Familiar Strangers: Lineage Connection and Diaspora Direct Investment in China

Fanghao Chen Jinan University

Ruichi Xiong University of Toronto

Xiaobo Zhang Peking University & IFPRI

October 10, 2022

Abstract

As a developing country, China became the top foreign direct investment (FDI) destination in just a few decades, defying the "Lucas Paradox." Using a unique administrative dataset of universe foreign firms in China, we document that initial FDIs were mainly driven by Chinese diaspora, while massive non-diaspora FDIs did not materialize until to a later stage. Leveraging the staggered opening-up of Chinese prefectures during 1981-1996 as an identification strategy, we find that following the opening-up, diaspora direct investments (DDIs) are more likely to enter prefectures with stronger lineage connections. These prefectures also witness a greater number of non-diaspora foreign and domestic private entrants later on.

Key Words: Diaspora, Lineage Network, Foreign Investment, FDI Spillover, China

JEL Codes: F21, F22, F23, O19

*Fanghao Chen: chenfanghao@jnu.edu.cn, School of Economics at Jinan University. Ruichi Xiong: ruichi.xiong@rotman.utoronto.ca, Rotman School of Management at University of Toronto. Xiaobo Zhang: x.zhang@gsm.pku.edu.cn, Guanghua School of Management at Peking University and International Food Policy Research Institute. Ruichi Xiong and Fanghao Chen acknowledge the financial support from China Research Initiative at the Rotman School of Management, University of Toronto. Xiaobo Zhang is grateful for the financial support from China Natural Science Foundation (project numbers 71874008, 71441008, 71873121, and 72192844). We thank Nathaniel Baum-Snow, April Franco, Kailin Gao, Ruixin Wang, Shuo Chen, Shilin Zheng as well as seminar participants at University of Toronto, Sun Yat-Sen University, Harbin Institute of Technology (Shenzhen), Jiaxing University, University of Nottingham Ningbo, Peking University, Southwestern University of Finance and Economics, Fudan University, Huaqiao University, University of International Business and Economics, European Winter Meetings of the Econometric Society, and ICCDS for helpful comments and discussions. We also thank Bohan Yang for excellent research assistance. All errors remain our own.

1 Introduction

In developing countries, capital is scarcer relative to labor, implying higher returns to capital. In principle capital should flow from developed countries to developing countries to chase the higher returns. Yet, in reality developing countries attract less amount of foreign direct investment (FDI) than their developed counterparts (World Bank, 2017). The puzzle was highlighted in Lucas (1990) and later is referred to as the "Lucas Paradox". A large body of subsequent literature has tried to explain the puzzle.

Human capital (Noorbakhsh, Paloni and Youssef, 2001), institutional quality (Alfaro, Kalemli-Ozcan and Volosovych, 2008), political risk (Julio and Yook, 2012), or a combination of these factors (World Bank, 2017) have been listed as key determinants of the puzzle. However, it is a long journey to improve human capital and build sound institutions, the two fundamental factors. If a developing country, like Peru, can manage to improve its human capital and institutional quality to the levels of Australia, it is already a developed country, to which the puzzle does not apply. Lucas (1990) (page 96) suspects that "[o]nly insofar as political risk is an important factor in limiting capital flows can we expect transfers of capital to speed the international equalization of factor prices." A research question arises: in the absence of high-quality fundamentals, can foreign investors overcome the political risk in the first place?

The recent development experience of China defies the Lucas Paradox, demonstrating that it is possible to attract FDIs in imperfect institutional environment. When China started its reform and opening-up in the late 1970s and 1980s, market institutions were far from adequate. For instance, private ownership was not even recognized by the constitution until 2004. Yet, in just several decades, China has transformed from an impoverished country to the second largest economy in the world, becoming a leading destination of FDIs. One salient feature is that the initial FDIs were primarily driven by Chinese diaspora from Hong Kong, Macao, Taiwan, and South Asian countries (Vogel, 1990; Lever-tracy, Ip and Tracy, 1996; Huang, Jin and Qian, 2013). The massive FDIs from multinationals in developed countries, which were not related to diaspora, did not come to fruition until the late 1990s.

Thanks to their lineage and cultural familiarity with local conditions, overseas Chinese investors were in a better position to navigate the uncertain political and economic environments

than non-diaspora foreign investors. After observing the success of diaspora direct investments (DDIs), other foreign and domestic firms followed suit. This is the story our paper aims to tell.

Based on the administrative business registration data maintained by the State Administration of Industry and Commerce of China (SAIC), we first identify foreign firms by ownership code. By examining the surname and ID of a foreign firm' legal representative, we can determine whether the foreign firm is controlled by a Chinese diaspora or not. Next, we measure the degree of lineage connection between the diaspora firm and each prefecture based on the probability of the surname of the firm's legal representative appearing in the general population of the prefecture.

By leveraging the staggered opening-up of Chinese prefectures in the 1980s and early 1990s, we conduct an exercise of event studies, showing that after a prefecture's opening-up, DDIs are more likely to fly to prefectures with stronger lineage connections. However, after China's accession to the WTO, the lineage bias disappears, suggesting that as the market environment improves, the role of lineage connection in helping to attract foreign direct investments is no longer as important as before.

We also find a strong long-term spillover effect of DDIs. Prefectures initially attracting more DDIs later witness a greater number of entry of foreign and domestic private firms. The economic effect is sizable. A seed of diaspora firm in 1996 at the prefecture level brings about additional 0.418 non-diaspora foreign firms and about 137 domestic private firms survived as of 2014.

Our paper is closely related to the literature on the role of social affinities in facilitating economic exchange across regions as reviewed by Rauch (2001). More recent literature has covered richer dimensions of social affinities, such as ethnic ties (Rauch and Trindade, 2002), bilateral trust (Guiso, Sapienza and Zingales, 2009), linguistic proximity (Melitz and Toubal, 2014), migrant network (Javorcik et al., 2011; Parsons and Vézina, 2018), and even Facebook relationships (Kuchler et al., 2020). We complement the literature by studying the effect of surname-based linage connections on inward FDIs.

While most studies in this strand of literature examine the impact on trade, the effect of social affinities on investment has been less examined. The existing studies mostly use cross-country

data. However, it is difficult to draw a casual conclusion due to the presence of unobserved confounding factors in cross-country settings. An alternative approach is to focus on a particular country or industry. Burchardi, Chaney and Hassan (2019), the closest study to ours, shows that ancestry compositions matter to FDI inflows into U.S. counties. Our article differs from their seminar paper in two aspects. First, we focus on China, a large developing country. Second, we study the effect of surname-based lineage connection of diaspora on investment back in their homeland.

Using survey data, Nanda and Khanna (2010) finds that diaspora networks play an important role in providing business leads and financing in Indian software industry. Saxenian (2007) documents a rising trend of highly skilled entrepreneurs in Silicon Valley launching new enterprises in their home countries. By comparison, our study covers the universe of diaspora firms in China in all the sectors rather than just the high-tech sectors. In developing countries, the labor-intensive low-tech sectors are more common and sizable than high-tech sectors. Yet, knowledge about DDI in these sectors in developing countries is scant. Our paper fills in this gap.

Another study close to ours is Ma (2018). Employing the geographical borders of Chinese linguistic dialect zones as identification strategy based on the Annual Survey of Industrial Firms in China (ASIF) with sales above five million RMB from 1998 to 2006, he finds a positive effect of linguistic similarity in attracting investment from Hong Kong, Macao, and Taiwan (HMT). Compared with Ma (2018), our sample covers all diaspora firms regardless of their origins and sizes. Moreover, we zoom in the earlier period prior to 1998, when the development of market institutions was still in the infancy stage. Finally, we show a strong seeding effect of the early lineage-driven DDIs on subsequent non-diaspora investment.

Our study also speaks to the literature on the spillovers of inward FDIs in China. Using the ASIF data, Huang, Jin and Qian (2013) find that the HMT investment is unperformed compared to FDIs from other countries. Ma (2018) shows that the presence of HMT investment has a negative effect on the performance of local private enterprises. Using the same dataset, Lu, Tao and Zhu (2017) concludes that the spillover effect of multinationals on local domestic firms is rather limited. Three key features distinguish our paper from these studies. First, our study

covers all DDIs, not just HMT investments used in previous studies. Second, the administrative universe firm registration database allows us to study the long-run seeding effect on the entry of both domestic and foreign firms. Third, our focus is on the extensive margin (i.e. firm entry), rather than the intensive margin (i.e. productivity) as discussed in the previous studies.

The findings of the paper may shed policy implications for other developing countries. In the absence of ideal institutional environment, developing countries can strategically tap diaspora investment in the beginning, while improving local institution quality along the way, a fundamental factor to attracting more subsequent non-diaspora FDIs. Our findings do not negate the importance of improving local institutions and competitive environment in attracting FDIs. As shown in Du, Lu and Tao (2008), U.S. multinationals prefer to invest in regions with better quality of economic institutions after China launched a nationwide opening-up policy and DDIs had demonstrated a success.

The rest of the paper is organized as following. Section 2 introduces the historical background. Section 3 describes the data and the measure of lineage connections. Section 4 presents our identification strategy and empirical findings. Section 5 estimates the long run spillover effect of early lineage-driven DDIs on subsequent entry of non-diaspora investment. Section 6 concludes.

2 Historical Background

China was largely a closed planned economy until the policy of ”reform and opening-up” was introduced in 1978. However, the process of opening-up was gradual and did not happen over night (Lever-tracy, Ip and Tracy, 1996; Branstetter and Lardy, 2006).

Table 1 summarizes the staggered opening process during the early opening-up period. In 1979, the *Law on Joint Ventures* was passed. It provided, for the first time in thirty years, a legal framework under which foreign firms were allowed to operate in Mainland China. Next year, four Special Economic Zones (SEZ), including Shenzhen, Zhuhai, Shantou, and Xiamen, were established. Foreign corporations enjoyed broad autonomy and preferential tax treatment in these SEZs. In 1984, additional 14 cities were granted the status of *Open Coastal City* for

attracting FDIs[1] . Then Yingkou prefecture, Weihai prefecture, and Hainan province were also added to the list of opening regions in 1985, 1987, and 1988, respectively. The political crisis in 1989 put a halt to the process of opening-up for a few years until Deng Xiaoping's visit to the southern China in 1992. The visit heralded a new era of opening-up to foreign capital at a much larger scale.

The top panel of Figure 1 plots the number of foreign entrants and the survival-adjusted number (existing for at least four years or still alive as of 2014) in Mainland China by entry year, aggregated from the administrative business registration database, in relation to the timing of staggered rolling-out opening policies. As shown in the panel, the number of foreign entrants rose steadily from 1980 to 1991 along with the opening of SEZs and *Open Coastal Cities*. The number of foreign entrants spiked in 1992 and 1993, thanks to Deng Xiaoping's southern tour in 1992, which showed the government's determination to further opening up. Another milestone is China's formal accession to the WTO in 2002. Following the WTO agreement, China eliminated most restrictions on foreign entry and ownership, rekindling a surge in foreign entrants during 2000-2005.

Noticeably, the entry of foreign firms was already very active even before China joined the WTO. In fact, by 1999, China had already become one of the most popular destinations for FDI flows, only second to the United States (Huang, 2003). This achievement is remarkable considering that at the time China still lacked solid market institutions for attracting foreign investment according to the conventional wisdom.

FDIs are not a homogeneous group. One defining feature of FDIs in China is the high concentration of diasporas direct investment. The bottom panel of Figure 1 decomposes foreign entrants into diaspora ones and non-diaspora ones. From 1980 to 2014, the share of diaspora entrants was over 90% from 1984 to 1994, highlighting the dominant role of diaspora direct investments in the early opening-up period. Since then, its relative importance has declined, but still accounting for more than 60% of total FDIs as of 2015.

[1]Throughout the paper, we use city and prefecture interchangeably. To be precise, "city" is not a well defined notion in China, and not generally comparable to that in western countries (such as Metropolitan Statistical Areas). The administrative unit closest to the size of a city is a prefecture in China (despite great variance in administered area), including those of a higher administrative status (such as Beijing, Shanghai, Chongqing, and Tianjin). Rural counties under the administration of a city are also included in the statistics for that city.

China has a long history of emigration dating back to as early as 1300s A.D. (Kuhn, 2008). According to the Global Migrant Origin Database, the stock of overseas Chinese was as high as 5.8 million in 2007, globally ranking the 6th place in terms of origin countries.[2] Among the vast number of overseas Chinese, many have become successful entrepreneurs in the host countries. According to the estimate by *The Economist.* in 2019, Chinese diasporas contributed to more than three quarters of the South East Asian billionaire wealth.[3]

A hiatus of China's closure to the outside world for three decades since 1949 did not fully cut off the bond between overseas Chinese and their ancestral land. Many overseas Chinese, though physically abroad, are emotionally attached to their ancestral hometowns. For example, they maintain some traditional practices, i.e. compiling genealogy books and worshiping ancestors in lineage temples at their adopted land (Clark, 2015); They remain socially in touch with group members of their lineage at their ancestral land in the form of sending remittance and writing letters (Tan, 2006; Kuhn, 2008).[4]

After China reopened its door in 1978 and strove to attract foreign investment, the persistent yet lately dormant networks between overseas Chinese and their ancestral land rekindled.[5] Thanks to the lineage ties, massive diaspora-led direct investments ventured into china, particularly to places with strong lineage connections, despite initial imperfect market and institutional environment, while most non-diaspora FDIs held an attitude of "wait and see".

3 Data Description

This section first describes the major data for empirical analyses and their sources. Second, we explain how to identify diaspora firms as foreign firms that are controlled by overseas Chinese.

[2] www.sussex.ac.uk/Units/SCMR/drc/about/index.html

[3] "Chinese Diaspora Inc: High-Wire Act," *The Economist*, May 30th, 2020

[4] In China, lineage refers to the group of descendants of one common patrilineal ancestor across multiple generations who share the same surname. For thousands of years in China, family clans have played an important role in providing local public goods and promoting within-group cooperative behavior as an informal institution (Greif and Tabellini, 2017)

[5] One may be concerned that over time, the emotional attachment of Chinese diasporas, especially younger and foreign-born ones, with their ancestral homelands might wane, weakening the strength of lineage connection. However, according to the interviews conducted by Tan (2006), many China-born parents would bring along their foreign-born children to ancestral hometowns to cultivate their self-identification as a member of the lineage group. Moreover, the second or third generations of diasporas largely embrace their parents' business networks after taking the reins, even if they no longer consider themselves Chinese.

Next we define the measure of lineage connection for each surname-prefecture pair. Last, the summary statistics of key variables used in the empirical analyses are provided.

