economic importance of mechanisms usually regarded as "soft factors" within organisations and markets.

The first chapter contributes to the literature by providing one of the first causal evidence on how the ability to communicate face-to-face (in addition to electronic communication) can increase organisational performance. The study focus on a large organisation where workers must complete a task that requires communicating electronically with their teammates. The teams have to answer emergency calls and based on the extracted information they allocate police officers to attend the incident place. A computerized queuing system that allocates calls to workers in a way that is orthogonal to the incidents creates a natural experiment where teammates often share the same room. Co-location of teammates is associated with a significant reduction in the allocation and response time of the police. An additional contribution of the chapter is to show that the benefits from face-to-face communication are largely contingent to the characteristics of the tasks, teams and working environment. In this sense, the increase in productivity is larger when the task is more urgent, the team is more homogenous and the workload is higher. Additionally, the chapter develops a theoretical framework that allows understanding the costs associated to face-to-face communication. An advantage of the empirical setting is that it allows measuring and comparing the benefits and the (operational) costs of face-to-face interaction. Results indicate that benefits are significantly larger than (operational) costs.

The second chapter exploits a natural experiment during the "age of mass migration" where thousand of immigrants travelled (mainly from Europe) to the US sharing the voyage with other individuals from different backgrounds and socioe-

conomic status. The study focus on the social interactions among individuals who met for the first time in the ship during the few days of the voyage. Using a novel dataset linking more than 300,000 immigrants to their ships of arrival, I study how the characteristics of the (previously unknown) shipmates affect labour and residential outcomes. I focus on two predetermined characteristics of the shipmates that proxy for the quality of their connections at destination. Namely, the average earnings and the total number of past emigrants from the same place of origin to the US. Both variables aim to measure how well connected is a shipmate. Results indicate that individuals travelling with better connected shipmates get higher quality jobs in the US. The chapter discusses several results suggesting that shipmates provide information or access to job opportunities. An important finding of the chapter is that the effects are stronger for those individuals with poor connections. The results of the study highlight the importance of brief social interactions and their persistent effects on labour outcomes.

The third chapter shows a causal link between executions and a local reduction in serious violent crime (homicides, rapes and assaults with weapon) in the days surrounding the event. The identification strategy exploits the high frequency of crime data (county x day) and the fact that executions are set many months in advance and at the state level. Focusing on the county where the capital crime occurred and using daily variation makes the identification of the effects credible and transparent. The chapter offers a theoretical explanation for the causal pattern observed in the data. The simple behavioural model is based on the idea that information updates about death penalty can (temporarily) increase the perceived costs of crime. The model has additional implications that are consistent with other findings of the chapter. In this sense, results indicate that executions have lower effects in places with high propensity to death penalty and when executions received lower media attention.

Bibliography

Abramitzky, R., Boustan, L.P, and Eriksson, K. (2014), "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration" *Journal of Political Economy*, 122, June: 467-506.

Abramitzky, R., Boustan, L.P, 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 (5): 1832-56.

Acemoglu, D., and Autor, D. (2011), "Skills, tasks and technologies: Implications for employment and earnings", *Handbook of Labor Economics*, 4, 1043-1171.

Aho, A. V., and Corasick, M. J. (1975), "Efficient String Matching: An Aid to Bibliographic Search". *Communications of the Association for Computing Machinery*. 18 (6): 333-340. doi:10.1145 /360825. 360855.

Ammermueller, S., and Pischke, J. S. (2009), "Peer Effects in European Primary Schools: Evidence from PIRLS," *Journal of Labor Economics* 27, July 2009, 315-348.

Angrist, J. D. (2014). "The perils of peer effects," *Labour Economics*, Elsevier, vol. 30(C), pages 98-108.

Angrist, J. D., and Lang, K. (2004) "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *American Economic Review*, 94 (5): 1613-1634.

Angrist, J. D., and Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Arrow, K. E. (1974), *The limits of organization.*, W. W. Norton Company, 17 Feb 1974.

Atack, J., and Bateman, F. (1992), "Matchmaker, Matchmaker, Make Me a Match: A General Personal Computer-Based Matching Program for Historical Research" *Historical Methods* 25, 2: 53-65.

Athey, S. (2016), "Machine Learning and Causal Inference for Policy Evaluation". KDD '15 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pages 5-6

Autor, D. H., Levy, F., and Murnane, R. J. (2003), "The skill content of recent technological change: An empirical exploration", *The Quarterly Journal of Economics*, 118(4), 1279-1333.

