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 20, 2022

Abstract

As a developing country, China became the top destination for foreign direct investment in just a few decades, defying the "Lucas paradox." Using a unique administrative dataset of the universe foreign firms in China, this paper documents that initial foreign direct investment was mainly driven by the Chinese diaspora; massive non-diaspora foreign direct investment did not materialize until a later stage. Leveraging the staggered opening up of Chinese prefectures during 1981-96 as an identification strategy, the paper finds that following the opening up, diaspora direct investment is more likely to enter prefectures with stronger lineage connections. These prefectures also witness a greater number of non-diaspora foreign and domestic private entrants in the later period.

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 the University of Toronto. Xiaobo Zhang: x.zhang@gsm.pku.edu.cn, Guanghua School of Management at Peking University and the International Food Policy Research Institute. Ruichi Xiong and Fanghao Chen acknowledge financial support from the 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, and Shilin Zheng as well as seminar participants at the University of Toronto, Sun Yat-Sen University, Harbin Institute of Technology (Shenzhen), YES Conference at Jiaxing University, University of Nottingham Ningbo, CCER Summer Institute at Peking University, Southwestern University of Finance and Economics, Fudan University, ICCDS at Shanghai Jiao Tong University, Huaqiao University, University of International Business and Economics, and European Winter Meetings of the Econometric Society for helpful comments and discussions. We 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 practice, developing countries attract less foreign direct investment (FDI) than their developed counterparts (World Bank, 2017). This puzzle was highlighted in Lucas (1990) and later 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), and 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, p. 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?

China's recent development experience defies the Lucas paradox, demonstrating that it is possible to attract FDI in an 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 for FDI. A salient feature is that the initial FDI was primarily driven by the 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 FDI from multinationals in developed countries, which were not related to the 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 environment than the non-diaspora foreign investors were. After observing the success of diaspora direct investment (DDI), 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's legal representative, we can determine whether the foreign firm is controlled by a member of the Chinese diaspora or not. Next, we measure the degree of lineage connections 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 event study exercise, showing that after a prefecture's opening up, DDI was more likely to go to prefectures with stronger lineage connections. We also find a strong, long-term spillover effect of DDI. Prefectures that initially attracted more DDI later witnessed a greater number of foreign and domestic private firms. The economic effect is sizable. The seed of a diaspora firm in 1996 at the prefecture level brought about an additional 0.418 non-diaspora foreign firm, 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 networks (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 FDI.

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 causal 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. In the closet study to ours, Burchardi, Chaney and Hassan (2019) shows that ancestry composition matters for FDI inflows into U.S. counties. Our paper differs from their seminal paper in two aspects. First, we focus on China, a large developing country. Second, we study the effect of the surname-based lineage connections of the diaspora on investment back in their homeland.

Using survey data, Nanda and Khanna (2010) find that diaspora networks play an important role in providing business leads and financing in the Indian software industry. Saxenian (2007) documents a rising trend of highly skilled immigrants 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 the high-tech sectors. Yet, knowledge about DDI in these sectors in developing countries is scant. Our paper fills this gap.

Another study that is close to ours is by Ma (2018). He employs the geographical borders of Chinese linguistic dialect zones as the identification strategy, based on the Annual Survey of Industrial Firms in China (ASIF) with sales above 5 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 on the 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 DDI on subsequent non-diaspora investment.

Our study also speaks to the literature on the spillovers of inward FDI in China. Using the ASIF data, Huang, Jin and Qian (2013) find that the HMT investment underperforms compared with FDI 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) conclude 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 DDI, not just the HMT investments used in previous studies. Second, the database of the administrative universe of firm registrations 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 previous studies.

The findings of the paper may have policy implications for other developing countries. In the absence of an ideal institutional environment, developing countries can strategically tap diaspora investment in the beginning, while improving the quality of local institution along the way, which is a fundamental factor for attracting more subsequent non-diaspora FDI. 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 have preferred to invest in regions with better quality of economic institutions after China launched the nationwide opening-up policy and DDI had demonstrated success.