3.1 Data Sources

The main dataset we use is the administrative business registration database maintained by the SAIC, which tracks the universe of firms ever registered in China. We choose 2014 as the end year of the sample, because the number of new entrants spiked after China launched a national business registration reform in the year. For each firm, we observe its entry date, exit date (left blank if the firm still survived as of 2014), latest-reported registered capital by the end of 2014, 4-digit industry, county-level location, ownership type, and list of immediate shareholders and registered personnel, including board members and senior executives.

The second dataset is the China Population Survey in 2005, which reports individual surnames of a 0.2% representative sample. Since the sample is representative at the prefecture level, We can compute the surname distributions for each prefecture. We then leverage the geographic variation in surnames to construct our measure of lineage connection for each surname-prefecture pair.

3.2 Identify Diaspora Firm

By definition, diaspora firms have to be foreign firms in the first place. There are two ways in the literature to define foreign firms. One approach relies on the administrative firm ownership code assigned by government agencies, while the other approach uses the shareholder information. In this paper we opt for the first approach, which is less complicated than the second one. In reality, a firm's immediate shareholder structure does not necessarily coincide with its structure of ultimate control. For example, investors can exert control over a firm through holding shells (Bai et al., 2020). If following the second approach, we have to go through many layers of ownership structure to determine the real control shareholders. For simplicity, we decide to adopt the more straightforward first approach by using the administrative ownership code

readily available in the dataset to identify foreign firms.[6]

Having pinned down foreign firms, the next step is to identify diaspora firms. It would be ideal to define diaspora firms as those with overseas Chinese as the controlling shareholder. Unfortunately, most of the shareholders of foreign firms are foreign entities registered outside China. Thus, it is impossible to trace the shareholders of those foreign entities (Bai et al., 2020). As a second best, we use the information of legal representatives to help identify diaspora firms.

According to the corporate law in China, legal representatives take the major legal responsibility of the registered firms. Legal representatives have been used as a proxy for entrepreneurs in the literature (Dai et al., 2019). One key advantage of this method is that the names of legal representatives are public available in the business registration database. We develop an algorithm as described in Appendix A.1 to extract the surnames of the legal representatives of foreign firms and examine whether they are ethnic Chinese or not.[7]

A legal representative is identified as an overseas Chinese if she or he has a Chinese surname and holds a non-mainland Chinese ID.[8] Diaspora firms refer to those foreign firms with overseas Chinese as their legal representatives. We present several pieces of evidence in support of this definition. First, as demonstrated in Appendix A.2, 97% of all foreign firms have unique legal representatives, but the list of directors, CEOs, and top-ranked executives is often incomplete. Using legal representatives as a proxy has the highest sample coverage. Second, because 92% of immediate shareholders of foreign firms are legal persons rather than natural persons, there are no surnames for most shareholders in these foreign firms. It is almost impossible to back out the ultimate nature-person owners of foreign firms through ownership chain, because the SAIC does not have information on the owners of foreign legal entities that are registered abroad (Bai et al., 2020). Third, Appendix A.1 reports that the chance of a legal representative overlapping

[6]One might be concerned that a firm's ownership might have changed as a result of privatization taking place during the late 1990s. However, privatization at the time primarily took the form of Management Buy-Outs (MBO) rather than selling shares to foreigners. Moreover, any change in ownership will trigger a change in firm identifier as well, creating a new legal entity (Chen et al., 2020). These new legal entities, including those with new foreign owners, are included in our sample. Thus it is unlikely that the potential change in ownership type would affect our identification of foreign firms. For these reasons, we stick to the first approach.

[7]There is a concern that in a few countries, such as Thailand and Indonesia, some Chinese diasporas abandoned their Chinese names in exchange for local names under the pressure of national assimilation policy. In the presence of name assimilation, our method may underestimate the scale of diaspora direct investments. The problem may not be that serious. As documented in Tan (2006), Chinese diasporas often use their original Chinese names to signal their Chinese identity when dealing with Chinese businesses.

[8]Non-mainland IDs include foreign passports and residency cards of Hong Kong, Macau, or Taiwan (HMT).

with a top-ranked executive within a foreign firm is 93%. This means that a legal representative is highly likely to be the person in charge of the business.

The definition is not perfect and subject to several caveats. First, the sample of diaspora firms excludes disguised foreign firms that are actually represented by individuals from mainland China. These seemingly diaspora firms are the so-called "round-trip FDIs" (Huang, 2003). Chen (2022) more rigorously defines "round-trip FDIs" as "the direct investment activities where a domestic resident makes investment in the territory of the People's Republic of China directly or through special-purpose vehicles, that is, establishes a foreign-invested enterprise or a project through new establishments, M&A and other modes, and acquires any ownership, right of control, right of business management, or other relevant rights and interests". We document in Figure A.8 that the levels and trends of round-trip FDIs identified using our method are close to the previous literature which uses aggregate statistics to gauge the presence of round-trip FDIs in China (Geng, 2019).

Since the purpose of round-trip FIDs is to seek the preferential treatment granted by the government to foreign firms, including this group of firms in the sample would exaggerate the presence of diaspora firms in China, making it more difficult for us to detect the seeding role of diaspora firms as found in the paper. Second, we cannot rule out the possibility that foreign citizens delegate the firm control rights to Chinese citizens. In this case, the number of diaspora firms would be under counted. As a robustness check, we include this group of foreign firms as diaspora firms. The effect becomes even stronger as shown in section 4.3.

Third, our main dataset excludes non-diaspora foreign firms since we want to keep our sample as comparable as possible. Including non-diaspora foreign firms would mechanically increase the size of the control group with no lineage connection. Our findings are robust to the inclusion of non-diaspora foreign firms in the sample. Last, since legal representatives are not necessarily the major shareholders of firms, it is possible that some labelled diaspora firms are actually not owned by overseas Chinese. However, the action of appointing overseas Chinese as legal representatives likely reveals the foreign firms' preference in taking advantage of lineage connection. Ignoring these observations would underestimate the effects of lineage connection

in facilitating the entry of other foreign firms. [9]

3.3 Measure Lineage Connection

For a given surname of a diaspora firm's legal representative, we use its geographic distribution, drawn from China Population Survey in 2005, to measure the strength of lineage connection between the firm and different locations. Specifically, we define the lineage connection m_{sp} between a surname s and a prefecture p as follows:[10]

$$m_{sp} = \frac{E_{sp}}{\sum_p E_{sp}} \tag{1}$$

where E_{sp} denotes the size of population with surname s in prefecture p, and the denominator stands for the total population with surname s in China. This measure can be interpreted as the probability of one overseas Chinese with surname s having prefecture p as the ancestry origin. Notice that our measure of lineage connection is size-free because it is normalized by the total population with the same surname for China as a whole.[11]

Using surnames instead of migration patterns to proxy lineage connections has a key advantage in data availability. Although there are relatively rich data on immigration into developed countries such as the United States (Burchardi, Chaney and Hassan, 2019; Sequeira, Nunn and Qian, 2020), data on emigration at the sub-national level from developing countries like China is largely unavailable to our knowledge. Given the paucity of emigration data, our surname-based measure serves as an alternative proxy for lineage connections, by making use of the fact that lineage groups in China are usually operating within surnames (Clark, 2015).

Figure 2 visualizes the geographic distributions of the 20 most populous surnames in China, which are ranked based on the China Population Survey in 2005, from three samples: emigration-

[9] There could be potential differences between diaspora-owned foreign firms and diaspora-managed foreign firms in incentives and performance. We leave the question for future research.

[10] We choose prefecture as the geographic unit of our empirical analysis mainly because the population survey data is representative at the prefecture level, but not at the lower county level.

[11] If we instead use E_{sp} as the measure of lineage connection, a spurious correlation between this measure and the entry of diaspora firms may occur even in the absence of lineage connections. Some populous surnames like "陈" ("Chen", "Chan", "Tan") among overseas Chinese are also widespread across China, the diaspora firms with these surnames are more likely to be observed in large prefectures where people with the same surnames are also more likely present. The positive correlation therefore does not necessarily mean lineage connections. By comparison, as shown in Appendix A.7, the lineage connection measure in equation 1 is mildly negatively correlated with the size of prefecture population by surname, ruling out the possibility of a positive spurious correlation.

intensive provinces (including Guangdong, Fujian, and Zhejiang), other provinces, and all legal representatives of diaspora firms registered from 1980 to 2014. Several salient features are apparent from the figure. Overall, the surname distribution among legal representatives of diaspora firms closely resembles that of the population in emigration-intensive provinces, but sharply differs from other provinces. For example, "陈" ("Chen", "Chan", "Tan")[12] is the most common surname both among overseas Chinese legal representatives and in emigration-intensive provinces. But it ranks only the fifth in popularity for Chinese population as a whole. This simple comparison suggests a possible lineage connection between diasporas and the destinations of their investments.

3.4 Summary Statistics

We restrict our attention to the early opening-up period from 1981 to 1996 before the Asian crisis and China's accession to WTO. Utilizing the entry and exit information in the SAIC database, we construct two measures of diaspora firm entry at surname-prefecture-year level: number of new entrants and survival-adjusted number of entrants (defined as those lasting for more than 4 years). We exclude four autonomous region，including Xinjiang, Tibet, Ningxia, and Inner Mongolia，which are mainly composed of ethnic minorities. Given that the surname-based lineage mostly operate among ethnic Han Chinese (Zhang, 2020), and most overseas Chinese are Han Chinese, our linage connection measure is not applicable there. We further exclude four mega cities in China, including Beijing, Shanghai, Guangzhou, and Shenzhen, because their massive economic sizes attract vast number of internal migrants, masking the historical distribution of local surnames.

Table 2 presents the summary statistics for all the variables used in empirical analyses. Panel A shows that only 5.2% of the surname-prefecture-year cells have at least one diaspora firm entry and the percentage for diaspora firms survived in 2014 is even smaller (0.9%). The lineage connection measure varies widely with a mean of 0.005 and a standard deviation of

[12]Due to Chinese-English translation, sometimes there are multiple English spellings for the same Chinese surname. Figure A.1 illustrates the complexity of spelling-character mapping between the two languages, using "陈" —the most common overseas Chinese surnames written in Chinese characters, and "Tan"—the most common English spelling that does not follow the regular *Pinyin* system for "陈", as an example. To address this challenge, we aggregate the overseas Chinese surnames to the Chinese character level using probabilistic weights, in cases a surname in Chinese character corresponds to multiple ways of English spelling. See Appendix A.1 for details.

0.011 as indicated in Panel B. Panel C and panel D summarize the main outcomes of interest at the prefecture-year or prefecture level which will be studied later in this paper.

4 Lineage Connection and Diaspora Firm Entry

4.1 Identification Strategy

We employ the following baseline specification to empirically identify the effect of lineage connection on the entry of diaspora firms:

$$Y_{spt} = \eta_{sp} + \theta_{st} + \delta_{pt} + \beta \times Open_{pt} \times m_{sp} + \lambda \times S_{spt} + \epsilon_{spt} \tag{2}$$

where s, p, and t denote surname, prefecture, and entry year, respectively. Y_{spt} represents the outcome variables at surname-prefecture-year level. One key outcome variable is the number of diaspora firm entrants. However, this outcome variable alone cannot capture potential variations in entry quality. As a robustness check, we supplement another measure to mitigate this concern: the survival-adjusted number of diaspora entrants, defined as the number of diaspora entrants that survive for no fewer than four years following Kerr and Nanda (2009).[13]

Under all specifications, we control for surname-prefecture fixed effects η_{sp}, surname-year fixed effects θ_{st}, and prefecture-year fixed effects δ_{pt} if not specifically mentioned. The broad set of fixed effects help us guard against a wide range of confounding factors, such as geographic advantages, place-based policies, and surname-specific expertise. Besides, we also control S_{spt}, the number of incumbent firms in the year prior to the entry year, in order to capture potential agglomeration or competition spillovers from incumbent firms.

The key variable of interest is the interaction term $Open_{pt} \times m_{sp}$. $Open_{pt}$ is a time-variant dummy indicating whether prefecture p was open to foreign capital at year t. $Open_{pt}$ equals 1 if prefecture p has been awarded the opening status since year t, and equals 0 otherwise.[14] m_{sp} is the measure of lineage connection between surname s and prefecture p as defined in equation

[13] Robustness checks using other thresholds are shown in Figure A.9 in the Appendix.

[14] See Table 1 for the time table of opening Chinese prefectures to foreign capital during the opening-up period. As Shanghai is excluded from the sample, the opening of Pudong district in 1990 is not captured in our analysis.

1. The error term ϵ_{spt} captures all the idiosyncratic disturbances. Standard errors are clustered at surname-prefecture level.

The coefficient of interest β measures the effects of the staggered opening of Chinese prefectures to foreign capital on the entry of diaspora firms in relation to their lineage connections with the prefectures. Our empirical design is essentially a staggered triple-difference strategy with varying treatment intensity across surname-prefecture pairs: (1) the differences in surnames within a prefecture before and after its opening; (2) the differences between opened and closed prefectures; (3) differences across surname-prefecture pairs with varying strengths of lineage connections.

The causal interpretation of our estimates depends upon one crucial assumption: the number of diaspora entrants for a given surname in a prefecture does not exhibit an existing trend prior to the prefecture's opening-up. We argue that this identification assumption is highly likely to hold for three reasons. First, our measure of lineage connection has been shown to be persistent (Bai, 2020). Second, foreign capital inflows were largely prohibited before a prefecture was granted the opening status. Third, various fixed effects have been controlled to reduce confounding factors that could contaminate our causal estimates.

To further check the validity of our identification strategy, we employ the following event-study framework:

$$Y_{spt} = \eta_{sp} + \theta_{st} + \delta_{pt} + \sum_{\tau=-4}^{4} \beta^{\tau} \times Open_{pt}^{\tau} \times m_{sp} + \lambda \times S_{spt} + \epsilon_{spt} \tag{3}$$

where $\tau = t - t_p$ refers to the time window relative to the opening shock of prefecture p. By employing this specification, we can examine the dynamic effects of lineage connection both before and after the local opening shocks. $Open_{pt}^{\tau}$ equals 1 if year t is τ years after the opening of prefecture p and 0 otherwise. The omitted benchmark group is $\tau = -1$. Hence all estimates of β^{τ} should be interpreted as being relative to one year prior to the opening shock. For our identification assumption to be valid, we expect β^{τ} to be statistically indifferent from zero for all $\tau < 0$.

As shown in the top two plots of Figure 3, β^{τ} is indifferent from zero when $\tau < 0$ no matter whether we use the number of diaspora entrants or the survival-adjusted number of diasporas

entrants as an outcome variable, validating the no pre-trend assumption.

4.2 Baseline Results

We first estimate equation 2 to evaluate how lineage connection affects the entry of diaspora firms in the face of opening-up at the prefecture level and present the results in columns (1) and (2) of Table 3. According to column (1), an increase in one standard deviation of lineage connection (0.011) is associated with 0.02 more diaspora entrants, accounting for one-third of the average number of diaspora entrants across all cells aggregated at the surname-prefecture-year level. When using the survival-adjusted number of diaspora entrants as an alternative outcome variable in Column (2), the coefficient for "Open $\times$ lineage Connection" variable remains highly positive and significant, suggesting that the lineage connection facilitates the entry of diaspora firms even if we adjust entry quality by survival.