Baeza-Yates, R. A., and Gonnet, G. H. (1996) Fast text searching for regular expressions or automaton searching on tries. *Journal of the Association for Computing Machinery*43, 6, November, 1996, 915-936.

Bailey, M., Cole, C. Henderson, M., and Massey, C. (2017), "How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth" *NBER Working Paper* No. 24019, November 2017.

Bailey, W. C. (1998). "Deterrence, brutalization, and the death penalty: Another examination of Oklahoma's return to capital punishment." *Criminology*, 36(4), 711-734.

Bandiera, O., Barankay, I., and Rasul, I. (2010), "Social incentives in the workplace", *The Review of Economic Studies*, 77(2), 417-458.

Bandiera, O., Barankay, I., and Rasul, I. (2013), "Team incentives: evidence from a firm level experiment", *Journal of the European Economic Association*, 11(5), 1079-1114.

Bandiera, O., Rasul, I., and Viarengo, M. (2016), "The Making of Modern America: Migratory Flows in the Age of Mass Migration", *Journal of Development Economics*, Vol. 102, May 2013, pp 23-47.

Bandura, A. (1982), "The psychology of chance encounters and life paths." *American Psychologist*, 37(7), 747-755.

Battisti, M., Peri, G. Romiti, A. (2016), "Dynamic Effects of Co-Ethnic Networks on Immigrants' Economic Success" , *Mimeo*, October 2016.

Bavelas, A., and Barrett, D. (1951), "An Experimental Approach to Organizational Communication", *Personnel* 27, 367-371.

Bayer, P., Ross, S. L., and Topa, G. (2008), "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, 116(6), 1150-1196.

Bayley, D. H. (1994), *Police for the Future.* New York, NY: Oxford University Press.

Beaman, L. (2012) "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S." *Review of Economic Studies* 79 (1), 128-161.

Becker, G. S. (1968) "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 1968, 76:2, 169-217.

Benson, B. L., and Rasmussen, D. W. (1991), "Relationship between illicit drug enforcement policy and property crimes", *Contemporary Economic Policy*, 9(4), 106-115.

Bentolilla, S., Michelacci, C., and Suarez, J. (2010), "Social Contacts and Occupational Choice," *Economica*, 77(305), 20-45.

Berk, R. (2005) "New claims about executions and general deterrence: Dja Vu all over again?", *Journal of Empirical Legal Studies*, 2(2), 303-330.

Bertrand, M., Luttmer, E. F. P., Mullainathan, S. (2000), "Network Effects and Welfare Cultures", *Quarterly Journal of Economics*, 115(3), pp. 1019-1055, August 2000.

Blanes i Vidal, J., and Kirchmaier, T. (2017), "The Effect of Police Response Time on Crime Clearance Rates", *Review of Economic Studies*, forthcoming.

Bleakley, H., and Chin, A. (2010), "Age at Arrival, English Proficiency, and Social Assimilation among U.S. Immigrants" *American Economic Journal: Applied Economics* 2 (1): 165-92.

Bloom, N., Garicano, L., Sadun, R., and Van Reenen, J. (2014), "The distinct effects of information technology and communication technology on firm organization", *Management Science*, 60(12), 2859-2885.

Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2015), "Does working from home work? Evidence from a Chinese experiment.", *The Quarterly Journal of Economics*, Oxford University Press, vol. 130(1), 165-218.

Blume, A., and Ortman, A. (2007), "The Effects of Costless Pre-Play Communication: Experimental Evidence from Games with Pareto-Ranked Equilibria", *Journal of Economic Theory*, 132, 274-290.

Borjas, G. J. (2015), "The Slowdown in the Economic Assimilation of Immigrants: Aging and Cohort Effects Revisited Again," *Journal of Human Capital* 9, no. 4: 483-517.

Borjas, G. J. (2000), "Ethnic Enclaves and Assimilation," *Swedish Economic Policy Review*, 7, 89-122.

Borjas, G. J. (2015), "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics* 3, October: 463-89.

Bramoullé Y., Galeotti A., and Rogers B. (eds.) (2016), *The Oxford Handbook of the Economics of Networks.* Oxford University Press.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45(1), 5-32, 2001

Breza, E. (2016) " Field Experiments, Social Networks, and Development." In: Y. Bramoull, A. Galeotti and B. W. Rogers (Hrsg.), *The Oxford Handbook of the economics of networks*, (Oxford Handbooks), Oxford: Oxford University Press, S. 649-671.