The rest of the paper is organized as follows. 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 DDI 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. For the first time in 30 years, it provided a legal framework under which foreign firms were allowed to operate in Mainland China. The next year, four special economic zones (SEZs) were established — Shenzhen, Zhuhai, Shantou, and Xiamen. Foreign corporations enjoyed broad autonomy and preferential tax treatment in these SEZs. In 1984, 14 additional cities were granted the status of *Open Coastal City* for attracting FDI[1] . Then Yingkou prefecture, Weihai prefecture, and Hainan province were also added to

[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

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 Southern China in 1992. The visit heralded a new era of opening up to foreign capital on a much larger scale.

The top panel in 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 the staggered roll-out of the opening policy. 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 implement further opening up. Another milestone was China's formal accession to the World Trade Organization (WTO) in 2002. Following the WTO agreement, China eliminated most restrictions on foreign entry and ownership, rekindling a surge in foreign entrants during 2000-05.

Noticeably, the entry of foreign firms was already very active even before China joined the WTO. By 1999, China had become one of the most popular destinations for FDI flows, second only 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.

FDI is not a homogeneous group. A defining feature of FDI in China is the high concentration of DDI. The bottom panel in Figure 1 decomposes foreign entrants into diaspora and non-diaspora ones from 1984 to 2014. The share of diaspora entrants was over 90% from 1984 to 1994, highlighting the dominant role of DDI in the early opening up period. Since then, its relative importance has declined, but it still accounted for more than 60% of total FDI as of 2014.

China has a long history of emigration dating back to as early as the 1300s (Kuhn, 2008). According to the Global Migrant Origin Database, the stock of overseas Chinese was as high as 5.8 million in 2007, ranking in sixth globally in terms of origin countries.[2] Among the vast

counties under the administration of a city are also included in the statistics for that city.

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

number of overseas Chinese, many have become successful entrepreneurs in their host countries. According to the estimate by *The Economist.*, in 2019, the Chinese diaspora contributed more than three-quarters of the Southeast Asian billionaire wealth.[3]

China's closure to the outside world for three decades, starting in 1949, did not fully cut off the bond between overseas Chinese and their ancestral land. Although they are physically abroad, many overseas Chinese are emotionally attached to their ancestral hometowns. For example, they maintain some traditional practices, such as compiling genealogy books and worshiping ancestors in lineage temples in their adopted land (Szonyi, 2002). They also remain in touch with group members of their lineage in their ancestral land in the forms of sending remittances and writing letters (Tan, 2006; Kuhn, 2008).[4]

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

3 Data Description

This section first describes the major data we use in the empirical analyses and their sources. Second, we explain how to identify diaspora firms as foreign firms that are controlled by overseas Chinese. Next we define the measure of lineage connection for each surname-prefecture pair. Last, summary statistics of the key variables used in the empirical analyses are provided.

[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 (Szonyi, 2002; Greif and Tabellini, 2017)

[5] There may be concern that over time the emotional attachment of members of the Chinese diaspora, especially younger and foreign-born individuals, to their ancestral homelands might wane, weakening the strength of the lineage connection. However, according to interviews conducted by Tan (2006), many China-born parents would bring their foreign-born children to their 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.

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 chose 2014 as the end year of the sample, because the number of new entrants spiked after China launched a national business registration reform in that year (Barwick et al., 2022). 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 of 2005, which reports the 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 Identification of Diaspora Firms

By definition, diaspora firms must be foreign firms in the first place. The literature suggests two ways to define foreign firms. One approach relies on the administrative firm ownership code assigned by government agencies, and the other approach uses shareholder information. In this paper, we opt for the first approach, which is less complicated than the second one. 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 we had followed the second approach, we would have had to go through many layers of ownership structure to determine the real control shareholders. For simplicity, we decided to adopt the more straightforward first approach, by using the administrative ownership code readily available in the dataset to identify foreign firms.[6]

[6] There might be concern 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 rather than selling shares to foreigners. Moreover, any change in ownership would trigger a change in the firm identifier as well, creating a new legal entity (Chen et al., 2021). 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

Having pinned down foreign firms, we proceed to identify diaspora firms. It would be ideal to define diaspora firms as those with overseas Chinese as the controlling shareholders. 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 information on the firms' 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). A key advantage of this method is that the names of legal representatives are publicly 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 an ID from somewhere other than Mainland China.[8] Diaspora firms refer to 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, chief executive officers, 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 of the shareholders in these foreign firms. It is almost impossible to back out the ultimate natural person owners of foreign firms through the 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 with a top-ranked executive within a foreign firm is 93%. This means that the legal representative is highly likely to be the person in charge

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 members of the Chinese diaspora 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 the DDI. The problem may not be that serious. As documented in Tan (2006), member of the Chinese diaspora often use their original Chinese names to signal their Chinese identity when dealing with Chinese businesses.