To examine the time-varying effect, the top two panels of Figure 3 plot the event-study estimate for the same two outcome variables as in Table 3 based on equation 3. Interestingly, the effect is highly statistically positive only in the first two years after opening up and then diminishes.

The recent development in the econometrics of Difference-in-Difference (DID) or more generally Two Way Fixed Effects (TWFE) model points out that biases could emerge from staggered treatments in the presence of heterogeneous treatment effects.[15] Following the methodology and package in De Chaisemartin and D'Haultfoeuille (2020) and De Chaisemartin, D'Haultfoeuille and Guyonvarch (2021), we check whether our baseline results are still robust or not after correcting potential biases. We divide the surname-prefecture pairs with continuous lineage connection into two discrete groups: the ones with above or below median lineage connection. It would be ideal to apply the estimation methods using the original continuous treatment variable. However, we find that it is computationally infeasible due to our large dataset. Thus, we adopt this second-best approach.

Columns (3) and (4) of Table 3 report the average treatment effects when replacing the continuous connection variable with a dummy variable and correcting the potential bias of staggered

[15] See a survey in De Chaisemartin and D'Haultfoeuille (2022).

DID. The coefficient for the interaction term between the opening dummy and the connection dummy remains highly significant. As shown in the bottom two plots of Figure 3, there are no parallel trends prior to a prefecture's opening-up, further validating our identification strategy.

One key difference between the bottom two plots and top two plots in Figure 3 is that the effect of lineage connection lasts beyond the first two years after correcting biases originated from the staggered DID and heterogeneous treatment effects. This is likely because the staggered DID estimations fail to capture the spillovers from the early diaspora entrants on later ones, after all prefectures opened to foreign investment. These spillovers not only affect the diaspora entrants, but also subsequent non-diaspora entrants, which will be discussed in Section 5.

4.3 Robustness Checks

This section presents various analyses to check the robustness of the main findings.

Sorting on the Qualities of Entrants. There is a possibility that early diaspora firms have better quality than later ones. In this case, the observed positive effect of lineage connection would likely be overestimated, because the estimation fails to take the quality premium of early entrants into account. To mitigate this concern, columns (1) and (2) of Table 4 further look at two proxy variables for entry quality, survival ratio and average registered capital of diaspora entrants survived as of 2014. The coefficients for the interaction terms are insignificant from zero, suggesting that sorting on the qualities of entrants is not at play in driving our key results.

Alternative Dependent Variables. We notice that the number of entry and survived diaspora firms are rare at surname-prefecture-year level (see Table 2). The OLS estimates may yield biases when outcome variables do not follow a normal distribution. We use two ways to address this concern. First, we replace our firm count measures with a dummy variable indicating whether there is at least one entrant in a surname-prefecture-year triplet or not. Columns (1) to (3) of Table A.4 show that the effect of lineage connection still holds at the extensive margin.

Second, considering the entry likely follows a Poisson distribution, we also run Poisson regressions. The variation of lineage connections across surname-prefecture pairs, however, is too small to make the maximum likelihood estimation converge. We instead use a dummy

variable indicating whether the value of lineage connection is higher than the median as what we do in columns (3) and (4) of Table 3. As displayed in Table A.7, the results of Poisson regressions are reassuringly robust.

Alternative Lineage Connection Measure. Since the lineage measure is drawn from China Population Survey in 2005, it is likely subject to domestic migration prior to 2005, which might have shaped the spatial distribution of surnames and the entry of diaspora firms at the same time. From 1949 to 1978, domestic migration was highly restricted under the stringent Hukou system. Even after China gradually loosened its grip on internal migration since 1984, mass migration did not happen until the late 1990s (Tombe and Zhu, 2019). Clark (2015) documents that geographic distribution of certain elite surnames is rather stable throughout history even after warfare and revolutions. Bai (2020) compares the surname distribution across prefectures in China Population Survey in 2005 with the surname distribution in Harvard's China Biographical Database (CBDB), confirming that the two distributions are highly correlated.

Nonetheless, we provide additional checks by constructing an alternative measure based on the surnames information of registered personnel who have worked in *domestic firms* established before 1992, available from the SAIC database. Column (1) of Table A.5 presents the estimation result using this alternative lineage measure . The effect of lineage connection remains highly significant, although our alternative measure is constructed based on the surnames of entrepreneurs and managers in domestic firms, which may not have the exact distribution as the total population.

Alternative Standard Errors. In the baseline regressions, standard errors are clustered at the surname-prefecture level to account for serial heteroskedasticity within a surname-prefecture pair. However, the error terms could still be correlated at a more aggregate level. Column (2) of Table A.5 reports the regression results with standard errors clustered at the prefecture level. The significance levels of key variables change little.

Including Round-trip Diaspora Entrants. In the main analyses, we have excluded round-trip diaspora firms. As a robustness check, we include them back in the sample and repeat the baseline regression in column (3) of Table A.5. The estimates for the interaction term between opening dummy and lineage connection becomes larger. The results are not surprising given

that including round-trip firms would inflate the number of diaspora firms, thus overstating the effect of lineage connections.

Permutation Tests. To check whether our results are influenced by other measurement errors in our lineage connection variable, we perform two permutation tests by randomly shuffling our lineage connection measures within and across prefectures. The left panel of Figure 4 plots the distribution of the estimates for our two main outcome variables from 100 simulations using reshuffled lineage connections across surnames within a prefecture. The right panel plots the distribution of estimated coefficients from 100 simulations with lineage connections being reshuffled across prefectures within a surname. Clearly, both distributions are centered around zero. If there are systematic measurement errors, the simulated distributions would overlap with the actual distributions to some degree and be centered around a positive value. The results suggest that the identified effects of lineage connection on diaspora firm entry are not driven by unobserved measurement noises in our lineage connection metric.

Subsample Regressions. There is a possibility that our findings are driven by a few dominant regions or surnames. For example, emigration-intensive provinces like Guangdong, Fujian, and Zhejiang disproportionately sent abroad more emigrants and attracted more diaspora firms at the same time than other regions. Furthermore, a few notable surnames are common in both FDI-intensive regions and among diaspora entrepreneurs.

To check how sensitive our results are to the dominant regions and surnames, we run two regressions in columns (1) and (2) of Table A.6. The sample in the regression of column (1) excludes emigration-intensive provinces, while column (2) includes only the emigration-intensive provinces. The coefficients for the key variable remain positive and statistically significant in both sub-sample regressions no matter whether the number of diaspora entrants or the survival-adjusted number of diaspora entrants is used as an outcome variable. The lineage effect exists not only in emigration-intensive province, but also outside these provinces. Not surprisingly, the magnitude is larger in column (2) than in column (1), suggesting that the effect is more pronounced in emigration-intensive provinces. Column (3) presents the estimation results excluding those emigration-intensive surnames. The coefficients for the key variable remain significantly positive. Finally, column (4) excludes FDI-intensive prefectures. The results barely

change.

4.4 WTO Placebo Test

We have demonstrated the importance of international lineage networks in facilitating the entry of diaspora firms in the early reform period when market environment was immature. A question arises: how would the effects of lineage connection change as market institutions improve? China's accession to the WTO in 2002 provides a natural setting to answer the question, since China had spent decades to improve its market institutions in order to qualify as a member of the WTO.

In Table 5, we run our baseline triple-difference regressions using the WTO accession of China in 2002 as a shock. As shown in column (1) and column (2) of Table 5, the WTO accession has no significant effect on the number of diaspora entrants and survival-adjusted number of diaspora entrants by lineage connection. As domestic market institutions became more aligned with international standards, the effect of lineage connection diminished.

5 Seeding Effects in the Long Run

Having shown that lineage connection is conducive to the entry of diaspora foreign firms in the early opening-up period when market institutions were yet to be fully developed, in this section we further investigate if the industrial "seeds" planted by the early DDI have lasting spillover effects on the entry of non-diaspora foreign firms and domestic private firms. We proceed in two steps: (1) showing that diaspora firms are more likely to be pioneers in 4-digit industries of prefectures than non-diaspora firms; (2) estimating the seeding multipliers of diaspora firms on the entry of non-diaspora foreign firms and domestic private firms separately using an IV strategy.

The literature on the spillover effect of FDIs in China generally use the above-scale and state-owned industrial firms from the late 1990s to the mid-2000s.[16] We complement the literature in several ways. First, we extend the sample period to the early 1980s when the seeds of FDIs

[16] See Appendix A.2 for detailed discussions on the advantages of our data.

were first planted. Second, our sample includes all the FDIs covering all sectors and firms of all sizes. Third, we pay particular attention to the lasting effect of early DDIs on subsequent entry of non-diaspora FDIs and domestic private firms, which has been largely ignored in the previous literature.

5.1 Diaspora Firms as Pioneers

Here we define a pioneering firm as the first entrant into any 4-digit industry in a prefecture. Figure 5 plots the share of diaspora pioneering firms among all diaspora entrants and the share of non-diaspora pioneering firms among all non-diaspora entrants from 1980 to 2014. Two patterns are apparent from the figure. First, the share of pioneering firms declines over time. This is not surprising given the total number of prefecture-industry cells is fixed. With the entry of pioneering firms, the available number of cells naturally becomes smaller over time. Second, in the early opening-up period from 1980 to the early 1990s, the share of diaspora pioneering firms in total diaspora entrants exceeds that of non-diaspora, suggesting that diaspora firms are more likely to be pioneers than the non-diaspora firms. After 1995, the gap in the likelihood of being pioneers between diaspora firms and non-diaspora firms has closed. These patterns indicate that diasporas played an important role in setting up pioneering firms in new industries and new places in the early reform period despite imperfect market environment.

5.2 Estimate Long Run Seeding Multipliers

Hausmann and Rodrik (2003) reckons that pioneering firms generate large positive externality for subsequent entrants in developing countries. Given the wide presence of diaspora pioneering firms in the early opening-up period, we would expect to observe a large spillover effect for later entrants if their hypothesis holds true. In this section we estimate the multipliers of earlier DDI on subsequent entrants.

Following Sequeira, Nunn and Qian (2020), we first isolate lineage-driven DDIs predicted by the interaction term between surname distributions and the staggered opening of Chinese prefectures in the early opening-up period based on the estimate of equation 2 (the "zero stage" regression). The lineage-driven DDI predicted by the "zero stage," $\hat{\beta}$, is then used to con-

struct an instrument for the observed diaspora firm stock in 1996. With the estimated coefficient, we obtain the predicted value of $\hat{\beta} \times Opening_{pc} \times m_{sp}$—the lineage-driven DDI at the surname-prefecture-year level. We then aggregate them over all surnames and entry years to get a predicted value of the cumulative lineage-driven DDIs for each prefecture in year 1996 as following:

$$\hat{DDI}_p = \sum_s \sum_{c=1981}^{1996} \hat{\beta} \times Opening_{pc} \times m_{sp} \tag{4}$$

Next, we use the predicted lineage-driven DDI as an instrument variable for the observed DDI in 1996 in the following two stage least square (2SLS) regression:

$$\pi_p = \alpha + \gamma \times DDI_p + \lambda \times X_p + \epsilon_p, \tag{5}$$

where π_p is our long-term outcome of interest; DDI_p is the observed number of diaspora firm stock in prefecture p in 1996. X_p denotes a set of controls at the prefecture level.

To test the validity of the instrument variable, we perform a balanced test as shown in Table A.8. We separately regress the instrument and the observed number of diaspora firms in 1996 on a set of prefecture-level characteristics correlated with local economic development, including the distance to sea, the slope and elevation of the prefecture's land, cultivated land area per capita in 1996, and the average wage in 1996. Although the observed diaspora firms in 1996 is correlated with these variables, our instrument is not, confirming that our instrument passes the balanced test.

Table 6 presents the estimation results of OLS, reduced form, 2SLS, and first stage regressions in four panels. The key long-run measures of non-diaspora investment include non-diaspora foreign firm stocks, domestic private firm stocks, and registered capital of these survived firms. Besides the controls variables in the balanced test, we also include province fixed effects. As shown in Table A.9, both the OLS and IV estimates without these controls are quite similar to those in Table 6. The stock of diaspora firms in 1996 is positively associated with the number of non-diaspora foreign and domestic private enterprises as of 2014. All the IV estimates are larger than the OLS estimates. According to the IV estimates shown in columns

(1) and (2) in Panel C, one additional diaspora firm in 1996 brings about 0.418 subsequent non-diaspora foreign and 137 domestic private firm entrants.

The positive effect is not only observed for the number of firm entry, but also for firm size, measured in average registered capital in 2014. One more diaspora firm in 1996 leads to 0.2-0.3 percent increase in registered capital for newly registered foreign and domestic enterprises as of 2014, as indicated in column (4). The emerging new and larger enterprises mean more local employment. Column (5) of Table 6 quantifies that one more diaspora firm in 1996 generate 4,099 more employment by 2015 for an average prefecture.

6 Conclusions

The inflow of massive diaspora investments in the early opening-up period is a salient feature of Chinese growth story. However, the destinations of DDIs are not random. Our paper shows that areas with stronger lineage connections have attracted more DDIs during the early opening-up period when market institutions were far from perfect. Moreover, we find that early DDIs, many of which are pioneering firms, facilitate the subsequent entry of non-diaspora foreign and domestic private enterprises.

