Essays in Development Economics and Economic History

by

Anna Baiardi

Thesis submitted to the University of Warwick

in partial fulfillment for the degree of

Doctor of Philosophy

in Economics

Department of Economics

March 2017

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Firstly, I would like to thank my supervisors Chris Woodruff and Rocco Macchiavello for their invaluable guidance and for helping me find my way through the Ph.D. Moreover, I am grateful to Sascha Becker, who provided crucial help for the second chapter. I would also like to thank Clément de Chaisemartin, Mirko Draca, Bishnu Gupta, Clément Imbert, Luigi Pascali, Roland Rathelot, Debraj Ray and Fabian Waldinger. A special thanks goes to Fabrice Defever for kindly sharing part of the data used in Chapter 3. I acknowledge the financial support of the ESRC.

I am grateful to have found so many great friends at Warwick: Giorgos Galanis, Andrea Naghi, Mahnaz Nazneen, Guillem Ordonez, Andis Sofianos, Raghul Venkatesh and Boromeus Wanengkirtyo; and for the support of my old friends, Giulia and Sara, despite the distance.

This could have not been possible without my friend and co-author Christina Ammon, with whom I have shared all the good and bad times of this journey, and without Spyros Terovitis, who has provided the encouragement I needed by always believing in me.

Lastly, I would like to thank my family: my parents, my sister and my grandmother, not only for their support but for having given me the opportunity to reach this important goal.

Anna Baiardi

March 2017

Declarations

The second chapter, *The Persistent Effect of Gender Division of Labour: African American Women After Slavery* is solely my own work.

The third chapter, *The Benefits of the Bamboo Network in International Trade*, is co-authored with Christina Ammon. We contributed in equal parts to the data preparation and development of the empirical strategy. The results and the interpretation were developed from joint discussion.

The fourth chapter, *Ethnic Chinese Networks and Technology Diffusion: The Chinese Exclusion Act*, is solely my own work.

I declare that the material contained in this thesis has not been used or published before, and has not been submitted for another degree or at another university.

Anna Baiardi

March 2017

Abstract

The first chapter provides an overview of the topics covered in this thesis.

The second chapter explores the effect of historic gender division of labour during slavery on African American women's performance in the labour market. Using census data from 1870 to 2010, I show that African American women living in areas with lower levels of gender division of labour were more likely to participate in the labour market and have higher occupation income scores after emancipation. The effects are persistent for at least 70 years after the end of slavery. I analyse the mechanisms driving the results, distinguishing between labour supply and demand channels, and I explore intergenerational transmission of gender roles.

The third chapter empirically assesses the importance of ethnic networks in facilitating international trade. In particular, it investigates the impact of ethnic Cantonese networks in the United States on the export performance of firms based in Southern China. The results indicate that exposure to ethnic networks has a positive effect on exports, both at the extensive and the intensive margin. We explore the mechanisms underlying the results, distinguishing between information flows, contract enforcement, foreign investment and technology diffusion.

The fourt chapter analyses the effect of ethnic Chinese networks in the United States on knowledge diffusion and innovation in China. I construct a proxy for the ethnic network based on historic Chinese settlements and current industry employment patterns, exploiting the migration restrictions imposed by the Chinese Exclusion Act of 1882. The results indicate that when innovation in the U.S. increases, industries that are more exposed to the ethnic network in the U.S. innovate more in China. This suggests that ethnic networks contribute to the diffusion of technology across countries.

Abbreviations

ATE	Average Treatment Effect
FDI	Foreign Direct Investment
GAEZ	Global Agro-Ecological Zones
IPC	International Patent Classification
IPUMS	Integrated Public Use Microdata Series
ISIC	International Standard Industrial Classification of All Economic Activities
IT	Information Technology
IV	Instrumental Variable
LATE	Local Average Treatment Effect
LFP	Labour Force Participation
NAICS	North American Industry Classification System
NAPP	North Atlantic Population Project
NBER	National Bureau of Economic Research
NHGIS	National Historical Geographic Information System
OECD	Organisation for Economic Co-operation and Development
OIS	Occupation Income Score
OLS	Ordinary Least Squares
R&D	Research and Development

RMB	Renminbi
SD	Standard Deviation
TFP	Total Factor Productivity
TFPR	Total Factor Productivity Revenue
SIC	Standard Industrial Classification
U.K.	United Kingdom
U.S.	United States
2SLS	Two-Stage Least Squares

Chapter 1

Introduction

A growing literature focuses on uncovering the persistent effect of historical factors on current economic outcomes. Since the work of Acemoglu et al. (2001), a number of papers have analysed the long run effects of institutions on economic development (Nunn, 2008; Nunn and Wantchekon, 2011; Dell, 2010). Numerous studies focused instead of the importance of sharing similar culture and history in the context of a number of economic activities: international trade (Rauch and Trindade, 2002), knowledge flows (Griffith et al., 2006; Kerr, 2008), labour markets Munshi (2003), firm growth (Banerjee and Munshi, 2004; Woodruff and Zenteno, 2007).

This thesis relates to both literatures, by investigating the long term economic consequences of slavery practices in regards to gender roles, and the effect of ethnic and kinship ties on economic transactions.

In Chapter 2 I analyse the long run effect of gender division of labour during slavery on female labour market choices and perfomance of African Americans. Goldin (1977, 1990, 1991) has studied extensively the differences between black and white female labour force participation and suggested that these may be the result of divergent cultural beliefs of the two ethnic groups regarding gender roles. Goldin proposes the idea that gender roles in the African American community might have been shaped by slavery: forced labour for both genders could mean that African American women did not experience the stigma associated with white female labour until the first half of the 20th century. I carry forward this idea by suggesting that differences in gender roles during slavery, rather than forced labour, may explain differences in African American women's labour market outcomes.

Kinship and ethnic networks have been the focus of numerous papers in the trade and economic development literature. They have been shown to ease economic transactions by helping overcome market frictions, particularly informational barriers and weak institutions such as contract enforcement (Greif, 1993; Rauch, 2001). Chinese diaspora networks are an example of strong cultural and ethnic ties which can affect economic outcomes Weidenbaum and Hughes (1996). Chapter 3 examines the importance of these ties in relation to international trade, by analysing the ethnic network of ethnic Cantonese in the United States and its effect on the export behaviour of Cantonese firms. The peculiarity of this network is that it was formed through a mass migration wave during the 18th century.

Chapter 4 analyses the more broad ethnic network of early ethnic Chinese in the U.S., focusing on its impact on knowledge diffusion and innovation. Saxenian (2002) has documented the extent of information sharing of recent migrants with their country of origin. Instead, I focus on a network of early migrants in order to shed light on the extent of technology spread from a frontier country to a less developed country.

Chapter 2

The Persistent Effect of Gender Division of Labour: African American Women After Slavery

2.1 Introduction

A number of recent studies have aimed at understanding the orgins and consequences of gender division of labour, by exploring their historical roots (Alesina et al., 2013), their effect in the labour markets and their intergenerational transmission mechanisms (Farré and Vella 2013; Kleven and Landais 2016). In the context of the United States, Goldin (1977) suggests that forced labour of both women and men during slavery may have shaped the gender roles of African Americans, resulting in high female labour force participation rates after emancipation compared to white women. This mechanism is further investigated by Boustan and Collins (2014), who find a positive link between black mothers' labour market status and the likelihood of their female children to be in the labour market.

In this paper I link the literature on the persistence of gender roles with that on on the long term effects of slavery in the U.S., by exploring whether gender roles adopted during slavery can explain differences in labour market outcomes of African American women after the end of this institution. I test this hypothesis by exploiting geographic variation in the prevalence of two crops, the cultivation of which involved different practices regarding gender division of labour during slavery. Cotton cultivation in the South of the U.S. was characterised by large plantations where female and male slaves worked side by side; therefore female slaves performing domestic tasks were less common. This is opposed to tobacco plantations, where the decline in production of the crop during the 19th century reduced the need for female labour in agriculture, which resulted in a sharper gender division of labour. I exploit variation in cotton prevalence relative to tobacco during slavery in order to explain differences in labour market outcomes of African American women between 1870 - five years after emancipation - and 2010. I use U.S. census microdata, which allow me to control for individual characteristics and county-specific effects.

I find that in areas where cotton was more prevalent relative to tobacco African American women were more likely to participate in the labour market and to have higher occupation income scores; this implies that, conditional on labour force participation, women in cotton regions were more likely to be employed in occupations paying higher wages. The effects are economically significant: in 1880, a one standard deviation increase in production of cotton relative to tobacco increases labour force participation of women relative to men by 8 percentage points, and occupation income scores by about 15 percent. The effects decline in magnitude over time but persist for 90 and 70 years after Emancipation respectively. I run robustness checks with different measures of gender division of labour, including instrumental variable regressions where I instrument crop production by county with crop suitability.

Furthermore, I investigate the mechanisms behind these results, distinguishing between the effect of gender roles and the possible effect of cotton prevalent regions having better labour market conditions, such as higher labour demand for agriculture.

To disentangle labour demand from labour supply channels, I analyse the labour market outcomes of migrants from cotton and tobacco regions. Because individual level data for county of origin is not available for the first decades of the 20th century, I construct a data set of migrants by making use of the fact that a significant proportion of African Americans after emancipation adopted the surname of the plantation owners they worked for. I infer the origin of African American women who have migrated after slavery by matching their mother's surname with the surnames of slaveholders in cotton or tobacco regions;[1] I focus on mother's surname rather than individuals' surname to limit the extent of measurement error arising from women adopting their husband's surname after marriage. Using data from 1930, I find that migrants from cotton counties are more likely to participate in the labour force compared to other mi-

[1] A recent paper by Ochsner and Roesel (2016) uses a similar empirical strategy, which relies on surnames in order to relate current political preferences for right-wing parties to past migration patterns.

grants, and more likely to have better occupations, conditional on labour force participation. These results are confirmed using data on migration from the 1940 census. I investigate intergenerational transmission of gender roles, by exploring heterogeneous effects for mother labour force status. I find that the latter has a larger effect on women's occupation income score in areas with higher relative cotton prevalence, suggesting that the higher labour market experience of mothers in cotton areas positively affects their daughter's probability of finding a better job.

I then focus on demand mechanisms. My results indicate that a larger proportion of African American women are employed in agriculture in regions where cotton prevalence is higher. However, labour market outcomes of other ethnic groups do not appear to exhibit the same patterns. Moreover, conditional on working in service and manufacturing, women in cotton regions have better occupations compared to women in other areas. While these results do not rule out demand effects, they are suggestive of other mechanisms being in place.

I investigate alternative channels which may explain the results: higher human capital, less discrimination and better access to social networks. Whereas I find no positive effect on wages or education levels, I show evidence pointing at lower discrimination towards African American women in cotton regions and of better access to social networks. Overall, my findings indicate that gender roles during slavery had long lasting effects on labour market outcomes of black women.

The paper is organized as follows. Section 2.2 introduces the related literature, while section 2.3 provides a summary of the historical background, describing the role of women in cotton and tobacco plantations during slavery. In section 2.4 I give a detailed description of the data used for the analysis, and section 2.5 outlines the empirical strategy. Section 2.6 focuses on the main results, whereas in section 2.7 I investigate the mechanisms. Finally, section 2.8 concludes.

2.2 Related Literature

This work relates to several strands of literature. Among these, the closest explores the long run effect of slavery on female labour. Starting with Goldin (1977), the economics literature has recognized the long term consequences of slavery on labour market outcomes of African Americans, in particular women. By analysing historical census data, Goldin (1977) was the first to document that forced labour during slavery could be one of the causes underlying the differences in labour force participation of African American and Caucasian women in the U.S. Up until 1980s, the percentage of black women in the labour force was consistently higher than that of white women (Boustan and Collins, 2014), and the difference survives even after controlling for income Goldin (1977, 1991). This suggests that poverty is not the only cause of higher female labour force participation rates in the African American community, but other factors such as preferences may play a role. Goldin (1977) outlines the hypothesis that the "stigma" associated with being employed outside the household that is prevalent within white communities in the U.S. up to the mid of the 20th century may not affect African Americans. She relates these differences to slavery: forced labour caused women and men to work side by side, therefore hindering the formation of cultural norms related to gender roles. This could have resulted in black women being more likely to work than white women, at the same level of income. Boustan and Collins (2014) test this idea with U.S. historic census data and find that black women whose mother is in the labour force are more likely to participate in the labour force, and even more so if the mother was born in a Southern State. This paper contributes to the literature by exploring the hypothesis that different practices regarding gender division of labour during slavery may explain differences in labour market outcomes of women within the African American community, analysing both their labour market status and their occupations.

This paper relates to the work of Bailey and Collins (2006) who document wage gains of the female African American population in the 1940s. Ager et al. (2016) also link slavery to labour force participation of African American women, by studying the negative effect of an income shock - the Boll Weevil Plague, which affected cotton plantations in the early 20th century - on female labour supply. They find that African American women were particularly affected by this pest, and their results indicate that cotton was a very important source of income for black women. The study of Chay and Munshi (2013) is relevant to this work as it suggests a positive relation between historical labour intensive plantations and the formation of networks in the post-slavery U.S.; their evidence demonstrates that networks played a role on migration patterns and voting behaviour of black communities after Emancipation.

Another strand of literature to which this paper contributes focuses on the transmission of gender roles over time, among which Alesina et al. (2013). They find that the use of the plough transformed agriculture into a mostly male dominated activity, and this had persistent implications on the shaping gender roles across countries. This study differs from their paper by providing evidence of the effects of gender division of labour arising from slavery, which caused an arguably exogenous change in gender division of labour for African Americans. Farré and Vella (2013) study the intergenerational transmission of gender roles from mothers to their children. Kleven and Landais (2016), using data from Denmark, find evidence of gender inequality being primarily driven by the dynamics of childbearing. Fernández and Wong (2014) instead explain changes in labour force participation over time in the U.S. with the shift from bilateral to unilateral divorce laws and the decrease in the gender gap.

Finally, this work relates to the more broad literature on the long run effects of slavery, both in origin and destination countries. Nunn (2008) and Nunn and Wantchekon (2011) find persistent effects of slavery on economic development and trust in African countries, whereas Bertocchi (2015) studies the persistent effects of slavery both in African countries (origin) and in North and Latin America (destination). More focused on the U.S. are the papers by Hornbeck and Naidu (2014), who analyse the consequences of the Great Mississipi flood of 1927 on agricultural mechanization in the South, and the work by Christian (2014), which explores the long run effects of lynching on labour market outcomes of African Americans.

2.3 Historical Background

2.3.1 Cotton and Tobacco Plantations During Slavery

Cotton and tobacco have been cultivated in the U.S. since the beginning of the 17th century and they were among the major crops cultivated in the U.S. during the slavery period.

Tobacco started in Virginia, in the Chesapeake bay and spread to North Carolina, Kentucky, Pennsylvania and Tennessee, whereas cotton was widespread mainly in the Southern regions. Figure 2.1 shows the geographic distribution of cotton and tobacco production by county: cotton cultivation was more prevalent in the South, whereas tobacco was mainly concentrated in the East. The cultivation of the two crops started as small farms, but later developed into large plantations, and slavery played a major role in this transformation (Walsh, 1989). Tobacco was the main crop produced in the U.S. until the 18th century, when prices dropped due to overproduction, and export demand for the crop was greatly reduced. This caused either a shift towards more cotton production (mostly in Southern regions, more suitable for this crop), or diversification in the Chesapeake; the latter implied an increased production of subsistence crops in addition to tobacco (Walsh, 1989). This shift forced tobacco planters to change their production processes: as tobacco's importance as a cash crop decreased, women were less needed in the fields. This resulted in males mostly working in the fields and women performing domestic work and other type of tasks. Moreover, as described by Walsh (1989), the change in production was characterized by the introduction of the plough, which may have contributed to the division of labour by gender.

The role of women in cotton cultivation during slavery has been described by several authors in the literature, in the fields of both economics and history. Metzer (1975), Shlomowitz (1979) and Jones (2009) suggest that women have a comparative advantage in picking cotton, the most labour intensive activity in cotton plantations, and Wayne (2007) describes the work in cotton fields, stating that women, men and children worked side by side. Goldin and Sokoloff (1984) suggest that women have a comparative advantage in crops that required extensive cultivation, such as cotton and tobacco. Whereas in principle cotton and tobacco cultivation are indeed characterized by extensive plantations, as described above some differences arose across the two crops in the way the plantations were organized, which lead to tobacco cultivation effectively promoting starker differences in the types of tasks performed by men and women.

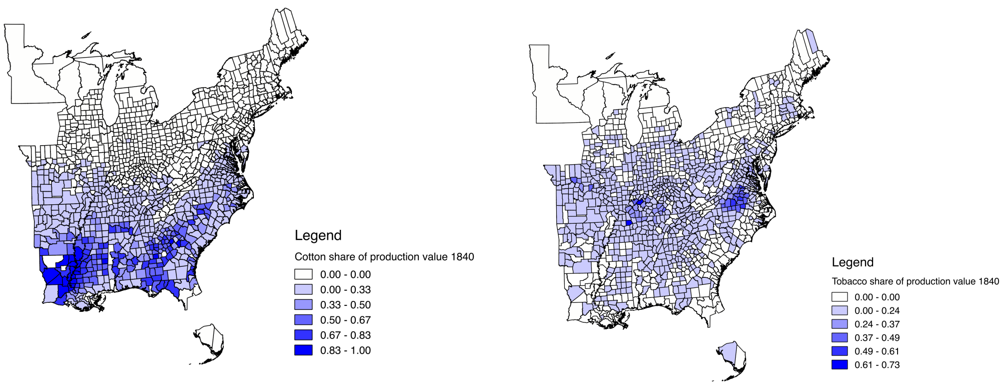


Figure 2.1. Cotton and Tobacco Shares of Total Production Value 1840

Notes: The two maps show the shares of cotton (left) and tobacco (right) production value to total production value of agricultural products by county in 1840, measured in U.S. dollars. Source: U.S. Agricultural Census 1840

2.3.2 After Slavery

Slavery ended with the Emancipation Proclamation in 1863. While tobacco cultivation had already decreased dramatically during the 19th century, cotton cultivation declined after Emancipation; the drop in cotton production, which was caused by declining prices alongside the spreading of the Boll Weevil, is regarded by some as one of the main causes of the Great Migration (Higgs, 1976; Lange et al., 2009). This term refers to the mass migration of African Americans from the agricultural South to the North, in particular to Northern cities, between 1910 and 1970. Migration of African Americans from the South to the North was instead very limited before the 1910s, partly because of lack of demand for labour for African Americans; also, as described by Ransom and Sutch (2001), it is believed to be a consequence of the low upward occupational mobility of Southern African Americans working in agriculture (Collins and Wanamaker, 2014).

Consequently to the end of slavery, African Americans, who typically did not possess formal surnames during slavery, suddenly had the need to adopt one for official documents. Although precise information about exact numbers is difficult to obtain, anecdotal evidence suggests that a significant percentage of African Americans decided to keep the surname of the slaveholder for whom they used to work. For instance Gutman et al. (1976), in his book "The Black Family in Slavery and Freedom, 1750-1925", which collects information about several families and descendants of ex-slaves, finds that 27% of the 181 individuals interviewed in South Carolina, and 36% of the 217 interviewed in Texas, claimed that their family retained the surnames of the last slaveholder. However, often those individuals who did not retain the surname of their last slaveholder decided to adopt the surname of a previous slaveholder.

2.4 Data

In this paper I use cross-sectional data from historical and recent U.S. Censuses. I merge these with measures of crop production during slavery times.
I use microdata from the U.S. population censuses from 1870 to 2000, and from the American Community Survey for year 2010. The samples used range from 1% to 100% of the total samples of each year, depending on availability. I use microdata from the 10% sample of the 1880 Population Census to perform the analysis on 1880. The advantage of this data set is that all minorities are oversampled: it contains 1 in 5 observations for African Americans (and other minorities) of those contained in the 100% Census. It includes individual-level information about age, sex, labour force status, detailed occupation and industry, occupation income score, marital status and family members. Occupation income scores are constructed by assigning a median earning to each occupation, measured in 1950 U.S. dollars. They have been used as a proxy for wages, since data about earnings is not available in censuses prior to 1940 (Biavaschi et al., 2013). I use the log of occupation income scores as one of the main dependent variables in my analysis.

Slaveholder surnames used to construct the data set of migrants are obtained from the 1860 sample of the slave schedules available from the Integrated Public Use Microdata Series (IPUMS). This is a 1-in-20 sample of the slaves enumerated in 1860, which includes the surname of slaveholders (the owners of the plantations where the slaves were registered), and limited information about the slaves themselves, including the total number of slaves in each holding.[2] All of the census data used in this paper are available from the IPUMS, North Atlantic Population Project (NAPP) or National Historical Geographic Information System (NHGIS) websites.

I obtain crop production and acreage variables at the county level from the Agriculture Census of years 1840 and 1880 respectively. To construct the instrument for crop production, I use detailed crop suitability data available from Global Agro-Ecological Zones (GAEZ), which I aggregate at the county level to match the individual data from the Population Censuses. Crop suitability indices determine how well soil conditions match crop requirements given specific levels of inputs and irrigation. The suitability indices used in this paper are calculated for low input levels and rain-fed production (no irrigation). This is to reduce to a minimum the possibility that the variation in suitability actually reflects differences in management or irrigation systems. Cotton and tobacco suitability indices range from 0 to 8, where higher numbers indicate higher suitability.

Table 2.1 shows descriptive statistics of the sample of African Americans in 1880, in counties where cotton prevalence is higher than median (left panel) and below median (right panel).[3] Notice that there are differences in individual characteristics across the two regions. A higher percentage of the population is female in regions where cotton prevalence is high, and the difference is over two percentage points. This may be a

[2] For a more detailed description of the slave schedules, see Ruggles et al. (2010).

[3] For this table, women and men are pooled together.

consequence of the fact that there was a higher share of female slaves in cotton regions[4]. Labour force participation is also higher in cotton regions, by almost 7 percentage points. However, notice that occupation income score is lower in counties with high relative cotton prevalence. There are very small differences in age, and share of married individuals; however, the share urban population is noticeably larger in counties with low cotton prevalence. The number of children in the household is higher in cotton regions, by 0.5 children.

	Relative Cotton Share>median			Relative Cotton Share <median				
	Mean	Max	Min	Mean	Max	Min	Mean Dif.	N
Female	.518	0	1	.496	0	1	.022***	1,608,221
Labour Force Participation	.741	0	1	.673	0	1	.068***	1,518,531
Occupation Income Score (100$)	13.767	3	80	14.942	3	80	-1.174***	1,094,914
Age	36.096	25	54	36.333	25	54	-.237***	1,608,221
Married	.782	0	1	.714	0	1	.068***	1,608,221
Urban	.116	0	1	.311	0	1	.195***	1,608,221
Literate	.835	0	1	.869	0	1	-.033***	1,608,213
N Children	2.588	0	9	2.068	0	9	.521***	1,608,221

Table 2.1. Descriptive Statistics - African Americans in 1880

Notes: I include individuals whose age is between 25 and 54, and split the sample into individuals who live in counties with *crop share* below and above median. The statistics on occupation income score are conditional on labour force participation.***p>0.01 **p>0.05 *p>0.10. Source: U.S. 1880 full Population Census.

Figure A.3.1 shows the top ten occupations by gender among the African American population. The vast majority - about 80% - of males are employed in agriculture, which is also the first industry by employment of women. However, a much larger percentage of females are employed in private households, as domestic help, compared to males: 35 versus 4 percent. The remaining share of employed individuals - about 15% of men and 6% of women - is employed in manufacturing, construction and other services.

[4]Data from the Slave Census of 1860.

2.2 shows the evolution of the share of African Americans by gender employed in agriculture over time, with data from the decennial U.S. censuses. Notice that whereas agriculture was one of the main activities at the beginning of the 20th century, its importance reduces dramatically starting from 1930, and the decrease is relatively larger for women than it is for men.

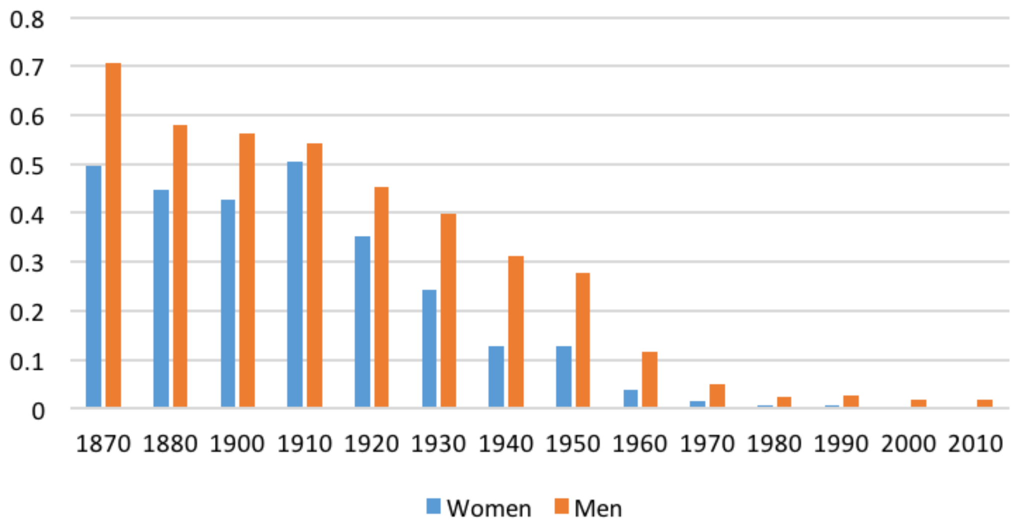


Figure 2.2. Share of African Americans Employed in Agriculture by Gender

Notes: This graph shows the share of African Americans by gender in the labour force who work in agriculture over time. Source: U.S. Census.

2.5 Empirical Strategy

2.5.1 Measuring Gender Division of Labour

In this subsection I outline the measures of gender division of labour used in this paper.

As I am interested in capturing variation in gender roles during slavery, the ideal measure would be the number of female slaves working in cotton and tobacco plantations across geographical regions in the U.S. To the best of my knowledge, this measure is not available; therefore, I construct a proxy using production data for cotton and tobacco prevalence during slavery. The earliest county-level data is available for year 1840; it reports the total vaule of crop production.

I construct an index of production of cotton relative to tobacco, which allows me to capture the intensity of the treatment, i.e. the average level of gender division of labour during slavery in a specific county. I calculate this measure as follows:

$$relative\, cotton\, production\, 1840_c = \frac{cotton\, value_c}{cotton\, value_c + tobacco\, value_c}$$

where $cotton\, value$ and $tobacco\, value$ measure the value of production of cotton and tobacco in county c as recorded in the 1840 Agriculture Census.

The higher the index of relative production, the more cotton was produced relative to tobacco, which implies less gender division of labour overall in that county. Note that this measure is only defined for counties where total cotton and tobacco production is greater than zero, and therefore allows to compare areas where at least one of the two crops was grown.

The measure outlined above has some limitations. Firstly, production value is affected by prices; however, this concern is mitigated by the fact that I only consider production at one point in time. The second concern is that production is affected by productivity: one county may produce more cotton compared to tobacco not because the first employs more workers, but because labour is more productive. Whereas this is a theoretical possibility, by comparing Figures 2.1 and A.3.4, which show crop production in 1840 and suitability respectively, production of both crops appears to be geographically concentrated in regions where suitability for that specific crop is higher. To tackle these issues I carry out robustness checks with a similar measure constructed with the number of acres of land cultivated with each crop in 1880, the first year for which this data is available at the county level. I construct this measure as follows:

$$relative\,cotton\,share\,1880_c = \frac{cotton\,acres_c}{cotton\,acres_c\ +\ tobacco\,acres_c}.$$

.

Figure A.3.5 shows cotton and tobacco shares of farmland in 1880.

One further concern with the measures outlined above is that they ignore the extent of the production of cotton and tobacco relative to the total agricultural production. This would not be a concern if one assumes that all other crops are gender neutral. To avoid making this assumption, I run robustness checks using a similar measure for relative production, which computes the relative production of the two crops as a share of total crop production by county:

$$relative\, cotton\, value\, 1840_c = \frac{cotton\, value_c - tobacco\, value_c}{total\, crop\, output\, value_c}.$$

To ease interpretation as a share of total production, I apply a linear transformation by adding one and dividing by two, in order to obtain a variable taking values between 0 and 1.

Table A.2.1 shows summary statistics for the different measures of relative cotton prevalence. Notice that the measures differ in the number of counties they are identified for. The main 1840 measure is available for 804 counties.

2.5.2 OLS Regression

In order to analyse the effect of cotton suitability on labour market outcomes of African American women I estimate the following equation, restricting the sample to African Americans:

$$\begin{aligned} y_{i,c} = \beta_0 + \beta_1 relative\, cotton\, prod_c \times female_i + \\ \beta_2 female_i + \beta_3 \mathbf{X}_i + \beta_4 \mathbf{Z}_c \times f_i + \theta_c + \epsilon_{i,c} \end{aligned} \tag{2.1}$$

where i is an African American individual; c indicates an American county; y is either a binary variable indicating whether the individual is in the labour force or a continuous variable measuring the log of occupation income score; *relative cotton prod* is the measure of relative prevalence of cotton and tobacco in 1840 as described in section 2.5.1; *female* is a binary variable taking value of 1 if the individual is a woman and 0 otherwise; $\boldsymbol{X}$ is a vector of individual controls (age, urban vs rural, marital status, income of spouse); $\boldsymbol{Z}$ is a vector of county controls fixed at year 1870 (number of manufacturing establishments, population, improved land in farms), which I interact with the *female* dummy; and θ represents U.S. county fixed effects. I estimate this equation with cross sectional data from years 1870 to 2010, and all regressions are restricted to individuals between 25 and 54 years old, to facilitate comparisons with the existing literature (Goldin, 1977, 1990).

I choose the above as my main specification, as opposed to only analysing the effect for African American women across counties, because the interaction term allows to control for county specific effects. An alternative specification would compare African American women with Caucasian women. However, for the beginning of the time period considered, white women do not provide a good comparison group for black women, as black females were mostly be employed in agriculture and private households, which were uncommon occupations among white women (Boustan and Collins, 2014). Nevertheless, for later years, particularly since the mid of the 20th century, it is interesting to compare black and white women's labour market outcomes. I analyse this in sections 2.6.3 and 2.6.4.

2.5.3 Instrumental Variable Regression

There are two main concerns with the above empirical strategy. One may be worried that cotton and tobacco production would be endogenous to cultural factors related to gender roles. Even though slaves could not take decisions about where to locate or what crops to grow, plantation owners may have different preferences in relation to gender roles which in turn may affect their decisions. Moreover, the OLS estimates could be biased due to measurement error: the prevalence of the two crops may have changed over time during slavery, and this cannot be accounted for using only the 1840 data. In addition, as described above, the measure may be affected by prices or productivity. Notice that the first concern would introduce a positive bias in the OLS estimates, whereas in the second case the bias would be towards zero.

Therefore, I run instrumental variable regressions where I instrument relative prevalence of cotton compared to tobacco with the relative suitability for the two crops. Figure A.3.4 shows that both cotton and tobacco are more suitable in the east, with tobacco suitability being mainly concentrate in the North, whereas cotton is more suitable in the South. Notice that in areas where African Americans are present in 1880 (see Figure A.3.9), crop shares and suitability follow similar patterns. I measure relative crop suitability by simply taking the ratio of the suitability for the two crops:

$$relative\,suitability_c = \frac{cotton\,suitabilty_c}{tobacco\,suitability_c}.$$

The first stage regression is:

$$\begin{aligned} relative\,cotton\,prod_c \times female_i = \beta_0 + \beta_1 relative\,suitability_c \times female_i + \\ \beta_2 female_i + \beta_3 \mathbf{X}_i + \beta_4 \mathbf{Z}_c \times female_i + \delta_c + \epsilon_{i,c} \end{aligned} \tag{2.2}$$

where *cotton share* and *relative suitability* are the measures of relative cotton prevalence described in section 2.5.1 and δ represents county fixed effects.