Brunner, B., and Kuhn A. (2009), "To Shape the Future: How Labor Market Entry Conditions Affect Individuals Long-Run Wage Profiles," *IZA Discussion Paper Series*, No. 4601.

Burgess, S., Propper, C., Ratto, M., Scholder, K., von Hinke, S., and Tominey, E. (2010), "Smarter Task Assignment or Greater Effort: the impact of incentives on team performance", *The Economic Journal*, 120(547), 968-989.

Cameron, A., Gelbach, J., and Miller, D. (2011), "Robust Inference With Multiway Clustering.", *Journal of Business and Economic Statistics*, 29(2), 238-249.

Caeyers, B., and Fafchamps, M. (2016), "Exclusion Bias in the Estimation of Peer Effects," *NBER Working Papers* 22565, National Bureau of Economic Research, Inc.

Card, D., and Dahl, G. B. (2011). "Family violence and football: the effect of unexpected emotional cues on violent behavior." *The Quarterly Journal of Economics*, 126 1, 103-43.

Catalini, C. (2016), "Microgeography and the direction of inventive activity", *Working paper.*

Chalfin, A., and McCrary, J. (2017), "Criminal Deterrence: A Review of the Literature." *Journal of Economic Literature*, 55 (1): 5-48.

Chan, D. C. (2016), "Teamwork and moral hazard: evidence from the emergency department", *Journal of Political Economy*, 124(3), 734-770.

Charles, K., and Durlauf, S. N. (2013) "Pitfalls in the use of time series methods to study deterrence and capital punishment", *Journal of Quantitative Criminology*, 29(1), 45-66.

Chiswick, B. R. (1978), The Effect of Americanization on the Earnings of Foreign-born Men, *Journal of Political Economy* 86.5: 897-921.

Christien P., and Churches T. (2005), "Febrl - Freely extensible biomedical record linkage.", *Manual*, release 0.3, edition 2005.

Cloninger, D. O., and Marchesini, R. (2001) "Execution and deterrence: a quasi-controlled group experiment", *Applied Economics*, 33(5), 569-576.

Cloninger, D. O., and Marchesini, R. (2006) "Execution Moratoriums, Commutations and Deterrence: the Case of Illinois", *Applied Economics*, 38(9), 967-973.

Cochran, J. K., Chamlin, M. B., and Seth, M. (1994). "Deterrence or Brutalization? An Impact Assessment of Oklahoma's Return to Capital Punishment." *Criminology*, 32: 107-134.

Cohen-Cole, E., Durlauf, S., Fagan, J., and Nagin, D. (2008) "Model Uncertainty and the Deterrent Effect of Capital Punishment", *American Law and Economics Review*, 11(2), 335-369.

Cooper, R. W., DeJong, D. V., Forsythe, R., and Ross, T. W. (1992), "Communication in Coordination Games", *Quarterly Journal of Economics*, 107, 739-771.

Curtis, I. (2015), "The use of targets in policing ", available at: https://www.gov .uk/govern ment/uploads/system/uploads/attachment_data/file/466058 /Review_Targets_2015.pdf

Daniels, R. (2002), *Coming to America: A History of Immigration and Ethnicity in American Life.*, Second Edition, Harper Collins, isbn=9780060505776.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999), "The economics of career concerns, part II: Application to missions and accountability of government agencies", *The Review of Economic Studies*, 66(1), 199-217.

Dezhbakhsh, H., and Rubin, P. H. (2007) "From the'Econometrics of Capital Punishment'to the'Capital Punishment'of Econometrics: On the Use and Abuse of Sensitivity Analysis", *Emory Law and Economics Research Paper*, (07-18), 07-21.

Dezhbakhsh, H., Rubin, P. H., and Shepherd, J. M. (2003) "Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data", *American Law and Economics Review*, 5(2), 344-376.

Dodd, T., and Simmons, J. (Eds.) (2002/03), "British Crime Survey", available at http://webarchive.nationalarchives.gov.uk/20110220105210 /rds.homeoffice.gov.uk/rds/pdfs2/hosb703.pdf

Donohue III, J. J., and Wolfers, J. (2005) "Uses and abuses of empirical evidence in the death penalty debate", *Stanford Law Review*, 58, 791-846.