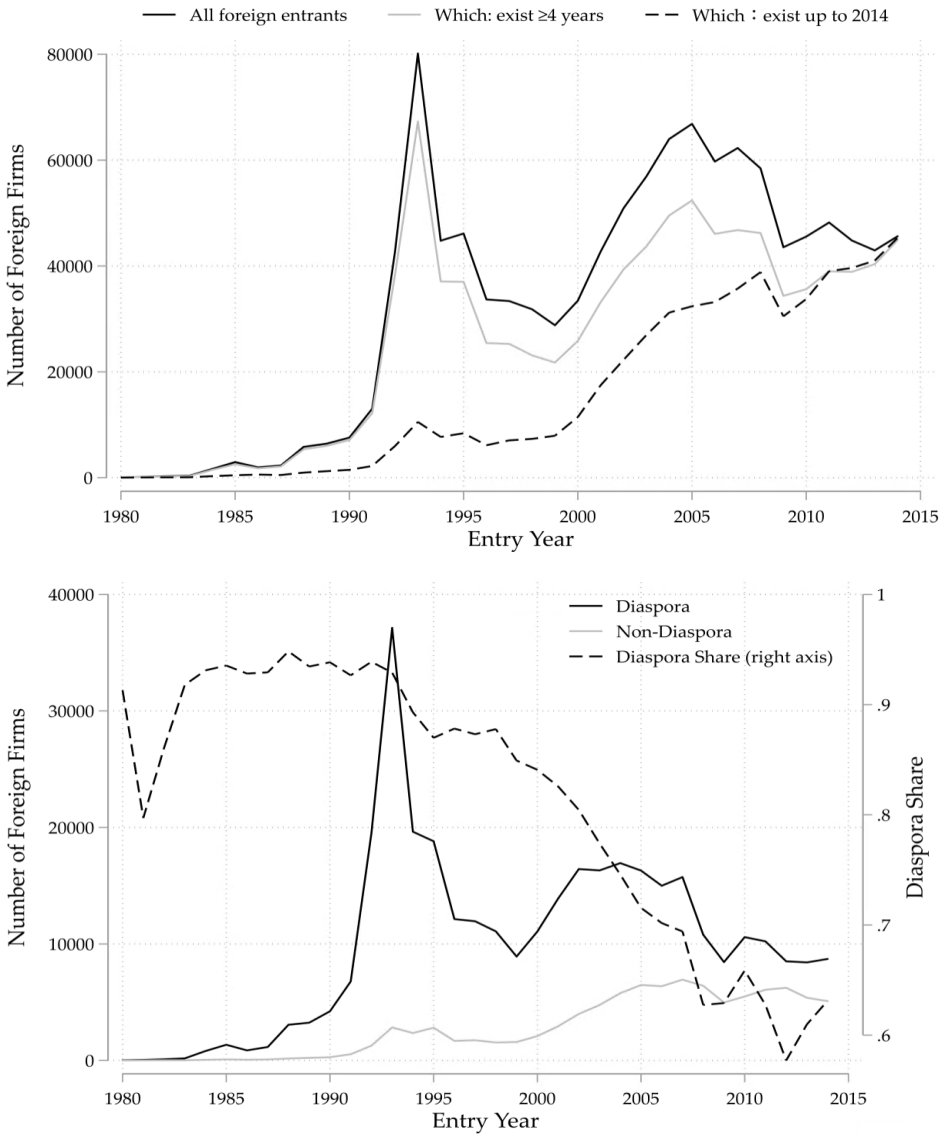
China's experience in tapping diaspora investments in the early stage of development may shed light on other developing countries. Plagued with informational problems, policy uncertainty, and poor contracting environment, developing countries often struggle with attracting FDIs, in particular pioneering foreign firms. As shown in China, a more practical solution is to tap the diaspora networks for their direct investment in the early stage of development.

In fact, the International Organization of Migration, in a 2005 survey, found that more than 90% of the countries had policies or programs targeting diaspora.[17] DDIs account for a significant proportion of the FDIs in other developing countries than China, for example 25% in Armenia during 1998-2014 (Riddle, Hrivnak and Nielsen, 2010) and 26% in India during 1991-2001 (Wei and Balasubramanyam, 2006). Collier, Gregory and Ragoussis (2019) calls for more active policies to attract pioneering firms in fragile and conflict-affected states. However, it would be extremely challenging for a multinational firm without local connections to

[17]The report is retrieved from https://publications.iom.int/system/files/pdf/wmr_2005_3.pdf

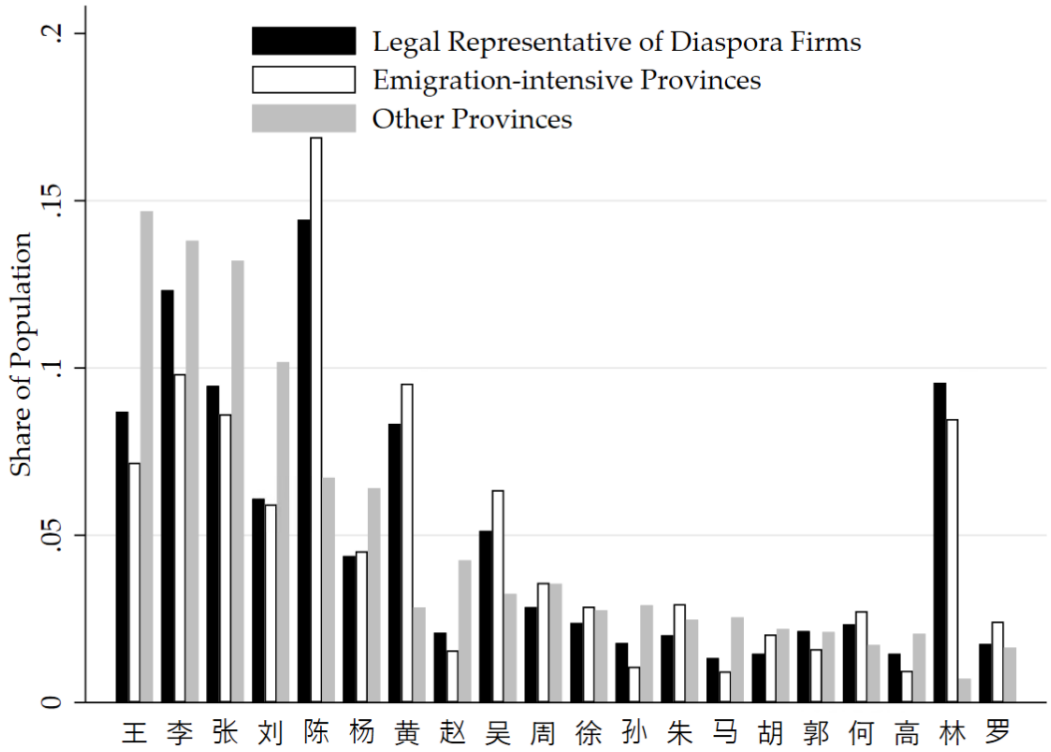
navigate the uncertain and even hostile environment in those states. A future research topic is to explore whether DDIs can serve as pioneering firms even in these countries.

Figure 1: Entry and Survival of Foreign Firms



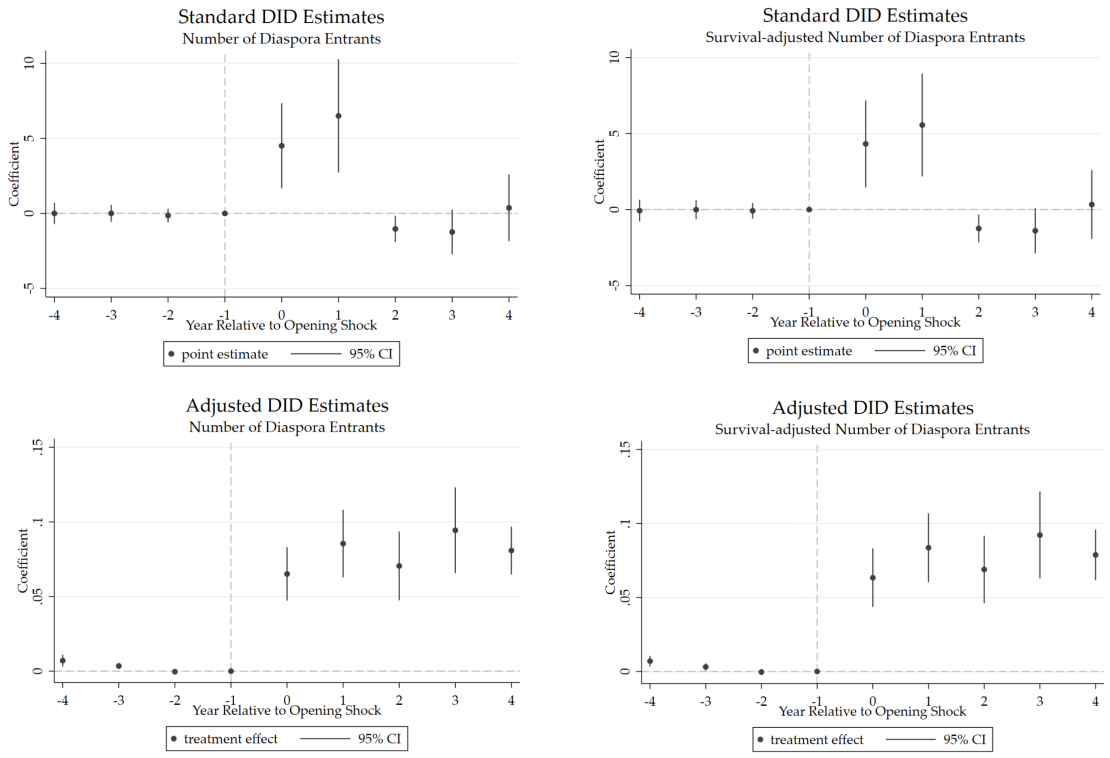
Note: This figure uses firm registration data from the SAIC database. The top panel plots the number (and survived number) of foreign firms by the year of entry during 1980-2014. The bottom panel further displays the number of diaspora and non-diaspora entrants on the left axis as well as the share of diaspora firms on the right axis.

Figure 2: Surname Distributions in Different Populations



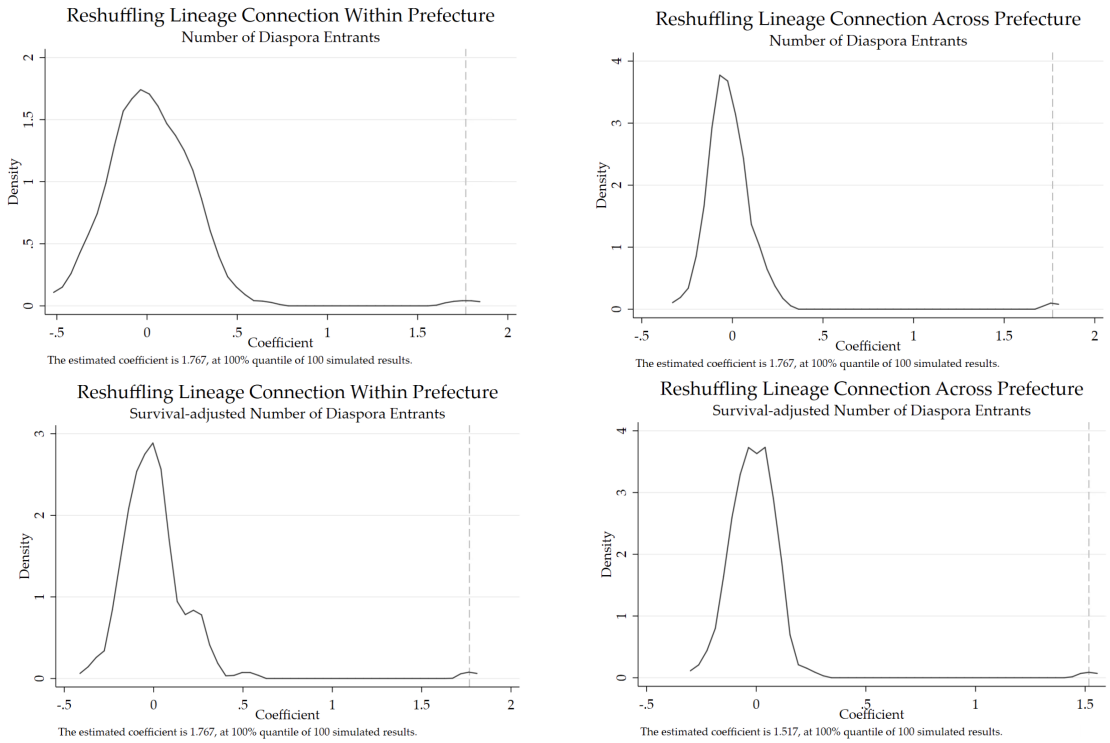
Note: This figure plots the shares of population with the top 20 surnames based on the China Population Survey in 2005 in three samples: the legal representatives of diaspora firms in the SAIC database, the emigration-intensive provinces, and other provinces. Emigration-intensive provinces include Guangdong, Fujian, and Zhejiang. The distribution of surnames among legal representatives of diaspora firms in the SAIC database closely mirrors that among the population of emigration-intensive provinces inferred from the China Population Survey in 2005, but sharply differs from other provinces.

Figure 3: Event Study



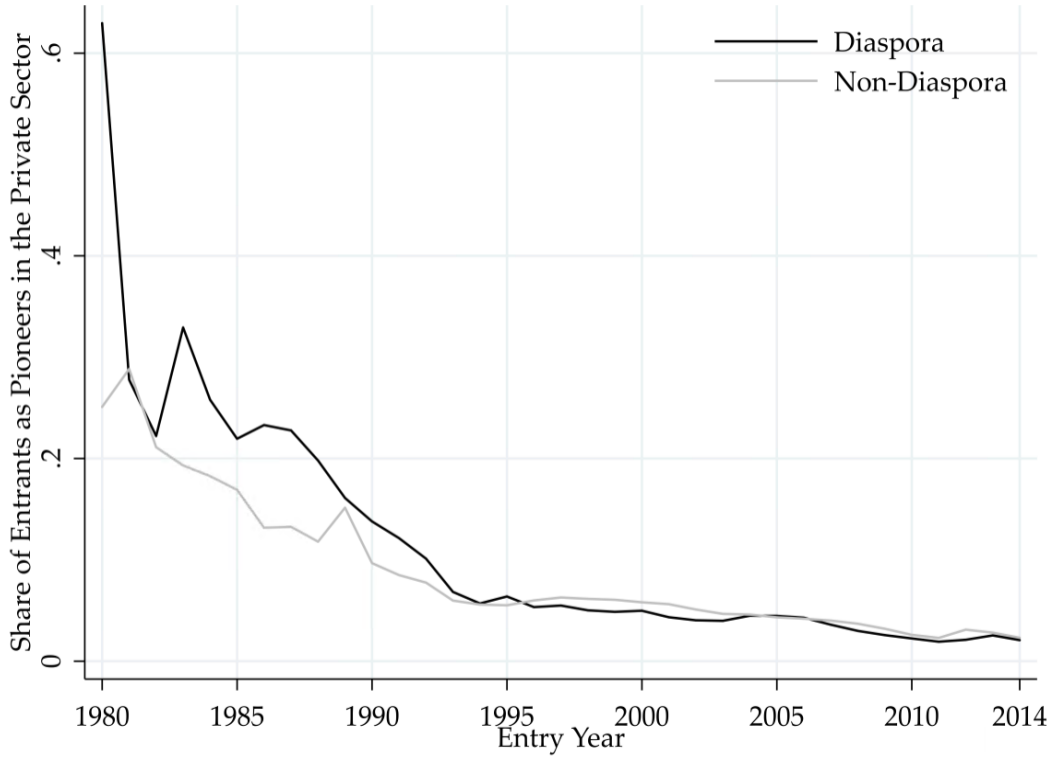
Note: This figure plots the coefficients β^τ obtained from standard event-study specification: $Y_{spt} = \eta_{sp} + \theta_{st} + \delta_{pt} + \sum_{\tau=-4}^{4} \beta^\tau \times Open_{pt}^\tau \times m_{sp} + \lambda \times S_{spt} + \epsilon_{spt}$ where $Open_{pt}^\tau$ indicates opening status in period τ years later than the observed shock and m_{sp} is our measure of lineage connection between surname s and prefecture p. We control for fixed effects η_{sp}, θ_{st}, δ_{pt} and stock of diaspora firms S_{spt}. The top two panels plot the coefficients β^τ obtained from event-study for standard DID estimation. The bottom panels display the coefficients for estimations that have corrected potential biases from the staggered DID using the Stata command "did_multiplegt.'(De Chaisemartin and D'Haultfoeuille, 2020; De Chaisemartin, D'Haultfoeuille and Guyonvarch, 2021). For the bottom panels, since it is computationally infeasible to apply the estimation methods to our large dataset, we replace the continuous measure with a dummy variable indicating whether the measure is above the median value or not, as a second best choice. The outcomes of interest on the left and right panels are the number of diaspora entrants and the survival-adjusted number of diaspora entrants, respectively.