2.5.4 Migrants

In order to understand whether the effects of gender division of labour on the labour market of African American women can be at least partly attributed to the shaping of gender roles in the form of attitudes towards the labour market, I compare labour market outcomes of African Americans who were born in areas where cotton was cultivated, but subsequently migrated to other areas. One major challenge is that information about migration is not available at the county level from the U.S. Census, except for the year 1940; however, since the main migration wave started in the 1910s, it is important to analyse migration for earlier years as well. To overcome the data limitations, I adopt two complementary identification strategies. The first approach implements surname matching of African Americans and white slaveholders and compares the geographic location of the former relative to the latter in order to obtain a measure of migration for 1930; the second exploits the migration information available in the 1940 census.

For the first strategy, I retrieve information about migration by making use of the fact that after Emancipation it was common for African Americans to adopt the surnames of their former slaveholders (Van Deburg, 1997), as one was needed for official documents. The reasoning behind this approach is the following: sharing the surname of a slaveholder that owned a plantation in a cotton (or tobacco) county increases the likelihood of the individual having been a slave herself, or beign a descendant of a slave who worked in a cotton (or tobacco) plantation. I obtain a proxy for migration by analysing the location of the individuals whose surname matches that of a former slaveholder: if the individual is located in a state where slavery was not prevalent, I treat her as a migrant.[5] However, an additional complication arises when trying to identify women by their surnames: most women adopted their husband's surname after marriage, and the Censuses do not report women maiden names. In order to reduce measurement error due to surname changes after marriage, and to avoid excluding married women from the analysis altogether, I match the surname of slaveholders with the surname of the mother of each individual. Although this strategy only deals with recent surname changes, because mass migration only started about 15 years before the census data was collected it is likely that a substantial proportion of the individuals would have encountered their spouse close to their place of origin. Appendix A.1 provides a detailed description of how I construct the data set.

Information about individuals' surnames is available until the census of 1930. I carry out the analysis of migrants in 1930 for two reasons. Firstly, until the 1920 census, migration to the North was very limited; secondly, the microdata available for the 1930 census covers 5% of the individuals, whereas the 1920 microdata only covers 1% of the sample, which does not provide sufficient observations for the analysis. To better capture the effect of gender roles, I focus on individuals living in urban areas, so I only consider locations where labour market conditions would differ from those of the agricultural South. To further ensure that the surname matching captures individuals from cotton and tobacco regions I only in-

[5] I will consider as slave states those States where slavery persisted until emancipation: Missouri, Kentucky, Virginia, Maryland, North Carolina, Tennessee, Arkansas, South Carolina, Georgia, Alabama, Mississippi, and Louisiana.

clude migrants that are registered in non-slave states and whose mother was born in a slave state.

I restrict the sample to African American migrants of age between 25 and 54. I estimate the following equation:

$$y_{i,c} = \gamma_0 + \gamma_1 cotton\,migrant \times female_i + \gamma_2 tobacco\,migrant \times female_i + \gamma_3 cotton\,migrant_i + \gamma_4 tobacco\,migrant_i + \gamma_5 female_i + \gamma_6 \mathbf{X}_i + \xi_c + \varepsilon_{i,c} \tag{2.3}$$

where *cotton migrant* and *tobacco migrant* are dummy variables taking value 1 if the surname of the mother of the individual matches the surname of a cotton or tobacco slaveholder respectively; $\boldsymbol{X}$ is a vector of individual controls (as in equation (2.1)); and ξ represents state fixed effects.[6]

It is clear that measuring migration by surname matching is subject to measurement error, the extent of which is very hard to estimate. In fact, not all ex-slaves decided to take the surnames of slaveholders; evidence of this are the organizations which arose after emancipation encouraging African Americans to choose different surnames (Van Deburg, 1997)[7]. In my data set I match around 75% of the individuals. Clearly, not all of the matched individuals adopted the surname of their latest slaveholder; I am therefore capturing an increase in the probability that the matched individual is a descendant of an ex-slave who worked in specific crops.

[6] I choose to include State fixed effects for the main specification because I restrict the sample to individuals in urban areas, which allows for little within county variation.

[7] See section 2.3 for further details.

As an indirect test of the extent to which the surname matching captures the origin of African Americans, I use the number of cotton and tobacco surnames by county as a predictor of my measure of cotton prevalence by county, and of female labour force participation. Table A.2.8 shows a positive and significant correlation with both outcomes, a negative correlation of tobacco surnames with the cotton prevalence measure, and no correlation of other surnames with the share of African American women to men in the labour force. The positive correlation between cotton surnames and the share of women in the labour force is illustrated in Figure A.3.11.

The second method follows closely from the main empirical strategy of this paper. For the year 1940, information is available about the county of residence 5 years prior to the census year. This allows for a direct study of migration: I estimate equation (2.1), where the main variable of interest is relative cotton prevalence in the county of origin of the individual, interacted with *female*. To make the analysis comparable to that of 1930, I only include migrants located in non-slave states and whose mother was born in a slave state.

One issue is that migrants may be a selected sample of the population, and hence that the differences observed are due to the different characteristics of migrants compared to non-migrants. As my samples are entirely composed of migrants, this is not a major concern.

Another concern is that the selection into migration differs across regions with higher cotton or tobacco prevalence, which results in migrants of different origins having different skills. In fact, ideally one should compare individuals with similar characteristics to avoid the results being driven by selection. Understanding migration patterns adds some insights about possible differences in selection of migrants across regions with different cotton and tobacco prevalence. For this, I use data from the 1940 census and regress the number of migrants by county in the previous 5 years on my measure of cotton prevalence and on county characteristics: the percentage of African American women and men in the labour market, the percentage of African American women and men who are literate, and total population. I use county characteristics from the 1930 census, as contemporaneous county-level variables may be affected by migration. Table A.2.9 shows that the number of migrants is higher in regions where cotton prevalence is higher: a one standard deviation increase in cotton prevalence increases the number of migrants by 30% of a standard deviation. Moreover, in counties where female labour force participation was higher, the number of migrants is also higher, but there is no significant association between male labour force participation and number of migrants. Also, both female and male literacy rates are positively associated with migration. I then split the sample into counties with relative cotton prevalence higher or lower than median. The results show that migration is higher in counties with higer female labour force participation and higher male labour force participation where cotton prevalence is above median; however, this is not true for counties with cotton prevalence below median. This suggests that more of the migrants were born in areas in the South with higher female labour force participation and higher cotton prevalence. The higher number of migrants in cotton regions might indicate negative selection of migrants from cotton areas, which can be explained with fixed costs associated with migration, in terms of travel, looking for a new job, and housing (Collins and Wanamaker, 2014). In fact, at the same level of skills, women in cotton areas may have on average higher income because of the higher likelihood to participate in the labour force; this may enable them to overcome these costs. Therefore, more lower skilled migrants would be able to afford migrating. Moreover, Higgs (1976) suggests the decrease in cotton demand as one of the main reasons for migration to the North; this would push individuals who were previously employed in agriculture in cotton regions to migrate. As I analyse migrants in urban areas, it is likely that the skills required there would be different from those required in agriculture. Finally, one should note that the analysis involves a comparison between women and men originating from the same areas, which reduces the possibility of capturing selection effects.

2.6 Results

In this section I describe the main results of the effect of cotton production relative to tobacco on labour market outcomes of African American women. Firstly, I document the effects on labour force participation and occupation income score using data from 1880. I then document the persistence of the effects, using census data from 1870 to 2010.

2.6.1 1880 Census

The first set of results are obtained analysing data from the 1880 census. Table 2.2 shows the effect of living in areas where cotton was a more important crop relative to tobacco on labour force participation (columns 1 and 2) and log of occupation income score of African American women (columns 3 and 4), conditional on being in the labour force. All reported results are produced including individual controls and county controls interacted with *female*[8]. State fixed effects are included in columns 1 and 3, and columns 2 and 4 add county fixed effects. The coefficients reported correspond to the effect of a one standard deviation increase in cotton production value relative to tobacco in 1840. Both specifications show a positive effect on labour force participation of African American Women and on their occupation income score. A one standard deviation increase in share of farmland cultivated with cotton relative to tobacco has an additional positive effect on labour force participation for females compared to males of around 8 percentage points. Notice that the coefficient is very similar when including State or county fixed effects; this suggests that the results are not driven by inherent differences in the labour market conditions of counties with a prevalence of cotton or tobacco.[9] Similarly, women's occupation income score increases with cotton prevalence. The effect is economically significant: a one standard deviation increase in relative cotton production increases the average occupation income score by 15% when including State fixed effects, and 16% with county fixed effects.

Column 1 and 3, which only include State fixed effects, allow for the estimation of the effect of cotton production for African American men, which is negative for both labour force participation and occupation income score. This implies that the higher the cotton production relative to tobacco, the lower the probability of men to be in the labour force, and the

[8]Excluding individual or county controls interacted with *female* produces qualitatively similar results. See section 2.5.2 for a description of the controls included.

[9]Note that with county fixed effects the variable *relative cotton share* is dropped from the regression, as it is measured at the county level, and thus the coefficient is not estimated.

lower their average occupation income score. The negative results for men may indicate some substitutability between the labour of African American men and women.

	Labour Force Participation		Occupation Income Score	
	(1)	(2)	(3)	(4))
Relative Cotton Production 1840×Female	.084***	.079***	.144***	.155***
	(.009)	(.009)	(.015)	(.015)
Relative Cotton Production 1840	-.021***		-.034***	
	(.006)		(.009)	
Female	-.825***	-.822***	-.593***	-.620***
	(.026)	(.027)	(.038)	(.041)
Individual Controls	Y	Y	Y	Y
County FE	N	Y	N	Y
State FE	Y	Y	Y	Y
County Controls*Female	Y	Y	Y	Y
Counties	747	747	746	746
N	186,563	186,563	134,798	134,798
R-Squared	0.4637	0.4857	0.2056	0.2512

Table 2.2. OLS Regressions - 1880

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variable *relative cotton production* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840. Data from the 1880 Population Census.

Overall, these results indicate that in areas where cotton is a more prevalent crop compared to tobacco women are more likely to participate in the labour market, and they are more likely to be employed in occupations that give higher wages. By looking at the type of occupations (Figure A.3.2), one can notice that in counties with higher cotton prevalence women are more likely to work in agriculture (about 55%), whereas in counties with lower cotton prevalence the most common occupation is services in private households (66%). A much smaller percentage of women (roughly 4%) are employed in other occupations - mostly manufacturing and services activities - in both regions.

I then analyse heterogeneous effects on occupation income score for women working in agriculture. Table A.2.2 shows the results of the regressions including the triple interaction of *relative cotton share*, *female* and a dummy variable indicating whether the individual is employed in agriculture. The results indicate that the increase in occupation income score is not driven by individuals working in agriculture: in fact, the coefficient of the main variable of interest remains positive and significant. This suggests that the results are not only driven by a mechanical effect deriving from higher labour demand for women in agriculture in cotton regions. I discuss this further in the section 2.7.

2.6.2 Instrumental Variable

In order to reduce endogeneity concerns of the OLS estimates caused by omitted variable bias or measurement error, I estimate instrumental variable regressions where I use cotton suitability relative to tobacco as an instrument for relative production.

Table 2.3 shows the two-stage least squares (2SLS) results. Panel A illustrates the results of the second stage, while panel B shows the first stage. The first stage results suggest that the measure of relative cotton suitability is a strong predictor of the relative cotton shares of farmland: a one standard deviation increase in relative cotton suitability increases relative cotton prevalence by .28-.32 standard deviations, depending on the specification. The correlation between the two measures is strong: the F-test of excluded instruments is always well above the critical value of 10 suggested by Stock et al. (2002).

The 2SLS estimates are always positive and significant, and larger

	Labour Force Participation		Occupation Income Score	
	(1)	(2)	(3)	(4)
	Panel A: Two-Stage Least Squares			
Relative Cotton Production 1840×Female	.175***	.177***	.188***	.259***
	(.035)	(.039)	(.050)	(.053)
	Panel B: First Stage			
Relative Cotton Suitability×Female	.328***	.286***	.264***	.259***
	(.039)	(.041)	(.042)	(.053)
T-test of Excluded Instruments	36.76	49.03	23.39	33.86
Individual Controls	Y	Y	Y	Y
County FE	N	Y	N	Y
State FE	Y	Y	Y	Y
County Controls*Female	Y	Y	Y	Y
Counties	747	747	746	746
N	186,563	186,563	134,798	134,798

Table 2.3. Instrumental Variables Regressions - 1880

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variables *relative cotton production* and *relative cotton suitability* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

than the OLS, both for labour force participation and for occupation income score. This is consistent with the OLS estimates being biased towards zero due to measurement error of cotton production. The coefficients are larger when including county fixed effects (columns 2 and 4).

2.6.3 Robustness Checks and Additional Results

In this section I describe a set of robustness checks that I perform using the 1880 sample. In addition, I report the results of estimating a model similar to equation (2.1), but restricting the sample to black and white women, and subsequently to only African American women.

Firstly, I estimate a probit model with labour force participation as a dependent variable. Column 1 of Table A.2.4 shows the results: the coefficient is positive and statistically significant, although smaller in magnitude

(3.3 percentage points).

One may be concerned that the results could be driven by African Americans migrating to areas with better job prospects after the end of slavery. To address this concern, I restrict the analysis to counties where slavery was still an institution in 1850 and 1860. Figure A.3.9 shows a map of the slave population in 1860 (top panel) and the African American population in 1880 (bottom panel) by county. It is evident from the two maps that slave population in 1860 predicts African American population in 1880 (major migration waves only started in 1910). However, some counties in the West show a very small percentage of African Americans in 1880 even if no slavery was present at the time, suggesting that, although limited, some migration took place between Emancipation and 1880. Column 2 and 4 of Table A.2.4 show a very similar coefficient for labour force participation compared to the baseline OLS, and a slightly smaller coefficient for occupation income score, but statistically and economically significant.

An additional concern is that the positive effect of cotton on female labour force participation may be partly or entirely driven by income: women might participate more in the labour market when family income is very low. As no measure of income at the individual level is available for the late 19th century censuses, I run regressions including married African American women and controlling for husband's occupation income score as a proxy of family income.[10] The results are summarized in columns 3 and 5 of Table A.2.4. The coefficients are positive and statistically significant, suggesting that spouse's income is not a main driver of the results.[11]

[10]Note that spouse income is only available for currently married individuals.

[11]Note that the sample size is significantly reduced when restricting the regression to married African American women.

Interestingly, spouse income has a negative effect on women's probability to participate in the labour market, but a positive effect on occupation income score. However, husband's income could be considered a "bad control"; in fact, it is likely to be correlated with individuals' own labour market outcomes and could therefore be an outcome itself.

I then perform robustness checks using other measures of cotton prevalence. Table A.2.5 shows the results. Columns 1 and 4 show the estimates when using the measure based on the share of farmland cultivated with cotton and tobacco in 1880; columns 2 and 5 summarize the reduced form estimates, using relative crop suitability; and columns 3 and 6 report the results using the measure of relative cotton prevalence as a share of total production[12]. All three measures produce qualitatively similar results.

One may be interested in testing how labour market outcomes of African American women differ from labour market outcomes of Caucasian women in cotton counties. Even though labour force participation of white women is very low until the mid of the 20th century (below 30% until the 1950s) (Boustan and Collins, 2014), a small percentage is reported to work in agriculture. I run regressions including black and white women where I interact the measure of relative cotton prevalence with a dummy variable taking value 1 if the woman is African American. Table A.2.3 shows that African American women in cotton regions are more likely to participate in the labour force compared to Caucasian women by about 6.7 percentage points (column 1), and their occupation income score is about 12% higher (column 3). Following Kleven and Landais (2016), I investigate whether there are significant differences between African American

[12] The measures are described in sections 2.5.1 and 2.5.3.

and Caucasian women in the number of children; contrary to their findings, here I observe a positive differential effect on the number of children of African Americans. Furthermore, I perform a similar exercise, but comparing African American women in cotton counties to other African American women; the variable of interest is now relative cotton production at the county level. Columns 2 and 4 of Table A.2.3 show that there are positive effects on labour force participation and occupation income score, but no effect on the number of children[13].

2.6.4 Persistence

The results described so far refer to the year 1880, 15 years after slavery was abolished in all States. In order to analyse persistence, I run regressions with cross sectional data from all census years between 1870 and 2010. Figure 2.3 plots the coefficients of the variable $relative\,cotton\,production \times female$ obtained by estimating equation (2.1) with ordinary least squares (OLS) including individual controls and county fixed effects, where the dependent variables are labour force participation (left panel) and the log of occupation income score (right panel).

Both graphs exhibit a similar pattern: the positive effect of cotton for females on both labour force participation and occupation income score is stronger than the effect for males, and it is particularly evident up to the beginning of the 20th century, when it starts to phases out. The occupation income score effect becomes negative in 1960 and 1970, only to phase out

[13]The results columns 2, 4 and 6 of Table A.2.3 are produced by controlling for State, and not county, fixed effects, as the measure of relative cotton prevalence would be collinear with county fixed effects.

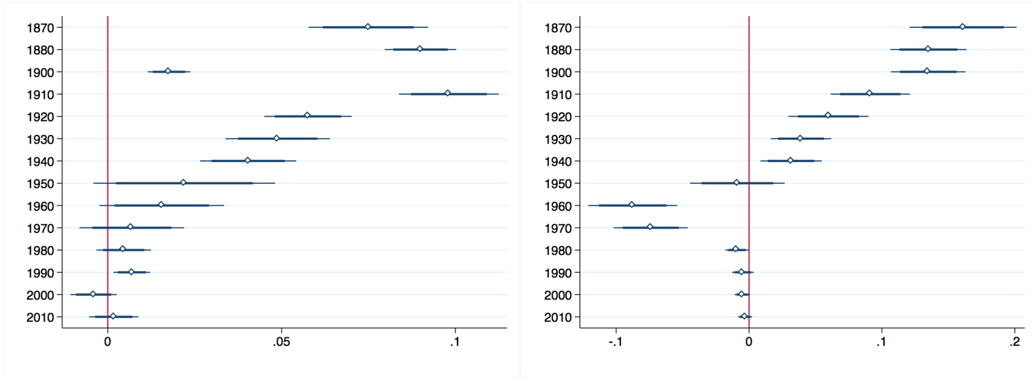


Figure 2.3. Persistence - African Americans

Notes: These two graphs show the coefficient plots of regressing labour force participation (left) and log of occupation income score (right) on the variable $relative\,cotton\,production \times female$, produced with OLS regressions using census data for years 1870-2010. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.

towards zero in the following decades. Persistence is stronger for labour force participation: the effect is positive and significant (at least at the 5% level) until 1960, and it approach zero at the end of the 20th century, whereas for occupation income score it is positive until 1940.

It is worth analysing how the labour market response of African American women compare to those of Caucasian women in cotton regions over time. Figure 2.4 shows the coefficient plot of a similar regression to equation (2.3), but where the variable of interest is the interaction of $relative\,cotton\,production$ and the dummy variable $black$. The left panel shows that labour force participation is again higher for African American women until 1940[14], until it phases out in the following decades, with the exception of 1970 and 2010. The differential effect on occupation income score for African Americans is also positive and significant at the beginning

[14]The coefficient is significantly different from zero at the 10% significance level.

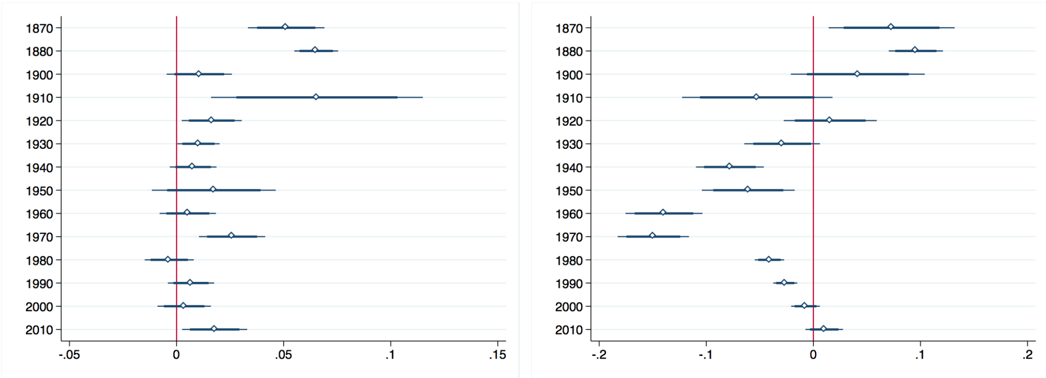


Figure 2.4. Persistence - Black and White Women

Notes: These two graphs show the coefficient plots of regressing labour force participation (left) and log of occupation income score (right) on the variable *relative cotton production* $\times$ *black*, produced with OLS regressions using census data for years 1870-2010. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.

of the period analysed, and it reaches the maximum size of about 10% in 1880. However, it becomes negative in the mid of the 20th century, and slowly approaches zero at the end of the century.

While discussing persistence, it is important to understand the possible role of migration. Firstly, migration from the South to the North may have contributed to the phasing out of the effects. Since the 1910s and up until the 1970s, the Southern States of the U.S., and especially rural areas, have been the subject of mass out-migration of African Americans, who settled in urban areas of the Midwest, Northeast and West (the Great Migration). In fact, I find higher migration rates from counties with higher cotton prevalence compared to tobacco. Secondly, migration within the South East cannot be excluded, even at the end of the 19th century. In fact, minor migration waves appear to have happened around 1880 from the South to Kansas, Oklahoma and Colorado (Johnson and Campbell, 1981). However, data from the 1900 census suggest that by that year 90% of African Americans still lived in the South. Finally, note that the high migration rates in cotton regions imply that the population in later decades may be a selected sample of the original population.

2.7 Mechanisms

In this section I describe how I attempt to disentangle the mechanisms behind the positive effect of cotton and tobacco on labour force participation and occupation income score.

Firstly, in order to disentangle the effect of differences gender division of labour from local labour market conditions in cotton and tobacco areas, I analyse the labour market outcomes on migrants originating from cotton and tobacco areas and locating in urban areas of different counties.

Secondly, I explore whether there are heterogeneous effects for women whose mother is in the labour force, which sheds light on the intergenerational transmission of gender roles.

Furthermore, I investigate whether the results can be explained with higher demand for labour in cotton areas, in particular for agriculture.

Finally, I analyse two alternative mechanisms: access to social networks and discrimination towards African American women.

2.7.1 Migrants

Analysing the labour market outcomes of migrants from cotton and tobacco areas is key for understanding whether the results described in the previous sections are driven by differences in gender roles in cotton and tobacco regions, or whether they are a consequence of differences in labour demand.

	Labour Force Participation		Log Occupation Income Score	
	(1)	(2)	(3)	(4)
Mother with Cotton Surname*Female	.033**	.033**	.150***	.150***
	(.014)	(.014)	(.039)	(.039)
Mother with Tobacco Surname*Female		-.031		-.019
		(.093)		(.068)
Mother with Cotton Surname	.002	.002	.027*	.027*
	(.008)	(.008)	(.015)	(.015)
Mother with Tobacco Surname		.019		.050
		(.027)		(.044)
Female	-.112***	-.112***	-.888***	-.888***
	(.038)	(.037)	(.058)	(.057)
Individual Controls	Y	Y	Y	Y
State FE	Y	Y	Y	Y
N	38,502	38,502	28,231	28,231
R-Squared	0.3726	0.3726	0.5168	0.5168

Table 2.4. Migrants to Urban Areas: 1930

Notes: The regressions include individuals of age between 25 and 54 who live in urban areas and in states where no slavery was present, and whose mother was born in a slave state. Data from the U.S. 1930 5% Population Census. Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10

Table 2.4 shows the results of estimating equation (2.3) with the 1930 sample. Columns 1 and 2 shows that there is a positive relation between having a mother with a surname which matches the surname of a cotton slaveholder and participating in the labour market, and that the relation does not change when adding tobacco migrants as a control. Note that the positive relation only appears to be significant for female migrants, and not for male migrants: this suggests that women with mothers who had a surname associated with cotton are significantly more likely to be in the labour force. The size of the coefficient is 3.3%. The results show no effect for individuals whose mothers match the surname of a tobacco slaveholder. The results for occupation income scores conditional on labour force participation are also positive and significant for females with cotton surnames. This suggests positive effects potentially due to female workers having higher labour market experience compared to other migrants.

Table 2.5 shows the results of the analysis of individuals who migrated in the 5 years previous to the census of 1940. I estimate equation (2.1) defining relative production of the two crops according to the values in the county of origin. The patterns are very similar to those described above: both labour force participation and occupation income scores are higher for female migrants from regions where cotton is relative more present. The coefficients are smaller in size, possibly reflecting the fact that the individuals included in this analysis are all recent migrants, whereas the 1930 migrants may be a combination of recent and longer-term migrants. Column 3 shows that the relative cotton measure does not affect years of education of migrants. This implies that education is not a main driver of the results, and it can be interpreted as evidence against differences in the selection of migrants from cotton and tobacco regions.

These results are suggestive of differences in gender roles between migrants from cotton and tobacco regions, indicating that the higher female labour force participation found in cotton regions does not solely reflect higher labour demand.

2.7.2 Intergenerational Transmission

In this section I analyse the intergenerational transmission mechanisms of labour market outcomes, and whether these are different in cotton and

	Labour Force Participation	Log Occupation Income Score	Years of Education
	(1)	(2)	(3)
Relative Cotton Share 1880×Female	.019**	.033***	-.003
	(.005)	(.011)	(.036)
Relative Cotton Share 1880	-.004	-.033***	-.066*
	(.005)	(.011)	(.035)
Female	-.815***	.276	.229***
	(.059)	(.094)	(.116)
Individual Controls	Y	Y	Y
County FE	Y	Y	Y
Counties	1,009	926	403
N	34,869	28,644	16,201
R-Squared	0.2675	0.3830	0.1623

Table 2.5. Migrants to Urban Areas: 1940

Notes: The variable *relative cotton share* measures the relative prevalence of cotton vs. tobacco in the county of origin of the individual. I include county of destination fixed effects. The sample is restricted to African Americans who migrated from a slave state to urban areas in non-slave states, whose mother was born in a slave state, and whose information about mother labour force status is known. Data from the U.S. 1940 100% Population Census. Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10

tobacco areas. In order to do this, I explore the presence of heterogeneous effects for individuals whose mother is in the labour force by including a triple interaction in the baseline the model.

The results are summarized in Table 2.6. Columns 1 and 2 show the heterogenous effects for mother's labour force participation in the 1880 sample, and columns 3 and 4 include the sample of migrants in 1940. The results show that in both samples mother labour force participation has an effect on occupation income score of daughters, but not on labour force participation. The increase in occupation income score is substantial and remarkably similar across the two samples: about 7 percent for the 1880 sample, and 6 percent for the 1940 migrants. In the 1880 data, mother labour force status cannot explain the entire increase in occupation income score for women in areas with relative high prevalence or cotton: the coefficient of cotton prevalence interacted with female has a positive effect, although smaller compared to the baseline results. However, in the case of migrants, the increase in labour force participation for women from high cotton prevalence counties can be entirely explained by mother labour force status.

	1880 Sample		Migrants 1940	
	Labour Force	Occupation Income Score	Labour Force	Occupation Income Score
	(1)	(2)	(3)	(4)
Relative Cotton Production 1840×Female	.027***	.106***	.047***	.015
	(.004)	(.014)	(.014)	(.023)
Relative Cotton Production 1840×Female ×Mother LFP	.001	.067***	.014	.059***
	(.005)	(.018)	(.011)	(.018)
Relative Cotton Production 1840×Mother LFP	-.001	.002	-.001	-.032*
	(.004)	(.008)	(.011)	(.017)
Mother LFP×Female	.027**	-.163***	.129	-.473***
	(.013)	(.021)	(.087)	(.148)
Mother LFP	-.001	.066***	.016	.268*
	(.005)	(.018)	(.041)	(.141)
Female	-.132***	-.295***	-.558***	.143*
	(.010)	(.034)	(.059)	(.084)
Individual Controls	Y	Y	Y	Y
County FE	Y	Y	Y	Y
Counties	720	684	403	746
N	115,528	76,509	16,201	13,652
R-Squared	0.5415	0.1596	0.3261	0.2517

Table 2.6. Heterogeneous Effects for Mother Labour Force Status

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variable *relative cotton production* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. LFP indicates labour force participation. Data from the 1880 Population Census.

The fact that mother labour force participation matters more in cotton areas might indicate higher labour market experience, which can be transmitted from mother to daughter. Labour market experience may include knowledge of specific jobs, or general knowledge of the labour market, which translates into higher occupation income score of their daughters.

Another possible interpretation regards access to networks: being in the labour force in cotton areas allow women to have better access to social networks in order to find better jobs for their daughters, which translate into higher occupation income scores. The differential effect may be a result of a denser network in cotton areas, due to the presence of more women. in the labour force Interestingly, mother labour force status has a negative effect on occupation income score of daughters in counties with no cotton production.

The increase in labour force participation can partly be explained with mother labour force status: in fact, the size of the coefficient of cotton prevalence interacted with female is reduced compared to the baseline results, and one can observe a positive and significant effect of the interaction of mother labour force participation with female. This is suggestive of gender roles regarding the choice to participate in the labour market transmitting from mother to daughter. While I find no significant difference in the transmission of these gender roles for areas with higer or lower cotton and tobacco prevalence, intergenerational transmission of gender roles may still explain the persistence of results, due to the higher female labour force participation rates in cotton areas.

2.7.3 Demand for Labour

In this section I explore whether the main results could be explained by higher demand for labour in cotton counties. Although analysing the effects for migrants allows to disentangle demand and supply mechanisms, labour demand, in particular for agriculture, may still partly explain why African Americans have better labour market outcomes in cotton areas. Women may have higher labour force participation rates and occupation income score due to the prevalence of agriculture in the South in the first half of the 20th century.

Firstly, differences in demand for labour are likely to exist between areas with high and low relative cotton prevalence. I indirectly test this channel by firstly analysing the effect for African American men, and secondly whether women of other ethnicities display similar labour market outcomes to African American women.

As discussed in section 2.6.1, African American males do not experience higher labour market outcomes in cotton areas compared to other areas.

I investigate the effects on women of other ethnicities using the 1880 census data to estimate equation (2.1), but restricting the sample to exclude African Americans. The remaining ethnicities are Caucasians, Native American, Chinese and Japanese. Table A.2.6 shows no correlation between female labour force participation and relative cotton prevalence, but a positive and significant effects on occupation income score. I explore this further by including in the model a triple interaction of the main explanatory variable of interest with a dummy variable indicating that the individual works in the agricultural sector. Column 3 shows that the entire increase in occupation income score is due to women working in agriculture, who on average have higher occupation income scores than men. Moreover, the triple interaction has a negative coefficient, which indicates that women working in agriculture where cotton is relative more prevalent than tobacco have a lower occupation income score. Note that the same analysis for African American women discussed in section 2.6.1 revealed higher average occupation income scores in counties with higher cotton prevalence even after controlling for the agriculture interaction term. In conclusion, relative cotton prevalence seem to only affect labour market outcomes of African American women, which is suggestive that the results are not only driven by higher labour demand for agriculture.