Donohue, J. J., and Wolfers, J. (2009) "Estimating the impact of the death penalty on murder", *American Law and Economics Review*, 11 (2), 249-309.

Durlauf, S. N., and Nagin, D. S. (2010) "The Deterrent Effect of Imprisonment. In Controlling Crime: Strategies and Tradeoffs", *University of Chicago Press*, 43-94.

Dustmann, C., Glitz, A., Schonberg, U., and Brucker, H. (2015), "Referral-based Job Search Networks," *forthcoming The Review of Economic Studies*.

Edin, P. A., Fredriksson, P., and Aslund, O. (2003), "Ethnic Enclaves and the Economic Success of Immigrants: Evidence from a Natural Experiment", *Quarterly Journal of Economics*, vol 118, s329-357

Ehrlich, I. (1975) "The Deterrent Effect of Capital Punishment: A Question of Life and Death", *The American Economic Review*, 65(3), 397-417.

Ekelund, R., Jackson, J., Ressler, R., and Tollison, R. (2006). "Marginal Deterrence and Multiple Murders." *Southern Economic Journal*, 72(3), 521-541. doi:10.2307/20111831.

Family Archive CD 354 (1998), "Passenger and Immigration Lists Index", *Gale Research*, Broderbund.

Fagan, J. (2006) "Death and Deterrence Redux: Science, Law and Causal Reasoning on Capital Punishment", *Ohio State Journal of Criminal Law*, 4, 255.

Feigenbaum, J. J. (2016), "Automated Census Record Linking: A Machine Learning Approach", *Working Paper*.

Feigenbaum, J. J. (2014), "Intergenerational Mobility during the Great Depression", *Working Paper.*

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 40, 1183-1210.

Ferenczi, L., and Wilcox, W. F. (1929). *International Migrations.* Vols. 1 and 2., New York, NBER.

Ferrie, J. (1996), "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29:4, 141-156.

Fitjar, R. D., and Rodriguez-Pose, A., (2016) "Nothing is in the Air" . *CEPR Discussion Paper* No. DP11067.

Frakes, M., and Harding, M. (2009) "The Deterrent Effect of Death Penalty Eligibility: Evidence from the Adoption of Child Murder Eligibility Factors", *American Law and Economics review*, 11(2), 451-497.

Friebel, G., Heinz, M., Krueger, M., and Zubanov, N. (2017), "Team Incentives and Performance: Evidence from a Retail Chain", *American Economic Review*, 107(8), 2168-2203.

Gant, J., Ichniowski, C., and Shaw, K. (2002), "Social Capital and Organizational Change in High? Involvement and Traditional Work Organizations", *Journal of Economics and Management Strategy*, 11(2), 289-328.

Genda, Y., Kondo, A., and Ohta, S.(2010), "Long-Term Effects of a Recession at Labor Market Entry in Japan and the United States," *Journal of Human Resources*, Vol. 45, No. 1, pp. 157-196.

Goel, D., and Lang, K. (2016), "Social Ties and the Job Search of Recent Immigrants," *IZA Discussion Papers* 9942, Institute for the Study of Labor (IZA).

Gibbs, J. (1989). "Conceptualization of Terrorism." *American Sociological Review*, 54(3), 329-340.

Gibbons, R., and Roberts, J. (Eds.) (2013), *The Handbook of Organisational Economics.* Princeton University Press.

Giulietti, C., Wahba, J., and Zenou, Y. (2014), "Strong versus Weak Ties in Migration," *IZA Discussion Papers* No. 8089.

Glaeser, E. L. (1999), "Learning in cities," *Journal of Urban Economics* 46, 254-277

Glitz, A., (2017), "Coworker networks in the labour market", *Labour Economics*, 44, issue C, p. 218-230.

Goeken, R., Huynh, L., Lenius, T., and Vick, R. (2011), "New Methods of Census Record Linking" *Historical Methods* 44:7-14.

Goldin, C. (1994), "The Political Economy of Immigration Restriction in the United States, 1890 to 1921," The Regulated Economy: A Historical Approach to Political Economy, C.Goldin and G.D.Libecap (eds.), *University of Chicago Press.*

Gottfredson, M. R., and Hirschi, T. (1990). *A general theory of crime.* Stanford University Press.