[8] Non-Mainland China IDs include foreign passports and residency cards of Hong Kong, Macau, or Taiwan.

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 provided so-called "round-trip FDI" (Huang, 2003). (Chen, 2022, p. 393-98) more rigorously defines round-trip FDI 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". Figure A.8 documents that the share of round-trip FDI in total FDI identified using our method are close to the previous literature, which uses aggregate statistics to gauge the presence of round-trip FDI in China (Geng, 2019). Since the purpose of round-trip FDI 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 connections. Our findings are robust to the inclusion of non-diaspora foreign firms in the sample.

Last, since the legal representatives are not necessarily the major shareholders of firms, it is possible that some of the firms that we label diaspora firms are not owned by overseas Chinese. However, the action of appointing overseas Chinese as legal representatives likely reveals the foreign firms' preference for taking advantage of the lineage connections. Ignoring these observations would underestimate the effects of lineage connections in facilitating the entry of other foreign firms. [9]

3.3 Measurement of Lineage Connections

For a given surname of a diaspora firm's legal representative, we use the surname's geographic distribution, drawn from the China Population Survey of 2005, to measure the strength of the 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 the 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 origin of their ancestry. 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 are 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,

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

[10] We chose 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 connections, a spurious correlation between this measure and the entry of diaspora firms may occur even in the absence of lineage connections. Some common 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 surname are also more likely to be present. Therefore, the positive correlation does not necessarily mean that there are lineage connections. By comparison, as shown in A.7 in the Appendix, the lineage connection measure in equation 1 is mildly negatively correlated with the size of the prefecture population by surname, ruling out the possibility of a positive spurious correlation.

which are ranked based on the China Population Survey of 2005, from three samples: emigration-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 it differs sharply from the distribution for 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. However, it ranks only the fifth among the Chinese population as a whole. This simple comparison suggests a possible lineage connection between the diaspora 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 the WTO. Utilizing the entry and exit information in the SAIC database, we construct two measures of diaspora firm entry at the surname-prefecture-year level: number of new entrants and the survival-adjusted number of entrants (defined as those lasting more than four years). We exclude four autonomous regions — Xinjiang, Tibet, Ningxia, and Inner Mongolia — which are mainly composed of ethnic minorities. Given that surname-based lineage mostly operates among ethnic Han Chinese (Zhang, 2020), and most overseas Chinese are Han Chinese, our lineage connection measure is not applicable there. We further exclude four mega cities in China — including Beijing, Shanghai, Guangzhou, and Shenzhen — because their massive economic size attracts a vast number of internal migrants, masking the historical distribution of local surnames.

Table 2 presents the summary statistics for all the variables used in the 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 that were still surviving in 2014 is even smaller

[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 for cases when a Chinese character surname corresponds to multiple English spellings. See Appendix A.1 for details.

(0.9%). The lineage connection measure varies widely, with a mean of 0.005 and a standard deviation of 0.011, as indicated in panel B. Panels C and D summarize the main outcomes of interest at the prefecture-year or prefecture level, which will be studied later in the paper.

4 Lineage Connection and Diaspora Firm Entry

4.1 Identification Strategy

We employ the following baseline specification to identify empirically the effect of lineage connections 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 the surname-prefecture-year level. A 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 the analysis with 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 the specifications, we control for surname-prefecture fixed effects, η_{sp}; surname-year fixed effects, θ_{st}, and prefecture-year fixed effects, δ_{pt}, unless otherwise 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. We also control S_{spt}, the number of incumbent firms in the year prior to the entry year, 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 in year t. $Open_{pt}$ equals 1 if prefecture p has been awarded opening status since year t, and it equals 0 otherwise.[14] m_{sp} is

[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 period of opening

the measure of lineage connection between surname s and prefecture p as defined in equation 1. The error term, ϵ_{spt}, captures all the idiosyncratic disturbances. Standard errors are clustered at the 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, and (3) differences across surname-prefecture pairs with lineage connections of varying strengths.