I then examine whether there are differences in human capital or productivity, which may imply differences in labour demand for black women. To test this hypothesis I estimate equation (2.1) with literacy (1880-1930) or years of education (1940-2010) and wages (1940-2010) as dependent variables. I find no conclusive evidence of differences in human capital or productivity. Figure A.3.12 shows the coefficient plots. We can notice positive and significant coefficients for literacy and education in some of the years (1870, 1940, 1950, 1960 and 2000); however, the positive results are not consistent throughout the sample, therefore bringing no definite evidence of human capital being behind the labour market outcomes results. Moreover the bottom panel of Figure A.3.12 shows a negative effect on wages of African American women in counties with high cotton prevalence.

2.7.4 Discrimination and Access to Social Networks

In this section I explore two other channels that may explain the positive effect of cotton on black women's labour market outcomes: lower discrimination towards African American women and better access to social networks.

Discrimination may be lower where cotton was more prevalent because cotton regions were characterized by a higher concentration of female slaves; hence it is likely that African American women had more opportunities to interact with Caucasians. In turn, this could result in lower discrimination, which may imply a higher probability to be hired by white employers. For an indirect test of discrimination, I investigate whether relative cotton production predict the ratio of mixed couples composed of white husband and African American wife. In this regression I control for State fixed effects, to reduce the possibility of capturing effects of laws or customs regarding mixed marriages. Column (1) of Table A.2.7 shows a positive and significant effect, which corresponds to a 30% increase in the ratio of mixed couples in counties with higher cotton prevalence. This is suggestive of lower discrimination in cotton counties; however, this channel cannot explain the results on migrants.

Moreover, because cotton areas displayed a higher concentration of African Americans women, they may benefit from better access to social networks. Chay and Munshi (2013) put forward a similar hypothesis: they find evidence of African Americans being able to better exploit social networks in areas where the crops grown were more labour intensive, due to the higher concentration of African American individuals. In order to test the social network channel I estimate equation (2.1) adding as explanatory variable the number of female slaves in a county in 1860 interacted with the female dummy. Columns (2) and (3) show that controlling for cotton prevalence, the number of female slaves is positively associated to both labour force participation and occupation income score. This may indicate that better access to social networks plays a role. However, the coefficient of cotton prevalence times female decreases in magnitude, but remain significant, signalling that although the two measures are correlated, cotton prevalence has a positive effect on labour market outcomes which is independent from the higher density of women.

2.8 Conclusion

Understanding the effect of gender roles on the labour market outcomes of women is an important policy question, particularly in countries where women are largely underrepresented in the formal labour market. This paper contributes to the literature on the persistent effect on gender roles on women's labour market outcomes, by analysing the case of African American women after slavery. I find higher labour force participation and higher occupation income scores of African American women after slavery in areas where gender division of labour was less well-defined during slavery, which I identify as areas producing cotton. I find persistent effects until at least 70 years after the end of slavery for occupation income scores, and at least 90 years for labour force participation. Due to the lack of migration data for the first decades of the 20th century, I construct a data set of migrants by matching surnames of African Americans in urban areas of free states in 1930 (during the Great Migration) with surnames of slaveholders in the 1860 census. I find a positive relation between having a cotton surname and having higher labour force participation and occupation income score for women, and this relation is stronger for women than for men. These results are confirmed when analysing direct data on migration from the 1940 census.

My findings are consistent with intergenerational transmission of labour market experience from women from cotton areas to their female children, which does not happen in tobacco regions. Although I find no evidence of higher human capital or productivity, higher occupation income scores may indicate that women from cotton regions have more labour market experience which, although uncorrelated with formal education, enables them to find better jobs.

Chapter 3

The Benefits of the Bamboo Network in International Trade

3.1 Introduction

Non-tariff trade costs have been the focus of the empirical trade literature since the seminal contribution of Obstfeld and Rogoff (2001). These trade costs appear to be large and important determinants of trade flows across countries; however, they are still not well understood. The key empirical challenge consists on the fact that non-tariff barriers are inherently unobservable. The study of the role of ethnic networks in facilitating trade allows not only to document the importance of unobservable frictions, but also to shed light on their nature; in fact, ethnic networks are helpful for overcoming only some of these barriers. Although the importance of ethnic networks in promoting trade has been widely documented (Rauch and Trindade, 2002; Rauch, 2001, 1999; Parsons et al., 2014), there is still little consensus on the exact channels at play, which is often caused by the aggregate nature of the data available.

In this paper, we add to the literature by studying the effect of ethnic networks using firm level data from Guangdong Province, China. The network in question was created subsequently to a mass migration wave of ethnic Cantonese people from said province in Southern China to the United States in the late 19th century. Therefore, we study the effect of improved access to the the ethnic network in the U.S., which we define as the network of American-born ethnic Cantonese residing in the United States.

Our identification strategy makes use of the diversity of ethnicities and languages in Southern China. More specifically, we calculate the degree of exposure to ethnic networks of firms in Guangdong as the interaction of two dimensions. Firstly, an industry dimension: we assume that a given firm's exposure to the network increases with the number of ethnic Cantonese workers employed in industries that it is likely to have trading relationships with. We propose that a Chinese firm is more likely to trade with U.S. firms operating in the same 4-digit industry, with downstream manufacturing buyers, or with retailers and wholesalers, which sell the goods it produce. Therefore, we calculate three different measures of network exposure at the industry level, corresponding to the number of ethnic Cantonese workers in each of those related industries.

Secondly, according to our measure, firms network exposure increases with cultural closeness to the ethnic network in the U.S. For this measure we exploit the geographic location of firms: "cultural exposure" to the network is higher for firms located in one of the sending counties of migrants in the 19th century. This flows through two channels. Firstly, there may still exist kinship ties, i.e. the descendants of the migrants in the U.S. may have maintained relations with individuals in the place of origin. The second channel is language, as the main languages spoken in the sending counties is a dialect belonging to the Cantonese group, which is not intelligible for individuals speaking dialects of other groups. In order to shed light on the relative importance of the two mechanisms, we also employ another measure of cultural closeness, namely a dummy variable indicating whether a firm is located in a Cantonese speaking area.

We then construct our final measure of exposure as an interaction of the two dimensions and our main analysis follows a difference-in-difference framework at the county-industry level, where we control for county and industry fixed effects.

Our main data set is the Annual Survey of Industrial Enterprises from the year 2004 conducted by the China's National Bureau of Statistics for the province Guangdong. The survey covers all manufacturing enterprises with output larger than 5 million RMB, which in Guangdong amounts to around 34,500 firms. The data set provides information about a number of firm variables in addition to exports. Furthermore, we have access to the Chinese Economic Census, which records employment and revenues for the universe of registered firms. We use this data set to analyse how the firm size distribution is affected and whether there were additional effects on smaller firms or firms in other sectors. Finally, we construct our measure of network exposure from U.S. census data.

The main findings indicate that more connected firms are more likely to export, and have higher export value. This means that the benefits of the ethnic network affect both the extensive and the intensive margin of trade. The effects are economically significant: for a one standard deviation increase in network exposure, the probability of exporting is about 3 percentage points higher for each of the three measures of industry network, and the value of exports conditional on the firm being an exporter is about 6 to 14 percent higher, depending on the measure of network exposure.

We conduct a number of exercises to investigate which -direct and indirect- channels may be driving our results. The literature has proposed two channels through which ethnic networks can directly impact exporting behaviour. Firstly, when formal contract enforcement mechanisms are weak or non-existent, ethnic networks can overcome this void through collective punishment, thus facilitating trade. Secondly, ethnic networks can overcome important information barriers that potential exporters face, such as lack of knowledge about the tastes of consumers in the destination market or play an important role in matching buyers and sellers. Moreover, ethnic networks can have indirect effects on exports through aiding the spread of technological knowledge or through promoting foreign direct investment.

We find little evidence that exports are driven by technological knowledge flows that are aided by the network, as profitability does not increase nor is the effect larger for high-tech firms. We do, however, find a positive effect on labour productivity for smaller manufacturing firms, which could be indicative of knowledge flows playing a role for this subset. Similarly, while foreign direct investment does seem to increase with exposure to the network, it cannot explain our results.

Concerning the direct channels, we further investigate whether the results can be explained by improvements in contract enforcement or by the flow of information through the network. Following Rauch (1999), we study whether the effect is driven by differentiated goods, for which the information channel should be of greater importance. We do find a larger effect for differentiated goods, supporting Rauch (2001)'s finding, that the information channel is of particular importance.

Finally, we also find effects on other firm-level variables: More connected firms have higher output and profits but lower domestic sales. They also have significantly higher fixed assets and capital, while keeping the total number of workers constant. Firms with more exposure to the network also have a higher share of workers with a university degree, pay higher wages and have higher management expenses.

While these findings are consistent with a number of hypothesis, one possibility is that ethnic networks lower informational barriers, which then allows firms to specialise in products for the U.S. market that are not demanded by Chinese consumers. Therefore, these firms sell less in the domestic market but exports are increased. The fact that more connected firms employ more high-skilled workers and more fixed assets in turn is consistent with quality upgrading that is likely to be required to access the American market.

Our contribution is twofold. Firstly, we estimate the effect of ethnic networks on international trade at the firm level. To the best of our knowledge, our paper is the first to distinguish between the effects of ethnic networks on the intensive and extensive margin of trade at firm level. Moreover, we have access to a rich data set of firm level variables, which allows us identify the channels though which ethnic networks affects trade.

Secondly, our identification strategy captures the effects of an ethnic network formed through historic migration. Other papers have measured ethnic networks with past migration patterns (Kerr and Lincoln, 2010; Kerr, 2013; Griffith et al., 2006; Parsons et al., 2014) to mitigate reverse causality concerns. However, our very narrow definition of network allows us to further overcome the possibility that recent migrants may selectively settle in U.S. counties with high prevalence of industries which are well performing in China.

The paper is organized as follows: section 3.2 describes the related literature; section 3.3 outlines the mechanisms through which ethnic networks can affect international trade; section 3.4 provides an overview of the migration of Cantonese people to the U.S.; section 3.5 describes the data sources and how we construct the data sets; section 3.6 describes in detail the empirical strategy. We illustrate the results on exports and other firm variables in sections 3.7 and 3.8, and finally section 3.9 concludes.

3.2 Related Literature

This work relates to several strands of literature, which analyse the effect of ethnic networks on several aspects of the economy: international trade, labour markets, knowledge diffusion and innovation, and firm growth in developing countries. Moreover, it relates to the literature on the impact of language and culture on international trade.

Firstly, it contributes to the literature on the role of co-ethnic networks on international trade. Rauch and Trindade (2002) are among the first to empirically document the role of Chinese ethnic networks in international trade in a cross country study; they find a higher volume of bilateral trade among China and countries with a higher share of ethnic Chinese population. Rauch (2001) provides a review of the literature on the effect of social networks on international trade focusing mainly on two channels: reduction of information barriers and better contract enforcement. A closely related paper in terms of research question and empirical strategy is the recent contribution by Parsons et al. (2014): the authors use a natural experiment - the migration of the Vietnamese Boat People to the U.S. - to provide empirical evidence of a link between ethnic migrant networks and exports of American firms, thus focusing on the benefits of ethnic networks for the receiving country of migrants. Their findings establish a causal link between the geographical location of Vietnamese people in the U.S. and the volume of exports to Vietnam, suggesting that networks play a role in international trade between the two countries.

A number of papers examine the effect of networks on trade, however with a focus on the role of social networks rather than ethnic networks. Combes et al. (2005) study how migrant and business networks affect trade across French regions, measuring the latter with the number of connections between plants which belong to the same business group. Moreover, a recent cross-country study by Head et al. (2010) analyse the long term impact of countries' independence on bilateral trade with the former colonizing country and the rest of the world; they find negative effects on trade flows, and a small but positive effect on the extensive margin (i.e. probability to export) with other former colonies of the same colonizer and with the rest of the world.

The importance of migration networks in the labour market has been reviewed by Montgomery (1991) and has been the focus of a number of recent empirical papers. The work by Munshi (2003) examines their role on labour market organizations: he empirically analyses the network of Mexican migrants in the U.S. and its effect on the labour market in the destination country, by measuring the size of network as the share of individuals in the community of origin who are located in the U.S. The findings indicate that more connected individuals have higher probability of being employed and have higher wages. McKenzie and Rapoport (2007) investigate the empirical relationship between migration and wealth using data from Mexico, highlighting that migration costs decrease with the size of the migrant network originating from the same community. They find an inverse U-shaped relationship between migration and inequality in rural Mexico, and that migration contributes to decreasing inequality in sending villages with high rates of historic migration.

The role of ethnic and migrants networks has been studied in the context of knowledge and technology diffusion. Griffith et al. (2006) present evidence for the existence of knowledge spillovers across countries; with firm-level data from the U.K., they test the effect of "technology sourcing" exploiting pre-1990s technology boom location of U.K. affiliated firms in the U.S. for identification. Kerr (2008) focuses on the role of the U.S. as a frontier country in the diffusion of scientific knowledge and recognises the role of ethnic networks in knowledge transfers across countries. In a more recent paper, the same author studies the extent to which comparative advantage is an determinant of trade, using the geographic overlay of the location of past migrants communities and innovation in the U.S. to identify differences in technology diffusion and therefore comparative advantage (Kerr, 2013). Kerr and Lincoln (2010) and Moser et al. (2014) study the effect of skilled migration on innovation in the U.S., the first exploiting changes in the H-1B visa regulations to overcome endogeneity, and the latter focusing on the contribution to innovation of Jewish migrants from Germany to the U.S. at the time the Nazi party was in power.

As we explore the effect of networks not only on exports but also on other firm variables, our work also relates to the work on co-ethnic networks and firm growth. Banerjee and Munshi (2004) demonstrate that social ties of businessmen to the local community play a role in the allocation of capital, which is not necessarily in favour of the most productive firms; they show that among textile firms in Tirupur (India), those owned by locals entrepreneurs have higher fixed capital and capital intensity of production compared to firms owned by outsiders. The paper by Woodruff and Zenteno (2007) analyses the impact of the migration networks in Mexico on the development of microenterprises, and find evidence of migration networks playing a role in alleviating capital constraints in the most capital intensive industries. Another example of this literature is Nanda and Khanna (2010), which find that in India entrepreneurs rely more on ethnic networks if firms are located outside software hubs, indicating that networks are most important in environments with limited access to information and financial institutions.

Focusing on foreign direct investment (FDI), the recent work by Javorcik et al. (2011) find that the presence of migrants in the U.S. increases FDI in the country of origin, and Burchardi et al. (2016) show similar results by instrumenting ancestry composition with measures of “push” and “pull” factors, which refer to factors causing migration from a country to the U.S. and migration from all countries to a specific county in the U.S. respectively.

Besides Rauch and Trindade (2002), others have documented the role of Chinese migrant networks on many aspects of the economy. Saxenian (2002) highlights the role of first generation migrants from China and India working in the Silicon Valley, for sharing both information about technology and investment in business partnerships with their counterparts residing in their country of origin. Felbermayr and Toubal (2012) revisit the evidence found by Rauch and Trindade (2002) finding more modest effects on international trade, although still positive. Rotunno and Vézina (2012) document the importance of Chinese networks on tariff evasion, measured by the difference between exports reported by other countries and imports reported in China. Finally, Zhang and Song (2002) document the role of FDI on Chinese exports since China's open door policy.

Furthermore, our paper relates to the literature documenting the effect of culture and language on economic activities. Guiso, Sapienza and Zingales (2009) empirically estimate the role of culture, including language and genetic similarities, on bilateral trade and investment across countries. Melitz (2008) investigates the importance of a common language for international trade and finds that direct communication is more effective than having to resort to translation. Specific to the Chinese context, Chen et al. (2014) find that, although nowadays Mandarin is commonly spoken in all provinces in China, fluency in the local dialects has positive consequences on individuals' income in both services and manufacturing, and particularly in sales jobs.

3.3 Mechanism

Obstfeld and Rogoff (2001) have put forward the hypothesis that the trade patterns empirically observed - for instance strong border effects - indicate the presence of large non-tariff trade costs. Ethnic networks can reduce non-visible trade costs in two ways, as discussed in Rauch and Trindade (2002): Firstly, by reducing informational barriers and secondly by providing enforcement mechanisms for contracts in the presence of weak institutions. However, ethnic networks can also increase trade flows indirectly, either through increasing the flow of technology and knowledge, which makes firms more competitive in the export market, or by facilitating foreign direct investment, which in turn might increase trade, for example by allowing firms to outsource production. Finally, migrant networks could affect trade simply by changing tastes and preferences across countries. In contrast to both the direct and indirect channels, which are welfare enhancing by decreasing frictions and distortions, the latter would not necessarily imply an improvement in welfare.

Information barriers are likely to play an important role in impeding trade as has been highlighted by Rauch (2001). Firms might hold imperfect information about, for instance, the taste of foreign consumers or the products sold by foreign competitors. Obtaining this information might be an important fixed costs that impedes firms from exporting. Furthermore, such information about tastes and markets might become more costly to obtain with both geographic as well as cultural and linguistic distance. Additionally, when trading specialised or customised goods, the detailed requirements might be harder to communicate across linguistic and cul-

tural hurdles. Ethnic migrant networks can facilitate information flow by reducing these costs, as connected individuals face lower cultural and linguistic impediments. In our case, Chinese-Americans might have a distinct advantage in communicating the preferences of American consumers and the market structure to Chinese producers. In addition networks can play an important role by acting as intermediaries and aid matching of importers and exporters. If we assume that there are search frictions due to asymmetric information, matching between exporters and importers can be costly (Ahn et al., 2011). Networks hold information about the types and trustworthiness of available producers and buyers and thus reduce search costs for their members (Rauch and Trindade, 2002).
The fact that this channel plays an important role in our setting is supported by qualitative evidence such as by Weidenbaum and Hughes (1996), who describe the role of the "bamboo network" - the ethnic Chinese networks overseas - in facilitating trade. They claim that *"[t]he leading business men know each other personally and do deals together, with information spreading through an informal network rather than through more conventional channels"*. Moreover, Kotkin (1993) writes that *"Chinese entrepreneurs remain in essence arbitrageurs, their widespread dispersion a critical means of identifying prime business opportunities"*, which indicates the importance of information flows.

The second direct channel through which ethnic networks can further trade is through enforcing informal and formal contracts. Contract enforcement may be particularly problematic in the context of international trade: Imperfections of the justice system are likely to be enlarged by distance and thus firms find it harder to enforce contracts across borders than domestically. Furthermore, contractual details are harder to pin point in the presence of geographic and linguistic barriers. Finally, in the case of international trade large sunk costs may have to be paid before it can be observed whether all parties complied to the agreement. These imperfections might be particularly salient if domestic institutions are weak, such is the case in developing countries, and thus particularly relevant for China (Anderson and Marcouiller, 2002). As a result, firms are likely to rely on informal and relational contracts. Ethnic networks can play an important role in enforcing such relational contracts by providing punishment mechanisms, as has been emphasised in the seminal contributions by Greif (1989, 1993). This channel has also been described by Weidenbaum and Hughes (1996), who write that *"[i]f a business owner violates an agreement, he is blacklisted. This is far worse than being sued, because the entire Chinese network will refrain from doing business with the guilty party"*. In addition, in many cases, if the quality of a product is hard to observe, network intermediaries often are willing reduce risk for the importer by guaranteeing quality and deepening reputational concerns (Chaney, 2014).

Another channel through which ethnic networks may affect international trade, albeit indirectly, is promoting technology diffusion across countries. The relationship between ethnic networks and knowledge flows has been analysed widely in the innovation literature (Kerr, 2008; Kerr and Lincoln, 2010), though rarely in relation to exporting behaviour (an exception is Kerr (2013)). We suppose that increased technology and knowledge diffusion allows connected firms to increase their productivity or product quality through imitation or innovation. This in turn, might make firms productive enough to make exporting profitable, thus indirectly leading to an increase in exports. It is important, however, to acknowledge that there might also be a reverse effect: For example, Keller (2002*b*, 2004) argues that trading differentiated intermediate goods, which embed innovation, facilitates technology flows across countries.

A second indirect channel through which ethnic networks can affect trade is through fostering foreign investment. This channel has been the focus of recent studies, which show that the presence of migration networks in a country increases FDI flows towards their origin country of migrants (Javorcik et al., 2011; Burchardi et al., 2016). This may happen because networks help information flow across countries, or because they enhance transactions in environments with weak institutions. High levels of FDI in turn have been hypothesised to promote international trade: Zhang and Song (2002), for example, explore this channel in the context of China, showing that increases in FDI over time can be linked to improvements in manufacturing export performance. In this way, an observed link between connectedness to the ethnic network and trade may be caused by increases in FDI.

The final mechanism highlighted in the literature is the so-called "preference channel", first discussed by Gould (1994). This channel supposes that a large migrant community in a destination country leads to an increase in trade simply by increasing the share of the population with tastes that favour goods from the sending country. We argue that this channel is less relevant in our setting, as we exploit industry-level differences in ethnic Chinese employment in the U.S. As a result, we do not capture aggregate effects caused by a large number of ethnic Cantonese from the sending counties in China in the U.S. but only differences across industries. Moreover, whereas the preference channel is generally consistent when analysing trade of differentiated and final goods (Rauch and Trindade, 2002), we explore network effects also for both intermediate and undifferentiated goods.

3.4 Historical Background

3.4.1 Emigration From China

Emigration from China has a long history starting in the 10th century with merchant emigrants building a trading network all over Southeast Asia. Mass emigration, however, only became prevalent in the middle of the 19th century. This first emigration wave started around 1842 with the loss of the first Opium War, when China was forced to open itself to Western influences. Emigration accelerated in the 1860's, when the Qing government lifted its ban on emigration, which had been imposed since the seventeenth century (Woon, 1990). By the outbreak of World War II, between 8.5 to 9 million Chinese were living outside Chinese borders all over the world from South East Asia to South and North America and Australasia. Migration only came to an abrupt halt in 1949 when the People's Republic of China was declared and the borders were closed.

During this first emigration wave, individuals emigrated mostly through informal networks in order to overcome credit and informational constraints. As a result the origin of migrants for a given destination country were extremely localised: While an estimated 90% of all emigrants during this period originated from Guangdong or Fujian (Woon, 1990), nearly all of the 19th century Chinese immigrants to the U.S. and the Kingdom of Hawaii originated from eight counties adjacent to Guangzhou. Historians estimate that 80-90% came from the Siyi districts, comprising Taishan, Xinhui, Kaiping, Heshan and Enping counties (Hsu 2000). Somewhere between 10-20% percent of immigrants came from Sanyi districts, comprised of Panyu, Nanhai and Shunde counties. These counties, which used to be seven counties at the time of emigration, are our main focus and will be referred to as the sending counties. A third, smaller group of immigrants came from Zhongshan, south of Guangzhou (for a Map of the counties see Figure 2). It is important to note that these counties, however, were not necessarily the ones that saw the most out-migration, but only saw the most migration *to the U.S.* Emigrants to other destination countries in South East Asia or Oceania originated from other counties. The general pattern that individuals originating from a given county migrated to the same country seemed to hold all over southern China (Voss and Allen, 2008).

Emigration was motivated by push-factors, such as overpopulation, and by the political instability that was spreading all over China, which caused the Taiping and the Boxer rebellions and the more localised Punti-Hakka Clan Wars, which culminated in around a million casualties. The two provinces in general, and the sending counties in particular, were some of the areas most characterised by over-population and food shortages, which were exacerbated by droughts and floods throughout the 19th century.

At the same time, there were a number of "treaty ports" established in both provinces, the most notable being Guangzhou (Kanton) in Guangdong. Treaty ports were ports ceded by the Qing dynasty to a number of

Western powers as well as Russia and Japan after the Opium wars, where free trade was allowed to take place. Thus shipping routes from these locations were already established (Faure and Siu, 1995).

The reasons why emigrant to the U.S. originated nearly exclusively from the sending counties are not well documented. The initial pattern could have originated because of the proximity to Guangzhou and Hong Kong, which were the main shipping destinations from San Fransisco. Initial patterns of migrations were likely to be predictive of future migration due to the need of a network in the destination country, coupled with the linguistically and culturally diverse landscape of Guangdong; the latter implied that individuals of the same linguistic group could benefit from a particularly close knit network, which others could not have accessed. Furthermore, the fact that emigration was banned until the end of the century could have contributed to informal networks being of particular importance.

Table 3.1 shows the number of Chinese migrants that arrived in Canada. As migrants from China arrived to the entirety of North America nearly exclusively via San Francisco, immigrants to Canada have the same origin as immigrants to the U.S. The data has been assembled by Yu (2011) using the Canadian Head Tax records from 1885 to 1949 and covers almost 100,000 individuals of Chinese origin. While such data exists for the U.S., to the best of our knowledge it has not been digitalised and made available. The Table shows that the sending counties identified above constitute around 98.3% of all the Canadian-Chinese in the sample.

	Number of Immigrants	Percentage
Taishan	44,131	47.7
Xinhui	13,858	15.0
Kaiping	13,350	14.4
Enping	3,754	4.1
Heshan	2,574	2.8
Siyi	**77,667**	**84.0**
Panyu	6,415	6.9
Nanhai	480	0.5
Shunde	420	0.5
Sanyi	**7,315**	**7.9**
Zhongshan	5,899	6.4
Other	1607	1.7
	Total: 92488	

Table 3.1. Origin of Canadian-Chinese by County in China

Notes: Data from Yu (2011).

3.4.2 Ethnic-Chinese in America

At the beginning of the first migration wave during the 19th century, the main pull factor was the gold rush. Like many Europeans, Chinese emigrants were drawn to the U.S. in the hope to make their fortunes extracting gold. Therefore, and for geographic convenience, most early Chinese immigrants settled in California. Later Chinese immigration was encouraged in order to provide labour for the construction of the Transcontinental Railway, as well as to a limited degree to replace former slaves in the southern plantations. By 1882, around 380 000 Chinese left China and settled in mainland U.S. and a further 46 000 settled in Hawaii (Voss and Allen, 2008).

The mass migration came to a stop because of the Chinese Exclusion

Act in 1882, which was subsequently extended under the name of Geary Act in 1892. It was the first major law that restricted immigration to the US and is seen a response to anti-Chinese sentiments that started from California, and then spread in the whole country, during the second half of the 19th century. The act prohibited migration of Chinese *"skilled and unskilled labourers and Chinese employed in mining"*, under penalty of imprisonment or deportation. Effectively, however, it impeded all kinds of Chinese migration, as it was hard for migrants to prove their intentions of not become labourers. There were few exceptions, for instance for wealthy merchants or students. Further State laws forbade interracial marriages, limited civil rights and restricted the possibility of employment by non-Chinese. Following the Act, more laws restricting migration were approved, until the National Origins Act of 1929, which capped immigration to 150,000 people per year and prohibited all immigration from Asia. It also prohibited Chinese from becoming U.S. citizens. In addition, other legislation excluded the Chinese from certain occupations, especially in California. The situation only improved with the repeal of the Exclusion Act in 1943, motivated in part by the alliance between China and the U.S. during World War II (Lee, 2003).

Upon their arrival, the ethnic Chinese had mostly settled in Chinatowns, among which the largest were in San Francisco and New York. Due to discrimination, the Chinese continued predominantly to live within the borders of the Chinatowns.

Even after the abolition of the Exclusion Act, the number of Chinese migrants was limited predominantly to family reunification and to 105 persons per year (Zhao, 2002). Large-scale Chinese immigration did not occur until 1965 when the Immigration and Nationality Act of 1965 lifted national origin quotas (Hing, 1993). The 2004 United States Census reports over 2.8 million Chinese in the United States, about 1% of the total population and 23% of the Asian population.

3.5 Data Sources and Summary Statistics

In this section we briefly outline the main data sources and provide some summary statistics.

3.5.1 Firm information

We have access to two data sets containing firm data in Guangdong. Firstly, the Annual Survey of Industrial Enterprises of 2004, which is a survey of all large manufacturing firms. Secondly, the Chinese Economic Census of the same year, which includes all registered enterprises, but only minimal information about each firm. We use the first data set for our main analysis as it includes information on exports, while we use the second data set to analyse spill-over and direct effects on small firms and for industries other than manufacturing.

The Annual Survey of Industrial Enterprises is an annual survey conducted by the China's National Bureau of Statistics. It collects information about all state-owned firms and all privately owned firms with revenues higher than 5 million RMB in the manufacturing sector. This also includes firms that are partially or fully foreign owned. Overall, around 5% of all firms in our sample are state owned, while around 12% are at least partially foreign owned. Our data set covers the Guangdong province only. We exclude state owned firms from our analysis, as their operation is generally regarded to be fundamentally different and access to the overseas Chinese network less important than for private firms.

The data set is very detailed in terms of firm-level variables observed. The main variable of interest is the value of total exports measured in thousands of RMB. We create two variables from this information: Firstly, a dummy variable indicating firms' export status; secondly, the log of export value. Around half of the firms in our sample are exporters. While this is a much higher share compared to other data sets (Bernard et al. (2009) for example find that in 2000 only 3.1% of all U.S. firms export), it is important to keep in mind that this is foremostly driven by the fact that our sample only includes large firms. Furthermore, Guangdong is one of the main exporting regions of China, and thus we would expect the share of exporters to be large. Unfortunately we do not have firm-level information on bilateral trade between China and the U.S.; thus as a robustness check we weigh reported exports by the average share of a given Chinese industry's total exports that is purchased by U.S. firms[1].

Other variables of interest for our analysis are: 4-digit Chinese Standard Industry Classification, total number of workers employed, capital by origin (domestic, foreign, state), profits, wages, age of the firm, R&D expenses and management expenses.

Table 3.2 shows some summary statistics of key variables in our sample. We show the summary statistics for the entire sample and for exporting firms separately. Regarding the whole sample, as mentioned

[1]For more details see section 3.7.2, which describes the robustness checks

previously, slightly less than 50% of all firms are exporting. Around 5% of the sample is state-owned and 10% of total capital is provided from foreign sources. In terms of employment, the average firm is medium sized, with 289 workers.

Comparing exporting and non-exporting firms, we find similarly to the general literature that exporting firms are significantly larger than non-exporters in terms of the number of workers, total wages paid and output. They have slightly more capital in total and fixed assets. They are, however, less likely to be state owned and have a smaller share of foreign owned capital. Furthermore, the average share of high-skilled workers is lower.

The Chinese Economic Census of 2004 is also compiled by the National Bureau of Statistics of the PRC and was accessed through the China Data Center at the University of Michigan. It contains information on the universe of all registered Chinese firms and thus contains a large number of observations. In contrast to the Survey of Industrial Enterprises, the census covers all industries and thus allows us to analyse also effects on services. Note that this data set nests the previous one. However, it contains only few firm-level variables, namely: employee number, revenues, 4 digit industry classification, ownership type and exact location. For privacy reasons, employee number and revenues are only given in bandwidths. We aggregate the location up to the county level to match it with our independent variable measuring network exposure. Again, we exclude all state-owned firms. This gives us 243,205 firm-level observations, out of which around 15% are at least partially foreign owned. Figure 3.1 shows the distribution of firms according to ownership. As can be seen, most for-

	Mean	Median	SD	Min	Max	N
Export Status	.500	0	.500	0	1	34,558
Export Value ('000 RMB)	37311.29	0	502314.9	0	1.50e+07	34,525
	All Firms					
Employment (N. workers)	288.530	121	816.850	1	71915	34,525
Share of High Skilled Workers(%)	.038	.008	.089	0	1	34,525
Total Wages ('000 RMB)	4470.353	1478	29928.25	6	4158034	34,525
Management Expenses ('000 RMB)	3580.832	979	39077.91	0	5078518	34,525
Profits ('000 RMB)	4108.157	226	83596.41	-278790	7728064	34,525
Output ('000 RMB)	85368.06	18540	938114.4	0	1.09e+08	34,525
Domestic Sales ('000 RMB)	83475.68	18008	925507.6	0	1.09e+08	34,525
Fixed Assets ('000 RMB)	24737.29	2932	551633.3	0	9.60e+07	34,525
Total Capital ('000 RMB)	19458.72	3600	277759.9	0	4.87e+07	34,525
Foreign Capital Share (%)	.104	0	.296	0	1	34,379
State Owned Dummy	.051	0	.219	0	1	34,525
	Exporting Firms					
Employment (N. workers)	439.233	200	1066.735	2	71915	17,232
Share of High Skilled Workers(%)	.032	.008	.070	0	1	17,232
Total Wages ('000 RMB)	6627.220	2398	27344.32	24	2204273	17,232
Management Expenses ('000 RMB)	5129.092	1387.5	54399.75	0	5078518	17,232
Profits ('000 RMB)	5509.688	260	94274.87	-246506	7728064	17,232
Output ('000 RMB)	119333.8	23501	975514.4	0	7.30e+07	17,232
Domestic Sales ('000 RMB)	42492.64	147.5	517657.9	1	3.97e+07	17,232
Fixed Assets ('000 RMB)	25285.13	4150.5	154693.3s	0	9111732	17,232
Total Capital ('000 RMB)	24297.66	6626.5	110130.8	0	8842600	17,232
Foreign Capital Share (%)	.050	0	.199	0	1	17,232
State Owned Dummy	.023	0	.153	0	1	17,232

Table 3.2. Descriptive Statistics: Firms Characteristics

Notes: This table shows the descriptive statistics of the main variables used in this paper. When applicable the data is shown in 2004 RMB, at which point 1USD =8.28 RMB. Source: 2004 Annual Survey of Industrial Enterprises.

eign owned firms are owned by individual or corporations based in Hong Kong, Macao or Taiwan. Table B.1.1 show the summary statistics for this data set.