Goyal, S., (2015), "Networks in Economics: A Perspective on the Literature", *Cambridge Working Papers in Economics*, Faculty of Economics, University of Cambridge.

Granovetter, M. S. (1973), "The strength of weak ties," *American Journal of Sociology*, 78, 1360-1380.

Granovetter, M. S. (1983), "The strength of weak ties: A network theory revisited," *Sociological Theory*, 1, 201-233.

Grogger, J. (1990) "The Deterrent Effect of Capital Punishment: an Analysis of Daily Homicide Counts", *Journal of the American Statistical Association*, 85(410), 295-303.

Grossman, L. (2009). "Frank McCourt, 'Angela's Ashes' Author, Dies". *TIME*, 19 July 2009.

Guetzkow, H., and Simon, H. A. (1955), "The Impact of Certain Nets upon Organization and Performance in Task-Oriented Groups", *Management Science*, 1, 233-250.

Guryan, J., Kroft, K., and Notowidigdo, M. J. (2009), "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," *American Economic Journal: Applied Economics*, American Economic Association, vol. 1(4), pages 34-68, October.

Gusfield, D. (1997), "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology." *Cambridge University Press.*

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2012), "Diversity and productivity in production teams.", In *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*, Emerald Group Publishing Limited.

Hatton, T. J., and Williamson, J. G. (1998), *The Age of Mass Migration: Causes and Economic Impact*, New York, Oxford University Press.

Hayek, F. A. (1945), "The Use of Knowledge in Society", *The American Economic Review*, 35(4), 519-530.

Hayes, R. M., Oyer, P., and Schaefer, S. (2006), "Coworker complementarity and the stability of top-management teams", *Journal of Law, Economics, and Organization*, 22(1), 184-212.

Hjalmarsson, R. (2009) "Crime and expected punishment: Changes in perceptions at the age of criminal majority", *American law and economics review*, 11(1), 209-248.

Hjalmarsson, R. (2012) "Can Executions Have a Short?Term Deterrence Effect on Non?Felony Homicides?", *Criminology and Public Policy*, 11(3), 565-571.

Hjort, J. (2014), "Ethnic divisions and production in firms", *The Quarterly Journal of Economics*, 129(4), 1899-1946.

HMIC (2012), "Policing in austerity: One year on." *Manuscript.*

Ho, T. K. (1995), "Random Decision Forests". *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14-16 August 1995. pp. 278-282

Hong, M., and Kleck, G. (2018), "The Short-Term Deterrent Effect of Executions: An Analysis of Daily Homicide Counts." *Crime and Delinquency*, 64(7), 939-970

Hopkins, A. A. (1910), *The Scientific American Handbook of Travel*, Munn and Co., New York, NY.

Hutchinson, E.P. (1981), *Legislative History of American Immigration Policy*, 1798-1965, Philadelphia: University of Pennsylvania Press, 410.

Imbens, G. W., and J. M. Wooldridge, (2009), "Recent developments in the econometrics of program evaluation", *Journal of Economic Literature*, 47(1), 5-86.

Ioannides, Y. (2012), *From neighborhoods to nations: the economics of social interaction.* Princeton, N.J.: Princeton University Press.

Ioannides, Y. M., and Loury, L. D.(2004), "Job Information Networks, Neighborhood Effects and Inequality," *Journal of Economic Literature*, 42 (4), 1056.1093.

Jackson, M. (2014) "Networks in the Understanding of Economic Behaviors." *Journal of Economic Perspectives*, 28 (4): 3-22.

Jackson, M. (2011) "An Overview of Social Networks and Economic Applications" in *The Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M.O. , North Holland Press 2011.

Jackson, M. (2009) "Networks and Economic Behavior", *Annual Review of Economics* 1:1, 489-511.

Jackson, M.., Rogers, B., and Zenou, Y. (2016) "Networks: An Economic Perspective". in *Oxford Handbook of Social Network Analysis*, R. Light and J. Moody (Eds.), Oxford, Oxford University Press, Forthcoming.

Jacobs, D., and Carmichael, J. T. (2001). "The politics of punishment across time and space: A pooled time-series analysis of imprisonment rates." *Social Forces*, 80(1), 61-89.

Jacobs, D., and Carmichael, J. T. (2002). "The political sociology of the death penalty: A pooled time-series analysis." *American Sociological Review*, 109-131.

Jacobs, J. (1969) *The Economy of Cities.* New York: Random House.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations", *The Quarterly Journal of Economics*, 108(3), 577-598.