Figure 4: Permutation Tests



Note: In this figure, we show estimates under permutations which randomly reshuffle our lineage connection measures within or across prefectures. The left panel plots the kernel density distributions of the estimates for our two main outcome variables from 100 simulations using the reshuffled lineage connections across surnames within the same prefecture. The right panel presents the kernel density distributions of estimated coefficients from 100 simulations based on lineage connections that are reshuffled across prefectures with the same surname.

Figure 5: Shares of Diaspora and Non-Diaspora Pioneering Firms



Note: This figure plots the share of diaspora pioneering firms among all diaspora entrants and the share of non-diaspora pioneering firms among all non-diaspora entrants by year of entry. A pioneering firms is defined as the first entrant in a prefecture for a 4-digit industry.

Table 1: China's Opening Process

Year	Opening Policy	Open Regions
1980	Special Economic Zone	Shenzhen, Zhuhai, Shantou, Xiamen
1984	Open Coastal City	Dalian, Qinhuangdao, Tianjin, Yantai, Qingdao, Lianyungang, Nantong, Shanghai, Ningbo, Wenzhou, Fuzhou, Guangzhou, Zhanjiang, Beihai
1985	Open Coastal City	Yingkou
1987	Open Coastal City	Weihai
1988	Special Economic Zone	Hainan Province
1990	Special Economic Zone	Shanghai Pudong District
1992	Deng Xiaoping's South Tour	All other regions

Source: https://en.wikipedia.org/wiki/Chinese_economic_reform.

Table 2: Summary Statistics

	N	Mean	Std
	(1)	(2)	(3)
Panel A: Surname-Prefecture-Year			
At Least One Entrant	1,345,024	0.021	0.142
Number of Entrants	1,345,024	0.060	1.437
Survival-Adjusted Number of Entrants	1,345,024	0.052	1.234
Survived Registered Capital in 2014 ($10^4 CNY$)	1,345,024	27.017	940.778
Survival Ratio of Entrants in 2014	27,846	0.168	0.338
Average Registered Capital per Survived Entrant	27,846	806.588	3,967.006
Panel B: Surname-Prefecture			
Lineage Connection (2005 Census)	48,179	0.005	0.011
Lineage Connection (SAIC)	57,802	0.005	0.012
Panel C: Prefecture-Year			
Number of Non-Diaspora Foreign Entrants	5,312	2.610	19.190
Number of Domestic Private Entrants	5,312	720.032	2,208.610
Panel D: Prefecture			
Diaspora Firm Stocks in 1996	287	372.111	1,216.842
Non-Diaspora Foreign Firm Stocks in 2014	296	114.693	408.289
Domestic Private Firm Stocks in 2014	332	56,581.290	116,000
Non-Diaspora Foreign Registered Capital in 2014 ($10^4 CNY$)	296	23.800	80.400
Domestic Private Registered Capital ($10^4 CNY$)	332	824.300	101,000
Employment Size in 2015 (10^3)	278	2,640.837	2,131.283
Distance to Sea	301	531.742	445.541
Wage in 1996 (CNY)	248	5,585.191	1,667.894
Slope	312	11.205	5.923
Elevation	312	657.812	813.129
Cultivated Land Area in 1996 (10^3 hectare)	256	0.750	0.674

Note: Panel A reports the summary statistics of diaspora firms at surname-prefecture-year level. The survival-adjusted entrant number is defined as the number of diaspora entrants which survive for no fewer than four years. In panel B, the measure of lineage connection is calculated based on the surnames of registered personnel working in domestic firms established prior to 1992. Panel C is at the prefecture-year level. Panel D reports summary statistics at the prefecture level. The first five variables in panel D are obtained from the SAIC database. The remaining six variables in panel D are retrieved from official documents or CSMAR database.

Table 3: Lineage Connection and Entry of Diaspora Firms

	Standard DID		Adjusted DID	
	Number of Diaspora Entrants	Survival-adjusted Number of Diaspora Entrants	Number of Diaspora Entrants	Survival-adjusted Number of Diaspora Entrants
	(1)	(2)	(3)	(4)
Mean of Dep. Var.	0.060	0.052	0.060	0.052
Open × Lineage Connection	1.767***	1.517***		
	(0.574)	(0.527)		
Open × High Dummy			0.079***	0.077***
			(0.007)	(0.010)
Adj.R^2	0.570	0.555		
N	1,344,421	1,344,421	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at surname-prefecture level. Survival-adjusted Number of Diaspora Entrants is the number of entrants that survive for more than 4 years (included). "Open" indicates whether the prefecture has been opened to foreign investment. "Lineage connection" is measured between surname and prefecture. Columns (3) and (4) follow De Chaisemartin and D'Haultfoeuille (2020) and use Stata command "did_multiplegt" with default settings to perform adjusted DID estimations (De Chaisemartin, D'Haultfoeuille and Guyonvarch, 2021). The variable "High Dummy" equals one if the lineage connection for a surname-prefecture pair is greater than the median across all pairs, and 0 otherwise. We use this dummy variable instead of continuous connection measure mainly because the estimation method used here is computationally infeasible on our dataset for the continuous measure.

Table 4: No Sorting on Entry Quality

	Entry Quality	
	Survival Ratio in 2014 Conditional on Entry	Log Average Registered Capital in 2014 Conditional on Entry
	(1)	(2)
Mean of Dep. Var.	0.168	6.693
Open × Lineage Connection	-0.333	9.291
	(0.825)	(6.529)
Adj.R^2	0.364	0.323
N	20633	20633
Number of Incumbent Firms	Y	Y
Surname-Prefecture FE	Y	Y
Surname-Year FE	Y	Y
Prefecture-Year FE	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at the surname-prefecture level. ”Open” indicates whether a prefecture opened to foreign investment. See equation 1 for the definition of “Lineage connection”. “Survival ratio in 2014 conditional on entry” is defined as the share of entrants that survives as of 2014 conditional on there is entry of diaspora firms. “Average registered capital in 2014 conditional on entry” refers to the average registered capital for those diaspora firms which were still active as of 2014.

Table 5: Placebo Test: the WTO Accession

	Number of Diaspora Entrants	Survival-Adjusted Number of Diaspora Entrants
	(1)	(2)
WTO Accession × Lineage Connection	-0.259	-0.051
	(0.197)	(0.162)
Adj.R^2	0.622	0.612
N	1,507,197	1,507,197
Number of Incumbent Firms	Y	Y
Surname-Prefecture FE	Y	Y
Surname-Year FE	Y	Y
Prefecture-Year FE	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at surname-prefecture level. Survival-adjusted Number of Entrants is the number of entrants that survive for more than 4 years (included). The sample period is from 1997 to 2014. Variable *WTO accession* takes value 1 for the period from 2002 onward, and 0 otherwise. Lineage connection is measured between surname and prefecture.

Table 6: Seeding Effects of DDIs

	Non-Diaspora Foreign Firm Stocks in 2014	Domestic Private Firm Stocks in 2014	Log Registered Capital of Non-Diaspora Foreign Firms in 2014	Log Registered Capital of Domestic Private Firms in 2014	Employment Size in 2015 (thousand)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
1996 Diaspora Firms	0.180**	57.987***	0.001***	0.000***	0.823***
	(0.082)	(6.309)	(0.000)	(0.000)	(0.183)
Panel B: Reduced Form					
Predicted Diaspora Firms	15.000*	5143.026***	0.129**	0.070**	154.461***
	(8.308)	(898.888)	(0.052)	(0.018)	(35.584)
Panel C: 2SLS					
1996 Diaspora Firms	0.418***	137.241***	0.003***	0.002***	4.099***
	(0.118)	(46.215)	(0.001)	(0.000)	(1.300)
Panel D: First Stage					
	Dependent Variable: 1996 Diaspora Firms				
Predicted Diaspora Firms			36.262**		
			(16.178)		
N	219	219	219	219	219
F statistics	31.088	31.088	31.088	31.088	31.088
Controls	Y	Y	Y	Y	Y
Province Fixed Effects	Y	Y	Y	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10%, respectively. Standard errors are clustered at province level and shown in parenthesis. The data on non-diaspora foreign firms and domestic private firms in 2014 are from the SAIC database. The data on employment size in 2015 is from the China Population Survey in 2015. Panel A presents the OLS estimates, while Panel B reports the reduced-form estimates from regressing outcomes of interest on our IV: the predicted diaspora firms in 1996 driven by the interaction of opening shocks and lineage connection in zero stage regression. Panel C shows the 2SLS estimates using the predicted diaspora firms in 1996 as IV for the observed diaspora firms in 1996. Panel D reports the first stage of the 2SLS estimation. Controls include province fixed effects, distance to the sea, log slope of the land, log elevation of the land, log cultivated land per capita in 1996, and log average wage in 1996 in each regression. The Cragg-Donald Wald F statistic is reported for IV regressions.

References

Alfaro, Laura, Sebnem Kalemli-Ozcan, and Vadym Volosovych. 2008. "Why doesn't Capital Flow from Rich to Poor Countries? An Empirical Investigation." *Review of Economics and Statistics*, 90(2): 347–368.

Bai, Chong-En, Chang-Tai Hsieh, Zheng Michael Song, and Xin Wang. 2020. "Special Deals from Special Investors: The Rise of State-Connected Private Owners in China." NBER Working Paper.

Bai, Ying. 2020. "Surname Distance and Technology Diffusion: The Case of the Adoption of Maize in Late Imperial China." Working Paper.

Branstetter, Lee, and Nicholas Lardy. 2006. "China's Embrace of Globalization." NBER Working Paper.

Burchardi, Konrad B, Thomas Chaney, and Tarek A Hassan. 2019. "Migrants, Ancestors, and Foreign Investments." *Review of Economic Studies*, 86(4): 1448–1486.

Chen, Fang. 2022. "What is "Round-Trip Investment"?" In *Essential Knowledge and Legal Practices for Establishing and Operating Companies in China*. 393–398. Springer.

Chen, Yuyu, Mitsuru Igami, Masayuki Sawada, and Mo Xiao. 2020. "Privatization and Productivity in China." Working Paper.

Clark, Gregory. 2015. *The Son Also Rises: Surnames and the History of Social Mobility.* Princeton University Press.

Collier, Paul, Neil Gregory, and Alexandros Ragoussis. 2019. "Pioneering Firms in Fragile and Conflict-Affected States." World Bank, Washington, DC.

Dai, Ruochen, Dilip Mookherjee, Kaivan Munshi, and Xiaobo Zhang. 2019. "The Community Origins of Private Enterprise in China." Working Paper.

De Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–96.

De Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." NBER Working Paper.

De Chaisemartin, Clément, Xavier D'Haultfoeuille, and Yannick Guyonvarch. 2021. "DID_MULTIPLEGT: Stata Module to Estimate Sharp Difference-in-Difference Designs with Multiple Groups and Periods." Stata Package.

Du, Julan, Yi Lu, and Zhigang Tao. 2008. "Economic Institutions and FDI Location Choice: Evidence from US Multinationals in China." *Journal of Comparative Economics*, 36(3): 412–429.

Geng, Xiao. 2019. "Round-Tripping Foreign Direct Investment in the People's Republic of China: Scale, Causes and Implications." Asian Development Bank.

Greif, Avner, and Guido Tabellini. 2017. "The Clan and the Corporation: Sustaining Cooperation in China and Europe." *Journal of Comparative Economics*, 45: 1–35.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2009. "Cultural Biases in Economic Exchange?" *Quarterly Journal of Economics*, 124(3): 1095–1131.

Hausmann, Ricardo, and Dani Rodrik. 2003. "Economic Development as Self-Discovery." *Journal of Development Economics*, 72(2): 603–633.

Huang, Yasheng. 2003. *Selling China: Foreign Direct Investment during the Reform Era.* Cambridge University Press.

Huang, Yasheng, Li Jin, and Yi Qian. 2013. "Does Ethnicity Pay? Evidence from Overseas Chinese FDI in China." *Review of Economics and Statistics*, 95(3): 868–883.

Javorcik, Beata S, Çağlar Özden, Mariana Spatareanu, and Cristina Neagu. 2011. "Migrant Networks and Foreign Direct Investment." *Journal of Development Economics*, 94(2): 231–241.

Julio, Brandon, and Youngsuk Yook. 2012. "Political Uncertainty and Corporate Investment Cycles." *Journal of Finance*, 67(1): 45–83.

Kerr, William, and Ramana Nanda. 2009. "Democratizing Entry: Banking Deregulation, Financing Constraints, and Entrepreneurship." *Journal of Financial Economics*, 94: 124–149.

Kuchler, Theresa, Yan Li, Lin Peng, Johannes Stroebel, and Dexin Zhou. 2020. "Social Proximity to Capital: Implications for Investors and Firms." NBER Working Paper.

Kuhn, Philip A. 2008. *Chinese among Others: Emigration in Modern Times.* Rowman & Littlefield.

Lever-tracy, Constance, David Ip, and Noel Tracy. 1996. *The Chinese Diaspora and Mainland China: An Emerging Economic Synergy.* Macmillan Press LTD.

Lin, Ping, Zhuomin Liu, and Yifan Zhang. 2009. "Do Chinese Domestic Firms Benefit from FDI Inflow?: Evidence of Horizontal and Vertical Spillovers." *China Economic Review*, 20(4): 677–691.

Lucas, Robert E. 1990. "Why doesn't Capital Flow from Rich to Poor Countries?" *American Economic Review*, 80(2): 92–96.

Lu, Yi, Zhigang Tao, and Lianming Zhu. 2017. "Identifying FDI Spillovers." *Journal of International Economics*, 107: 75–90.

Ma, Sen. 2018. "The Effects of Cultural Similarity on Foreign Direct Investment and Productivity of Domestic Firms: Identification from Borders of Chinese Dialect Zones." Working Paper.

Melitz, Jacques, and Farid Toubal. 2014. "Native Language, Spoken Language, Translation, and Trade." *Journal of International Economics*, 93(2): 351–363.