The causal interpretation of our estimates depends on a 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 and Kung, 2022). Second, foreign capital inflows were largely prohibited before a prefecture was granted opening status. Third, various fixed effects have been controlled to reduce confounding factors that could contaminate our causal estimates.

To 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 the identification assumption to be valid, we expect β^{τ} to be statistically indifferent from zero for

up period. As Shanghai is excluded from the sample, the opening of Pudong district in 1990 is not captured in our analysis.

all $\tau < 0$.

As shown in the top two plots in 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 diaspora entrants as the outcome variable, validating the no pre-trend assumption.

4.2 Baseline Results

We first estimate equation 2 to evaluate how lineage connections affect the entry of diaspora firms in the face of opening up at the prefecture level. We present the results in columns (1) and (2) in Table 3. According to column (1), an increase of one standard deviation in lineage connection (0.011) is associated with an increase in diaspora entrants of 0.02, accounting for one-third of the average number of diaspora entrants across all the cells aggregated at the surname-prefecture-year level. When using the survival-adjusted number of diaspora entrants as the alternative outcome variable in column (2), the coefficient for the ”Open × 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 in 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.

A recent development in the econometrics of the difference-in-difference (DID) or, more generally, two way fixed effects (TWFE) model points out that biases may 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 after correcting for potential biases. We divide the surname-prefecture pairs with continuous lineage connection into two discrete groups: those above or below the median lineage connection. It would be ideal to apply the estimation methods using the original continuous treatment variable. However, we find that this is computationally infeasible due to our large dataset. Thus, we adopt this second

[15] For a survey on this topic, see De Chaisemartin and D’Haultfoeuille (2022).

best approach.

Columns (3) and (4) in Table 3 report the average treatment effects when replacing the continuous connection variable with a dummy variable and correcting the potential bias of staggered 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 in Figure 3, there are no parallel trends prior to a prefecture's opening up, further validating our identification strategy.

A key difference between the bottom two plots and the top two plots in Figure 3 is that the effect of lineage connection lasts beyond the first two years after correcting for the biases originating 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 the prefectures opened to foreign investment. These spillovers affect not only 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 the early diaspora firms were of better quality than the later ones. In this case, the observed positive effect of lineage connections 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) in Table 4 further look at two proxy variables for entry quality, the survival ratio and average registered capital of the diaspora entrants that were still surviving as of 2014. The coefficients for the interaction terms are insignificantly different from zero, suggesting that sorting on the qualities of entrants is not at play in driving our key results.

Alternative dependent variables. Relatively few diaspora firms entered and survived at the surname-prefecture-year level (see Table 2). Ordinary Least Squares (OLS) regressions may yield biased estimates when the outcome variables do not follow a normal distribution. We address this concern in two ways. 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) in Table A.4 show that the effect of lineage connections still holds at the extensive margin.

Second, considering that entry likely follows a Poisson distribution, we also run Poisson regressions. However, the variation in lineage connections across surname-prefecture pairs is too small to make the maximum likelihood estimation converge. We instead use a dummy variable indicating whether the value of the lineage connection is greater than the median, as we do in columns (3) and (4) in Table 3. As displayed in Table A.7, the results of the Poisson regressions are reassuringly robust.

Alternative lineage connection measure. Since the lineage measure is drawn from the China Population Survey of 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, ever starting in 1984, mass migration did not happen until the late 1990s (Tombe and Zhu, 2019). Clark (2015) documents that the geographic distribution of certain elite surnames has been rather stable throughout history, even after warfare and revolution. Bai and Kung (2022) compares the surname distribution across prefectures in the China Population Survey of 2005 with the surname distribution in Harvard's China Biographical Database, confirming that the two distributions are highly correlated.

Nonetheless, we provide additional checks by constructing an alternative measure based on the surname information of registered personnel who worked in *domestic firms* established before 1992, from information available from the SAIC database. Column (1) in Table A.5 presents the estimation results using this alternative lineage measure. The effect of lineage connections 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) in Table A.5 reports the regression results with standard errors clustered at the prefecture level. There is little change in the significance levels of the key variables.