Jaravel, X., Petkova, N., and Bell, A. (2016), "Team-specific capital and innovation", *Working paper.*

Katz, L., Levitt, S. D., and Shustorovich, E. (2003) "Prison conditions, capital punishment, and deterrence", *American Law and Economics Review*, 5(2), 318-343.

Kovandzic, T. V., Vieraitis, L. M., and Boots, D. P. (2009) "Does the Death Penalty Save Lives?", *Criminology and Public Policy*, 8(4), 803-843.

Kramarz, F., and Skans, O. N. (2014) "When Strong Ties are Strong: Networks and Youth Labour Market Entry", *The Review of Economic Studies*, Volume 81, Issue 3, 1 July 2014, Pages 1164-1200.

Kugler, A. (2003), "Employee Referrals and Efficiency Wages," *Labour Economics* 10, 531-556.

Laibson, D. (2001), "A Cue-Theory of Consumption." *The Quarterly Journal of Economics*, Volume 116, Issue 1, 1 February 2001, Pages 81-119.

Lafortune, J., and Tessada, J., (2012), "Smooth(er) Landing? The Dynamic Role of Networks in the Location and Occupational Choice of Immigrants", *Working Papers ClioLab*, No 14, EH Clio Lab. Instituto de Economa. Pontificia Universidad Catlica de Chile.

Land, K. C., Teske, R. H., and Zheng, H. (2009) "The Short-Term Effects of Executions on Homicides: Deterrence, Displacement, or Both?", *Criminology*, 47(4), 1009-1043.

Land, K. C., Teske, R. H., and Zheng, H. (2012) "The Differential Short-Term Impacts of Executions on Felony and Non-Felony Homicides", *Criminology and Public Policy*, 11(3), 541-563.

Laxton, E. (1996) *The famine ships: The Irish exodus to America, 1846-51.* London: Bloomsbury.

Leavitt, H. J. (1951), "Some Effects of Certain Communication Patterns on Groups Performance", *Journal of Abnormal Psychology*, 46, 38-50.

Levenshtein, V. I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady.* 10 (8): 707-710.

Lindquist, M. J., Sauermann, J., and Zenou, Y. (2017), "Peer Effects in the Workplace: A Network Approach", *Working paper.*

Liu, Z. (2004) "Capital Punishment and the Deterrence Hypothesis: Some New Insights and Empirical Evidence", *Eastern Economic Journal*, 30(2), 237-258.

Loewenstein, G. (2000), "Emotions in Economic Theory and Economic Behavior." *American Economic Review*, 90 (2): 426-432.

Loewenstein, G. (1996), "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes*, Volume 65, Issue 3, 1996, Pages 272-292, ISSN 0749-5978.

Lynch, M. P., and Winkler, W. E. (1994), "Improved String Comparator," Technical Report, Statistical Research Division, Washington, DC: *U.S. Bureau of the Census.*

Lyons, E. (2016), "Team production in international labor markets: Experimental evidence from the field", *American Economic Journal: Applied Economics*, 9 (3): 70-104.

Marmaros, D., and Sacerdote, B. (2002), "Peer and Social Networks in Job Search," *European Economic Review* 46, 870-879.

Mas, A., and Moretti, E. (2009), "Peers at work", *The American Economic Review*, 99(1), 112-145.

Massey, D., Condran, G. A., Denton, N. A. (1987), "The Effect of Residential Segregation on Black Social and Economic Well-Being,", *Social Forces*, Volume 66, Issue 1, 1 September 1987, Pages 29-56.

Matza, D. (1964). *Delinquency and Drift.* New York, John Wiley & Sons, Inc., 1964.

Maurer, S. E., and Potlogea, A. V. (2017), "Male-biased Demand Shocks and Womens Labor Force Participation: Evidence from Large Oil Field Discoveries," *Working Paper Series of the Department of Economics, University of Konstanz* 2017-08, Department of Economics, University of Konstanz.

McKenzie, D., and H. Rapoport (2007), "Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico," *Journal of Development Economics* 84, 1-24.

McKenzie, D., and H. Rapoport (2010), "Self-selection patterns in MexicoUS migration: The role of migration networks," *Review of Economics and Statistics* 92, 811-821.