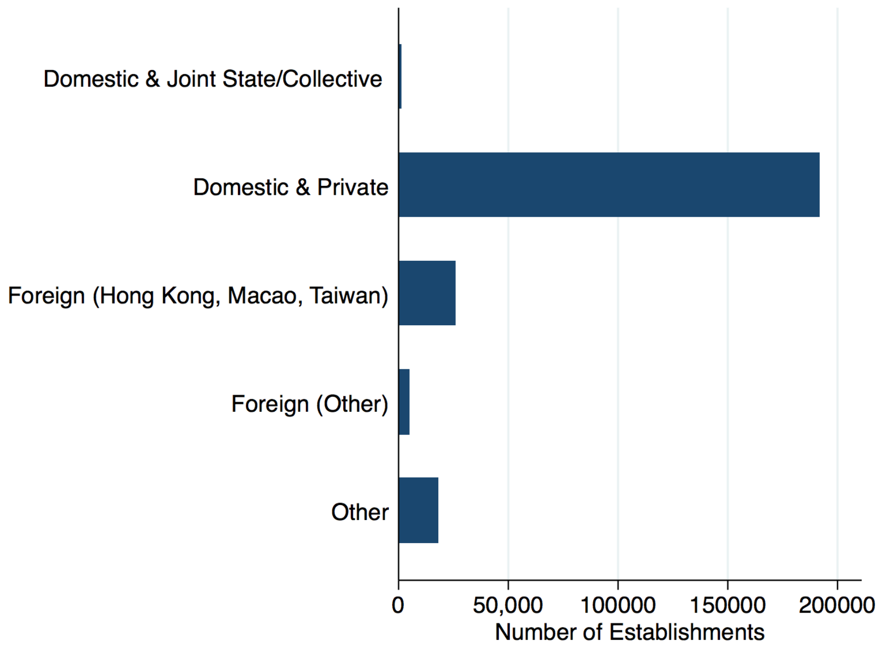


Figure 3.1. Distribution of Firms According to Ownership

Notes: Data from the 2004 Chinese Economic Census.

Profitability (TFPR)

In addition to analysing the effect of better access to the American Chinese network on exports, we also investigate the effect on a number of other firm variables. While most of these are easily observed in the Annual Survey of Industrial Enterprises, we need to construct our own measure of total factor productivity, as described in this section.

We do not have information about quantities produced, and instead rely on revenue data; thus, we are unable to accurately measure total factor productivity and instead are limited to estimating total *revenue* productivity. Revenue productivity however is only an imperfect approximation of actual total factor productivity: firms can have a high factor productivity because they are more efficient, or because they have higher idiosyncratic demand and thus are able to charge higher prices.

We model production with a simple Cobb-Douglas production function:

$$Y_{f,i,c} = A_{f,i,c} K_{f,i,c}^{\alpha_i^k} L_{f,i,c}^{\alpha_i^l} M_{f,i,c}^{\alpha_i^i} \tag{3.1}$$

Where $Y_{f,i,c}$ is output produced by firm f in industry i in county c, $K_{f,i,c}$ is capital, $L_{f,i,c}$ is labour and $M_{f,i,c}$ are all intermediate products used in production. We then follow the Foster et al. (2008) and calculate firm level total factor productivity as:

$$tfpr_{f,i,c} = y_{f,i,c} - \alpha_i^k k_{f,i,c} - \alpha_i^l l_{f,i,c} - \alpha_i^i m_{f,i,c} \tag{3.2}$$

where the lower-case letters indicate logarithms of establishment-level TFPR, value of output, labor inputs, capital stock and intermediate inputs respectively, and $\alpha_j (j = \{l, k, m\})$ are the factor elasticities for the relevant inputs. Output is measured as total value of output produced; capital is measured as the net value of all fixed assets, while labour input is measured as the number of employees. Intermediate inputs are measured as the total value of all intermediate industrial inputs.

It is important to note that the amount of inputs used is clearly endogenous to the firm specific productivity level. Thus we cannot identify the input elasticities for each firm. Instead we use average cost shares of inputs in each industry to calculate the elasticities. The cost shares for labour and materials are calculated using expenditure from the survey at the industry level, while the cost share of capital is calculated as the average amount of capital stocks times the respective capital rental rates for each firm's corresponding two digit industry, taken from US data in the absence such data being available for China.

3.5.2 Cultural Exposure Measure

The information about language distribution in Southern China is obtained from a coding scheme developed on the basis of the *Language Atlas of China*, which was compiled jointly by the Australian Academy of the Humanities and the Chinese Academy of Social Sciences in 1987. The Language Atlas organises the language varieties into hierarchical subgroups: phyla (e.g. Sino-Tibetan), stocks (e.g. Sinetic), super-groups (e.g. Mandarin), groups, such as Yue or Hakka languages, and subgroups. For a subset of locations also clusters and local dialects are recorded. The data we use in this paper has been encoded by Lavely (2001) at the county level, such that it provides for each county on the Chinese mainland the five most widely spoken language groups and sub-groups. Lavely (2001) defines language groups on the basis of whether they are mutually intelligible, while sub-groups may be mutually understandable, though this is not always the case. The coding scheme limits itself to Sinetic languages only and thus only applies to Han Chinese languages. Table 3.3 lists the most common Sinetic language groups spoken in China as well as the percentage of counties they are spoken in. The main language group of interest, Yue (referred to as Cantonese), is highlighted in bold. Guangdong is dominated by three different language groups: Cantonese, Hakka and Min.

Language Group	% of Counties (all)	% of Counties (Guangdong)
Mandarin	55.49	0.81
Wu	7.08	0
Jin	5.90	0
Min	5.69	19.51
Yue (Cantonese)	**4.85**	**52.85**
Gan	3.20	0.81
Hakka	1.95	25.20

Table 3.3. Distribution of The Most Common Language Groups Across Counties

Notes: % of Counties displays the percentage of counties (of either total or Guangdong) where the language group in question is the dominant language group spoken. Source: Lavely (2001).

In this work, we are investigating the effect of having better access to the American-Cantonese network, which was formed in the late 19th and early 20th century. As explained in section 3.4.1, this network's native language is two subgroups of the Cantonese language group: Siyi and Sanyi. Thus, our main measure of access to the American-Cantonese network, which measures whether the firm is located in one of the sending counties, coincides with whether a county's dominant language is either of the two dialects. However, we use the distribution of the Cantonese language group across counties as a robustness check. Figure 3.2 shows the geographical distribution of both the two dialects of interest (Siyi and Sanyi) as well as the distribution of Cantonese.

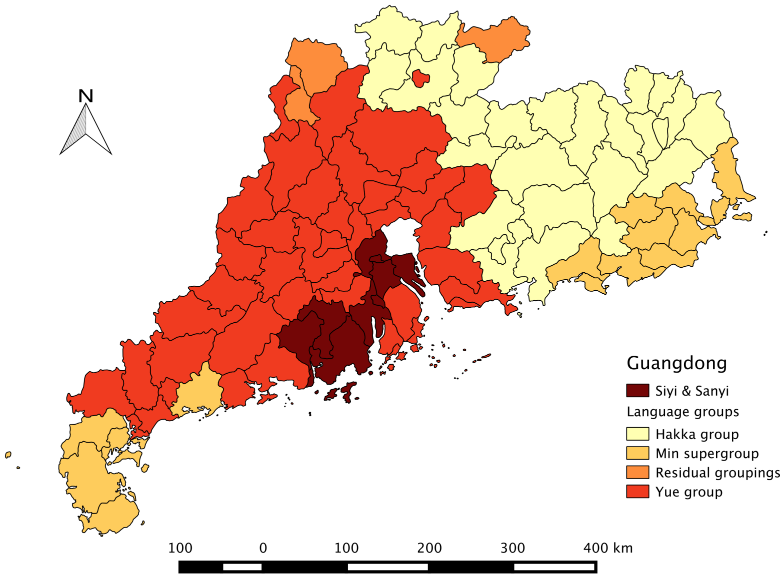


Figure 3.2. Distribution of Languages in Guangdong

Notes: Data from Lavely (2001).

3.5.3 Industry Exposure: Cantonese Workers by Industry in the US

We obtain the data on the number of workers Cantonese from the 5 percent sample of the 2000 Population Census, which gives us information at the four digit industry level. We consider only workers who claim to be of Cantonese origin, speak Cantonese at home and were not born in China. We impose the first two restrictions to accurately measure the network that we are interested in, i.e. the descendants of the early Cantonese migrants in the late 19th and early 20th century. Further, we exclude individuals who were born in China in order to mitigate the endogeneity concerns caused by possible selective settlements of more recent migrants. This means that we reduce the possibility that our effects are driven by Cantonese workers moving to the U.S. in order to work in industry that are particularly strong in their place of origin in China. Note that we use the absolute number of workers and not percentages, because our aim is to capture the probability that a given firm in industry i that has access to the Cantonese network has at least one contact in the same (or downstream) industry.

Figure 3.3 shows the number of Cantonese workers across industrial sectors according to the U.S. census 2000 industry classification. The Cantonese workers are distributed across a wide range of sectors. The sectors with the largest number of Cantonese workers are education, health and social services, manufacturing, professional services and retail trade, while relatively few Cantonese work in fishing, mining, utilities and the armed forces.

Tables B.1.3 and B.1.4 show the ten 6 digit NAICS industries with the highest and lowest number of Cantonese workers. Table B.1.3 reassures us that we are capturing a wide range of industries, which employ both high skill workers (e.g. in universities, IT related manufacturing and services) as well as low skill workers (e.g. in restaurants, grocery stores).

3.5.4 Linking Industries in China to the U.S.

In this paper, we suppose that a Chinese firm f in industry i can be exposed to the Cantonese U.S. network in three different ways: through American-Cantonese workers employed in the same industry i (referred to as "same industry"), through workers employed in a retail or wholesale sector that sell goods produced by industry i ("retail and wholesale") and finally though workers active in other downstream manufacturing sectors

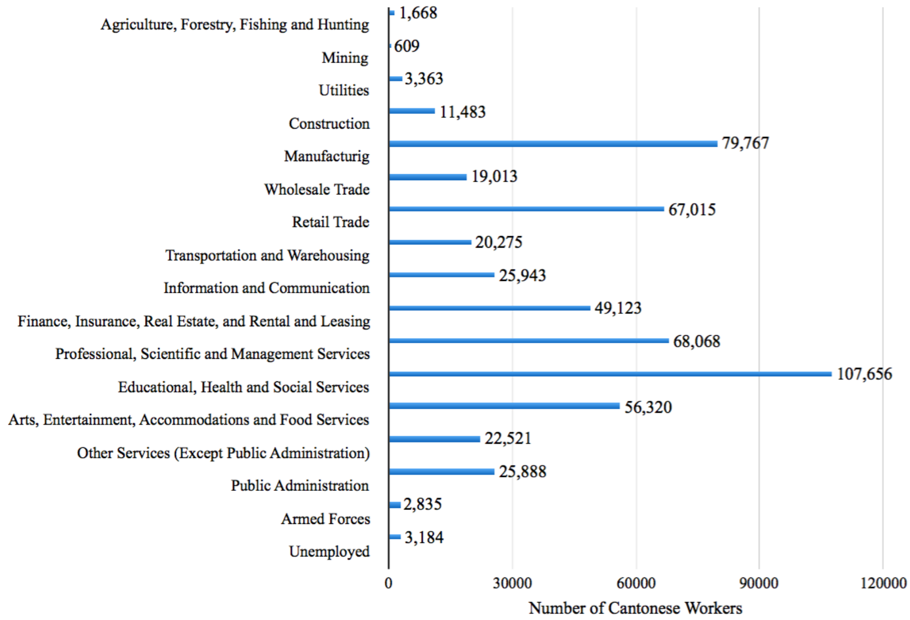


Figure 3.3. Distribution of Cantonese Workers Across Sectors

Notes: This graph shows the distribution of ethnic Cantonese workers by sector in the U.S. in 2000. Data from the U.S. Census.

that use industry i's products as intermediate inputs ("manufacturing").

Calculating the number of ethnically Cantonese American workers employed in the *same industry* is relatively straightforward. The main challenge is converting the industry classification used in the 2000 U.S. Census to the classifications used in our Chinese data. We start by converting the U.S. data to the International Standard Industrial Classification (ISIC), which allows us to identify industries up to the four-digit level, using conversion tables provided by the United States Census Bureau. However, while the Chinese industry classification scheme is modeled after the ISIC system, it is not identical and, to our knowledge, there is no official conversion. Thus we develop our own conversion scheme, by matching industries by their description. Despite also being reported at the four-digit level, the number of industries in the Chinese classification is larger than in the ISIC. Overall, we are able to match 687 out of 709 Chinese industries to 189 ISIC industries.

In order to calculate our second measure of industry exposure, the number of workers in a downstream retail or wholesale industry, we need to link Chinese industries to American retail industries. Ideally, we would make use of information how much each American retail or wholesale industry imports from each Chinese industry. As we do not have access to this information, we instead match manufacturing to retail and wholesale industries using industry descriptions. For more details on this process see Appendix B.3.

Thirdly, we match industries in China with American manufacturing industries that are likely to use the Cantonese firms' products as intermediate inputs. In order to establish which firm purchases which inputs and thus linking Chinese industries to related downstream industries, we use the information provided by the Input-Output tables compiled by the Bureau of Economic Analysis for the year 2002. These tables give detailed information of the inputs purchased for each six digit industry according to the North American Industry Classification. We thus calculate for each industry the weighted average of Cantonese workers that purchases its products according to the share of total sales of industry *i*.

Table B.1.2 shows the average exposure to our three measures at the industry level as well as the cultural exposure to the network.

3.6 Empirical Strategy

Our empirical strategy exploits the interaction between two measures of network exposure, one at the industry level and the other at the geographic level.

The treatment variable for all regressions is *network exposure*, which measures the exposure or access of each firm to the network of overseas ethnic Cantonese in the United States at the county industry level. In our framework, a firm can be potentially connected along two dimensions: an industry dimension and a cultural dimension. Thus our final measure is the interaction of industry exposure and cultural exposure:

$$network\, exposure_{i,c} = industry\ exposure_i \times cultural\ exposure_c \tag{3.3}$$

where i stands for four-digit industry and c stands for Chinese county. Firms are grouped according to the Chinese industry classification system. In China, counties are the third level of administrative divisions, after provinces and prefectures. As of September 2016, there are 2,852 county divisions. In our data set we observe firms in all 123 counties in Guangdong.

3.6.1 Industry Exposure

Industry exposure captures the degree of exposure of a Chinese firm to its ethnic network in the U.S. We analyse three ways in which a Chinese firm is linked to a U.S. industry.

Firstly, we analyse the link between Chinese manufacturing firms and related retailers and wholesalers in the U.S. We assume that, if a Chinese firm in Guangdong operates in an industry that produces, say, leather jackets, it is likely to be more exposed to the network if more workers of Cantonese origin (as the early migrants to the U.S.) work in an industry that sells leather jackets. Thus, if ethnic networks do increase exports for Chinese firms a major effect should be seen between manufacturing firms in China and wholesale and retail firms of the same products in the U.S. We measure this effect by calculating the number of workers of Cantonese origin in the U.S. that work in wholesale and retail industries which are likely to sell products manufactured by Cantonese or Chinese firms:

$$industry\,exposure\,retail_i = number\,of\,Cantonese\,workers_j$$

where i denotes the main 4-digit industry in which a Chinese firm operates and j is a retail or wholesale industry in the U.S. who is likely to buy the final goods produced by industry i.

Secondly, we analyse the link between Chinese firms and related downstream manufacturing industries in the U.S. These are manufacturing industries which use goods of the same category as the ones produced by Chinese firms in their manufacturing process, and are therefore potential buyers of their products. To illustrate the relation of this specific type of network to exports with our previous example, we assume here that a producer of raw or processed leather in China benefits more from the network if more ethnic Cantonese in the U.S. are employed in industries that use leather for production. We measure the effect of this specific network with the number of workers that are employed in related downstream manufacturing industries, which are likely to purchase products such as those produced by the Chinese firm. This measure is calculated as follows:

$$industry\, exposure\, manufacturing_i = number\, of\, Cantonese\, workers_k$$

where k is 4-digit a industry in the U.S. which uses the category of goods produced by industry i in its manufacturing process.

The third measure of industry-level network which we analyse measures the potential benefits for exports of being more connected to the same industry in the U.S. A Chinese firm is considered more connected to its ethnic network if it operates in a four-digit industry that employs a larger number of ethnically Cantonese workers in the U.S. Specifically, this is calculated as:

$$industry\, exposure\, same\, industry_i = number\, of\, Cantonese\, workers_i.$$

In summary, the three definitions of ethnic network exposure at the industry level capture different effects. The first and the second measure have a direct effect on exports: they focus on the importance of networks in order to sell goods to a foreign country. However, the effect of the third measure, which computes exposure to ethnic networks within the same industry, is both direct and indirect; the indirect effect is related to the fact that networks flowing along the same industry may help knowledge and technology flow to firms across countries. Note that the three different definitions of industry exposure to the Cantonese ethnic network may not necessarily be defined for the same number of industries, as firm may operate in an industry which includes only final goods, or only intermediate goods. Conversely, an firm can operate in an industry which includes both intermediate and final goods.

3.6.2 Cultural Exposure

The other dimension according to which a firm can be connected is its cultural similarity with the network of ethnic Cantonese workers in the U.S. This relies on the assumption that a firm located in a county in China whose population shares close cultural ties with early immigrants in the U.S. has better access to the American Chinese network.

To identify cultural closeness, we exploit the fact that the emigration to the U.S. until the 1980's was extremely localised around the southern coast of Guangdong, whose population mainly spoke, and still speaks, a particular dialect of Cantonese.[2] We define *cultural exposure* as a dummy variable indicating whether a firm is located in one of the sending counties of migrants to the U.S. in the 19th century.

With this measure of cultural exposure we are aiming at estimating the effect of both language similarity and kinship ties. It captures the effect of language because the linguistic landscape of South China is very diverse, such that even within few kilometers the local dialects are mutually unintelligible; therefore speaking the same or a similar language as the diaspora communities can improve the interaction between the two differ-

[2]Sections 4.1 and 3.5.2 give a detailed overview of the migration wave from Guangdong to the U.S. and the language landscape in Guangdong.

ent groups. While nowadays communication between mainland Chinese of all areas of origin is facilitated by the near universal knowledge of Mandarin Chinese, many American-Cantonese emigrated before the teaching of Mandarin became widespread.[3]

In addition, our measure captures the effect of networks developed through kinship ties because we focus on the exact counties where the early migrants to the U.S. originated from. Even after generations, it is possible that early migrants will still have ties with the relatives left behind in China. Moreover, anecdotal evidence shows that family ties are still an important component of ethnic networks, particularly in China (Gomez and Cheung, 2009).

As a robustness check, we use an alternative measure of cultural exposure: a dummy variable indicating whether a Chinese firm is located in a county where the main language spoken is Cantonese. Comparing the results when using these two different measures can shed some light on the relative importance of language versus kinship ties for international trade.

3.6.3 Difference-in-Difference Regression

We run difference-in-difference style regressions with a continuous variable (industry exposure) interacted with a dummy variable (cultural expo-

[3]That language still plays an important role in China today is demonstrated by for example Chen et al. (2014), who show that immigrants in Shanghai that are more likely to speak and understand the local dialect are more successful on a number of dimensions: for example they are more likely to be self-employed and have higher hourly earnings.

sure) of the following form:

$$y_{f,i,c} = \beta_0 + \beta_1 network\, exposure_{i,c} + \beta_2 age_{f,i,c} + \theta_i + \pi_c + \epsilon_{f,i,c} \qquad (3.4)$$

where f indicates firm, i indicates 4-digit industry, c indicates county in China, y is a firm-level outcome of interest, *age* indicates the age of the firm, θ and π are four-digit industry and Chinese county fixed effects. As our specification does not allow to control for industry-county fixed effects, we interact county controls with industry fixed effects as a robustness check, where the county controls included are dummy variables indicating whether average education, migration and total population of the individual counties are below (=0) or above (=1) median.

For most of our analysis we restrict to privately owned, domestic firms, but we also investigate how results change when we do not apply these restrictions: we run robustness checks where we include foreign owned firms, controlling for the amount of foreign capital of each firm. We cluster standard errors at the four-digit industry level, according to the Chinese industry classification. It should be noted that throughout the whole analysis we rely on cross-sectional data, as we have firm level information about Chinese firms in year 2004 only.

Note that both the industry and the geographic exposure, if interpreted individually, might be endogenous; thus, we focus on their interaction. As we essentially conduct a difference-in-difference analysis across industries and counties with different levels of exposure, our identifying assumption is that industries in China do not differ systematically across

Cantonese and non-Cantonese speaking areas.

Although we analyse the effect of the Cantonese network on exports, there may be network effects on imports from the U.S. to China as well. We focus on exports rather than on bilateral trade mainly because our data set does not contain firm level information about imports. However, it is important to note that the effect of the ethnic networks may be larger when considering bilateral trade rather than focusing on exports only.

3.6.4 Identification Concerns

This identification strategy may give rise to concerns related to the potential endogeneity of the network we estimate. Firstly, endogeneity may occur if more recent migrants decide to migrate to the U.S. to work in specific industries. This could occur if ethnic Cantonese migrants connected to certain industries decided to move to the U.S. in order to create export possibilities. As a result, those recent migrants would most likely work in the same industry, or related related downstream industries, in the U.S. While this is a potential issue, we mitigate the selective migration problem by excluding from our measure of industry exposure those ethnic Cantonese workers who are not born in the U.S. Given that emigration from China to the U.S. in recent times was very limited between 1949 and 1977, this strategy should rule out most of the recent migrants.

A further concern is that our result may be driven by non-random allocation of Cantonese workers across industries due to ethnic Cantonese people being skilled in particular industries; for this reason, these industries would thrive in both countries. If one assumes that those skills can persist for several generations, this potential issue would be stronger for the measure of industry network exposure based on Cantonese workers in the same industry rather than for the two other measures, which are based on connections across industries. Moreover, if this were happening our estimates are likely to be biased downwards, as part of the control group (the Cantonese speaking counties) would be affected by the network. As an indirect test, we estimate equation (3.4) restricting the sample to firms located in Cantonese counties, thus comparing the sending counties to other counties within the Cantonese area of Guangdong. Moreover, adding Chinese county controls interacted with fixed effects should to some extent take care of county-specific factors affecting exports in different industries.

Additionally, one may worry that the sending counties have different characteristics compared to the other counties in the province, or even compared to other counties within the Cantonese speaking area. In fact, the province of Guangdong hosts few "special economic zones" which are aimed at promoting trade; one of this is the largest city and port, Guangzhou. Although potential differences cannot be excluded, it is important to note that none of the special economic zones are located within the sending counties. Moreover, the county of Guangzhou is excluded from the analysis.

One remaining concern is the possibility that American firms choose to hire ethnic Cantonese workers because of the influence, or pressure, of their already existing trading partners in the sending counties or in Cantonese areas, because of their kinship ties with Cantonese workers in the U.S. In this case, we would still be capturing a network effect, but with reverse causality: overseas networks in China would affect the hiring choices of American firms in related industries. Although we cannot completely rule out this concern, it is unlikely that this channel would explain the entire differential in export flow from Cantonese firms to the U.S.

3.7 Results: Exports

3.7.1 Baseline Results

	Retail and Wholesale		Manufacturing		Same Industry	
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.045*	.029**				
	(.026)	(.012)				
Workers Related Retail	.094***					
	(.013)					
Network Downstream Manufacturing			.057***	.030***		
			(.008)	(.004)		
Workers Downstream Manufacturing			.039***			
			(.007)			
Network Same Industry					.061***	.030***
					(.017)	(.009)
Workers Same Industry					.042**	
					(.017)	
Sending Counties	-.177***		-.184***		-.175***	
	(.022)		(.019)		(.019)	
Age	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	Y	N	Y
County FE	N	Y	N	Y	N	Y
N of observations	22,413	22,413	24,069	24,069	24,043	24,043
R-Squared	0.0762	0.2878	0.0458	0.2927	0.0443	0.2918

Table 3.4. Extensive Margin: Probability of Exporting

Notes: The dependent variable is a binary variable indicating firms export status. Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The variables denoted as *workers* refers to the number of Cantonese workers in each 4-digit industries in the U.S. The coefficients of the variables *network* and *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

In this section we describe the main results, obtained by estimating equation (3.4). Table 3.4 shows the results when regressing export status on the three different measures of network exposure, estimated with linear probability model. Columns 1, 3 and 5, show the estimates without Chinese county and 4-digit industry fixed effects; we include them in columns 2, 4 and 6. For the variables measuring network and industry exposure, we report coefficients indicating the average increase in the dependent variable corresponding to a one standard deviation increase in the independent variable. First, notice that industry exposure, whether measured as the number of Cantonese workers in related retail, downstream manufacturing or same industry, is positively correlated with the probability of the firm being an exporter. Being located in one of the sending counties of migrants, however, is associated with a 17-18 percentage point decrease in the probability of exporting, which reflects underlying differences in factors affecting firms' export status across counties. However, we are mainly interested in the coefficients of the network exposure variables, which we define as the interaction of the two dimensions.

All of our measures of ethnic network exposure have a positive and significant effect on the probability of exporting. With fixed effects, a one standard deviation increase in exposure to retail and wholesale network is associated with a 2.9 percentage point increase in the probability of being an exporter, whereas exposure to downstream manufacturing network and same industry network are both associated with a 3 percentage point increase.[4]

Similarly, Table 3.5 shows the results on export value conditional on being an exporter. Here our measures of industry exposure are not consistently positive and significant, but being located in one of the sending

[4]Note that when estimating equation (3.4) with industry and county fixed effects, the coefficients of the variables *industry exposure* and *cultural exposure* are not estimated, as they vary at the industry and county level respectively.

	Retail and Wholesale		Manufacturing		Same Industry	
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.148***	.138***				
	(.041)	(.037)				
Workers Related Retail	.050**					
	(.024)					
Network Downstream Manufacturing			.086***	.057***		
			(.021)	(.020)		
Workers Downstream Manufacturing			-.019			
			(.020)			
Network Same Industry					.052	.079**
					(.051)	(.032)
Workers Same Industry					.054	
					(.039)	
Sending Counties	-.328***		-.250***		-.210***	
	(.094)		(.079)		(.080)	
Age	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	Y	N	Y	Y
County FE	N	Y	Y	N	Y	Y
N of observations	11,314	11,314	11,834	11,834	11,834	11,834
R-Squared	0.0185	0.1561	0.0146	0.1599	0.0157	0.1599

Table 3.5. Intensive Margin: Export Value

Notes: The dependent variable is log of export value. Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The variables denoted as *workers* refers to the number of Cantonese workers in each 4-digit industries in the U.S. The coefficients of the variables *network* and *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

counties always has a large negative effect on export value. The interaction of the two measures is positive and significant across all specifications except column 5 (without fixed effects), and the effects are large: they range from the 5.7% increase in export value with a one standard deviation increase in exposure to downstream manufacturing network, to 13.8% for the retail and wholesale network.

These results indicate that firms having access to a larger network in the U.S. can gain positive effects in terms of trade, both at the extensive and the intensive margin.

3.7.2 Robustness

In this section we present a set of robustness checks and additional results regarding exports.

Firstly, we explore how the results are affected when we measure cultural exposure including all counties where the most widely spoken language is Cantonese, instead of only the sending counties. Table B.1.5 shows the results: both the retail and downstream manufacturing network have a positive effect, whereas the effect of the same industry network is not statistically different from zero. The magnitude of the effect is smaller, which suggests a stronger effect of the network in the sending counties compared to other Cantonese counties.

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Workers Related Retail×Sending	.022*			.112***		
	(.013)			(.039)		
Workers Downstream Manufacturing×Sending		.025***			.042*	
		(.004)			(.021)	
Workers Same Industry×Sending			.029***			.076**
			(.008)			(.035)
Workers Related Retail×Cantonese	.019***			.061*		
	(.007)			(.032)		
Workers Downstream Manufacturing×Cantonese		.013***			.039**	
		(.004)			(.020)	
Workers Same Industry×Cantonese			.002			.011
			(.010)			(.051)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	22,393	24,041	24,017	11,259	11,777	11,777
R-Squared	0.2874	0.2920	0.2911	0.1505	0.1546	0.1544

Table 3.6. Exports - Heterogeneous Effects for Sending Counties

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables defined as *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

Furthermore, we analyse the presence of heterogeneous effects for firms located in the sending counties compared to firms in the Cantonese speaking part of Guangdong. We conduct this exercise for two reasons. Firstly, to address the identification concern that our results are driven by ethnic Cantonese having a comparative advantage in certain economic activities, due to human capital formation or preferences. Secondly, this allows us to distinguish whether the main driver of the results is the ethnic and linguistic similarity, which would be shared by all ethnic Cantonese, or kinship ties, which would only link the American-Cantonese to the sending counties. The results are shown in Table 3.6. We are able to test for heterogeneous effects by including in one regression the two network exposure variables, calculated with the measure of cultural exposure being either a Cantonese or a sending county dummy. Note that as the sending counties are all within the Cantonese speaking part of Guangdong, the network exposure variable constructed with the sending county dummy captures the differential effect. The results show that the differential effect for firms in the sending counties is always positive and significant, which indicates that firms located there benefit additionally from the network compared to the other Cantonese firms. We do however, also find a positive effect for firms in Cantonese counties for two of the network measures (downstream manufacturing and related retail and wholesale industries), though smaller in magnitude. This suggests that firms sharing a similar language and ethnicity with the U.S. network can experience beneficial effects even in the absence of close kinship ties.

The possibility that industry-specific skills of ethnic Cantonese (or of people originating from the sending counties) might drive the results would be further reduced if we could include industry-county fixed effects in our regressions; these would capture specific effects of each industry within a county. Although our specification does not allow it, we control for county characteristics interacted with industry fixed effects. The results are shown in Table B.1.6, and they indicate that Chinese county characteristics are not driving our results: the size of the coefficient is larger compared to the baseline specification.