Including round-trip diaspora entrants. In the main analyses, we excluded round-trip diaspora firms. As a robustness check, we include them back in the sample and repeat the baseline regression in column (3) in Table A.5. The estimate for the interaction term between the opening dummy and lineage connections becomes larger. This result is 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 in 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 reshuffled across prefectures within a surname. Clearly, both distributions are centered around zero. If there were 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 connections on diaspora firm entry are not driven by unobserved measurement noise 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 more emigrants abroad and attracted more diaspora firms, compared with other regions. Furthermore, a few notable surnames are common in 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) in Table A.6. The sample in the regression in 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 the outcome variable. The lineage effect exists not only in emigration-intensive provinces, 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 emigration-intensive surnames. The coefficient for the key variable remains significantly positive. Finally, column (4) excludes FDI-intensive prefectures. The results barely change.

5 Seeding Effects in the Long Run

We have shown that lineage connections were 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 investigate whether the industrial "seeds" planted by the early DDI have had lasting spillover effects on the entry of non-diaspora foreign firms and domestic private firms. We proceed in two steps: (1) showing that compared with non-diaspora firms, diaspora firms are more likely to be pioneers in 4-digit industries in the prefectures; (2) estimating the seeding multipliers of diaspora firms on the entry of non-diaspora foreign firms and domestic private firms separately using an instrumental variable (IV) strategy.

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

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

5.1 Diaspora Firms as Pioneers

Here we define a pioneering firm as the first private entrant into any 4-digit industry in a prefecture.[17] 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 that 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 exceeded that of non-diaspora pioneering firms, suggesting that diaspora firms were more likely to be pioneers than the non-diaspora firms. After 1995, the gap in the likelihood of being a pioneer between diaspora firms and non-diaspora firms had closed. These patterns indicate that the diaspora played an important role in setting up pioneering firms in new industries and new places in the early opening-up period despite the imperfect market environment.

5.2 Estimation of the Long-Run Seeding Multipliers

Hausmann and Rodrik (2003) reckon that pioneering firms generate a 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 this 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 the lineage-driven DDI predicted by the interaction term between the surname distribution and the staggered opening of Chinese prefectures in the early opening-up period based on the estimates of equation 2 (the "zero stage" regression). The lineage-driven DDI predicted by the "zero stage," $\hat{\beta}$, is then used

[17] Private entrants include entering firms of private, collective, and foreign ownership. Notice that we include collective firms as a group of private firms because township-village enterprises (TVEs) were essentially private enterprises wearing "red hats" before private ownership was officially recognized (Xu and Zhang, 2009). We exclude state-owned enterprises (SOEs) when defining pioneering firms because the entry of SOEs into a prefecture-sector cell was largely mandated by governments and did not truly reflect the match between the firm and the local environment.

to construct 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 the estimates over all the surnames and entry years to get the predicted value of the cumulative lineage-driven DDI for each prefecture in 1996 as follows:

$$D\hat{D}I_p = \sum_{s} \sum_{c=1981}^{1996} \hat{\beta} \times Opening_{pc} \times m_{sp} \tag{4}$$

Next, we use the predicted lineage-driven DDI as an IV for the observed DDI in 1996 in the following two-stage least squares (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 firms in prefecture p in 1996, and X_p denotes a set of controls at the prefecture level.

To test the validity of the IV, we perform a balance 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 that are correlated with local economic development, including distance to the sea, slope and elevation of the prefecture's land, cultivated land area per capita in 1996, and the average wage in 1996. Although the number of observed diaspora firms in 1996 is correlated with these variables, our instrument is not, confirming that our instrument passes the balance test.

Table 5 presents the estimation results of the 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 the surviving firms. In addition to the control variables in the balance 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 5. 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 an increase of 0.418 subsequent non-diaspora foreign firms and 137 domestic private firm entrants.

The positive effect is observed not only for the number of firms entering, but also for firm size, measured by registered capital in 2014. One more diaspora firm in 1996 leads to an increase of 0.2-0.3 percent in total registered capital for newly registered foreign and domestic enterprises as of 2014, as indicated in column (4) in Table 5. The emerging new and larger enterprises led to greater local employment. Column (5) in Table 5 quantifies that one more diaspora firm in 1996 generated 4,099 more jobs 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 China's growth story. However, the destinations of DDI were not random. Our paper has shown that areas with stronger lineage connections attracted more DDI during the early opening up period when market institutions were far from perfect. Moreover, we found that early DDI, much of which was pioneering firms, facilitated the subsequent entry of non-diaspora foreign and domestic private enterprises.