Nanda, Ramana, and Tarun Khanna. 2010. "Diasporas and Domestic Entrepreneurs: Evidence from the Indian Software Industry." *Journal of Economics & Management Strategy*, 19(4): 991–1012.

Noorbakhsh, Farhad, Alberto Paloni, and Ali Youssef. 2001. "Human Capital and FDI Inflows to Developing Countries: New Empirical Evidence." *World Development*, 29(9): 1593–1610.

Parsons, Christopher, and Pierre-Louis Vézina. 2018. "Migrant Networks and Trade: The Vietnamese Boat People as a Natural Experiment." *Economic Journal*, 128(612): F210–F234.

Rauch, James E. 2001. "Business and Social Networks in International Trade." *Journal of Economic Literature*, 39(4): 1177–1203.

Rauch, James E, and Vitor Trindade. 2002. "Ethnic Chinese Networks in International Trade." *Review of Economics and Statistics*, 84(1): 116–130.

Riddle, Liesl, George A Hrivnak, and Tjai M Nielsen. 2010. "Transnational Diaspora Entrepreneurship in Emerging Markets: Bridging Institutional Divides." *Journal of International Management*, 16(4): 398–411.

Saxenian, AnnaLee. 2007. *The New Argonauts: Regional Advantage in a Global Economy.* Harvard University Press.

Sequeira, Sandra, Nathan Nunn, and Nancy Qian. 2020. "Immigrants and the Making of America." *Review of Economic Studies*, 87(1): 382–419.

Tan, Chee-Beng. 2006. *Chinese Transnational Networks.* Routledge.

Tombe, Trevor, and Xiaodong Zhu. 2019. "Trade, Migration, and Productivity: A Quantitative Analysis of China." *American Economic Review*, 109(5): 1843–72.

Vogel, Ezra F. 1990. *One Step Ahead in China: Guangdong under Reform.* Harvard University Press.

Wei, Yingqi, and Vudayagiri N Balasubramanyam. 2006. "Diaspora and Development." *World Economy*, 29(11): 1599–1609.

World Bank. 2017. *Global Investment Competitiveness Report 2017/2018: Foreign Investor Perspectives and Policy Implications.* The World Bank.

Zhang, Chuanchuan. 2020. "Clans, Entrepreneurship, and Development of the Private Sector in China." *Journal of Comparative Economics*, 48(1): 100–123.

Online Appendix

A.1 Algorithm to Identify Overseas Chinese and Their Surnames

This section introduces the algorithm we use to identify overseas Chinese and their Chinese surnames among the registered personnel working in foreign companies. We extract from the SAIC database all the registered personnel working in foreign firms according to the firm's ownership code, with following variables available for each person: name, ID type, registry address, executive position, a dummy for the legal representative status.

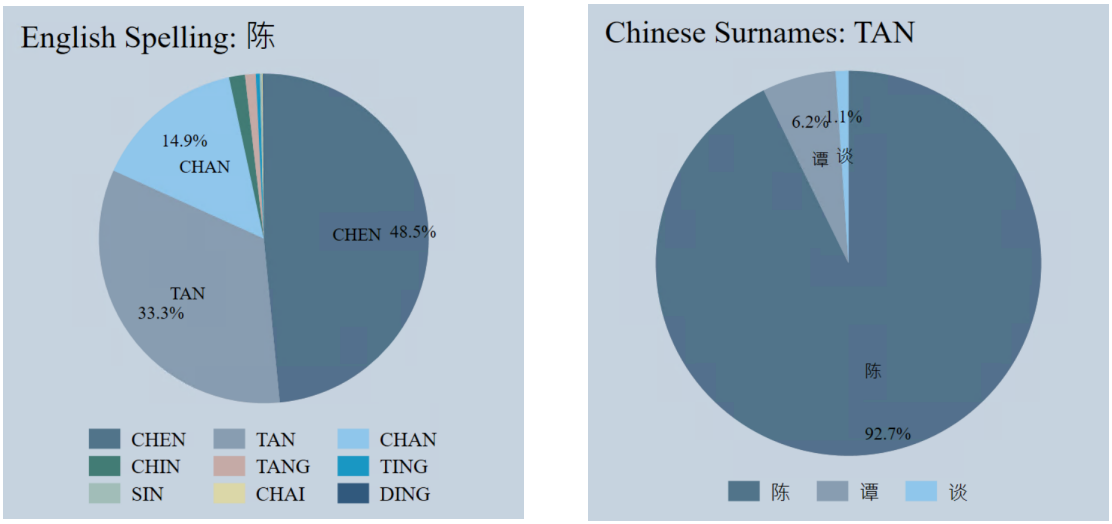
Before formally executing our algorithm, we perform a preparatory cleaning procedure to leave out all symbols (such as comma, periods, semicolon) and word content that is unrelated to names (such as titles,"Mr.", "Dr.", "appointed by the parent company"), since many name entries are unstructured in the raw data we acquire. This step breaks down the raw variable "name" into strings of pure Chinese or English characters. Thus name strings can be further categorized into three kinds: names written in pure Chinese (李小龙), names written in pure English (Bruce Lee), and names written in both Chinese and English (李小龙 Bruce Lee).

It is fairly straightforward to identify the Chinese surnames for name strings written in pure Chinese and in both Chinese and English, because both groups contain names written in Chinese. But it is relatively hard to identify surnames for the names written in English because English spellings of Chinese surnames unnecessarily map one-to-one into Chinese characters. See Figure A.1 for an illustrative example. To lessen this issue, we construct an English-Chinese mapping based on the third group of strings, the names written in both Chinese and English. We then break down each English spelling into Chinese surnames in proportion based on its relative presence in mixed entries. For example, a "TAN" is equivalent to 0.927 "陈", 0.062 "谭", and 0.011 "谈" as in Figure A.1.

Below we provide detailed introductions of the algorithm we execute to identify overseas Chinese and their surnames for each type of strings.

Names Written in Pure Chinese. Foreigners, typically Japanese and Korean, are also possible to register their names using Chinese characters. This further complicates the analysis. In light of these concerns, we apply the following procedures to identify ethnic Chinese and

Figure A.1: An Example of Spelling-Character Mapping between English and Chinese



Note: This figure illustrates the fact that the mapping between English spelling and Chinese character is not necessarily one-to-one for Chinese surnames. On the left panel we show the proportion of each possible English spelling of Chinese surname "陈" among registered personnel who have name entries with both English characters and Chinese characters in the SAIC database. On the right panel, we instead present the proportions of each possible Chinese characters which can be associated with the English spelling TAN. We use these empirical probabilities to map Chinese surname entries with only English strings into Chinese characters for calculating the lineage connections of diaspora firms.

their surnames:

1. For each string of a name entry, if the length of the string is greater than 4 Chinese characters, we tag the string as non-ethnic Chinese, since common ethnic Han Chinese names hardly contain more than 4 Chinese characters. If the length of the string is exactly 4 Chinese characters, we check whether the first two characters of the string match with double-character Chinese surnames (such as 欧阳, 司马). If the match is successful, we tag the name as ethnic Chinese, otherwise ethnic non-Chinese. If the length of the string is less than 4 Chinese characters, we match the first Chinese character with the Chinese surname dictionary. If the match is successful, we tag the name as ethnic Chinese. Otherwise, the name will be tagged as ethnic non-Chinese. This step produces a tag of ethnic Chinese status for each string of any name entry.

2. For each string of a name entry, we match the first one/two/three Chinese characters with the most common 1000 Japanese surnames obtained from Wikipedia. If the following three conditions are satisfied simultaneously, we tag the string as a Japanese surname (otherwise a non-Japanese surname): (1) the match with the most common 1000 Japanese

surnames is successful; (2) the ID type of the personnel is foreign passport; (3) the registry address starts with typical locations in Japan. This step produces a tag of Japanese status for each string of any name entry.

3. For each string of a name entry, we match the first Chinese character with the most common 100 Korean surnames obtained from Wikipedia. If the following three conditions are satisfied simultaneously, we tag the string as a Korean surname (otherwise a non-Korean surname): (1) the match with the most common 100 Korean surnames is successful; (2) the ID type of the personnel is foreign passport (3) the registry address starts with typical locations in South Korea. This step produces a tag of Korean status for each string of any name entry.

4. For each string of a name entry, if it is tagged ethnic Chinese, non-Japanese, and non-Korean, we deem the person an ethnic Chinese.

5. If a person is deemed an ethnic Chinese, we extract the leftest character of the first string of the name entry as the surname for the ethnic Chinese, given that the length of the first string of the name entry is shorter than 4. We extract the leftest two characters of the first sting of the name entry as the surname for the ethnic Chinese , if the length of first string of the name entry is exactly 4.

Names Written in Pure English. In the subgroup of name strings that are written in pure English, what is noticeable is that the surname can be placed in either the leftest string or rightest string. Besides, some irregular name entries that fail to insert blank space between surnames and given names make it infeasible to match name strings directly with the Chinese surname dictionary. In light of these concerns, we apply the following procedures to identify ethnic Chinese and their surnames:

1. We first divide the name entries into two groups: one with multiple strings (with blank space in the name entry), and the other with a single string (without blank space in the name entry).

2. For the group with multiple strings, we match the leftest and rightest character with the Chinese surname dictionary. If the match is successful for either the leftest character or the rightest character, we tag the name as ethnic Chinese. If only one of them is matched, the successfully matched surname is chosen to be the surname of the ethnic Chinese. If both characters are successfully matched, we keep the leftest character as the surname by default (in our database, surnames are more likely to be identified at the leftest position). Otherwise, a name entry is tagged as being ethnic non-Chinese.

3. For the group with a single string, we manually determine name entries' ethnic Chinese status and surnames.

4. We assign each English-spelling surname into Chinese characters in a probabilistic way. We use the observed empirical mapping between Chinese surnames and English spellings in the mixed entries as bootstrapped weights in the probabilistic assignment.

Names Written in Both Chinese and English. The group of names written in both Chinese and English serve as a "bridge" between English spellings and Chinese characters. We construct an English-Chinese surname mapping based on the group of names written in both Chinese and English. The mapping enables us to execute the probabilistic assignment for each English-spelling into Chinese characters.

1. We break down each name entry into two parts: the part of Chinese strings, and the part of English strings.

2. For the part of Chinese strings, we apply the same procedure as for the names written in pure Chinese. This step produces a temporary ethnic Chinese tag. We also extract the surname of ethnic Chinese written in Chinese character.

3. For the part of English strings, we apply the same procedure as for names written in pure English. This step produces another temporary ethnic Chinese tag. We also extract leftest English string and rightest English string.

4. If the temporary ethnic Chinese tags from both parts are negative, we tag the person as non-Chinese. If not, we proceed in following steps. We first match sequentially the leftest

English string and rightest English string with the identified Chinese characters from Chinese strings, based on the Chinese surname dictionary. If either match is successful, we tag this person as ethnic Chinese. We further deem the identified Chinese character as the ethnic Chinese's surname. We also identify the matched English string as the legitimate spelling for the Chinese character. If both matches are successful, we keep the leftest spelling as default.

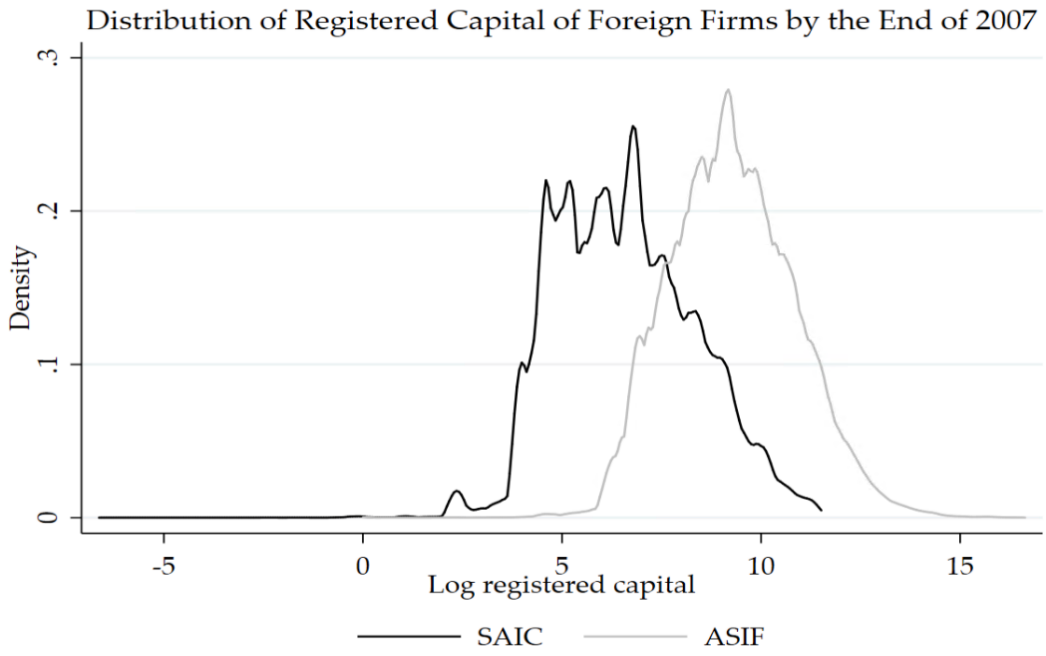
5. This procedures produces not only the surname for each ethnic Chinese, but also an English-Chinese mapping that can be used to randomly assign English-spellings into Chinese characters with empirical weights bootstrapped from the sample of overseas Chinese.

Determining Overseas Chinese. The previous steps identify whether a person is ethnic Chinese or not. We further separate the mainlander and the overseas Chinese according to the ID type associated with each registered personnel. An overseas Chinese is an ethnic Chinese that holds a non-mainland Chinese ID—including passport of a foreign country and travel permit for Hong Kong, Macau, Taiwan (HMT) residents.

A.2 Comparison with Other Data Sources on Chinese FDI

In this section, we illustrate the advantages of our comprehensive foreign firm dataset and cross-validate our data with other data sources.