The export data at the firm level measures total value of exports; we do not have information about the destination of those goods. In order to have a more precise estimate of how networks affect export to the U.S., rather than to any country, we weigh export value by the share of total industry exports which is imported by the U.S. We would expect larger coefficients if those industries which are more connected according to our industry exposure variable also sell a larger share of the total exports to the U.S., because the export value of firms in those industries would be multiplied by a larger share. The results of these regressions are show in Table B.1.8 Note that the coefficients are positive and significant, and they are also larger in magnitude compared to the baseline specification (columns 1, 3 and 5) and compared to Table B.1.6 (columns 2, 4 and 6).

We address one additional concern, namely the possibility that our measures of industry exposure identify spurious relationships rather than network effects. This would occur if larger industries were also employing a higher number of Cantonese workers; therefore the higher trade flows would be a result of the size of the industry, and not of the network. To tackle this issue we control for the number of total workers by industry (net of the ethnic Cantonese) interacted with cultural exposure. Table B.1.9 shows the results: the coefficients on our ethnic network exposure variables remain positive and statistically significant, with the exception of the effect of the retail network on the extensive margin, which is not statistically significant at the conventional levels (the p-value is .101).

Finally, we analyse foreign firms further by examining the effect of networks only on exports of foreign owned firms. Table B.1.13 shows that networks increase the value of exports and leave the probability of exporting unaffected. The size of coefficients for all three measures (larger than those for domestic firms) indicates a potential effect of foreign investment on the intensive margin; however, there does not seem to be an effect on the extensive margin. This suggests that foreign investment does not help overcoming fixed costs associated with trade, which should predominantly affect the probability of exporting.

3.7.3 Channels

Information vs. Contract Enforcement

As discussed in section 3.3, ethnic networks can facilitate trade by reducing information barriers or by improving contract enforcement. In order to further identify the relevant channels at play, we conduct the exercise suggested by Rauch and Trindade (2002) and investigate whether there are heterogeneous effects for differentiated goods. Rauch and Trindade (2002) distinguish between differentiated goods and goods with a reference price or that are traded on central markets. They argue that informational barriers should play a more important role for the decision to export differentiated goods, as for these goods - in contrast to the reference priced goods - the price is not a sufficient indication of the profitability of exporting. Networks might reduce informational barriers by sharing knowledge about consumer tastes or by matching buyers and sellers, both of which are more difficult to ascertain for Chinese firms in the case of differentiated goods. On the other hand, they claim that contract enforcement barriers should have the same role for both types of goods. Rauch and Trindade (2002) do, however, admit that there is an important caveat to this approach: the complex nature of differentiated goods means that aspects such as quality might be non-contractible. Thus the ethnic network might play a bigger role by ensuring that informal contracts are enforced. Furthermore, particularly in the intermediate goods sector, differentiated goods are more likely to be customised for each buyer and therefore hold-up problems are more severe.

	Probability of Exporting			Log Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	-.007			.281		
	(.022)			(.207)		
Network Related Retail × Differentiated	.043*			-.142		
	(.025)			(.211)		
Network Downstream Manufacturing		.005			.058***	
		(.039)			(.021)	
Network Downstream Manufacturing × Differentiated		.023			.036**	
		(.040)			(.016)	
Network Same Industry			-.030			.130
			(.024)			(.367)
Network Same Industry × Differentiated			.060**			-.038
			(.026)			(.368)
Sending × Differentiated	.004	.014	.004	.203	.157	.176
	(.024)	(.029)	(.024)	(.244)	(.139)	(.183)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	25,650	27,473	27,449	13,947	14,576	14,576
R-Squared	0.2751	0.2813	0.2805	0.1446	0.1473	0.1472

Table 3.7. Heterogeneous Effects for Differentiated Goods

Notes: Differentiated goods are identified using the classification scheme published by Rauch (1999), using the conservative classification. Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

We use the classification published by Rauch (1999) to identify industries that produce differentiated products; for our analysis we choose the most conservative classification. Table 3.7 shows the results for the three measures of network exposure on both the probability of exporting and the value of exports. Column (1) shows that in related retail industries, the effect of network exposure on the probability of exporting is driven by differentiated goods. However, we do not find such differential effect on the value of exports in column (4). In the case of exposure to the same industry network a similar pattern can be discerned (columns (3) and (6)). However, when we consider the network of downstream manufacturing firms, we observe the opposite pattern, i.e. no additional effect on the probability of exporting, but a large positive additional effect on the value of exports of around 3.6%. These findings indicate that overall there is a larger effect for differentiated goods, suggesting that the information channel may play a dominant role.

This is further supported by the fact that we find a similarly sized effect for all three measures of network exposure. In fact, while information could flow equally through the network in downstream manufacturing, related retail and wholesale industries as well as the same industry, the network in downstream industries (retail and manufacturing) should have a stronger effect if contract enforcement issues are driving our results. The fact that we find no such difference supports the interpretation of information flows being a main driver of the results.

Technological Knowledge Flows

Another channel highlighted in the literature is that networks facilitate the flow of technological knowledge. While this is clearly related to what we referred to as "information flows" in section 3.7.3, it has been often regarded as a separate channel, as it focuses on information about technology and production processes that could be more easily shared through the network. Even though technological knowledge is not directly related to trade, it could nevertheless be an important driver of our results.

We investigate this channel in two ways. We firstly analyse directly the effect of network exposure on total factor profitability (TFPR), which is the product of prices and total factor productivity (TFP). The rationale is the following: If technological knowledge is being transmitted from the U.S. to China, we should see a positive effect on firms' TFP. Although we do not observe this directly, as long as prices do not decrease more for exposed firms, this positive effect should translate into a higher TFPR. Table B.1.11 shows the results for all domestic firms, as well as for exporting and non-exporting firms. Columns (1) to (3) show that neither of our three measures of network exposure have a significant effect on profitability. As industry-specific technological knowledge flows are likely to be most affected by access to the overseas Chinese network in the *same* industry, we would expect this measure to have a larger effect. However, column (3) shows no significant effect. When analysing exporting and non-exporting firms separately (columns (5) and (6)), we find instead that average TFPR is decreased for exporting firms with greater access to the network in downstream manufacturing industries and in the same industry; the corresponding effects on TFPR are 1.6% and 2.0%. The coefficients for non-exporting firms are positive but not statistically significant. While these results seem counter-intuitive, they are consistent with the hypothesis that networks decrease the costs associated with exporting. If this were the case, we would expect the cut-off productivity level for which firms find it profitable to start exporting to be lower for connected than non-connected firms. Thus, in connected industry-county combinations, less productive firms would select into exporting, resulting in lower profitability for exporting firms. A similar effect might also exist for firms entry decisions. As we only observe private firms that have output greater than 5 million RMB, it is possible that despite a positive effect on a given firm's TFP, less productive firms are able to overcome the size threshold and thus average TFPR remains unchanged. We investigate this further in section 3.8.2, where we analyse the effect on labour productivity for the universe of firms in Guangdong.

As a second exercise, we explore whether the effects are larger for high-tech industries, where the technological gap between American and Chinese firms is the largest, and thus knowledge flows through the network are likely to be strongest. We classify industries according to the technology intensity definition of the OECD, which is based largely on R&D spending intensity. Table B.1.12, shows no additional effect for high tech industries.

Overall, we find little evidence that technological knowledge flows are driving our results, though we are unable to establish whether they have an additional effect on connected firms. Instead, our results are consistent with networks reducing the profitability threshold of exporting.

Foreign Investment

In this section we investigate whether foreign investment has an effect on exports. As outlined in section 3.3, ethnic networks might encourage foreign direct investment (FDI). This might itself cause firms to export more, or increase knowledge transfers. In this case, we should expect the observed increase in exports to be (at least partly) explained by an increase in FDI. Although our main analysis focuses on domestically owned firms, we are also interested in understanding whether the increase in exports for other firms is driven by FDI.

We explore this question by analysing all privately owned firms while controlling for the share of total equity that comes from a foreign source. Table B.1.7 shows the results on both the probability of exporting in columns (1) to (3) as well as on the logarithm of the value of exports in columns (4) to (6). The share of foreign capital itself has a positive effect on both the probability of exporting and the value of exports conditional on the firm being an exporter, documenting a positive relationship between foreign ownership and exporting. However, the coefficients of all three measures of network exposure are similar in magnitude to the baseline model and highly significant, indicating that the positive effect of higher exposure to the network cannot be explained by an increase in foreign capital share.

Moreover, we regress the share of foreign capital on network exposure (figures B.2.1 and B.2.2), which is discussed in more detail in section 3.8. We can only observe a positive effect on non-exporting firms, further indicating that while networks might increase foreign direct investment, this is not the most important channel.

3.8 Results: Other Firm Variables

3.8.1 Survey of Industrial Enterprises

In this section, we analyse how exposure to the American Chinese network affects other firm variables in addition to exports.

Figure B.2.1 shows the effect of our three measures of network exposure on total output, domestic sales, profits, fixed assets, capital and the share of capital that is foreign owned. While not all measures have a significant effect, they nevertheless follow the same pattern. There is a positive effect on output of 4.9% -5.8% and an increase in profits of 4.3-5.5%. Similarly, fixed assets increase by 5.6-7.1% and total capital by 8.2-10.7%. On the other hand, domestic sales seem to decrease: both the retail and same industry networks have a large negative effect, of around 13.1 to 15.6%. The effect of the downstream manufacturing network is similar in magnitude, but the estimate displays very large standard errors. The negative effect on domestic sales, together with a positive effect on exports, provide suggestive evidence of specialization of connected firms in exportable varieties, which are less demanded in the domestic market. The effect on the share of capital that is foreign owned is not statistically significant.

Figure B.2.2 splits the sample into exporting and non-exporting firms and shows that most of the effects are driven by exporting firms. It is important to keep in mind that exposure to the overseas network affects selection of firms into exporting, and therefore these results cannot be interpreted as causal. Output, profits, fixed assets and total capital of connected exporting firms are higher. The decrease in domestic sales is very large: a one standard deviation increase in our measure of exposure to the network in related retail and wholesale industries reduces domestic sales by 26.1%. The fact that the effect is the largest for retail and wholesale exposure is consistent with Chinese firms specialising in goods that appeal to American consumers but are not tailored to Chinese consumers' tastes. We can also observe that the effect for non-exporting firms is close to zero or negative in terms of output and domestic sales, despite being positive (although not always statistically significant) in terms of fixed assets and total capital. This could be caused by compositional effects, as access to the network may reduce the average productivity of non-exporting firms by driving more firms to export. Finally, we also observe a positive effect on the share of foreign capital of non-exporting firms.

We further investigate how our measures of exposure affect variables regarding employment. Figure B.2.3 shows that while total employment is unaffected, both the share of high skilled workers as a percentage of the total workforce and management expenses are higher for connected firms; the size is around 1 percentage point for the first and up to 8% for the latter. Further, wages per worker are between 1.1 and 1.6% higher. Figure B.2.4 shows that the effect is again nearly exclusively driven by exporting firms. These findings are consistent with connected exporting firms specialising in products of higher quality or with more complex production processes, which require higher skilled workers. However, higher management expenses could also be explained by the fact that exporting requires a greater number of managers as operations become more complex. Non-exporting firms exhibit a very different pattern on the other hand. While most effects are not statistically significant from zero, exposure to the network in downstream manufacturing industries is associated with a decrease in employment and total wages, while exposure to the network in the same industry is associated with an increase of 1 percentage point in the share of high skilled workers. This may be suggestive of knowledge diffusion about production processes through the network, which allows firms to complete more complex operative processes and which in turn require more high skilled workers.

3.8.2 Economic Census Data

In addition to analysing the effect on other firm variables for large manufacturing firms, we analyse data from the Economic Census of China to document the effects on small firms and firms in other sectors. Table B.1.14 shows the effect on the log number of employees by sector. Note that while we can construct a measure of network exposure within the same industry for all industries in our sample, we are only able to calculate the number of workers in related retail and downstream manufacturing industries for manufacturing plants, as the other industries do not produce tradable goods. Column (1) shows that a one standard deviation increase in exposure to the network in the same industry increases employment for all firms by 2.7%. Column (4), (5) and (6) display the effect for manufacturing, services and retail firms respectively. The effect is significant only for the manufacturing sector, in which it is an around 7.5% increase. Column (2) and (3) show that our other measures of network have a similar effect on manufacturing firms. Figure B.2.5 delves deeper into the effect of network exposure on the size distribution of firms, and shows that the main increase in average size appears to be driven by a decrease in the share of firms with less than 20 workers and an increase in the share of firms that have between 20-50 workers.

Furthermore, the effects on revenues per worker are shown in Table B.1.15. Here, we find no effect of an increased exposure to the retail and downstream manufacturing industry network. We do however observe a negative overall effect of increased access to the network in the same industry in the US, and column (4) shows a positive effect for manufacturing firms: a one standard deviation increase in the measure of exposure increases revenue per worker by 3.5%. However, this effect is evened out by the negative effect on firms in the retail sector, as is shown in column (6).

Figure B.2.6 shows the effect of all three measures on manufacturing firms of different sizes. While exposure to the same industry network increases the revenue per worker for small and medium sized manufacturing firms, exposure to the network in retail and wholesale industries has no significant effects, and in the case of downstream manufacturing, there is only a significant positive effect for firms employing between 50 and 500 workers. While there seems no significant effect for large firms, it is important to keep in mind that the bandwidths in which revenues and employment are noted become less precise at higher levels.

3.9 Conclusion

In this paper, we analyse how exports of Chinese firms are affected by access to the ethnic Chinese network in the U.S. We exploit an historic migration wave that lasted from the mid-19th century until 1949. The localised nature of this migration wave, together with the linguistic heterogeneity of southern China allows us to implement a difference-in-difference strategy with continuous treatment intensity: The main explanatory variable of interest is a dummy variable which equals one if a given firm in Southern China is located in a sending county, i.e. if the firm has a higher "cultural" exposure to the network, interacted with the number of American born Cantonese employed in a related industry in the U.S., which measures the "industry exposure". For the latter, we define 'related industries' in three different ways: firstly, as the same industry, secondly, as related retail and wholesale industries and thirdly, as downstream manufacturing industries.

We find that using all three measures of industry exposure, greater access to the American-Chinese network translates in significantly higher exports, both in terms of the probability of exporting as well as in the value of exports conditional on exporting. Moreover, we find a smaller but significant effect of the interaction of industry exposure with a dummy indicating whether a firm is located in a Cantonese speaking area. The fact that we find results both for the extensive and intensive margin indicates that the effects cannot purely be driven by a reduction in fixed costs. The results remain robust to a number of specifications. Furthermore, we find evidence that the effects are driven by differentiated goods, pointing towards networks alleviating information constraints.

One of the main contributions of paper is to analyse how other firm variables are affected by higher exposure to the ethnic network. We find that total factor productivity remains unchanged, and is even reduced for non-exporting firms, possibly due to a greater mass of firms selecting into exporting. Moreover, we observe that connected firms are larger in terms of output, employment, capital and assets, employ a larger share of high-skilled workers and have higher managerial expenses. However, we also find that domestic sales decrease with higher exposure to the network, indicating overall that connected firms are more likely to specialise in product varieties aimed at the export market but that are not suitable for domestic consumption. In so far that these export varieties are of higher quality and require more complex production processes, this hypothesis could also help explain the increase in the share of high-skilled employees and management expenses.

Chapter 4

Ethnic Chinese Networks and Technology Diffusion: The Chinese Exclusion Act

4.1 Introduction

The benefits of ethnic migrant networks for the origin country of migrants have been studied in the context of a range of economic activities: international trade (Rauch, 2001; Parsons et al., 2014), labour markets (Munshi, 2003), technology diffusion (Griffith et al., 2006; Kerr, 2008). Ethnic networks have been found to be especially important in environments with weak institutions, such as developing countries, as they help overcoming market frictions that might otherwise impede economic transactions. In fact, sharing the same roots, language and culture, can facilitate information flows across countries, improve contract enforcement (Greif, 1989, 1993; Rauch and Trindade, 2002), and promote technology diffusion by overcoming physical barriers such as geographical distance (Keller, 2002*a*). Among ethnic networks, Chinese networks have been shown to be particularly important for promoting innovation (Kerr, 2008), and for information sharing about technology (Saxenian, 2002).

In this paper, I focus on the role of ethnic networks on knowledge flows from country at the technology frontier - the U.S. - to a less developed country - China. I study whether Chinese ethnic networks foster technology flows of Chinese industries, using patent-class level panel data for China and the United States in the time period between 1985 and 2006. I measure knowledge flows as changes in innovation in China which happen in correspondence of changes in innovation in the U.S. I propose a novel instrument to measure ethnic networks, which predicts the number of ethnic Chinese workers in a given industry. This measure exploits restrictions to Chinese migration between the end of the 19th century and the mid of the 20th century - the Chinese Exclusion Act of 1882 - and captures the Chinese ethnic network formed through historic migration waves.

To construct the instrument, I interact the geographic variation in Chinese settlements in the U.S. just after the Chinese Exclusion Act with current industry employment patterns. The main explanatory variable of interest in my analysis is the interaction of the measure of the ethnic network and the number of U.S. patents in a given patent class; I interpret the coefficient of this variable as the effect of knowledge flows from the U.S. to China due to the network. The panel structure of the data allows to control for four-digit patent class fixed effects and time fixed effects; therefore, in my estimation I use variation in patents in the U.S. within industries over time. I measure innovation in China with the number of patents in a given patent class.

The findings suggest a significant increase in innovation in China in industries in which increases in patenting in the U.S. coincides with a larger ethnic Chinese industry network. These results confirm the existence of knowledge flows through ethnic networks. Because the effect of ethnic networks need not be contemporaneous, I aggregate patents into three-year time periods. I do not find evidence of lagged effects beyond the three year time period. Moreover, I investigate whether the effects are driven by high-tech industries. In contrast to the literature (Kerr, 2008), I find no evidence of this when I estimate a reduced form or instrumental variable (IV) equation; however, when measuring the network as the number of ethnic Chinese workers by industries, there is a positive differential effects for high-tech industries.

Ethnic networks have often been identified as the number of ethnic workers in a given country or geographic location (Rauch and Trindade, 2002); however, this measure raises reverse causality concerns. This issue involves the selective location of recent migrants to countries or regions with high prevalence of industries that are particularly advanced in their origin countries. Thus, one is likely to capture the effect of well trained migrants contributing to the development of the receiving country in addition to the network effect for the origin country. A number of recent studies have suggested measures of network based on past ethnic settlements and migration restrictions (Kerr and Lincoln, 2010; Kerr, 2013; Griffith et al., 2006; Parsons et al., 2014), which mitigate potential reverse causality concerns. The main contribution of this paper is to suggest an instrument for ethnic networks at the industry level constructed using historic settlements of over one century before the time period analysed; this reduces the possibility that migrants location decisions are related to current industrial patterns.

The closest strand of literature to this work focuses on the effect of ethnic networks on innovation, where the role of ethnic networks is to help knowledge flow from a technologically more advanced country (typically the U.S.) to a less advanced economy. Data from a recent survey reported by Saxenian (2002) shows that the majority of Chinese and Indian migrants working in the Silicon Valley tend to share information about technology with co-workers in their country of origin, and to enter in partnerships with them (Kerr, 2008). The author consequently highlights a great potential for growth, because these ties can promote entrepreneurship and innovation (Saxenian, 2006). This indicates strong links in terms of information sharing between first generation migrants and their network in the origin country. Kerr (2008) studies international technology diffusion from a frontier country - the United States - to other countries, by matching ethnicity of inventors in the U.S. with innovation in the country of origin; by looking of patent citations in the origin countries, they find that technology diffusion is strongest through ethnic Chinese networks. Similarly, Agrawal et al. (2011) develop a model in which countries benefit through sending skilled individuals to technologically more advanced countries, and estimate it by employing surname matching of Indian inventors with patent data, and using patent citations to measure knowledge. In a related paper, Agrawal et al. (2008) examine the contribution of spatial and social proximity on knowledge diffusion: their result indicate that while both are important, the marginal effect of spatial proximity is higher when individuals are less socially close. In the context of the United Kingdom, Griffith et al. (2006) establish a link between U.K. firms having R&D activities in the U.S. and firms' total factor productivity.

Ethnic networks have also been linked to activities of multinational companies, such as foreign direct investment, in the origin country. An example of this literature related to innovation is the work of Foley and Kerr (2013); they find a positive relation between the ethnicity of inventors working for multinational firms in the U.S. and the level of activity of these firms in the origin country of the inventors. Keller (2002*b*, 2004) highlights the importance of international trade on technology diffusion, focusing in particular on the key role of differentiated intermediate goods.

Besides contributing to the formation of overseas networks, skilled migration can provide substantial benefits in the destination countries, particularly in terms of innovation. Recent examples of this literature are the work by Kerr and Lincoln (2010), who have studied this question by taking advantage of exogenous changes in regulations for the H-1B visa over time to mitigate endogeneity issues, and Moser et al. (2014), who instead overcome endogeneity by analysing the innovation due to Jewish Germans emigrating to the United States during Nazi Germany.

The structure of the paper is as follows: section 4.2 provides information about the historical background on the Chinese Exclusion Act; section 4.3 outlines the empirical strategy; section 4.4 provides a description of the results and finally section 4.5 concludes.

4.2 Chinese Migration to the U.S. and the Exclusion Act

Migration from China to the United States happened in two waves. The first took place during the Gold Rush in the mid of the 19th century. Similarly to migrants from Europe and Central America, Chinese emigrants were drawn to the U.S., and in particular to California, in order to work in gold mines. A second wave corresponded with the timing of the construction of the First Transcontinental Railroad, which was built between 1863 and 1869, and in particular the Central Pacific Railroad. Migration was exacerbated by political unrest (including the Taiping Rebellion between 1850 and 1864) and poor economic conditions in the southern regions of China.

The number of Chinese migrants in this period is very large: whereas the U.S. census of 1890 reports over 100,000 ethnic Chinese throughout the U.S., estimates suggest that between 1850 and 1889 about 300,000 Chinese left China and settled in mainland U.S. and a further 46,000 settled in Hawaii (Voss and Allen, 2008). However, about half of these are thought to have been temporary workers who eventually returned to China. At the time, migration relationships between the two countries were regulated by Burlingame Treaty of 1868, which encouraged emigration from China to the U.S. The two mass migration waves fostered anti-Chinese sentiments throughout the country; consequently, efforts were made by to restrict Chinese migration.

The first response by the U.S. government to restrict Chinese migration was the Page Act of 1875, which forbade the entry of women and forced labourers. Subsequently, the Angell Treaty of 1880 suspended migration of "skilled and unskilled labourers and Chinese employed in mining", but allowed white collar workers to settle in the U.S. These restrictions continued under the Chinese Exclusion Act, which was expected to only be in place for 10 years, but was renewed with the Geary Act of 1992 and subsequently was made permanent until its abolotion in 1943.

The Chinese Exclusion Act was passed in 1882. It was the first major law banning citizens of a particular country to settle in the U.S. Although explicitly impeding migration of labourers, it effectively impeded all kinds of Chinese migration, as to be granted entry in the country one needed to prove her intentions not to become a labourer. During this period, the number of Chinese migrants was limited mostly to family reunifications, and to only 105 persons per year overall (Zhao, 2002).

The Act also affected the ethnic Chinese who already resided in the U.S., as it prevented them to become U.S. citizens. Moreover, due to the continuing anti-Chinese attitudes of the rest of the population, many of the early Chinese migrants generally settled in Chinatowns, of which the largest were San Francisco and New York. In addition, other legislation excluded them from certain occupations by imposing extra taxes of individuals of Chinese origin, particularly in California. This set of anti-immigration laws culminated with the National Origins Act of 1929, which capped immigration to 150,000 people per year and prohibited all immigration from Asia.

Figure C.0.1 shows Chinese settlements in the U.S. in 1890 on the left, just after the Act was passed, compared to the location of Chinese people in 1990. Although the numbers are much larger in 1990, the geographic distribution of ethnic Chinese today is similar to that of the early migrants.

4.2.1 Data

Patents

I use patents as a measure of innovation both in China and the U.S. The data set includes all published utility patents in the U.S. and in China for the time period between 1985 and 2006. The Chinese patent data for this time period is collected by the State Intellectual Property Office of China and has been made available and described by Holmes et al. (2013). The U.S. patent data used in this study comes from the NBER Patent Data Project, originally collected by Hall et al. (2001), which assembles all patent data from the United States from 1976 to 2006. The data is classified according to the International Patent Classification (IPC) at the 6 digit level; however, I aggregate them to the 4-digit level in order to match them with the index of network exposure, which varies at the 4-digit industry level.

Table 4.1 shows the descriptive statistics of the patent data for the two countries over the whole time period, aggregated at the IPC level. I obtain information on 812 different patent classes, including those classes for which one of the two countries record zero patents over the time period analysed. The average number of patents across all classes is larger in the U.S. compared to China.

Exposure to Ethnic Chinese Network

In order to create the index of Chinese employment at the industry level I use data from the US Census, specifically the 1890 U.S. Population Cen-

	Mean	Median	SD	Min	Max	N
			Patents			
U.S. Patents	4919.66	919.5	14068.89	0	183085	812
Chinese Patents	1236.459	221.5	2632.274	0	27738	812
			Ethnic Network			
Index of Chinese Employment	4.266	.304	25.736	0	575.671	812
Chinese Employment by Industry	16.892	10.443	18.950	0	102.509	812

Table 4.1. Descriptive Statistics

Notes: This table shows the descriptive statistics of the main variables used in this paper. I sum the number of patents by IPC class, over the years 1985-2006. The ethnic Chinese employment by industry is for year 1990.

sus and the 1995 U.S. Economic Census. From the first I obtain information about Chinese people living in the U.S. by county; from the second I gather data about employment by four-digit industry by Standard Industrial Classification (SIC) codes in each county. I then convert the employment information from SIC to IPC classification using the concordance developed by Silverman (2003), in order to merge it with the patent data. I use 1995 employment data, which is the median year in the data set, in order to obtain a measure which is valid for the time frame considered. I use microdata from the 1990 5% Population Census available from IPUMS in order to measure ethnic Chinese employment at the industry level in the U.S., because of the lack of availability of industry-level data on ethnic employment in 1995. Following Griffith et al. (2006), I measure the Chinese network in 1990 to avoid capturing Chinese migration corresponding to the U.S. technology boom.

Table 1 shows the summary statistics for the index of Chinese employment based on historical Chinese settlements. The next section explains in detail how this is constructed.

4.3 Empirical Strategy

In this section I describe the empirical strategy used to estimate the effect of Chinese networks on the diffusion of knowledge.

Firstly, I outline the ordinary least squares (OLS) regression. In the following section I describe the index of exposure to the Chinese network and the instrumental variable (IV) regression.

4.3.1 OLS Regression

I use a panel of patent data at the 4-digit IPC class level to analyse the relation between ethnic networks and innovation. I estimate a fixed effect regressions including time and industry fixed effects, therefore exploiting within industry variation in number of patents in the U.S. across time. Because knowledge transfers need not happen contemporaneously, and in order to reduce the chance of zero outcomes in each patent class, I aggregate the number of patents in each class into 3 year periods.

I firstly estimate the following equation with ordinary least squares (OLS):

$$ln(y)_{i,t} = \gamma_0 + \gamma_1 ln(cn\, employment)_i \times ln(x)_{i,t} + \gamma_2 ln(x)_{i,t} + \lambda_i + \xi_t + \nu_i \tag{4.1}$$

where $ln(y)$ is the log of Chinese patents in industry i and three-year time period t; $ln(cn\, employment)$ denotes the log of ethnic Chinese workers by industry in the U.S. in year 1990; $ln(x)$ denotes U.S. patents in industry i and three-year time period t; and λ and ξ indicate industry and three-year time period fixed effects. Note that the coefficient γ_2 is not estimated, as it is collinear with the patent class fixed effects.

4.3.2 Instrumental Variable Regression

The main concern with estimating network effects using the number of ethnic Chinese workers in U.S. industries is that this measure may be subject to reverse causality. This would happen if Chinese migrants were particularly well trained in industries that are innovative in China, and decided to purposely settle in locations in the U.S. where the same innovative industries are prevalent. Thus, by measuring the network with the number of ethnic Chinese workers in a given industry, the estimates are likely to capture the effect of ethnic Chinese contributing to the development of industries in the U.S. in addition to the network effect.

To mitigate reverse causality issues, I construct an index which predicts the number of ethnic Chinese workers in each 4-digit industry in the U.S. This measure exploits the geographic overlay between historical Chinese settlements in 1890 and current location of industries in each county.

This measure takes advantage of the fact that entry of ethnic Chinese to the U.S. was prohibited under the Chinese Exclusion Act, and due to discrimination internal mobility of existing Chinese migrants was limited until the Act was repealed in 1943.

I construct the variable as follows: for each county, I calculate the product of the employment share of a given industry in that county in 1995, multiplied by the number of ethnic Chinese in 1890. This is then summed across counties for each industry to give a single value for each industry. The index of Chinese network is then:

$$Chinese\,Network_i = \sum_c \left(\frac{Employment_{i,c}}{Total\,Employment_c} \times Chinese\,Population\,1890_c \right) \tag{4.2}$$

where *employment* and *total employment* represent the number of workers in county c who are employed in industry i, and the total employment in each county respectively, both measured in 1995. This index is then aggregated at the IPC level, to obtain a measure of the ethnic network which varies at the same level as the patent data. Higher values of this measure imply a higher probability of ethnic Chinese being employed in a given industry nowadays.

As my aim is to obtain an exogenous measure of Chinese employment, I use the geographic overlay of *historic* Chinese settlements and *current* industry location; this is to reduce the chance of capturing the effect of ethnic Chinese purposely settling in a specific county due to the presence of certain industries.

A graphic representation of the correlation between the predicted measure of Chinese employment based on historical settlements and current Chinese employment by industry is depicted in Figure C.0.2.

I then estimate a two-stage least square (2SLS) model where I instrument current Chinese employment with the measure described above. The first and second stage equations are stated below.

First Stage:

$$ln(cn\,employment)_i \times ln(x)_{i,t} = \alpha_0 + \alpha_1 ln(network\,1890)_{i,t} \times ln(y)_{i,t} + \alpha_2 ln(x)_{i,t} + \phi_i + \omega_t + \varepsilon_{i,t} \tag{4.3}$$

Second Stage:

$$ln(y)_{i,t} = \beta_0 + \beta_1 ln(cn\,employ\widehat{ment})_i \times ln(x)_{i,t} + \beta_2 ln(x)_{i,t} + \theta_i + \delta_t + \varepsilon_i \tag{4.4}$$

where $ln(cn\,employ\widehat{ment}) \times ln(x)$ are the predicted values of Chinese employment by industry obtained in the first stage; $ln(network\,1890)$ is the log index of Chinese network described above; ϕ and θ denote 4-digit patent class fixed effects whereas ω and δ denote 3-year time period fixed effects in the first and the second stage regressions respectively.

In addition, I estimate a model which includes lagged values of U.S. patents - $ln(x)_{t-1}$ - instead of contemporaneous values, interacted with the measures of Chinese employment and individually.

4.3.3 Identification Concerns

In spite of mitigating reverse causality, the empirical strategy outlined above raises three main concerns. Firstly, as shown in Figure C.0.1, a large number of Chinese migrants before the Exclusion Act settled along the West Coast and in particular in California, which hosts particularly innovative industries in the U.S. Therefore, the instrument may be simply capturing the fact that early Chinese migrants happened to settle in one of the most innovative regions in the U.S. I tackle this issue with three different approaches. Firstly, in all regressions I exclude the top 1% patent classes in terms of number of patents in both countries, to avoid the results being driven by outliers. Secondly, for robustness I calculate the index of ethnic Chinese network by excluding the migrants who settled in California; similarly, I calculate the number of ethnic Chinese workers net of those who live in California. Thirdly, I investigate whether the effects are driven by high-tech patent classes, which are prevalent on the West Coast. I describe the results in the next section.