Mocan, H. N., and Gittings, R. K. (2003) "Getting off death row: Commuted sentences and the deterrent effect of capital punishment", *Journal of Law and Economics*, 46, 453.

Mocan, N. H., and Gittings, R. K. (2006) "The impact of incentives on human behavior: can we make it disappear? The case of the death penalty", *NBER Working Paper* 12631, National Bureau of Economic Research.

Montgomery, (1991), "Social Networks and Labor-Market Outcomes - Toward an Economic Analysis" *American Economic Review*, vol 81, No 5, 1408-1418.

Mullainathan, S. (2002), "A Memory-Based Model of Bounded Rationality", *The Quarterly Journal of Economics*, 117, issue 3, p. 735-774.

Munshi, K. (2014) "Community Networks and Migration." *Journal of Economic Perspectives*, Volume 28, Number 4 - Fall 2014: 49-76

Munshi, K. (2003), "Networks in the Modern Economy: Mexican Migrants in the US Labor Market", *Quarterly Journal of Economics*, 549-599

Nagin, D. S. (2013) "Deterrence: A Review of the Evidence by a Criminologist for Economists", *Annual Review of Economics*, 5(1), 83-105.

Nam, C. B., and Boyd, M. (2004), "Occupational Status in 2000: Over a Century of Census-based Measurement," *Population Research and Policy Review* 23, 2004: 327-358.

National Research Council. (2012) "Deterrence and the Death Penalty", *National Academies Press.*

Oreopoulos, P., von Wachter, T., and Heisz A. (2006), "The Short- and Long-Term Career Effects of Graduating in a Recession: Hysteresis and Heterogeneity in the Market for College Graduates," *NBER Working Paper*, No. 12159.

Oyer, P. (2006), "Initial Labor Market Conditions and Long-Term Outcomes for Economists." *Journal of Economic Perspectives*, 20 (3): 143-160.

Palacios-Huerta, I., and Prat, A. (2012), "Measuring the Impact Factor of Agents within an Organization Using Communication Patterns", *Working paper.*

Patel, K., and Vella, F. (2013), "Immigrant Networks and their implications for Occupational Choice and Wages," *The Review of Economics and Statistics*, 95(4).

Pilling, K. (2010), "Police to publish call response times on website", available at: http://www.independent.co.uk/news/uk/crime/police-to-publish-call-response-times-on-website-1949967.html

Phillips, D. P. (1980) "The Deterrent Effect of Capital Punishment: New Evidence on an Old Controversy", *American Journal of Sociology*, 86(1), 139-148.

Phillips, D. P. (1982) "The Fluctuation of Homicides After Publicized Executions: Reply to Kobbervig, Inverarity, and Lauderdale." *American Journal of Sociology*, Vol. 88, No. 1 (Jul., 1982), pp. 165-167.

Reisler, M. (1976), *By the Sweat of Their Brow: Mexican Immigrant Labor in the United States, 1900-1940*, Westport, CT: Greenwood Press, 1976, 56.

Rubin, P. H. (2009) "Don't Scrap the Death Penalty", *Criminology and Public Policy*, 8(4), 853-859.

Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M. (2015) "Integrated Public Use Microdata Series: Version 6.0 [dataset]". Minneapolis: University of Minnesota.

Sacerdote, B. (2001) "Peer Effects With Random Assignment: Results For Dartmouth Roommates", *Quarterly Journal of Economics*, 2001, v116(2,May), 681-704

Sato, Y., and Zenou, Y. (2015), "How Urbanization Affect Employment and Social Interactions" *European Economic Review* 75:131-155.

Schulz, K. U., and Mihov, S. (2002), "Fast String Correction with Levenshtein-Automata". *International Journal of Document Analysis and Recognition.* 5 (1): 67-85.

Scruggs, O. M. (1988) *Braceros, Wetbacks, and the Farm Labor Problem: Mexican Agricultural Labor in the United States, 1942-1954.* New York: Garland Publishing.

Sedgewick, R., and Wayne, K. (2001) *Algorithms*, 4th Edition, Addison-Wesley, 2011.

Shepherd, J. (2005) "Deterrence versus Brutalization: Capital Punishment's Differing Impacts among States", 104, *Michigan Law Review* 203.

Shore, L. M., Chung-Herrera, B. G., Dean, M. A., Ehrhart, K. H., Jung, D. I., Randel, A. E., and Singh, G. (2009), "Diversity in organizations: Where are we now and where are we going?", *Human Resource Management Review*, 19(2), 117-133.