Figure A.2: SAIC versus ASIF: Firm Size

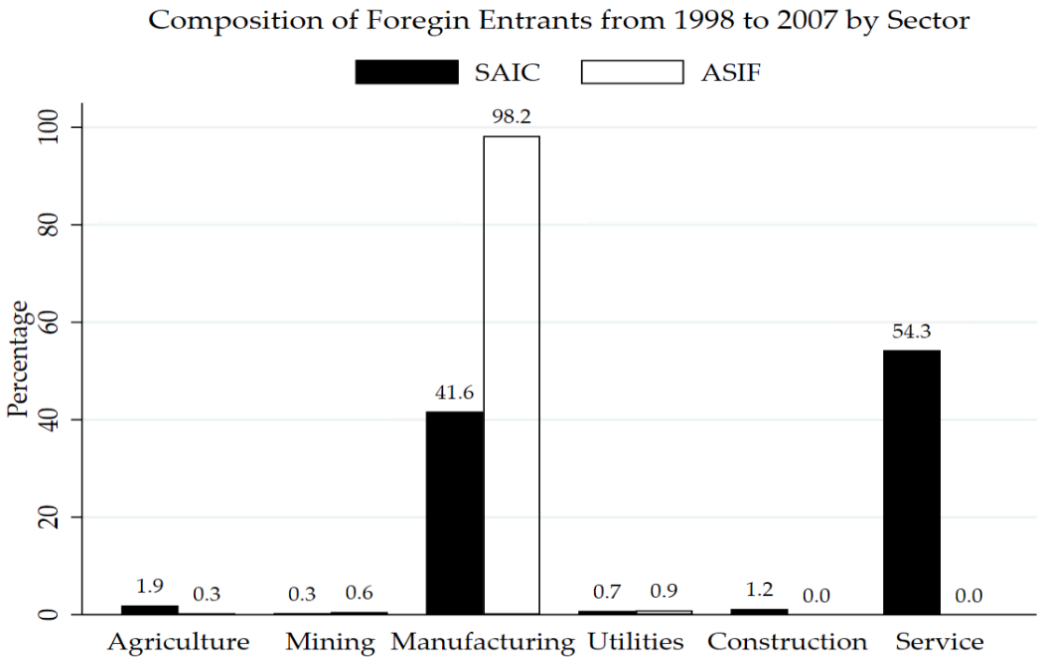


Note: In this figure, we plot the distribution of the registered capital of foreign firms observed in the SAIC database and ASIF respectively. We can immediately see that the ASIF firms are much larger than the firms in the SAIC database in terms of registered capital. Therefore using ASIF firms to study foreign firms in China would miss a tremendous amount of firms of smaller sizes.

Foreign firms covered by the ASIF data versus by the SAIC data. Another frequently used firm-level dataset is the Annual Survey of Industrial Firms (ASIF). We show that our SAIC database provides additional strengths for studying foreign firms in China, compared to the ASIF data. First, the SAIC database is representative for foreign firms of all sizes while the ASIF daa only covers large firms with sales above 5 million CNY. Figure A.2 plots the distribution of the registered capital of foreign firms by the end of 2007, constructed from the SAIC data and the ASIF data respectively. It clearly shows that the SAIC data covers smaller firms while the ASIF data does not. Second, Figure A.3 also depicts that while 98% of firms included in the ASIF data are manufacturing firms, our full-sample SAIC database suggests that manufacturing firms only account for 41.3% of the population of foreign firms. These two biases of the ASIF data in sample coverage could result in the large discrepancy between the

two databases in terms of the number of entrants over time, as shown in Figure A.4.

Figure A.3: SAIC versus ASIF: Sector

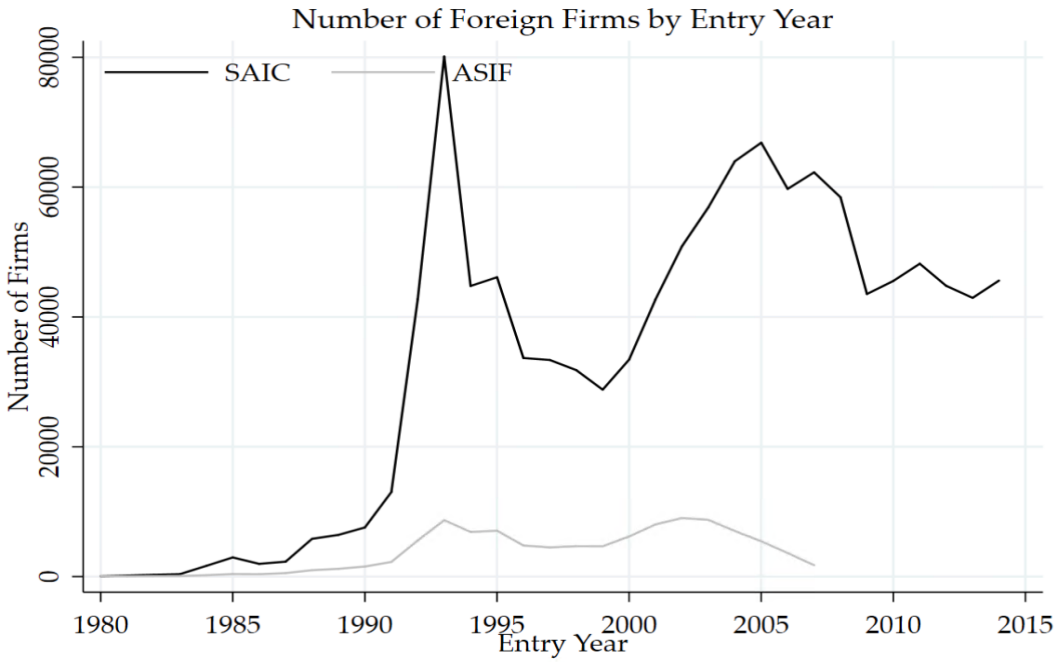


Note: In this figure, we plot the compositions of foreign firms in the SAIC database and the ASIF data. As expected, the ASIF data only covers manufacturing firms while in the SAIC database manufacturing firms can only account for about 40% of all firms in counts. Therefore using the ASIF data to study foreign firms in China would miss a great number of non-manufacturing firms which exerts great influences on the economic development of China.

Diaspora firms versus HMT firms.. Due to data limitation, prior literature often uses HMT firms to proxy diaspora firms (Lin, Liu and Zhang, 2009; Huang, Jin and Qian, 2013). There are two biases embodied in such approach. First, a lot of Europe-based and US-based multinationals invest in mainland China through Hong Kong as a conduit but are actually not diaspora firms; Second, residents in HMT only make up a small fraction of worldwide overseas Chinese. In Figure A.5, the number of diaspora entrants is always greater than the number of HMT entrants over time. Furthermore the gap widens during the process of China's gradual accession to WTO from 1995 to 2007.

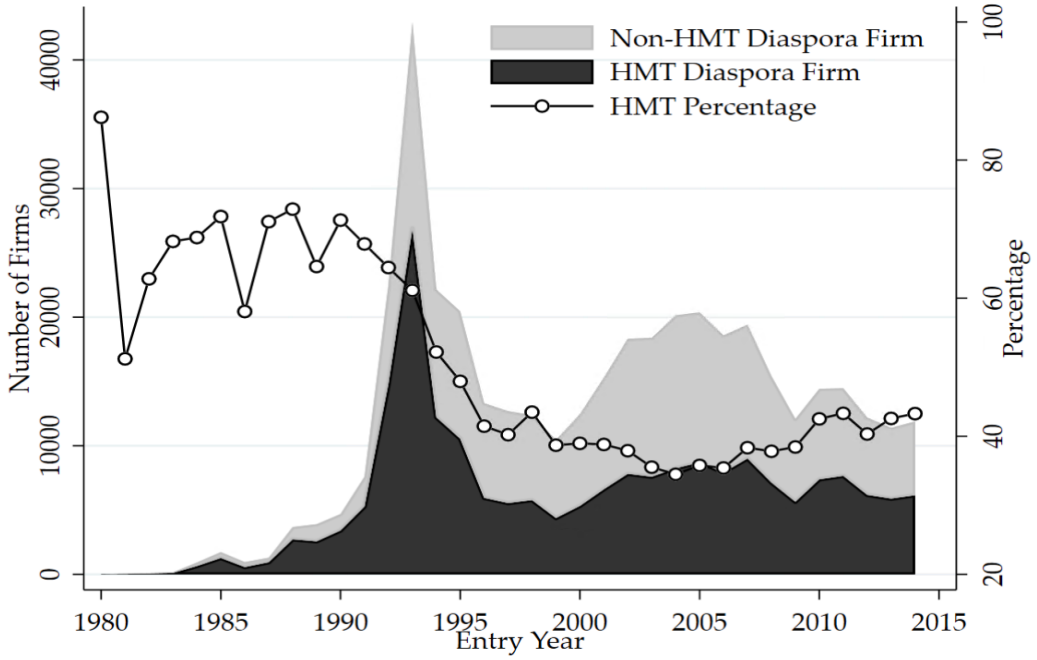
Foreign firms versus FDI. As a monetary concept, FDI includes both the initial and follow-up investment from a foreign-owned entity. We cross-validate our dataset of foreign firms both in counts and volumes with the official FDI data provided by the Ministry of Commerce of China. Figure A.6 suggests that the contracted FDI, measured in number of cases or USD, is highly correlated with the foreign firm entry we observe in our dataset. Besides, the realized FDI measured in USD is also highly correlated with the survived registered capital of foreign

Figure A.4: SAIC versus ASIF: Number of Foreign Entrants



Note: In this figure, we show the discrepancy between the SAIC database and the ASIF data over time by plotting the number of foreign entrants in the two datasets. We can see that using the ASIF data to study foreign firms would only cover a selective sample of all foreign firms and the gap is growing over time.

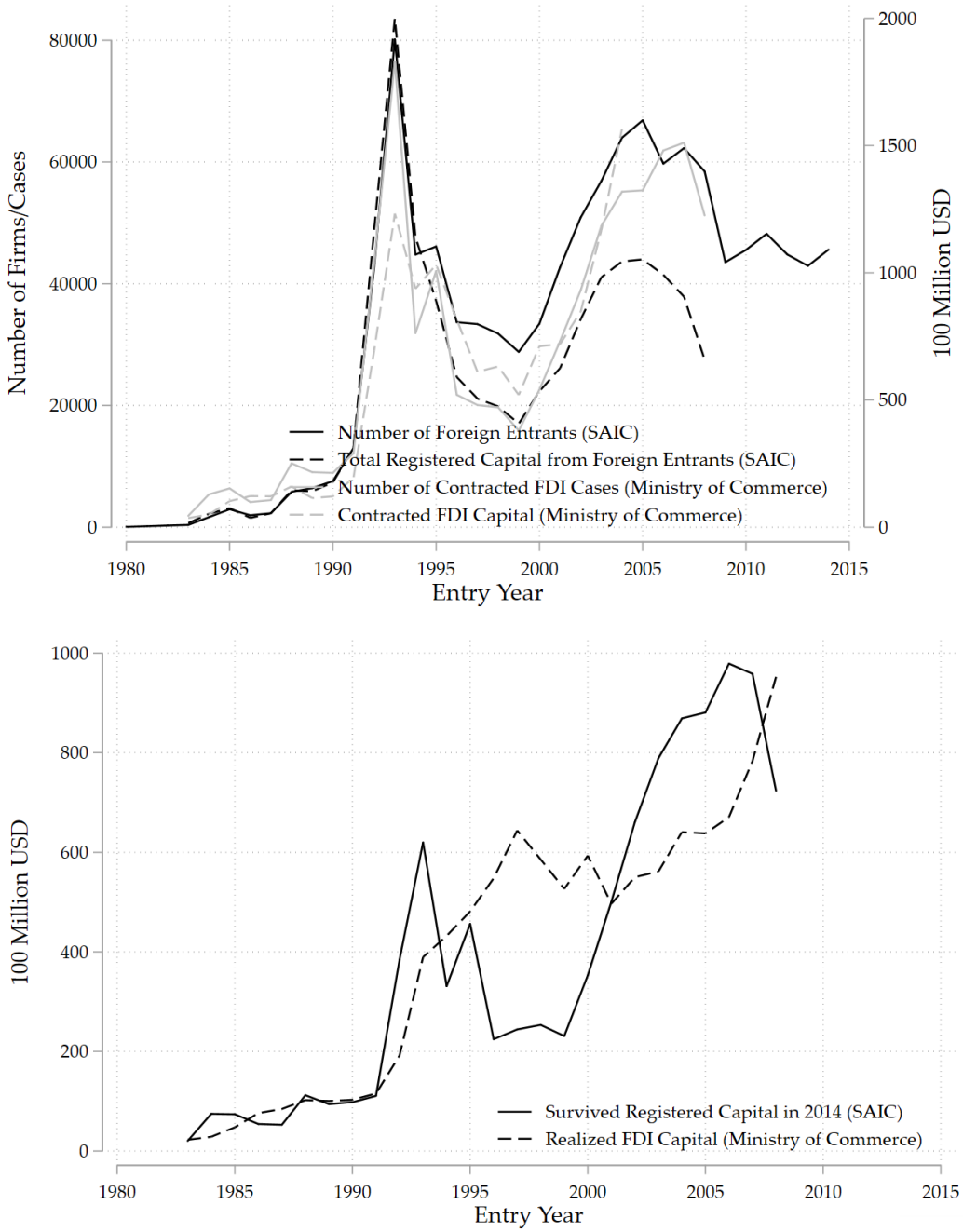
Figure A.5: Diaspora Firms versus HMT Firms



Note: In this figure, we separate HMT diaspora firms from the non-HMT diaspora firms. HMT firms refer to those diaspora firms from Hongkong, Macau, and Taiwan. We can see that the percentage of HMT firms among all diaspora firms has dropped to 40% even before 2000. This pattern indicates that HMT firms can not be used as a good proxy for diaspora firms.

firms by 2014 calculated from our dataset. Our dataset therefore is consistent with official aggregate FDI statistics both in counts and volumes.

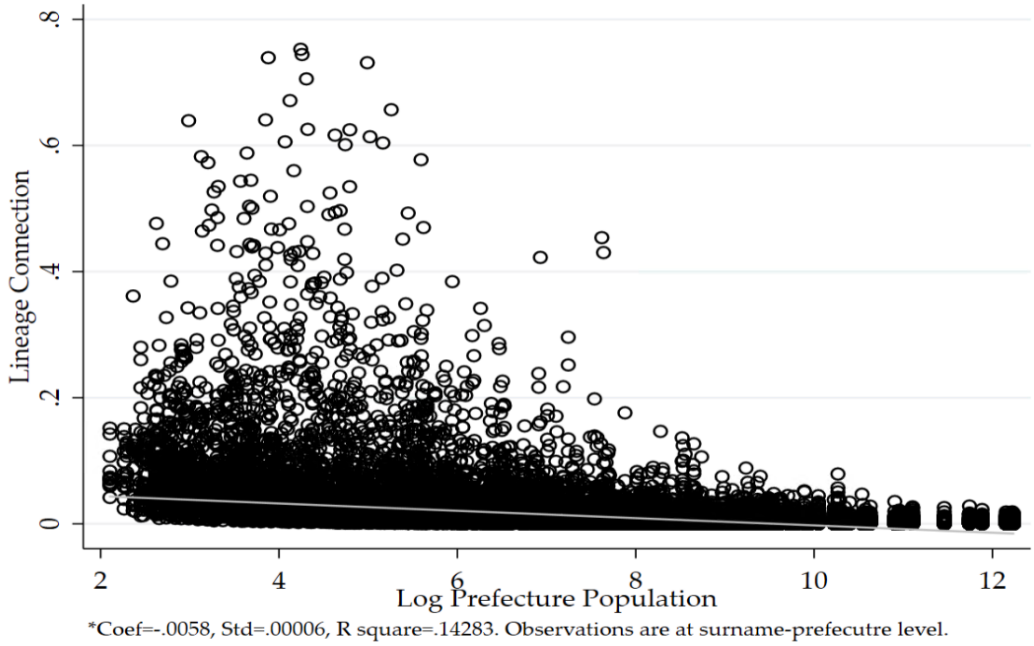
Figure A.6: Cross-validate with Official FDI Statistics in Aggregate



Note: In this figure, we try to cross-validate our use of the SAIC database to characterize foreign investments by comparing the entry and the registered capital of foreign firms calculated from the SAIC database with the number of contracted FDI cases and the contracted FDI capital observed in the data provided by the Ministry of Commerce.