The second concern regards the possibility that ethnic Chinese have inherently higher abilities in certain industries, and that these specific skills may lead industries in the U.S. to located in counties with a higher Chinese population. These industries would then outperform others in both countries. Consequently, my empirical strategy might be partly capturing the inherent ability of a certain population in addition to the network effect. Even though I cannot rule out this channel with the aforementioned identification strategy, it should be noted that upon arrival early migrants were mostly employed in manual jobs. Moreover, ethnic Chinese make up a relatively small share of an industry's total employment, and similarly early Chinese settlements only constitute a small share of total population in 1890. Therefore it seems unlikely that the location of the Chinese population would be the decisive factor for the location of industries.

Finally, using a regression model with logarithmic transformations implies that I am able to analyse only patent classes including a positive number of patents in both countries, and for which the measures of network is greater than zero. For the index of ethnic Chinese network, this means excluding industries which are only present in counties with no

Chinese settlements after the Exclusion Act. Although I analyse a selected sample of patent classes, I choose this specification because the distribution of patents and of Chinese workers by industry is positively skewed; thus a regression in levels may capture effects which are driven by patent classes with large numbers of patents or of Chinese workers.

4.4 Results

	OLS	Reduced Form	OLS	Reduced Form
	All Workers		Excluding California	
	(1)	(2)	(3)	(4)
Chinese Employment × Log U.S. Patents	.270***		.273***	
	(.064)		(.064)	
Ethnic Chinese Network × Log U.S. Patents		.058*		.067**
		(.034)		(.034)
Log U.S. Patents	-.541***	.223***	-.503***	.243***
	(.196)	(.057)	(.186)	(.056)
Patent Class FE	Y	Y	Y	Y
3-Year Time FE	Y	Y	Y	Y
N of observations	2736	2736	2736	2736
R-Squared	0.4426	0.4748	0.4411	0.4531

Table 4.2. OLS and Reduced Form

Notes: Standard errors in parenthesis and clustered at the 4-digit patent class level. The dependent variable is log of Chinese patents. ***p>0.01 **p>0.05 *p>0.10.

In this section I describe the effect of ethnic Chinese networks in the U.S. on innovation in China. Firstly, I show the OLS and reduced form estimates; in the latter model the main independent variable of interest is the interaction of log U.S. patents and log of the index of Chinese network (Table 4.2). The results indicate a larger increase in patenting in China corresponding to increases in U.S. patenting in industries which employ a larger number of ethnic Chinese workers: when measuring the ethnic network as the number of Chinese workers by industry, the differential increase in Chinese patenting for a one percent increase in the network is about .3% (column 1). Column 3 shows similar OLS estimates when excluding ethnic Chinese workers in California. The reduced form estimates are smaller in magnitude, but positive and statistically significant: the differential effect is about 0.06% when calculating the measure of ethnic networks including all ethnic Chinese in 1890, and it is larger (0.07%) when excluding Chinese settlements California. Figure C.0.3 illustrates the reduced form effect across patent classes, showing the correlation between U.S. and Chinese patents for classes with network exposure above and below median. For this figure, I aggregate the data at the patent class level, by summing patents across years. The graph displays a higher effect on Chinese innovation for patent classes with higher network exposure.

	All Workers (1)	Excluding California (2)
	Panel A: Two-Stage Least Squares	
Log Chinese Employment × Log U.S. Patents	.461**	.544**
	(.231)	(.222)
Log U.S. Patents	-1.080	-1.226**
	(.661)	(.605)
R-Squared	0.9212	0.9205
	Panel B: First Stage	
Log Chinese Network × Log U.S. Patents	.125***	.123***
	(.036)	(.036)
F-test of Excluded Instruments	11.80	11.66
R-Squared	0.9313	0.9257
Patent Class FE	Y	Y
3-Year Time FE	Y	Y
N of observations	2736	2736

Table 4.3. Instrumental Variable

Notes: Standard errors in parenthesis and clustered at the 4-digit patent class level. The dependent variable is log of Chinese patents. ***p>0.01 **p>0.05 *p>0.10.

Table 4.3 shows the two-stage least squares estimates using the index of Chinese network as an instrument for the number of ethnic Chinese workers. The results confirm the positive effect of the ethnic network on innovation, and the coefficient is larger in magnitude compared to the OLS estimates. This result is counter-intuitive if one expects positively biased OLS estimates induced by simultaneity. Nevertheless, I consider two possible explanations for obtaining IV estimates which are larger than the OLS. The first is the possibility that the index of ethnic network is a weak instrument for ethnic Chinese workers, i.e. that the correlation between the two variables is not very strong. This would cause the IV coefficients to be biased (Bound, Jaeger and Baker, 1995). However, the first stage estimates in Table 4.3 (panel B) indicate that the two measures are correlated, and the F-test of excluded instruments is above the cut-off value of 10 proposed by Stock, Wright and Yogo (2002) for one endogenous regressor. The second explanation involves the local average treatment effect (LATE) - the treatment effect in the subsample which is affected by the instrument - being higher than the average treatment effect. In fact, the "treatment" induced by the instrument, i.e. the access to an ethnic Chinese network composed of descendants of early migrants, may not affect all industries: looking at Figure C.0.1, the location of early settlements appears to be less evenly spread across the U.S. compared to more recent migrants. A larger LATE would occur if the instrument affects industries which benefit more from the network; this could happen if in these industries information were more easily spread across countries. Because the historic settlements are more concentrated on the Pacific coast, which is geographically closer to China, one may expect the ethnic network to be more effective for those industries.

	Instrumental Variable	OLS	Reduced Form
	(1)	(2)	(3)
Log Chinese Employment × Log U.S. Patents$_{t-1}$	-.259	-.003	
	(.200)	(.056)	
Log Chinese Network × Log U.S. Patents$_{t-1}$			-.033
			(.026)
Log U.S. Patents$_{t-1}$	.782	.045	.039
	(.570)	(.170)	(.042)
Patent Class FE	Y	Y	Y
3-Year Time FE	Y	Y	Y
N of observations	2344	2344	2344
R-Squared	0.9186	0.4253	0.3814

Table 4.4. Lag of U.S. Patents

Notes: Standard errors in parenthesis and clustered at the 4-digit patent class level. The dependent variable is log of Chinese patents. ***p>0.01 **p>0.05 *p>0.10.

I then estimate a model where I replace current values of U.S. patents with past values, i.e. for the previous three-year time period. Table 4.4 summarizes the results . I find no statistically significant effects for either of the three specifications. This suggests that the benefits of the ethnic network on innovation do not materialize with substantial delay, which confirms the results by Kerr (2008), who finds that patent citations due to ethnic networks are higher in the first few years after innovation takes place in the U.S.

Furthermore, I investigate the presence of heterogeneous effects for high tech patent classes, where I include the interaction term between U.S. patents, Chinese employment and a dummy variable taking value 1 for high-tech industries, and zero otherwise[1]. This allows to test whether the relation is driven by the most innovative industries in the U.S., or whether

[1]To construct this dummy variable, I use the Eurostat classification of high-tech patent classes.

	Instrumental Variable	OLS	Reduced Form
	(1)	(2)	(3)
Log Chinese Employment × Log U.S. Patents × High Tech	.936	.300*	
	(.848)	(.160)	
Log Chinese Network × Log U.S. Patents × High Tech			-.246**
			(.117)
Log Chinese Employment × Log U.S. Patents	.508**	.148**	
	(.227)	(.062)	
Log Chinese Network × Log U.S. Patents			.065*
			(.034)
Log U.S. Patents × High Tech	-3.292	-.675	.588***
	(3.004)	(.560)	(.153)
Log U.S. Patents	-1.220*	-.189	.185***
	(.637)	(.176)	(.056)
Patent Class FE	Y	Y	Y
3-Year Time FE	Y	Y	Y
N of observations	2736	2736	2736
R-Squared	0.9200	0.4139	0.3359

Table 4.5. Heterogeneous Effects for High Tech Patent Classes

Notes: Standard errors in parenthesis and clustered at the 4-digit patent class level. The dependent variable is log of Chinese patents. ***p>0.01 **p>0.05 *p>0.10.

network effects are stronger in other sectors. Table 4.5 presents the results. Whereas in all specifications the effect does not appear to be driven by high-tech industries (the coefficient of the interaction of log U.S. patents and the measures of ethnic network is always positive and large), the OLS and the reduced form estimates show opposite effects, while the IV estimates are positive but present very large standard errors. The OLS regression produces a positive and statistically significant effect of the triple interaction (column 2), which indicates an positive differential effect for the industries which are considered high-tech. However, the reduced form estimates are negative (column 3), which suggest that less innovative industries benefit more from the ethnic network, i.e. that knowledge flows which happen through ethnic networks are stronger for less innovative industries. Notice that the reduced form estimates are in contrast to what previously found by Kerr (2008), who measures the effect of more recent ethnic networks. This difference may be due to the fact that more recent migrants tend to locate in areas of the U.S. with high prevalence of high-tech industries. Moreover, it may be a result of the simultaneity bias of the OLS estimates.

4.5 Conclusion

This paper investigates the effect of ethnic migrant networks on knowledge flows from a country at the technology frontier to a less developed country. I focus on ethnic Chinese networks, and I propose a measure of the network which exploits geographic variation of early migrants settlements in the U.S. in 1890, just after the Chinese Exclusion Act was implemented. I construct an instrument for the ethnic network interacting early Chinese settlements with the share of industry employment by county, and then summing across counties to obtain an industry-level measure. I use a panel of patent data for China and the U.S. to test whether technology diffusion is fostered by networks, and I find that Chinese patents are higher in industries where higher patenting in the U.S. is combined with a larger ethnic network. This work confirms the results found in the literature, which indicate that networks help technology flow from a country at the frontier to less advanced countries (Griffith et al., 2006; Kerr, 2008). In contrast to the literature, IV and reduced form estimates suggest that the results are not driven by the high-tech sector.

Kerr (2008) suggests that knowledge flows through networks might increase manufacturing output in the sending country of migrants. This work could be extended by analysing the consequences of technology flows in China, in terms of entrepreneurship, foreign investment, and welfare.

Appendix A

The Persistent Effect of Gender Division of Labour: African American Women After Slavery

A.1 Construction of the 1930 Data Set of Migrants

To overcome data limitations about migration in the early 20th century, I match surnames of cotton and tobacco slaveholders in order to construct a data set of migrants in year 1930. In this section I outline in detail the procedure I use to construct the data set.

I use data from the 1860 slave schedules to obtain slaveholders' surnames and I associate a crop to each surname based on the value of production of crops in each county in 1860. To achieve this, I determine the most common crop in each county in terms of crop value, and match this crop to the surname of each slaveholder registered in that county. However, in order to determine the origin of individuals whose surname match that of a slaveholder, I need to obtain a data set where each surname is associated to only one crop. Therefore, for those surnames that appear in more than one county, I only consider the slaveholder owning the largest number of slaves.

To decrease the extent of masurement error, I drop the most common surnames from the data set, and those which are associated with famous persons at the time, such as the surnames of current and former Presidents. Finally, I match African Americans whose mother share a surname with a slaveholder, and I identify as "migrants" those individuals who are not registered in a slave state, but whose mother was born in a slave state.

Note that the slave schedules may report more than one slaveholder for some of the slaves; in the case of multiple slaveholders, I only consider the surname of the first slaveholder.

A.2 Tables

	Mean	Median	SD	Min	Max	N Counties
Relative Cotton Production 1840	.491	.457	.466	0	1	804
Relative Cotton Share of Farmland 1880	.420	0	.484	0	1	1,736
Relative Cotton Production as a Share of Total Production 1840	.546	.5	.128	.133	1	1,094
Relative Cotton Suitability	.888	.982	.499	.125	2.738	2,999

Table A.2.1. Relative Cotton Prevalence by County

Notes: Data from 1840 and 1880 Agricultural Censuses, and Gaez Crop Suitability. *SD* indicates standard deviation. The number of observations represents the number of U.S. counties for which information about crops is available.

	Log Occupation Income Score
Relative Cotton Share 1880×female	.167***
	(.014)
Relative Cotton Share 1880×female×agriculture	-.201***
	(.014)
Relative Cotton Share 1880×agriculture	.034***
	(.006)
Female×agriculture	.794***
	(.025)
Female	-.900***
	(.030)
Agriculture	-.651***
	(.011)
Individual Controls	Y
County FE	Y
County Controls*Female	Y
Counties	746
N	134,798
R-Squared	0.4846

Table A.2.2. Occupation Income Score, Heterogeneous Effects for Agriculture

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variable *relative cotton production* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Labour Force Participation		Log Occupation Income Score		N Children	
	(1)	(2)	(3)	(4)	(5)	(6)
Relative Cotton Production 1840 × Black	.067***		.119***		.117***	
	(.007)		(.014)		(.020)	
Relative Cotton Production 1840		.037***		.029***		.003*
		(.009)		(.017)		(.027)
Black	.059***		-.416***		-.146***	
	(.020)		(.038)		(.056)	
Individual Controls	Y	Y	Y	Y	Y	Y
County FE	Y	N	Y	N	Y	N
State FE	Y	Y	Y	Y	Y	Y
County Controls×*black*	Y	Y	Y	Y	Y	Y
Counties	779	735	777	647	779	737
N	300,841	93,884	65,243	42,625	342,879	102,473
R-Squared	0.4217	0.2407	0.1625	0.0782	0.2168	0.1430

Table A.2.3. Women: African American and Caucasian - 1880

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variable *relative cotton production* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Labour Force Participation			Log Occupation Income Score	
	Probit	Slave States	Control for Income	Slave States	Control for Income
	(1)	(2)	(3)	(4)	(5)
Relative Cotton Production 1840×female	.033**	.083***		. 106***	
	(.013)	(.008)		(.013)	
Relative Cotton Production 1840			.107***		.028***
			(.009)		(.011)
Spouse OIS			-.033***		.259***
			(.006)		(.016)
Female	-.771	-.843***		-.541***	
	(.058)	(.024)		(.037)	
Individual Controls	Y	Y	Y	Y	Y
County FE	N	Y	N	Y	N
State FE	N	Y	Y	Y	Y
County Controls*Female	Y	Y	Y	Y	Y
Counties	747	584	1,559	584	856
N	186,563	177,043	88,408	127,972	26,554
R-Squared	0.4731	0.4784	0.1540	0.2521	0.2661

Table A.2.4. Probit, Restricting to Slave States and Controling for Income

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. Columns 3 and 5 only include African American women. Spouse OIS refers to the spouse's occupation income score, which is used as a proxy for family income. As county FE cannot be included in columns 3 and 5, State FE and county controls are included. The coefficients of the variables *relative cotton share* and *relative cotton suitability* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Labour Force Participation			Log Occupation Income Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Relative Cotton Share 1880×female	.083***			.153***		
	(.008)			(.012)		
Relative Cotton Suitability		.052***			.066***	
		(.009)			(.013)	
Relative Cotton Production ×female			.111***			.134***
As a Share of Tot. Production 1840			(.008)			(.014)
Female	-.805***	-.825***	-1.011***	-.668***	-.584***	-.839***
	(.023)	(.033)	(.032)	(.041)	(.056)	(.067)
Individual Controls	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
County Controls*Female	Y	Y	Y	Y	Y	Y
Counties	1,558	1,730	1,021	1,541	1,710	1,015
N	251,559	260,106	204,701	180,365	186,444	147,380
R-Squared	0.4996	0.4974	0.4979	0.2648	0.2597	0.2674

Table A.2.5. Alternative Measures of Relative Cotton Prevalence

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variables *relative cotton share* and *relative cotton suitability* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Labour Force Participation	Log Occupation Income Score	
	(1)	(2)	(3)
Relative Cotton Production 1840 ×female	.005	.062***	-.009
	(.003)	(.011)	(.016)
Relative Cotton Production 1840 ×female×agriculture			-.042***
			(.014)
Relative Cotton Production 1840 ×agriculture			-.034***
			(.003)
Female×agriculture			.708***
			(.024)
Agriculture			-.707***
			(.004)
Female	-.920***	-.349***	-.672***
	(.007)	(.047)	(.050)
Individual Controls	Y	Y	Y
County FE	Y	Y	Y
Y			
County Controls*Female	Y	Y	Y
Counties	828	828	828
N	515,006	273,877	273,877
R-Squared	0.7731	0.2644	0.5863

Table A.2.6. Other Ethnicities

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The coefficients of the variable *relative cotton prduction* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Ratio of Mixed Race Couples	Labour Force Part.	Log Occ. Income Score
	(1)	(2)	(3)
Relative Cotton Production 1840 $\times$ female	-	.069***	.141***
		(.009)	(.017)
Relative Cotton Production 1840	.003*		
	(.001)		
Number of Female Slaves 1860 $\times$ female		.076***	.090***
		(.019)	(.022)
Female	-	-.817***	-.616***
		(.023)	(.046)
Individual Controls	-	Y	Y
County FE	-	Y	Y
STATE FE	Y	Y	Y
Industry FE	-	Y	Y
Counties	820	747	746
N	820	186,563	134,798
R-Squared	0.4298	0.4886	0.2547

Table A.2.7. Discrimination and Social Networks

Notes: Standard errors clustered at the county level.***p>0.01 **p>0.05 *p>0.10. The regression estimates in column 1 are at the county level and refer to mixed race couples where the wife is African American and the husband is Caucasian. The coefficients of the variable *relative cotton production* correspond to the change in the dependent variable due to a 1 standard deviation change in relative cotton production in 1840 or in relative suitability. Data from the 1880 Population Census.

	Relative Cotton Production 1840	Females/Males in the Labour Force
	(1)	(2)
Numbmer of Cotton Surnames	.079***	.011***
	(.009)	(.004)
Number of Tobacco Surnames	-.095***	-.003
	(.021)	(.003)
Number of Wheat Surnames	-.006	-.000
	(.007)	(.001)
Number of Rice Surnames	-.031**	-.001
	(.013)	(.000)
Number of Indian Corn Surnames	-.043	-.011
	(.039)	(.008)
State FE	Y	Y
N Counties	804	931

Table A.2.8. Surnames and Cotton Prevalence

Notes: The dependent variable in column 2 is the ratio of African American females to males in the labour force Standard errors clustered at the state level.***p>0.01 **p>0.05 *p>0.10

	Number of Migrants by County		
	All counties	Relative cotton prevalence>median	Relative cotton prevalence<median
	(1)	(2)	(3)
Crop Share 1880	.307*	-	-
	(.147)		
Female LFP 1930	.757**	1.538***	.185
	(.281)	(.346)	(.145)
Male LFP 1930	.292	.948*	.042
	(.185)	(.491)	(.081)
%Female Literate 1930	.332*	.214	.048
	(.162)	(.435)	(.079)
% Male Literate 1930	.505**	.660	.064
	(.223)	(.601)	(.119)
Total Population	.510*	.188	.037
	(.242)	(.280)	(.044)
State FE	Y	Y	Y
N Counties	749	486	263

Table A.2.9. Predictors of Migration by County: 1940

Notes: The dependent variable is the number of migrants from slave states by county in 1940, divided by its standard deviation. The variable *LFP* indicates labour force participation. ***p>0.01 **p>0.05 *p>0.10

A.3 Figures

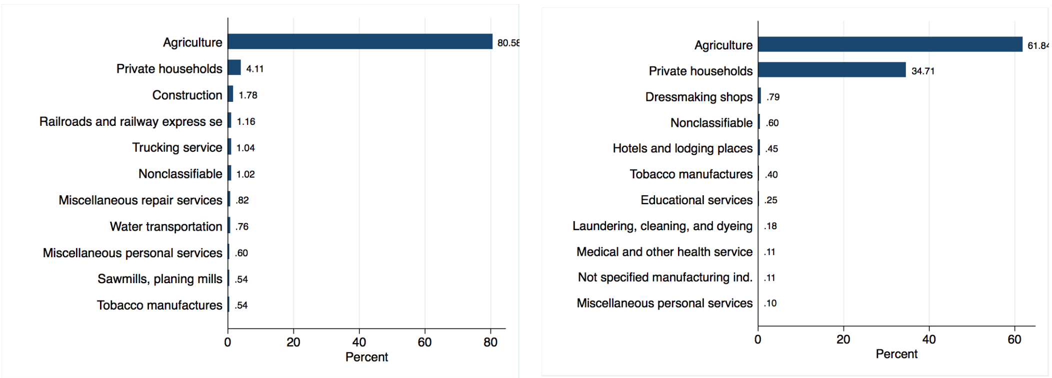


Figure A.3.1. Main Industries by Gender 1880

Notes: This graph shows the main 10 industries by employment of African American men (left) and women (right) in 1880. Source: U.S. Population Census 1880.

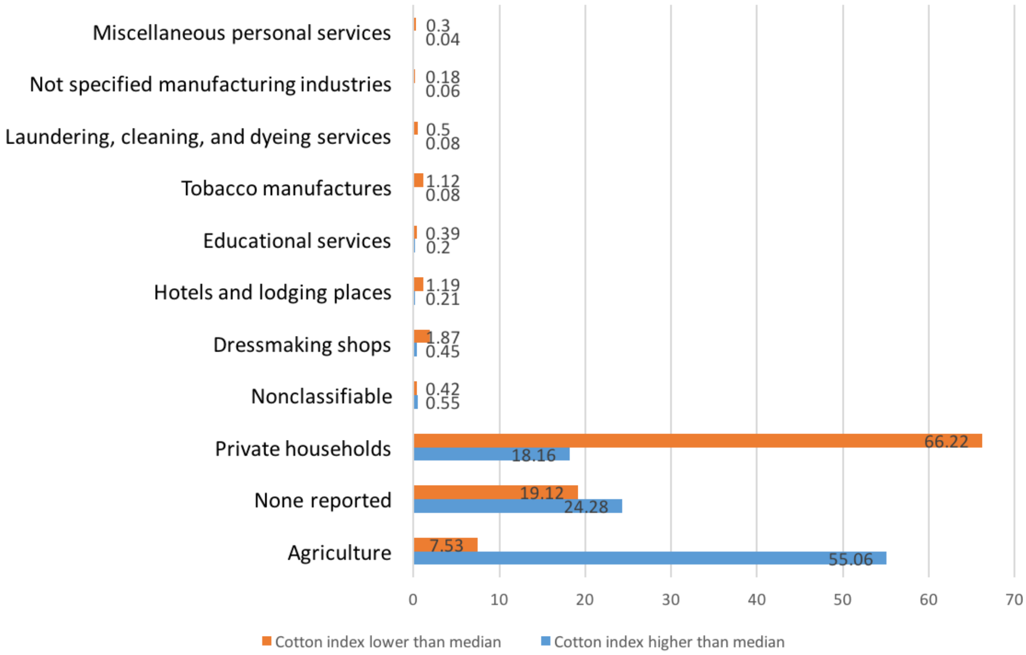


Figure A.3.2. African American Women's Occupations by Cotton Prevalence

Note: This graph shows the share of African American women in the labour force who work in the top 10 industries in 1880 in counties with cotton index higher than median (blue bars) and below median (orange bars). Source: U.S. Population Census 1880.

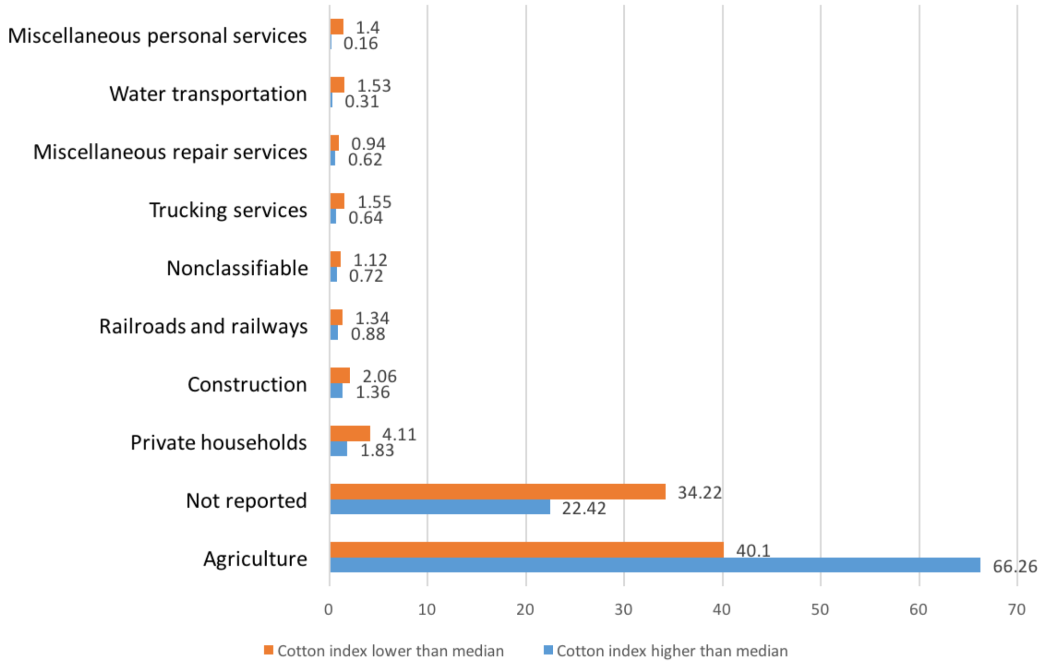
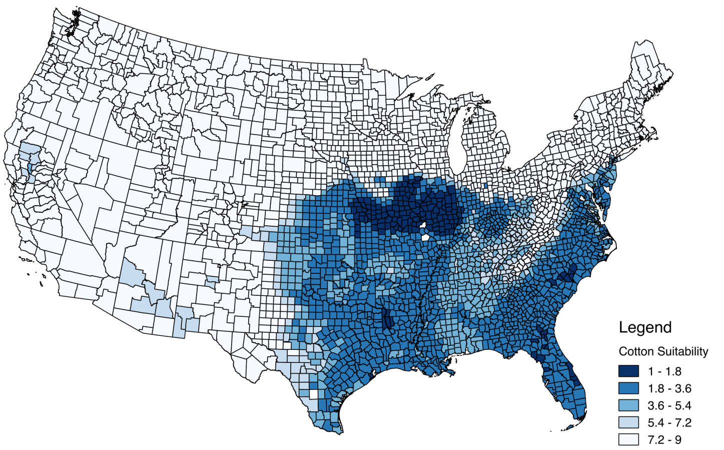


Figure A.3.3. African American Men's Occupations by Cotton Prevalence

Notes: This graph shows the share of African American men in the labour force who work in the top 10 industries in 1880 in counties with cotton index higher than median (blue bars) and below median (orange bars). Source: U.S. Population Census 1880.



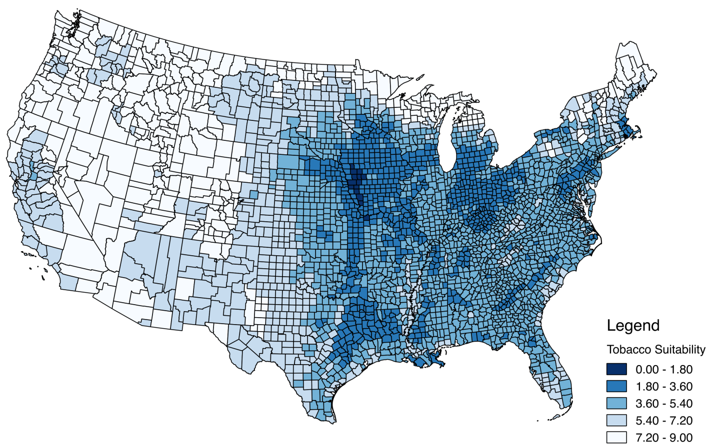
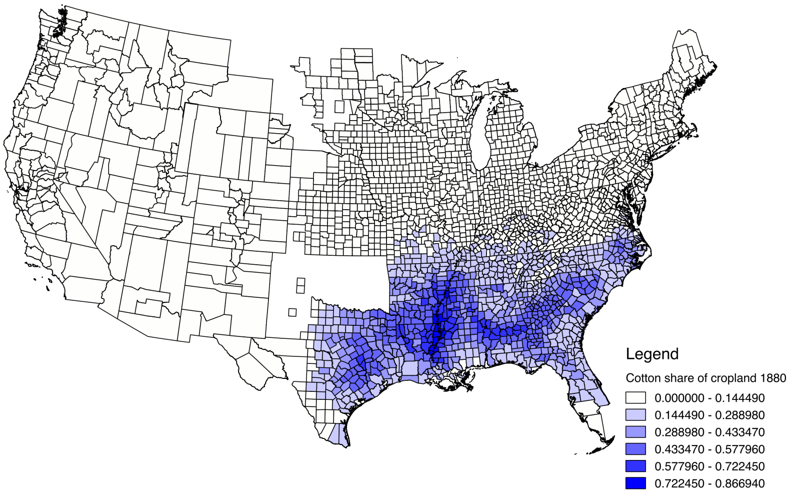


Figure A.3.4. Cotton and Tobacco Suitability

Notes: These two maps show the geographic distribution of cotton suitability (top) and tobacco suitability (bottom) by county in the U.S. Source: GAEZ-FAO Crop Suitability Data



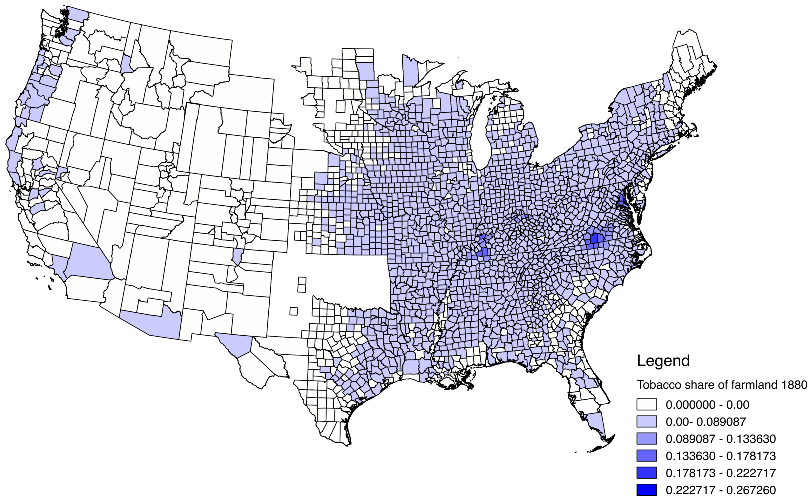


Figure A.3.5. Cotton and Tobacco Shares of Production Value 1880

Notes: These two maps show the shares of cotton (top) and tobacco (bottom) acreage to total crop acreage by county in 1880, measured in acres. Source: U.S. Agricultural Census 1880

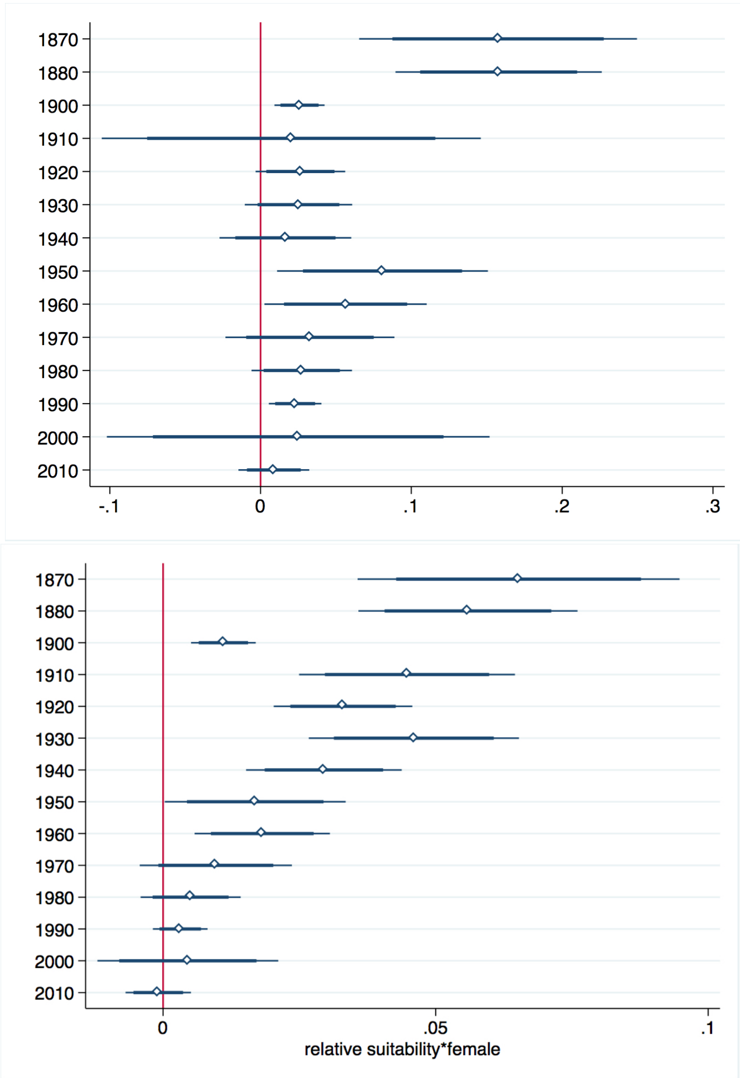


Figure A.3.6. Labour Force Participation - IV and Reduced Form Estimates

Notes: These two graphs show the coefficient plots of regressing labour force participation on the variable *relative cotton production* $\times$ *female* using census data for years 1870-2010, estimated with instrumental variable (top graph) and reduced form (bottom graph) regressions. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.