Simon, H. A. (1957), *Administrative behavior; a study of decision-making processes in administrative organization.* Oxford, England: Macmillan.

Simpson, H. (2009), "Productivity in public services.", *Journal of Economic Surveys*, 23(2), 250-276.

Sojourner, A. (2013), "Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR" *Economic Journal.* 123(569): 574-605

Sorensen, J., Wrinkle, R., Brewer, V., and Marquart, J. (1999) "Capital Punishment and Deterrence: Examining the Effect of Executions on Murder in Texas", *Crime and Delinquency*, 45(4), 481-493.

Stolzenberg, L., and D'Alessio, S. J. (2004) "Capital Punishment, Execution Publicity and Murder in Houston, Texas", *The Journal of Criminal Law and Criminology (1973-)*, 94(2), 351-380.

Storper, M., and Venables, A. J. (2004), "Buzz: face-to-face contact and the urban economy", *Journal of Economic Geography*, 4(4), 351-370.

Taylor, N. F. (2010), *Grandma, Tell Us a Story.*, Xlibris Corporation LLC, 9781450071819

Thaler, R. H., and Shefrin, H. M. (1981). "An Economic Theory of Self-Control," *Journal of Political Economy*, University of Chicago Press, vol. 89(2), pages 392-406, April.

Thompson, P., and Fox-Kean, M. (2005), "Patent citations and the geography of knowledge spillovers: A reassessment", *American Economic Review*, 450-460.

Topa, G. (2001), "Social Interactions, Local Spillovers and Unemployment," *The Review of Economic Studies*, 68, 261.295.

Topa, G.(2011) "Labor Markets and Referrals", in the *The Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M.O. , North Holland Press, 2011

Topel, R. (1991), "Specific capital, mobility, and wages: Wages rise with job seniority", *Journal of Political Economy*, 99(1), 145-176.

US, Bureau of the Census (1975) *Historical Statistics of the United States, Colonial Times to 1970.*, U.S. Bureau of the Census, Pt.1

Waber, B., Magnolfi, J., and Lindsay, G. (2014) "Workspaces That Move People" , *Harvard Business Review*, October 2014 Issue (https://hbr.org /2014/10/workspaces-that-move-people).

Wagner, R. A., and Fischer, M. J. (1974), "The string-to-string correction problem. *Journal of the Association for Computing Machinery*", 21, 168-173.

Watcher, T. v., and Bender, S. (2008) "Do Initial Conditions Persist Between Firms? An Analysis of Firm-Entry Cohort Effects and Job Losers using Matched Employer-Employee Data" in S. Bender, J. Lane, K. Shaw, F. Andersson, and T. von Wachter (eds), *The Analysis of Firms and Employees: Quantitative and Qualitative Approaches*, (University Chicago Press)

Weber, R. A., and Camerer, C. F. (2003), "Cultural Conflict and Merger Failure: An Experimental Approach", *Management Science*, 49, 400-415.

Wegge, S. A. (1998), "Chain Migration and Information Networks: Evidence from Nineteenth-Century Hesse-Cassel" *Journal of Economic History* 58 (4): 957-86

Weintraub, J., and Point, D. (2017) "The Ellis Island Name Change Myth", JewishGen, *Manuscript.*

Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* American Statistical Association: 354-359.

Yakubovich, V. (2005), "Weak ties, information, and influence: How workers find jobs in a local Russian labor market," *American Sociological Review*, 70, 408-421.

Zeisel, H., and Phillips, D. P. (1982) A Comment on "The Deterrent Effect of Capital Punishment"by Phillips. *American Journal of Sociology*, 88, no. 1 (Jul., 1982): 167-169.

Zimmerman, P. R. (2003) "State executions, deterrence, and the incidence of murder", *Journal of Applied Economic*, 7, 163-193.

Zimmerman, P. R. (2006) "Estimates of the deterrent effect of alternative execution methods in the United States: 1978-2000", *American Journal of Economics and Sociology*, 65(4), 909-941.

Zimmerman, P. R. (2009) "Statistical variability and the deterrent effect of the death penalty", *American Law and Economics Review*, 11 (2), 370-398.

Zimring, F. E., Fagan, J., and Johnson, D. T. (2010) "Executions, deterrence, and homicide: a tale of two cities", *Journal of Empirical Legal Studies*, 7(1), 1-29.