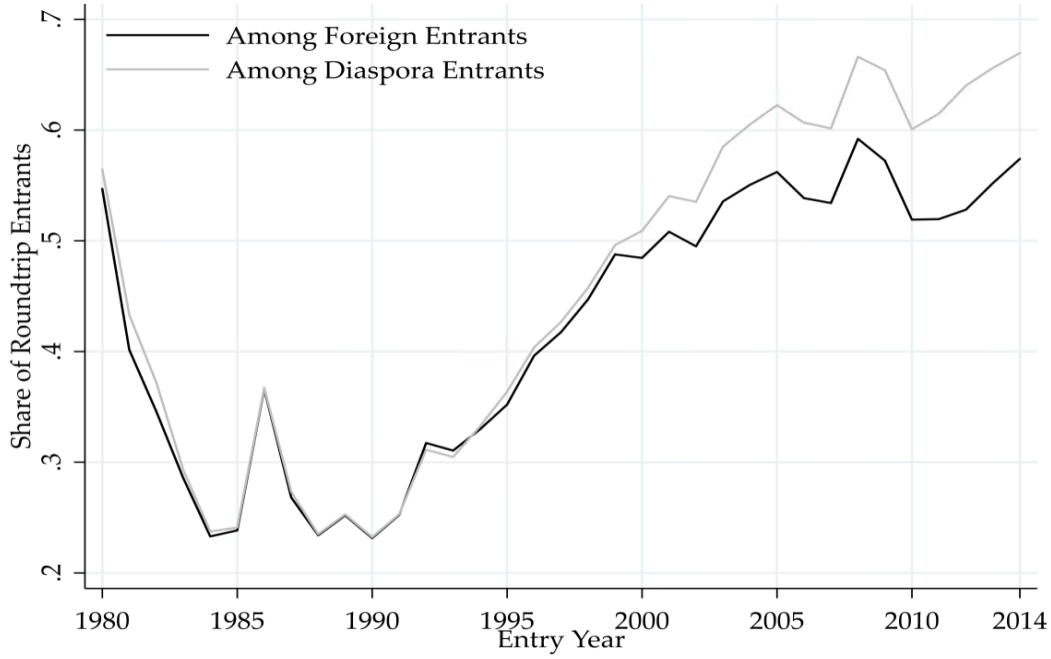
A.3 Additional Figures and Tables for Supporting the Measurement

Figure A.7: Lineage Connection Measure versus Population Size



Note: In this figure, we plot our lineage connection measures between surnames and prefectures over log prefecture population sizes. The pattern shows that our measure is size free in the sense that there is a mild negative correlation between lineage connections and log prefecture population sizes.

Figure A.8: Share of Roundtrip FDI



Note: In this figure, we plot the shares of roundtrip foreign entrants among all foreign entrants and among just diaspora entrants. Here roundtrip foreign entrant is defined as foreign entrants who have legal representatives holding Chinese national ID.

Table A.1: Legal Representative v.s. Top Executive Position in Foreign Firms

	Top Executive	Not Top Executive	Total
Legal Representative	45.18%	3.22%	48.40%
Not Legal Representative	15.54%	36.06%	51.60%
Total	60.72%	39.28%	100%

Prob(Top Executive/Legal Representative) =45.18%/48.40% =93.36%
Prob(Legal Representative/Top Executive) =45.18%/60.72% =74.41%

Note: The sample includes all registered personnel working in foreign firms ever existed from 1981 to 2014. The number represents the percentage of personnel in each category. Conditional on being a legal representative, a person's chance of holding a top executive position within a foreign firm is 93.36%. Conversely, the chance is reduced to 74.41% for one's being a legal representative given that he or she holds a top executive position.

Table A.2: Personnel Structure of Foreign Firms

	Percentage
Has a legal representative	96.84%
Has a chairman on the board	23.90%
Has a CEO	20.57%
Has more than one legal representative	1.35%
Has more than one chairman	4.78%
Has more than one CEO	1.24%

Note: The sample includes all registered personnel working in foreign firms ever existed from 1981 to 2014.

Table A.3: Registered Capital v.s. Other Economic Outcomes

	Log Registered Capital	
	(1)	(2)
Log Employment	0.025***	0.018***
	(0.003)	(0.004)
Log Assets	0.976***	0.416***
	(0.003)	(0.004)
Log Sales	-0.124***	-0.005
	(0.003)	(0.003)
R^2	0.710	0.946
N	150065	124964
Year Fixed Effects	Y	Y
Industry Fixed Effects	Y	N
Firm Fixed Effects	N	Y

Note: The sample includes all foreign firms according to the ownership code in Annual Survey of Industrial Firms (ASIF) during 1998-2007. Industry is at the 2-digit level. ***,**,* denote significance level at 1%, 5%, and 10%, respectively. Standard errors are clustered at the firm level and shown in parenthesis. This table validates our use of registered capital as the measure of entry quality since it is highly correlated with other important firm performance measures such as employment, assets, and sales.

A.4 Additional Robustness Checks

Table A.4: Robustness Checks I: Alternative Dependent Variables

	At Least One Diaspora Entrant	At Least One Survived Diaspora Entrant in 4 Years	At Least One Survived Diaspora Entrant in 2014	Log Survived Diaspora Registered Capital in 2014
	(1)	(2)	(3)	(4)
Open × Lineage Connection	0.766***	0.672***	0.165***	1.229***
	(0.112)	(0.102)	(0.053)	(0.422)
Adj.R^2	0.383	0.373	0.255	0.271
N	1,344,421	1,344,421	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at the surname-prefecture level. Open indicates whether the prefecture has been opened to foreign capital while lineage connection is between surname and prefecture calculated from equation 1. We deal with log zero by inverse hyperbolic transformation.

Table A.5: Robustness Checks II: Alternative Setups

	Open × Lineage Connection		
	Alternative Lineage Connection Measure	Alternative Standard Error	Include Roundtrip Diaspora Entrants
	(1)	(2)	(3)
Number of Diaspora Entrants	2.874***	1.767***	3.821***
	(0.946)	(0.493)	(0.820)
Survival-Adjusted Number	2.511***	1.517***	3.253***
	(0.903)	(0.456)	(0.743)
Number of Incumbent Firms	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y
Surname-Year FE	Y	Y	Y
Prefecture-Year FE	Y	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at the surname-prefecture level except in column (2). Survival-adjusted Number of Diaspora Entrants is the number of entrants that survive for more than 4 years (included). Open indicates whether the prefecture has been opened to foreign capital while lineage connection is between surname and prefecture calculated from equation 1. In column (1), we use alternative lineage connection measure calculated from registered personnel of diaspora firms founded before 1992 in the SAIC database. In column (2), we cluster the standard errors at prefecture level instead of surname-prefecture level. In column (3), we include also the potential roundtrip diaspora entrants which are defined as entered foreign firms represented by citizens of People's republic of China.

Table A.6: Robustness Checks III: Subsample Regressions

	Open × Lineage Connection			
	Excluding Emigration-Intensive Provinces	Within Emigration-Intensive Provinces	Excluding Emigration-Intensive Surnames	Excluding FDI-Intensive Prefectures
	(1)	(2)	(3)	(4)
Number of Diaspora Entrants	0.642***	6.001***	1.733***	1.534***
	(0.233)	(2.476)	(0.332)	(0.440)
Survival-Adjusted Number	0.482***	5.152***	1.480***	1.308***
	(0.167)	(2.186)	(0.292)	(0.380)
Number of Incumbent Firms	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y

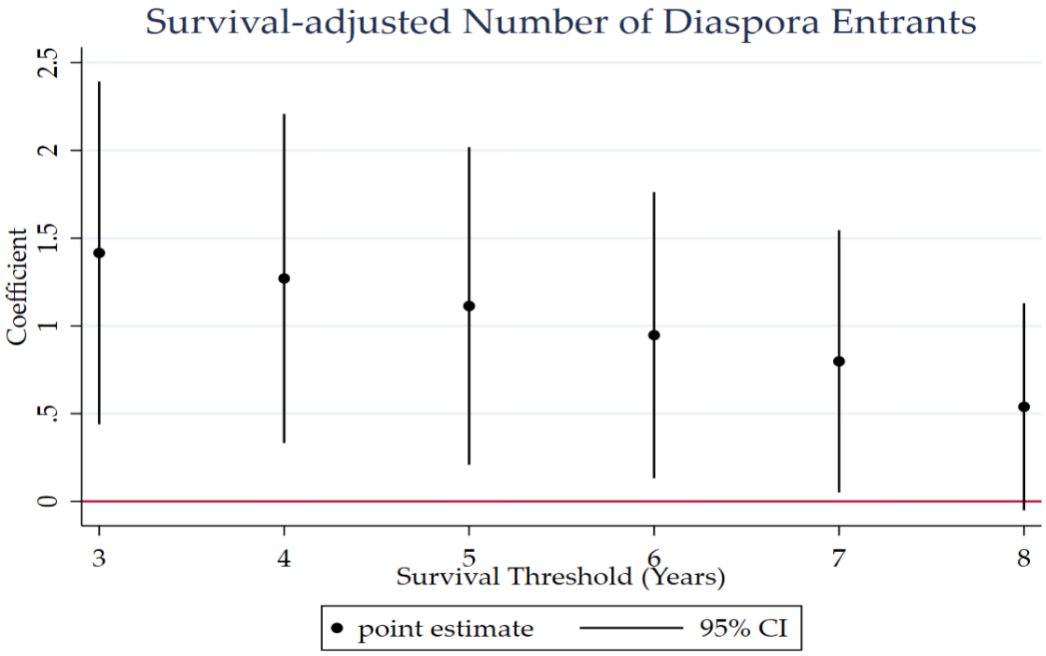
Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at the surname-prefecture level. Survival-adjusted Number of Diaspora Entrants is the number of entrants that survive for more than 4 years (included). Open indicates whether the prefecture has been opened to foreign capital while lineage connection is between surname and prefecture calculated from equation 1. Emigration-intensive provinces include Guangdong, Fujian, Zhejiang. Emigration-intensive surnames refer to the 20 most populous surnames among all overseas Chinese legal representatives from 1981 to 2014. FDI-intensive prefectures are those whose cumulative number of hosted foreign firms rank top 20 among all prefectures during 1981 to 2014.

Table A.7: Poisson Model

	Number of Diaspora Entrants	Survival-adjusted Number of Diaspora Entrants
	(1)	(2)
Open × High Connection Dummy	0.438*	0.431*
	(0.250)	(0.259)
Pseudo R^2	0.833	0.830
N	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y
Surname-Prefecture FE	Y	Y
Year FE	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. Standard errors are clustered at the surname-prefecture level. Survival-adjusted Number of Diaspora Entrants is the number of entrants that survive for more than 4 years (included). Open indicates whether the prefecture has been opened to foreign capital while lineage connection is between surname and prefecture calculated from equation 1. Estimation is based on a Poisson regression model. High Connection Dummy equals one if the lineage connection for a surname-prefecture pair is greater than the median across all surname-prefecture pairs and zero otherwise.

Figure A.9: Adjust Entry by Survival with Different Thresholds



Note: In this figure, we plot the coefficients obtained from running the regression in the baseline specification in equation 2 with different thresholds for survival-adjusted number of diaspora entrants.

A.5 Additional IV Results

Table A.8: Balanced Test

	Predicted Number of Diaspora Firms in 1996	Number of Diaspora Firms in 1996
	(1)	(2)
Distance to Sea	0.001	0.089
	(0.003)	(0.390)
Log Slope	-1.816	-404.623
	(2.325)	(334.211)
Log Elevation	0.466	74.096
	(1.208)	(170.632)
Log Cultivated Land per capita 1996	-1.181	-636.134***
	(0.911)	(125.156)
Log Average Wage 1996	3.311	1537.116
	(2.151)	(280.826)
R^2	0.325	0.672
N	231	231

Note: ***, **, * denote significance level at 1%, 5%, and 10% respectively. We regress the observed number of diaspora firms survived in 1996 and the predicted number of diaspora firms in 1996 separately on a set of prefecture-level controls in this table. In column (1), we see our instrument is orthogonal to these prefecture characteristics. In column (2), the observed number of diaspora firms in 1996, however, is more correlated with these prefecture characteristics in terms of more significant and larger coefficients and doubled R square.

Table A.9: Seeding Effects of DDIs: Without Controls

	Non-Diaspora Foreign Firm Stocks in 2014	Domestic Private Firm Stocks in 2014	Log Registered Capital of Non-Diaspora Foreign Firms in 2014	Log Registered Capital of Domestic Private Firms in 2014	Size of Employ-ment in 2015 (thousand)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
1996 Diaspora Firms	0.237***	72.352***	0.002***	0.000***	0.870***
	(0.077)	(12.126)	(0.000)	(0.000)	(0.114)
Panel B: Reduced Form					
Predicted Diaspora Firms	26.763***	6990.513***	0.262***	0.109***	195.872***
	(5.324)	(800.221)	(0.044)	(0.014)	(38.996)
Panel C: 2SLS					
1996 Diaspora Firms	0.635***	162.255***	0.005***	0.002***	4.476***
	(0.159)	(64.060)	(0.002)	(0.001)	(1.700)
N	266	266	266	266	266
F statistics	48.741	48.741	48.741	48.741	48.741
Controls	N	N	N	N	N
Province Fixed Effects	Y	Y	Y	Y	Y

Note: ***, **, * denote significance level at 1%, 5%, and 10%, respectively. Standard errors are clustered at province level and shown in parenthesis. The data on non-diaspora foreign firms and domestic private firms in 2014 are from the SAIC database. The Data on employment size in 2015 is from the 2015 Population Census of China. Panel A presents the OLS estimates, while Panel B presents the reduced-form estimates from regressing outcomes of interest on our IV: the predicted diaspora firms in 1996 driven by the interaction of opening shocks and lineage connection in zero stage regression. Panel C presents the 2SLS estimates using the predicted diaspora firms in 1996 as IV for the observed diaspora firms in 1996. The Cragg-Donald Wald F statistic is reported for IV regressions.