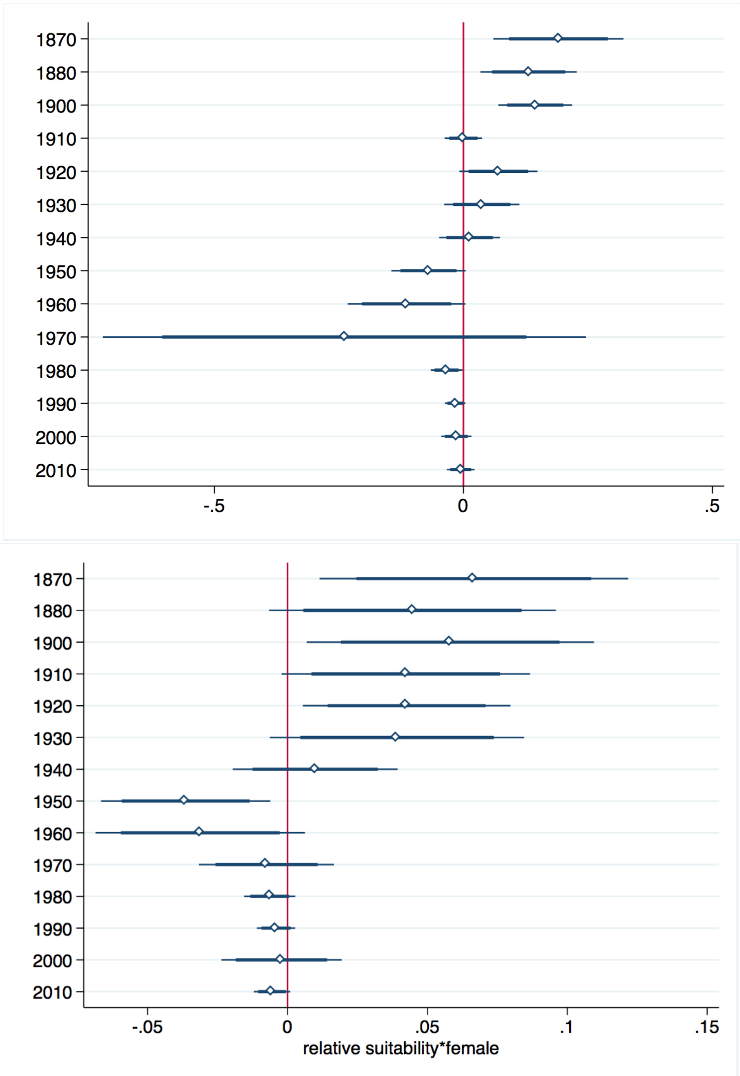


Figure A.3.7. Occupation Income Score - IV and Reduced Form Estimates

Notes: These two graphs show the coefficient plots of regressing occupation income score on the variable $relative\, cotton\, production \times female$ using census data for years 1870-2010, estimated with instrumental variable (top graph) and reduced form (bottom graph) regressions. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.

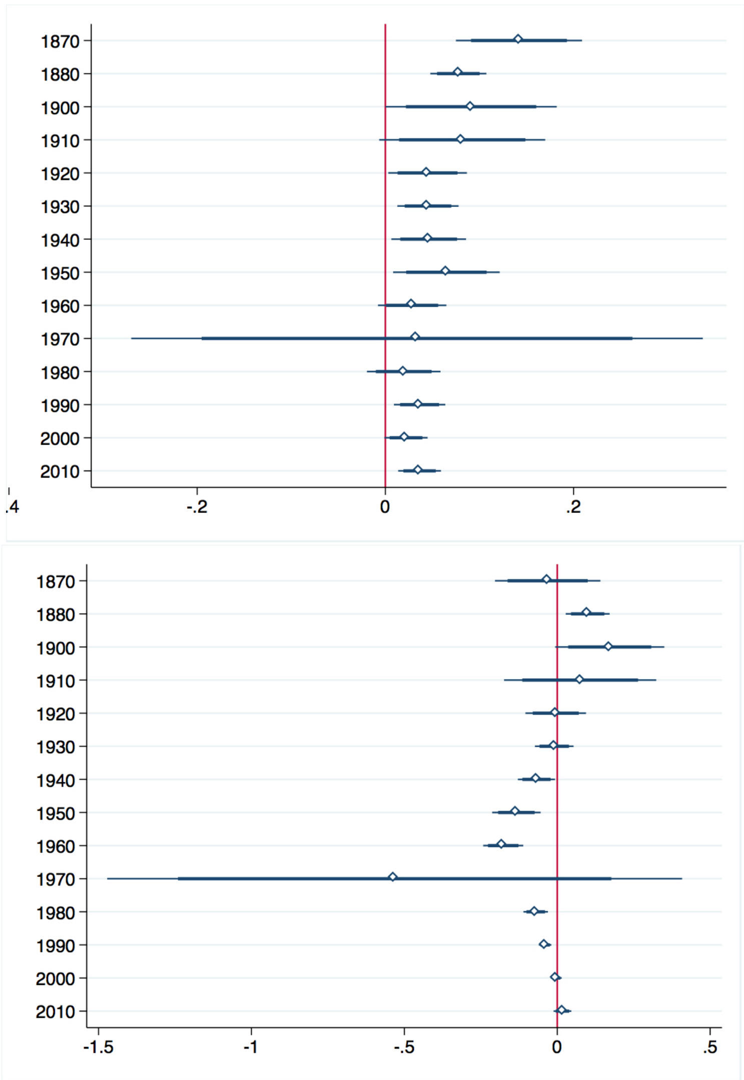
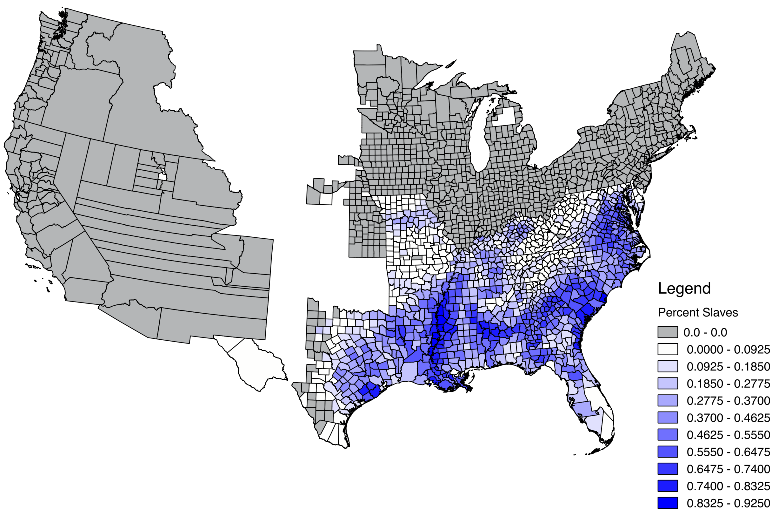


Figure A.3.8. African American and White Women - Labour Force Participation and Occupation Income Score

Notes: These two graphs show the coefficient plots of regressing labour force participation (top) and occupation income score (bottom) on the variable *relative cotton production* $\times$ *black*, produced with IV regressions using census data for years 1870-2010. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.



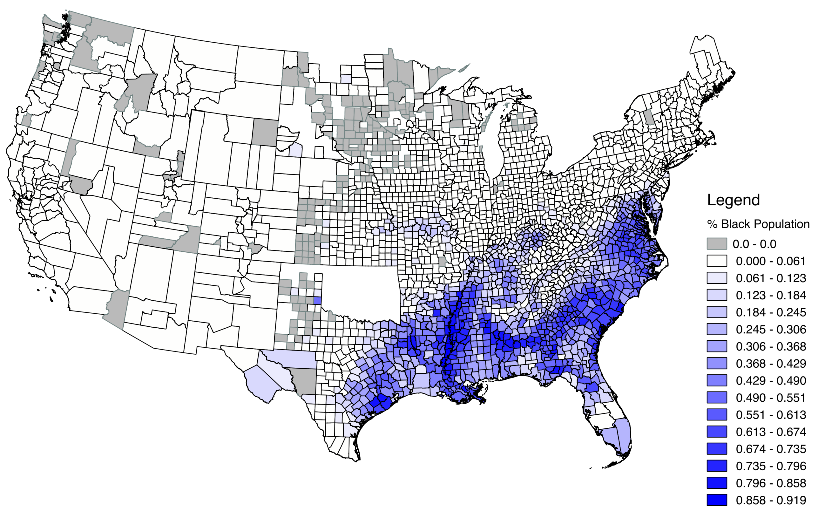


Figure A.3.9. Slavery 1860 and Black Population 1880

Notes: These two maps show the percentage of slaves by county in 1860 (top) and the percentage of African Americans by county in 1880 (bottom). Source: U.S. Historical Censuses

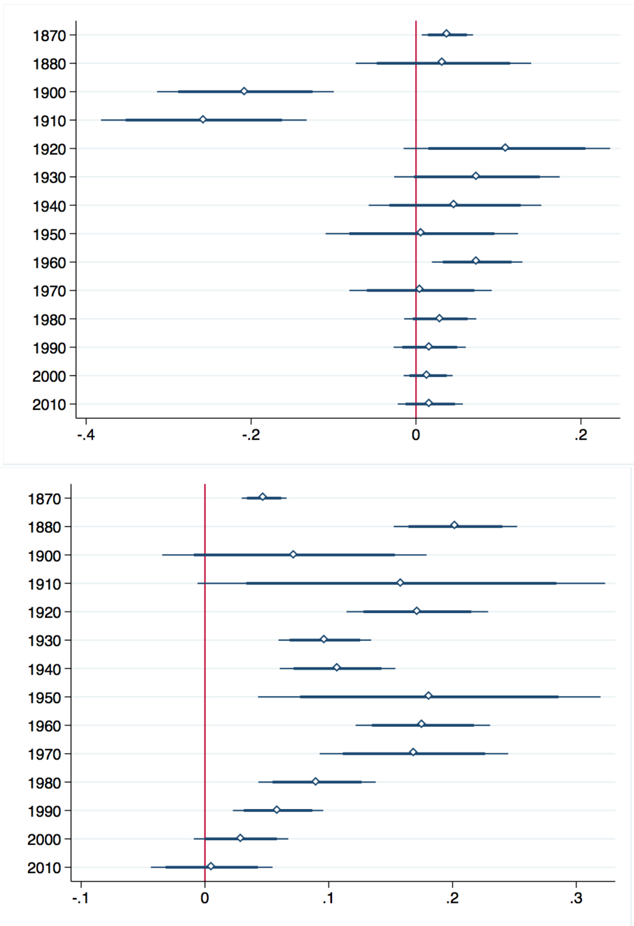


Figure A.3.10. Number of Children

Notes: These two graphs show the coefficient plots of regressing number of children on the variables *relative cotton production* (top) and *relative cotton production* × *black* (bottom) using census data for years 1870-2010. The size corresponds to the effect of a one standard deviation change in the dependent variable. The confidence intervals reported are 95 and 99%. I include individual controls and county fixed effects. Standard errors are clustered at the county level.

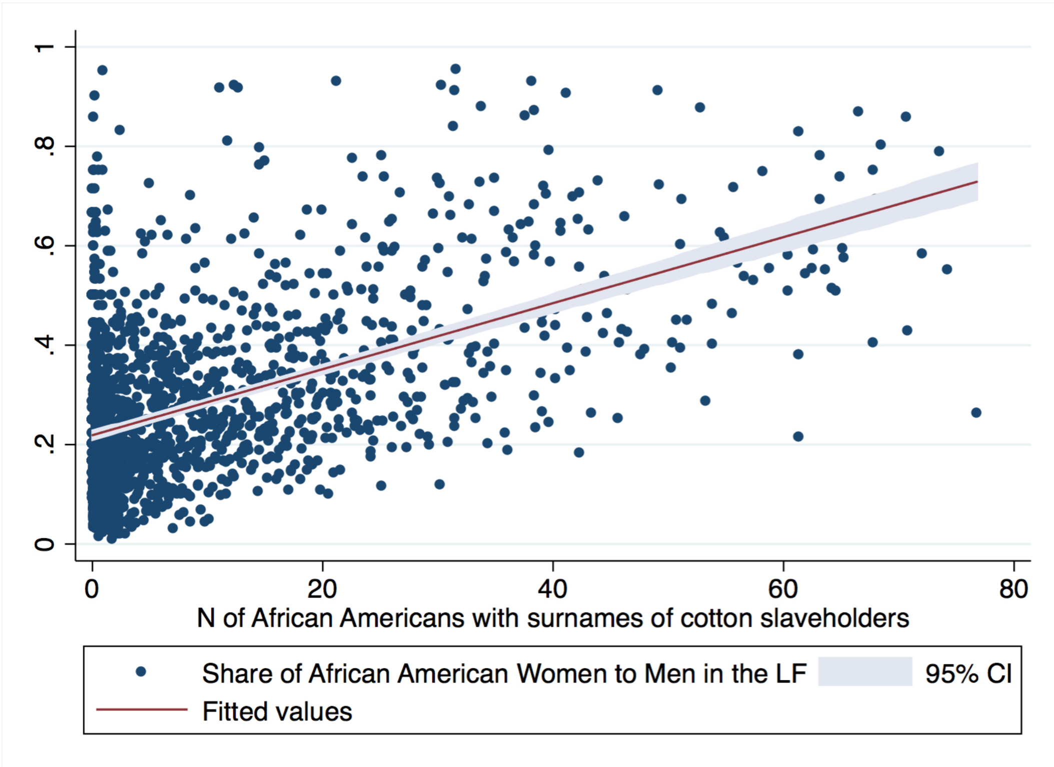


Figure A.3.11. Surnames of Cotton Slaveholders and Women in the Labour force by County 1880

Notes: This graph shows the correlation between the number of African American individuals with surnames that match those of cotton slaveholders (divided by its standard deviation) and the ratio of African American females to males in the labour force in 1880. I exclude the top and bottom 1% of both measures. Data source: U.S. census.

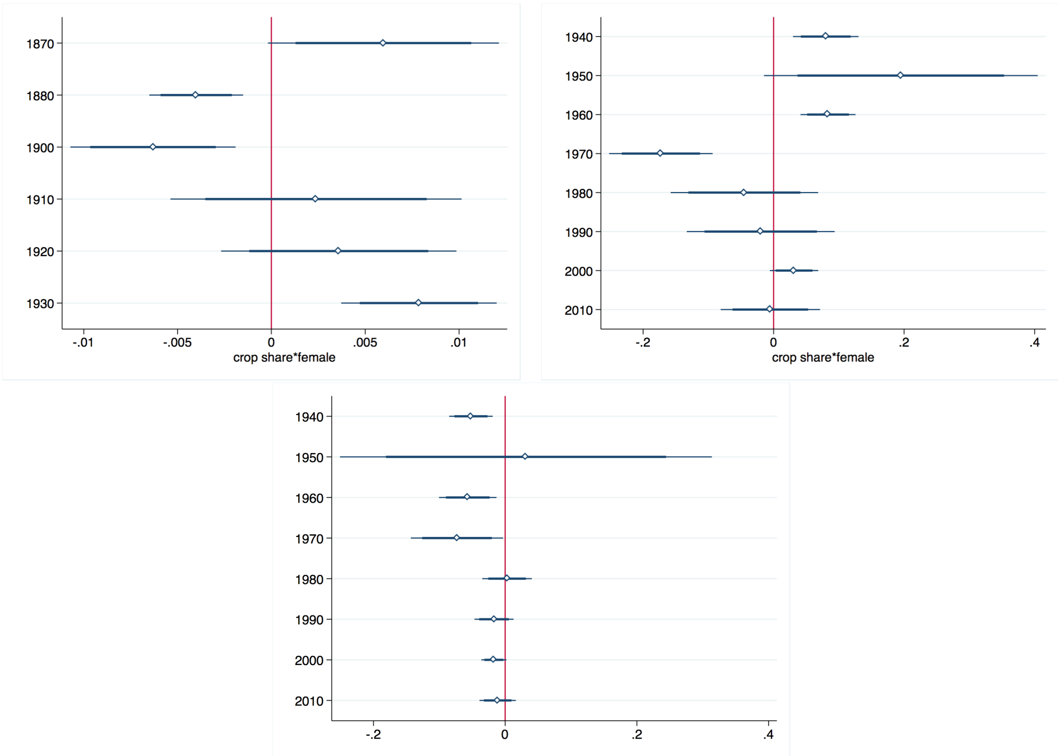


Figure A.3.12. Literacy, Education and Wages

Notes: These graphs represent the coefficients of the variable *relative cotton production×female*. The dependent variables are literacy at the top left panel (years 1880-1930), years of education at the top right panel (1940-2010) and log wages (years 1940-2010) at the bottom. The regressions include county fixed effects and individual controls, as described in equation (1). Standard errors are clustered at the county level.

Appendix B

The Benefits of the Bamboo Network in International Trade

B.1 Tables

	Mean	Median	SD	Min	Max	N
	Firm Level Statistics					
Employee Number	275.4835	10	3153.054	10	50000	243205
Revenue	1393.8407	75	8194.665	15	200000	243205
	County Level Statistics					
Number of Firms	13.55497	3	47.00418	1	2440	17945
% Firms $<$ 10 Workers	57.38828	66.66667	40.94633	0	100	17945
% Firms $>$ 10, $<$ 250 Workers	28.86719	11.20332	35.32179	0	100	17945
% Firms $>$ 250 Workers	13.74453	0	27.65033	0	100	17945

Table B.1.1. Descriptive Statistics: Chinese Economic Census

Notes: This table shows the descriptive statistics of the main variables used in this paper. Data from the Chinese Economic Census of 2004.

	Mean	Median	SD	Min	Max	N
			Cultural Exposure			
Sending Counties Dummy	.144	0	.351	0	1	34,525
Cantonese Counties Dummy	.681	1	.466	0	1	34,525
			Industry Exposure			
Cantonese Workers Retail and Wholesale	253.311	144.763	285.397	2.367	2789.528	31,026
Cantonese Workers Downstream Manufacturing	4112.293	2109.71	6210.863	22	25698.23	34,083
Cantonese Workers Same Industry	283.737	118.306	412.163	.322	8462.154	34,112

Table B.1.2. Descriptive Statistics: Cultural and Industry Exposure

Notes: This table shows the descriptive statistics of the main independent variables variables used in this paper. Sources: 2004 Annual Survey of Industrial Enterprises.

	Cantonese Workers
Restaurants and Other Food Services	47,199
Colleges and Universities	26,132
Hospitals	24,494
Elementary and Secondary Schools	20,436
Electronic Components and Products Manufacturing	14,070
Other Information Services	13,396
Computer systems design and related services	12,262
Construction	11,483
Grocery Stores	11,372
Securities, commodities, funds, trusts, and other financial investments	10,739

Table B.1.3. Top 10 U.S. 4 Digit Industries by Number of Cantonese Workers

Notes: U.S. 2000 Population Census data.

	Cantonese Workers
Carpets and rugs manufacturing	14
Logging	18
Tobacco Manufacturing	22
Railroad rolling stock manufacturing	22
Support activities for agriculture and forestry	24
Tire manufacturing	24
Farm supplies wholesalers	25
Not specified metal industries	37
Other transportation equipment manufacturing	38
Structural clay product manufacturing	41

Table B.1.4. Last 10 U.S. 4 Digit Industries by Number of Cantonese Workers

Notes: U.S. 2000 Population Census data.

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Workers Related Retail×Cantonese	.023***			.080***		
	(.008)			(.030)		
Workers Downstream Manufacturing×Cantonese		.017***			.047**	
		(.003)			(.019)	
Workers Same Industry×Cantonese			.007			.022
			(.010)			(.044)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	22,393	24,041	24,017	11,259	11,777	11,777
R-Squared	0.2872	0.2917	0.2908	0.1500	0.1545	0.1542

Table B.1.5. Exports - Network Exposure for Cantonese Counties

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. In these regressions, network exposure is constructed as *industry exposure* times a dummy variable which indicates whether a county's main language belongs to the Cantonese language group. The coefficients of the variables defined as *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.038**			.159***		
	(.015)			(.043)		
Network Downstream Manufacturing		.036***			.062***	
		(.004)			(.024)	
Network Same Industry			.037***			.084**
			(.010)			(.035)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
Controls×Industry FE	Y	Y	Y	Y	Y	Y
N of observations	22,234	23,861	23,837	11,083	11,572	11,572
R-Squared	0.3235	0.3301	0.3293	0.2028	0.2092	0.2093

Table B.1.6. Baseline Specification with County Controls Interacted with Industry Fixed Effects

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. Controls are at the county level and include total county population, emigration (%), and university attendance (%). The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.032***			.146***		
	(.012)			(.034)		
Network Downstream Manufacturing		.029***			.072***	
		(.004)			(.019)	
Network Same Industry			.028***			.102***
			(.009)			(.030)
Foreign Capital (%)	.035***	.035***	.035***	.320***	.317***	.317***
	(.006)	(.006)	(.006)	(.026)	(.025)	(.025)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	25,650	27,473	27,449	13,947	14,576	14,576
R-Squared	0.2794	0.2857	0.2849	0.1901	0.1910	0.1911

Table B.1.7. Control for Foreign Capital

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables defined as *network* and *foreign capital* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Export Value					
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.146***	.180***				
	(.037)	(.040)				
Network Downstream Manufacturing			.058***	.082***		
			(.020)	(.023)		
Network Same Industry					.087***	.121***
					(.032)	(.035)
Foreign Capital (%)		.026***		.320***		.319***
		(.027)		(.027)		(.027)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
County Controls×Industry FE	N	Y	N	Y	N	Y
Domestic Firms Only	Y	N	Y	N	Y	N
N of observations	11,641	14,176	11,641	14,176	11,641	14,176
R-Squared	0.6150	0.6442	0.6184	0.6473	0.6246	0.6088

Table B.1.8. Exports Adjusted by Share of Industry Exports to the US Vs. World

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. Controls are at the county level and include total county population, emigration (%), and university attendance (%). The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.028			.129***		
	(.017)			(.041)		
Network Downstream Manufacturing		.036***			.082***	
		(.004)			(.020)	
Network Same Industry			.038***			.097***
			(.008)			(.031)
Non-Cantonese Workers Related Retail×Sending	.002			.015		
	(.010)			(.035)		
Non-Cantonese Workers Downstream Manufacturing×Sending		-.025***			-.160***	
		(.008)			(.053)	
Non-Cantonese Workers Same Industry×Sending			-.029*			-.139
			(.017)			(.086)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	22,393	24,041	24,017	11,259	11,777	11,777
R-Squared	0.2871	0.2920	0.2912	0.1502	0.1548	0.1545

Table B.1.9. Control for Industry Size

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. In these regressions, network exposure is constructed as *industry exposure* times a dummy variable which indicates whether a county's main language belongs to the Cantonese language group. The coefficients of the variables defined as *network* and *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail×Large	.011			-.021		
	(.022)			(.073)		
Network Downstream Manufacturing×Large		-.020**			.012	
		(.010)			(.037)	
Network Same Industry×Large			-.033***			-.084
			(.012)			(.066)
Network Related Retail	.022			.157***		
	(.021)			(.061)		
Network Downstream Manufacturing		.040***			.043	
		(.008)			(.032)	
Network Same Industry			.046***			.143***
			(.014)			(.054)
Workers Related Retail×Large	-.014***			-.121***		
	(.007)			(.038)		
Workers Downstream Manufacturing×Large		-.003			-.119***	
		(.004)			(.017)	
Workers Same Industry×Large			.020			-.034
			(.015)			(.087)
Large Firm×Sending	.004	.041*	.052***	.086	.073	.125
	(.026)	(.021)	(.019)	(.147)	(.112)	(.114)
Large Firm	.233***	.219***	.202***	1.177***	1.129***	1.068***
	(.012)	(.010)	(.012)	(.057)	(.048)	(.056)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	22,393	24,041	24,017	11,259	11,777	11,777
R-Squared	0.3257	0.3301	0.3296	0.2363	0.2388	0.2377

Table B.1.10. Heterogeneous Effects for Large Firms

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. *Large* is a dummy variable = 1 if the number of employee is above median and zero otherwise. The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Domestic Firms			Exporting Firms			Non-exporting Firms		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Network Related Retail	-.003			-.009			.005		
	(.007)			(.011)			(.015)		
Network Downstream Manufacturing		-.005			-.016***			.010	
		(.004)			(.005)			(.012)	
Network Same Industry			-.003			-.020***			.015
			(.007)			(.007)			(.013)
Age	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
N of observations	22,147	23,627	23,627	11,196	11,731	11,731	10,886	11,829	11,829
R-Squared	0.1479	0.1495	0.1495	0.1372	0.1392	0.1392	0.1767	0.1764	0.1764

Table B.1.11. TFPR

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Log Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.032**			.131***		
	(.015)			(.042)		
Network Related Retail × High Tech	-.009			.062		
	(.021)			(.122)		
Network Downstream Manufacturing		.036***			.051**	
		(.004)			(.023)	
Network Downstream Manufacturing × High Tech		.031			.184	
		(.060)			(.400)	
Network Same Industry			.033***			.066**
			(.011)			(.032)
Network Same Industry × High Tech			.004			.122
			(.032)			(.193)
Sending × High Tech	.039	.043	.041*	-.063	-.107	-.108
	(.026)	(.029)	(.024)	(.206)	(.217)	(.157)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	22,393	24,041	24,017	11,259	11,777	11,777
R-Squared	0.2872	0.2921	0.2805	0.1502	0.1544	0.1545

Table B.1.12. Heterogeneous Effects for High Tech Industry

Notes: *High-tech* is a dummy variable equal to 1 for ISIC 3.1 industry codes 24, 29, 30, 31, 32, 33, 34, 352, 353, 359, 2423, based on the technology intensity definition published by the OECD. Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Probability of Exporting			Export Value		
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail	.020			.209**		
	(.022)			(.089)		
Network Downstream Manufacturing		.009			.169*	
		(.016)			(.094)	
Network Same Industry			.017			.214*
			(.024)			(.119)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	3,172	3,332	3,332	2,593	2,694	2,694
R-Squared	0.2174	0.2343	0.2343	0.2607	0.1544	0.2623

Table B.1.13. Exports - Foreign Owned Firms

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The coefficients of the variables defined as *network* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Log Number of Employees					
	All Firms	Manufacturing			Services	Retail
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail		.069*				
		(.038)				
Network Downstream Manufacturing			.094***			
			(.015)			
Network Same Industry	.027***			.075**	.004	.002
	(.007)			(.036)	(.018)	(.004)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	234,542	108,304	146,724	109,755	30,205	94,394
R-Squared	0.3389	0.2155	0.2853	0.2147	0.1032	0.1187

Table B.1.14. Log Employment by Sectors - Economic Census Data

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The variables denoted as *workers* refers to the number of Cantonese workers in each 4-digit industries in the U.S. The coefficients of the variables *network* and *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

	Log Revenue per Workers					
	All Firms	Manufacturing			Services	Retail
	(1)	(2)	(3)	(4)	(5)	(6)
Network Related Retail		.018				
		(.020)				
Network Downstream Manufacturing			.016			
			(.013)			
Network Same Industry	-.030**			.035**	.005	-.029*
	(.013)			(.014)	(.015)	(.016)
Age	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
N of observations	234,542	109,755	109,755	109,755	30,205	94,394
R-Squared	0.3389	0.2155	0.2853	0.2147	0.1032	0.1187

Table B.1.15. Log Revenue per Worker by Sectors - Economic Census Data

Notes: Standard errors in parenthesis and clustered at the four-digit Chinese industry level. The variables denoted as *workers* refers to the number of Cantonese workers in each 4-digit industries in the U.S. The coefficients of the variables *network* and *workers* correspond to the changes in the dependent variable caused by a 1 SD change in the independent variable. ***p>0.01 **p>0.05 *p>0.10

B.2 Figures

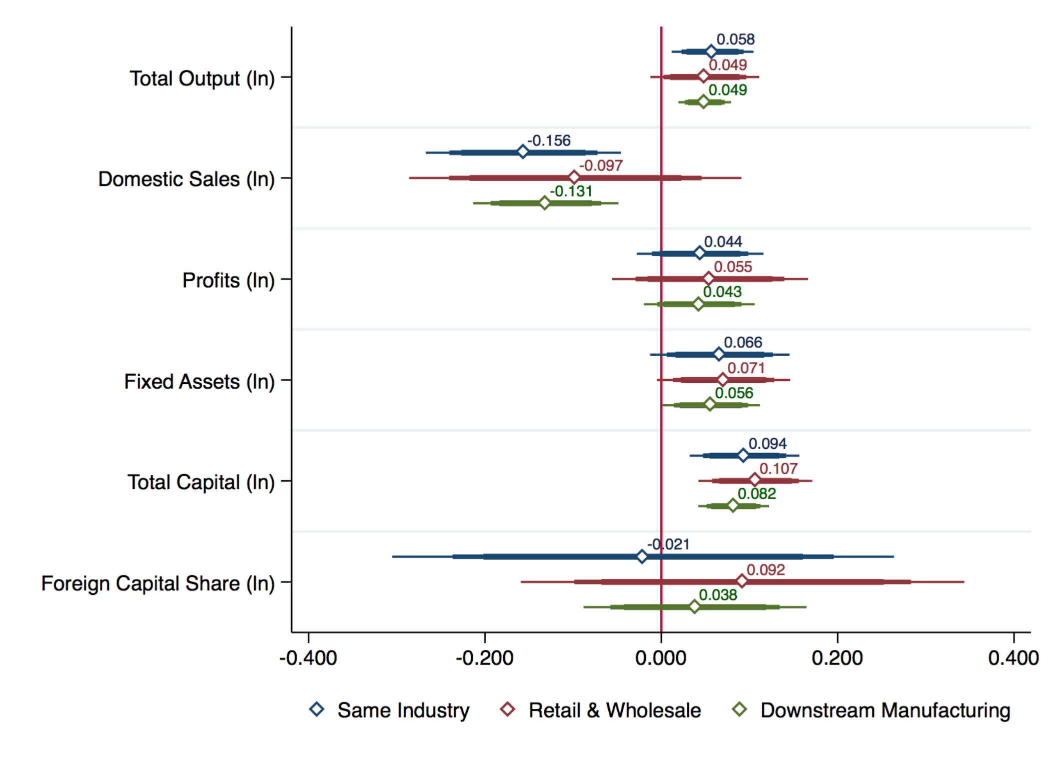


Figure B.2.1. Firm Outcomes: Output, Assets and Capital

Notes: Standard errors are clustered at the four-digit Chinese industry level. This graph plots the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The confidence intervals given are at the 99%, 95% and 90% level.

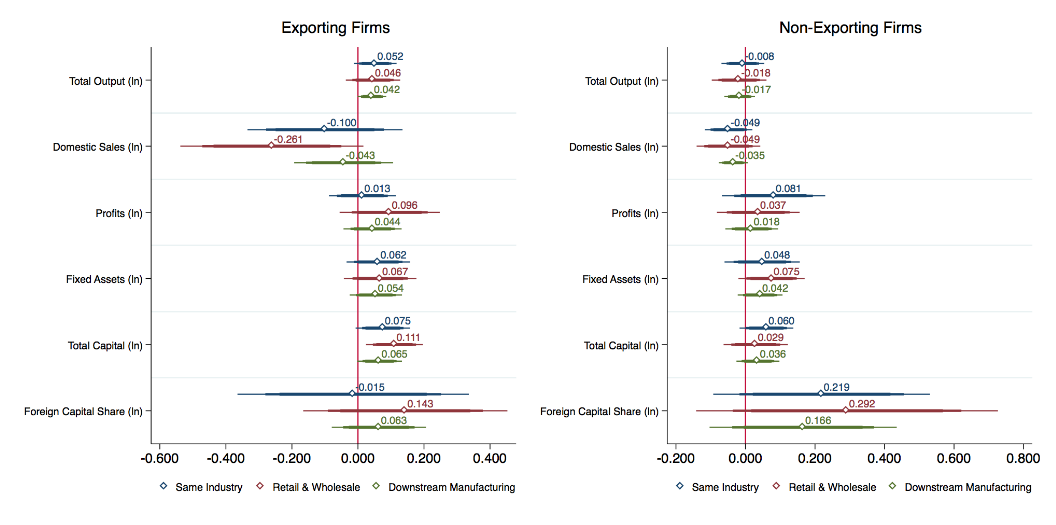


Figure B.2.2. Firm Outcomes: Output, Assets and Capital by Export Status

Notes: Standard errors are clustered at the four-digit Chinese industry level. These graphs plot the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The confidence intervals given are at the 99%, 95% and 90% level.

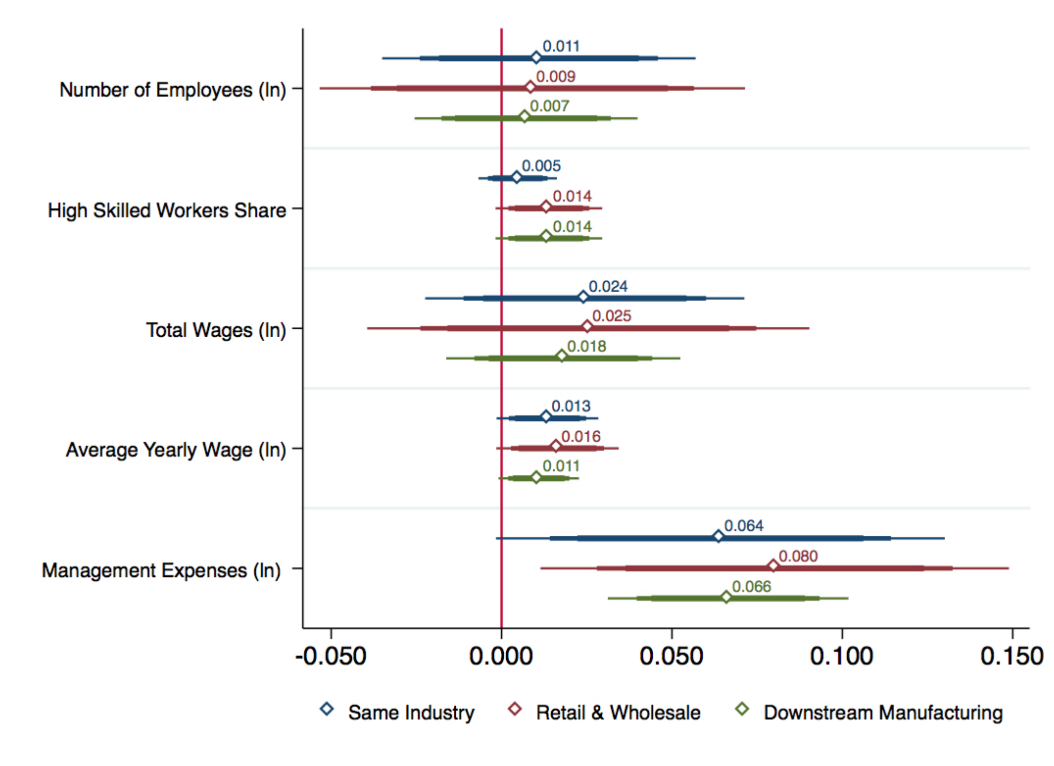


Figure B.2.3. Firm Outcomes: Employment and Expenses

Notes: Standard errors are clustered at the four-digit Chinese industry level. This graph plots the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The variable *average yearly wage* indicates the yearly average wage paid per worker. The confidence intervals given are at the 99%, 95% and 90% level.

Figure B.2.4. Firm Outcomes: Employment and Expenses by Export Status

Notes: Standard errors are clustered at the four-digit Chinese industry level. These graphs plot the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The variable *average yearly wage* indicates the yearly average wage paid per worker. The confidence intervals given are at the 99%, 95% and 90% level.

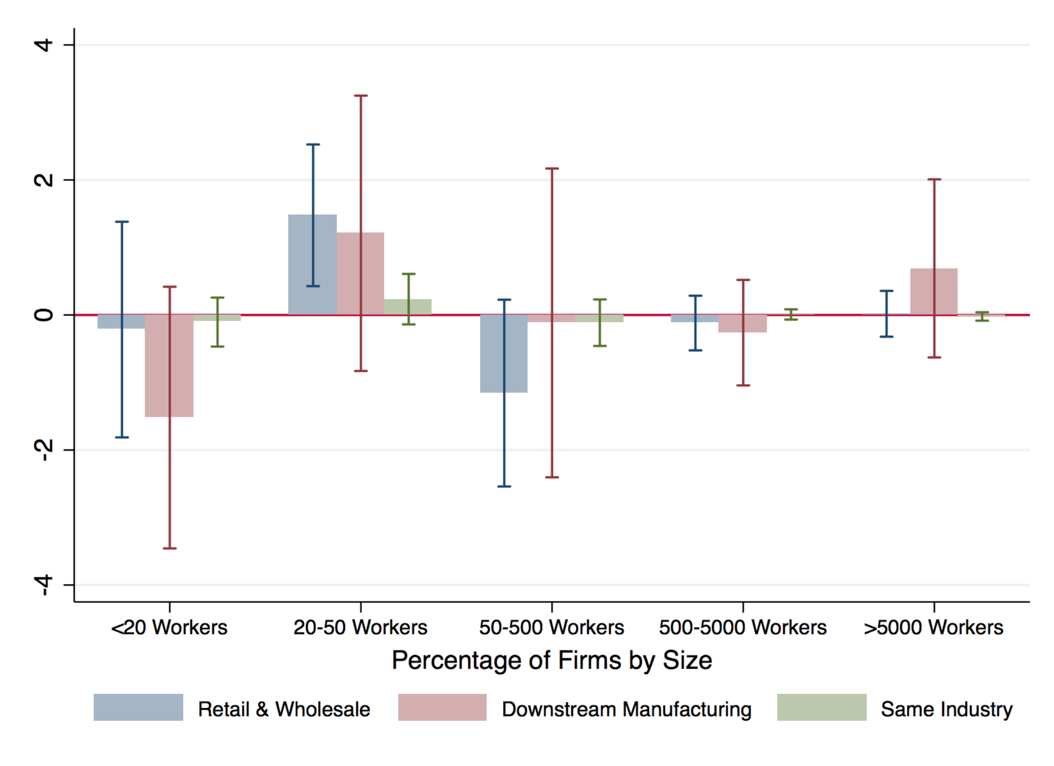


Figure B.2.5. Economic Census: Effect on Size Distribution - Manufacturing

Notes: Standard errors are clustered at the four-digit Chinese industry level. This graph plots the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The confidence intervals given are at 90% level.

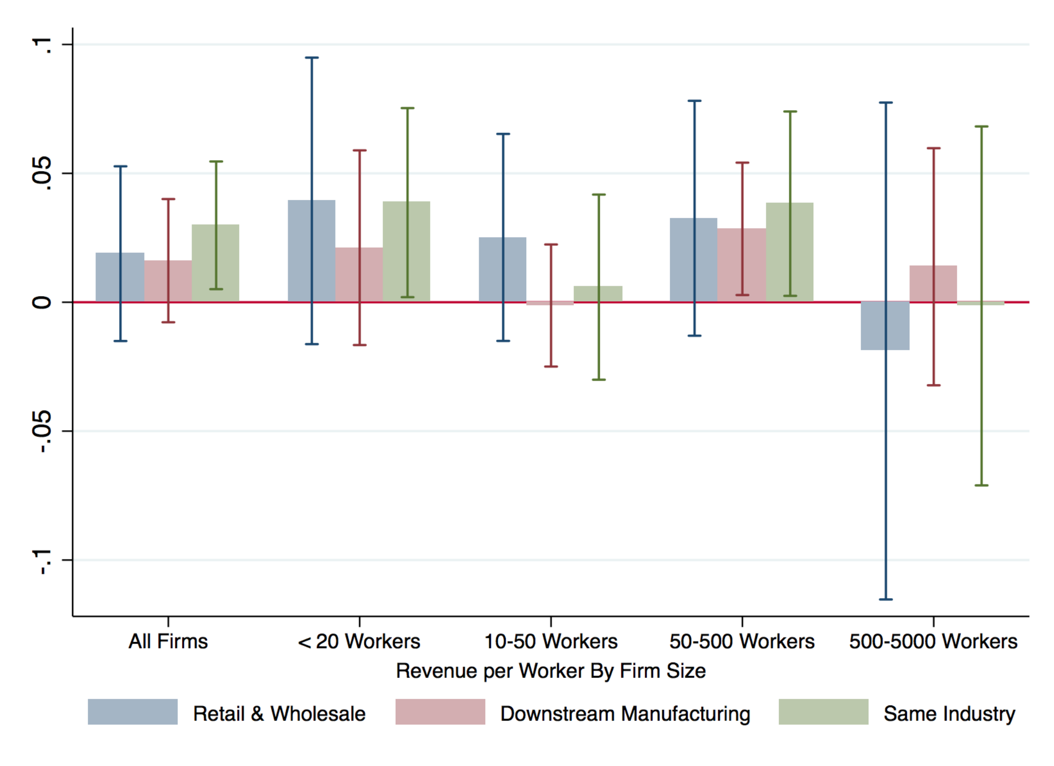


Figure B.2.6. Economic Census: Effect on Revenue per Worker by Firm Size - Manufacturing

Notes: Standard errors are clustered at the four-digit Chinese industry level. This graph plots the coefficients of the variable *network exposure* estimated in equation (4) for the three different measures of industry exposure; the coefficients correspond to the effect of a 1 SD change in the measure of network exposure. The confidence intervals given are at 90% level.

B.3 Matching Chinese Manufacturing Industries to U.S. Wholesale and Retail Industries

To calculate the measure of industry exposure described in Section 4.1 we need to match manufacturing industries in China with their likely wholesale and retail counterpart in the U.S. Ideally, to perform this matching we would have information about imports and exports that reveal the top industries which import manufacturing data.

In the absence of that, we need to impute which manufacturing industries match with which retail industries. We achieve this by first isolating key words in the description of the manufacturing industry in China. As a next step, we then match these key words to the description of the corresponding NAICS retail and wholesale industry. For example, in the case of a tobacco and cigarette manufacturer in China, we then conduct a search for the key words in the NAICS database for retail and wholesale industries whose description contained related words to Tobacco. In this case we link tobacco production to the wholesale industry 424940 "Tobacco and Tobacco Product Merchant Wholesalers" and 424590 "Other Farm Product Raw Material Merchant Wholesalers" which includes "Auction markets for tobacco" in their description, as well as the retail industry 453991 "Tobacco Stores". Of course these are imperfect measures of the retailers and wholesalers truly stocking these products, as we usually exclude generalised retailers.

In this way our matching may be subject to measurement error. In an attempt to reduce it, when we construct the industry exposure measure we take the average number of Chinese workers that work in the industries which are matched to each Chinese industry code.

Appendix C

Ethnic Chinese Networks and Technology Diffusion: The Chinese Exclusion Act

Figures

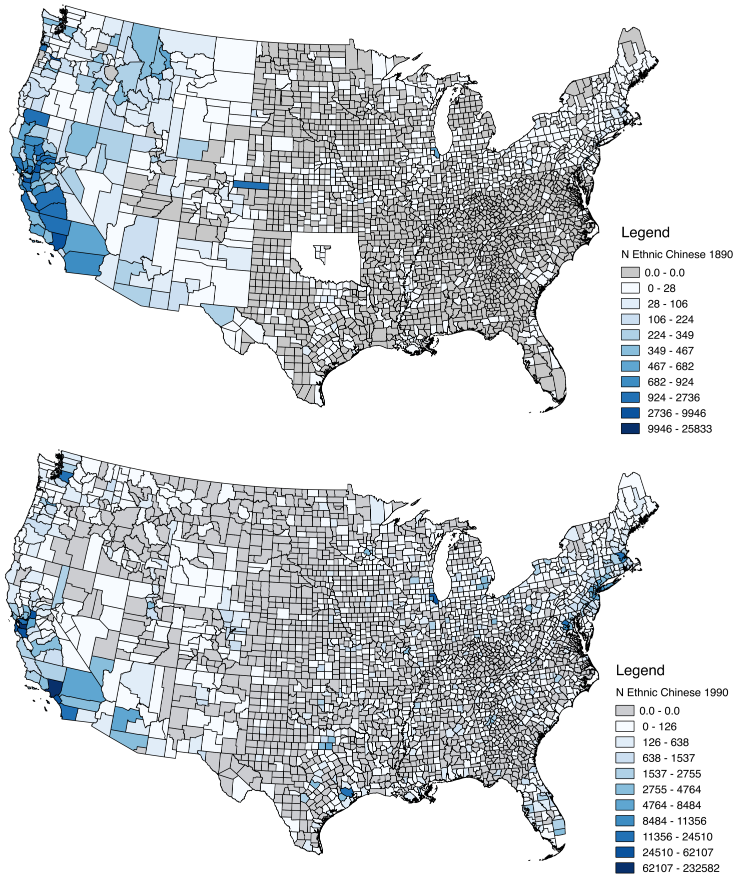


Figure C.0.1. Chinese Settlements 1890 and 1990 in the U.S.

Notes: These maps shows the geographic distribution of ethnic Chinese people in 1890 (top) and 1990 (bottom). Data source: U.S Population Census.

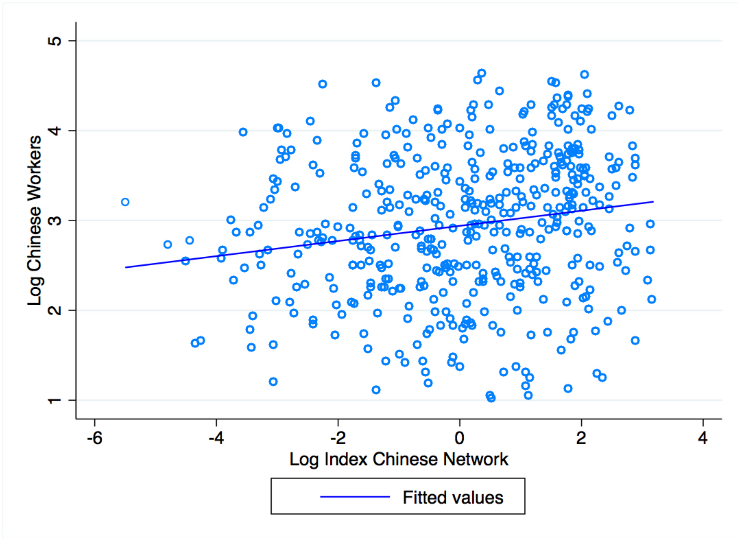


Figure C.0.2. Index of Chinese Network and Chinese Employment by Industry

Notes: This graph represents the correlation between the index of ethnic network exposure and number of employees by industry in the U.S. in 1990.

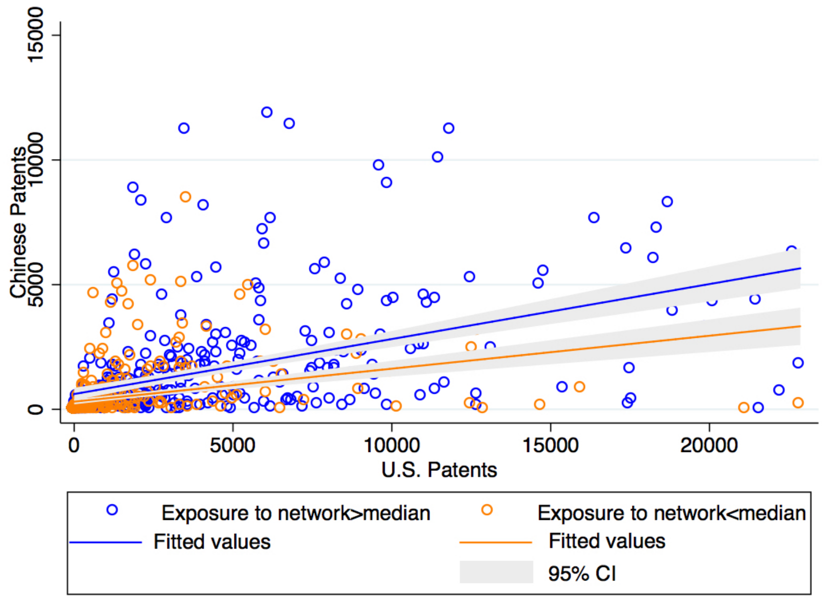


Figure C.0.3. Chinese and U.S. Patents by Network Exposure

Notes: This figure represents the correlation between Chinese and U.S. patents by four digit industry classification in industries with level of ethnic Chinese network exposure in the U.S. above (blue) and below (orange) median. For this figure I pool data over the years 1985-2006, and to avoid outliers I exclude the patent classes with the top 5% number of patents.

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001), 'The colonial origins of comparative development: An empirical investigation', *American Economic Review* pp. 1369–1401.

Ager, P., Brueckner, M. and Herz, B. (2016), 'The boll weevil plague and its effect on the southern agricultural sector, 1889–1929', *Explorations in Economic History* .

Agrawal, A., Kapur, D. and McHale, J. (2008), 'How do spatial and social proximity influence knowledge flows? evidence from patent data', *Journal of urban economics* **64**(2), 258–269.

Agrawal, A., Kapur, D., McHale, J. and Oettl, A. (2011), 'Brain drain or brain bank? the impact of skilled emigration on poor-country innovation', *Journal of Urban Economics* **69**(1), 43–55.

Ahn, J., Khandelwal, A. K. and Wei, S.-J. (2011), 'The role of intermediaries in facilitating trade', *Journal of International Economics* **84**(1), 73–85.

Alesina, A., Giuliano, P. and Nunn, N. (2013), 'On the origins of gender roles: Women and the plough', *The Quarterly journal of economics* **128**(2), 469–530.

Anderson, J. E. and Marcouiller, D. (2002), 'Insecurity and the pattern of trade: An empirical investigation', *Review of Economics and statistics* **84**(2), 342–352.

Bailey, M. J. and Collins, W. J. (2006), 'The wage gains of african-american women in the 1940s', *The Journal of Economic History* **66**(03), 737–777.

Banerjee, A. and Munshi, K. (2004), 'How efficiently is capital allocated? evidence from the knitted garment industry in tirupur', *The Review of Economic Studies* **71**(1), 19–42.

Bernard, A. B., Jensen, J. B. and Schott, P. K. (2009), Importers, exporters and multinationals: a portrait of firms in the us that trade goods, *in* 'Producer dynamics: New evidence from micro data', University of Chicago Press, pp. 513–552.

Bertocchi, G. (2015), 'The legacies of slavery in and out of africa'.

Biavaschi, C., Giulietti, C. and Siddique, Z. (2013), 'The economic payoff of name americanization'.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995), 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American statistical association* **90**(430), 443–450.

Boustan, L. P. and Collins, W. J. (2014), The origin and persistence of black-white differences in women's labor force participation, *in* 'Human Capital in History: The American Record', University of Chicago Press, pp. 205–240.

Burchardi, K., Chaney, T. and Hassan, T. A. (2016), 'Migrants, ancestors, and investments', *NBER Working Paper* (w21847).

Chaney, T. (2014), 'The network structure of international trade', *The American Economic Review* **104**(11), 3600–3634.

Chay, K. and Munshi, K. (2013), 'Black networks after emancipation: Evidence from reconstruction and the great migration'.

Chen, Z., Lu, M. and Xu, L. (2014), 'Returns to dialect: Identity exposure through language in the chinese labor market', *China Economic Review* **30**, 27–43.

Christian, C. (2014), 'Lynchings, labour, and cotton in the us south'.

Collins, W. J. and Wanamaker, M. H. (2014), 'Selection and economic gains in the great migration of african americans: New evidence from linked census data', *American Economic Journal: Applied Economics* **6**(1), 220–252.

Combes, P.-P., Lafourcade, M. and Mayer, T. (2005), 'The trade-creating effects of business and social networks: evidence from france', *Journal of international Economics* **66**(1), 1–29.

Dell, M. (2010), 'The persistent effects of peru's mining mita', *Econometrica* **78**(6), 1863–1903.

Farré, L. and Vella, F. (2013), 'The intergenerational transmission of gender role attitudes and its implications for female labour force participation', *Economica* **80**(318), 219–247.

Faure, D. and Siu, H. F. (1995), *Down to Earth: The Territorial Bond in South China*, Stanford University Press.

Felbermayr, G. J. and Toubal, F. (2012), 'Revisiting the trade-migration nexus: Evidence from new oecd data', *World Development* **40**(5), 928–937.

Fernández, R. and Wong, J. (2014), 'Unilateral divorce, the decreasing gender gap, and married women's labor force participation', *The American Economic Review* **104**(5), 342–347.

Foley, C. F. and Kerr, W. R. (2013), 'Ethnic innovation and us multinational firm activity', *Management Science* **59**(7), 1529–1544.

Foster, L., Haltiwanger, J. and Syverson, C. (2008), 'Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?', *The American Economic Review* **98**(1), 394–425.

Goldin, C. (1977), 'Female labor force participation: The origin of black and white differences, 1870 and 1880', *The Journal of Economic History* **37**(01), 87–108.

Goldin, C. (1990), *Understanding the gender gap: An economic history of American women*, New York: Cambridge University Press.

Goldin, C. D. (1991), 'The role of world war ii in the rise of women's employment', *The American Economic Review* pp. 741–756.

Goldin, C. and Sokoloff, K. (1984), 'The relative productivity hypothesis of industrialization: The american case, 1820 to 1850', *The Quarterly Journal of Economics* **99**(3), 461–487.

Gomez, E. T. and Cheung, G. C. (2009), 'Family firms, networks and ethnic enterprise: Chinese food industry in britain', *East Asia* **26**(2), 133–157.

Gould, D. M. (1994), 'Immigrant links to the home country: empirical implications for us bilateral trade flows', *The Review of Economics and Statistics* pp. 302–316.

Greif, A. (1989), 'Reputation and coalitions in medieval trade: evidence on the maghribi traders', *The journal of economic history* **49**(04), 857–882.

Greif, A. (1993), 'Contract enforceability and economic institutions in early trade: The maghribi traders' coalition', *The American economic review* pp. 525–548.

Griffith, R., Harrison, R. and Van Reenen, J. (2006), 'How special is the special relationship? using the impact of us r&d spillovers on uk firms as a test of technology sourcing', *The American Economic Review* **96**(5), 1859–1875.

Gutman, H. G. et al. (1976), *Black family in slavery and freedom, 1750-1925*, Pantheon Books.

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2001), The nber patent citation data file: Lessons, insights and methodological tools, Technical report, National Bureau of Economic Research.

Head, K., Mayer, T. and Ries, J. (2010), 'The erosion of colonial trade linkages after independence', *Journal of international Economics* **81**(1), 1–14.

Higgs, R. (1976), 'The boll weevil, the cotton economy, and black migration 1910-1930', *Agricultural History* **50**(2), 335–350.

Hing, B. O. (1993), *Making and remaking Asian America through immigration policy: 1850-1990*, Stanford University Press.

Holmes, T. J., McGrattan, E. R., Prescott, E. C. et al. (2013), Technical appendix for quid pro quo: Technology capital transfers for market access in china, Technical report, Federal Reserve Bank of Minneapolis.

Hornbeck, R. and Naidu, S. (2014), 'When the levee breaks: black migration and economic development in the american south', *The American Economic Review* **104**(3), 963–990.

Javorcik, B. S., Özden, Ç., Spatareanu, M. and Neagu, C. (2011), 'Migrant networks and foreign direct investment', *Journal of Development Economics* **94**(2), 231–241.

Johnson, D. M. and Campbell, R. R. (1981), 'Black migration in america: A social demographic history', *Studies in Social and Economic Demography* (4).

Jones, J. (2009), *Labor of love, labor of sorrow: Black women, work, and the family, from slavery to the present*, Basic Books.

Keller, W. (2002*a*), 'Geographic localization of international technology diffusion', *The American Economic Review* **92**(1), 120–142.

Keller, W. (2002*b*), 'Trade and the transmission of technology', *Journal of Economic growth* **7**(1), 5–24.

Keller, W. (2004), 'International technology diffusion', *Journal of economic literature* **42**(3), 752–782.

Kerr, W. R. (2008), 'Ethnic scientific communities and international technology diffusion', *The Review of Economics and Statistics* **90**(3), 518–537.

Kerr, W. R. (2013), Heterogeneous technology diffusion and ricardian trade patterns, Technical report, National Bureau of Economic Research.

Kerr, W. R. and Lincoln, W. F. (2010), The supply side of innovation: H-1b visa reforms and us ethnic invention, Technical report, National Bureau of Economic Research.

Kleven, H. J. and Landais, C. (2016), 'Children and gender inequality: Evidence from denmark'.

Kotkin, J. (1993), *Tribes: How race, religion, and identity determine success in the new global economy*, Random House Incorporated.

Lange, F., Olmstead, A. L. and Rhode, P. W. (2009), 'The impact of the boll weevil, 1892–1932', *The Journal of Economic History* **69**(03), 685–718.

Lavely, W. (2001), Coding scheme for the language atlas of china.

Lee, E. (2003), *At America's gates: Chinese immigration during the exclusion era, 1882-1943*, Univ of North Carolina Press.

McKenzie, D. and Rapoport, H. (2007), 'Network effects and the dynamics of migration and inequality: theory and evidence from mexico', *Journal of development Economics* **84**(1), 1–24.

Melitz, J. (2008), 'Language and foreign trade', *European Economic Review* **52**(4), 667–699.

Metzer, J. (1975), 'Rational management, modern business practices, and economies of scale in the ante-bellum southern plantations', *Explorations in Economic History* **12**(2), 123–150.

Montgomery, J. D. (1991), 'Social networks and labor-market outcomes: Toward an economic analysis', *The American economic review* **81**(5), 1408–1418.

Moser, P., Voena, A. and Waldinger, F. (2014), 'German jewish émigrés and us invention', *The American Economic Review* **104**(10), 3222–3255.

Munshi, K. (2003), 'Networks in the modern economy: Mexican migrants in the us labor market', *The Quarterly Journal of Economics* pp. 549–599.

Nanda, R. and Khanna, T. (2010), 'Diasporas and domestic entrepreneurs: Evidence from the indian software industry', *Journal of Economics & Management Strategy* **19**(4), 991–1012.

Nunn, N. (2008), 'The long-term effects of africa's slave trades', *The Quarterly Journal of Economics* **123**(1), 139–176.

Nunn, N. and Wantchekon, L. (2011), 'The slave trade and the origins of mistrust in africa', *The American Economic Review* **101**(7), 3221–3252.

Obstfeld, M. and Rogoff, K. (2001), The six major puzzles in international macroeconomics: is there a common cause?, *in* 'NBER Macroeconomics Annual 2000, Volume 15', MIT press, pp. 339–412.

Ochsner, C. and Roesel, F. (2016), 'Migrating extremists'.

Parsons, C., Vézina, P.-L. et al. (2014), 'Migrant networks and trade: The vietnamese boat people as a natural experiment', *University of Oxford, Department of Economics: Economics Series Working Papers* (705).

Ransom, R. L. and Sutch, R. (2001), *One kind of freedom: The economic consequences of emancipation*, Cambridge University Press.

Rauch, J. E. (1999), 'Networks versus markets in international trade', *Journal of international Economics* **48**(1), 7–35.

Rauch, J. E. (2001), 'Business and social networks in international trade', *Journal of economic literature* pp. 1177–1203.

Rauch, J. E. and Trindade, V. (2002), 'Ethnic chinese networks in international trade', *Review of Economics and Statistics* **84**(1), 116–130.

Rotunno, L. and Vézina, P.-L. (2012), 'Chinese networks and tariff evasion', *The World Economy* **35**(12), 1772–1794.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., King, M. and Ronnander, C. (2010), 'Integrated public use microdata series: Version 3.0 [machinereadable database]. minneapolis: Minnesota population centerdistributor], 2004'.

Saxenian, A. (2002), 'Silicon valleyâĂŹs new immigrant high-growth entrepreneurs', *Economic development quarterly* **16**(1), 20–31.

Saxenian, A. (2006), 'The new argonauts'.

Shlomowitz, R. (1979), 'The origins of southern sharecropping', *Agricultural History* **53**(3), 557–575.

Silverman, B. S. (2003), *Technological resources and the logic of corporate diversification*, Routledge.

Stock, J. H., Wright, J. H. and Yogo, M. (2002), 'A survey of weak instruments and weak identification in generalized method of moments', *Journal of Business & Economic Statistics* **20**(4), 518–530.

Van Deburg, W. L. (1997), *Modern Black Nationalism: From Marcus Garvey to Louis Farrakhan*, number 2, NYU Press.

Voss, B. L. and Allen, R. (2008), 'Overseas chinese archaeology: Historical foundations, current reflections, and new directions', *Historical Archaeology* pp. 5–28.

Walsh, L. S. (1989), 'Plantation management in the chesapeake, 1620–1820', *The Journal of Economic History* **49**(02), 393–406.

Wayne, T. K. (2007), *Women's roles in nineteenth-century America*, Greenwood Publishing Group.

Weidenbaum, M. L. and Hughes, S. (1996), *The bamboo network: How expatriate Chinese entrepreneurs are creating a new economic superpower in Asia*, Simon and Schuster.

Woodruff, C. and Zenteno, R. (2007), 'Migration networks and microenterprises in mexico', *Journal of development economics* **82**(2), 509–528.

Woon, Y.-F. (1990), 'International links and the socioeconomic development of rural china: an emigrant community in guangdong', *Modern China* **16**(2), 139–172.

Yu, H. (2011), The intermittent rhythms of the cantonese pacific, *in* 'Connecting Seas and Connected Ocean Rims', Brill, pp. 393–414.

Zhang, K. H. and Song, S. (2002), 'Promoting exports: the role of inward fdi in china', *China Economic Review* **11**(4), 385–396.

Zhao, X. (2002), *Remaking Chinese America: immigration, family, and community, 1940-1965*, Rutgers University Press.