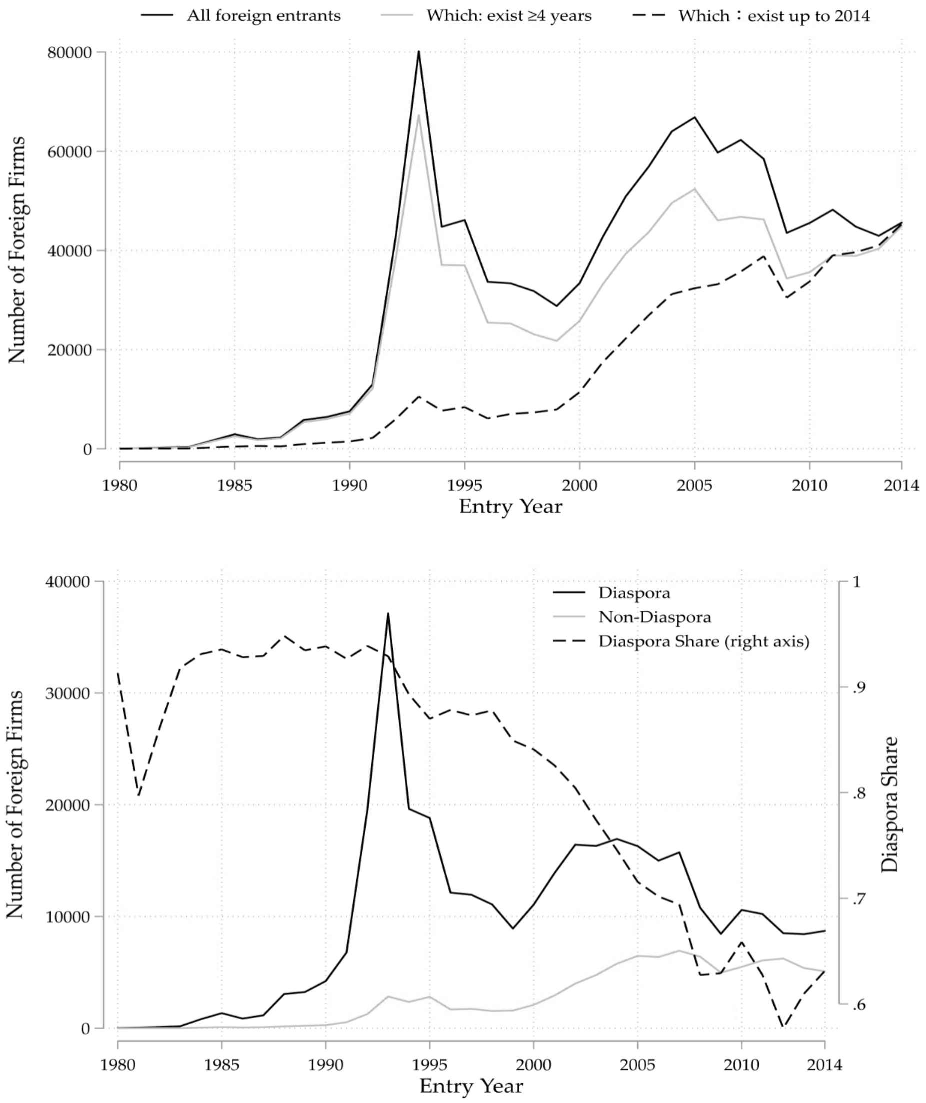
China's experience in tapping diaspora investments in the early stage of development may shed light on other developing countries. Plagued with informational problems, policy uncertainty, and a poor contracting environment, developing countries often struggle to attract FDI, in particular pioneering 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.

From a survey in 2005, the International Organization for Migration, found that more than 90% of the countries had policies or programs targeting their diaspora.[18] DDI accounts for a significant proportion of the FDI in developing countries other 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) call 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 navigate

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

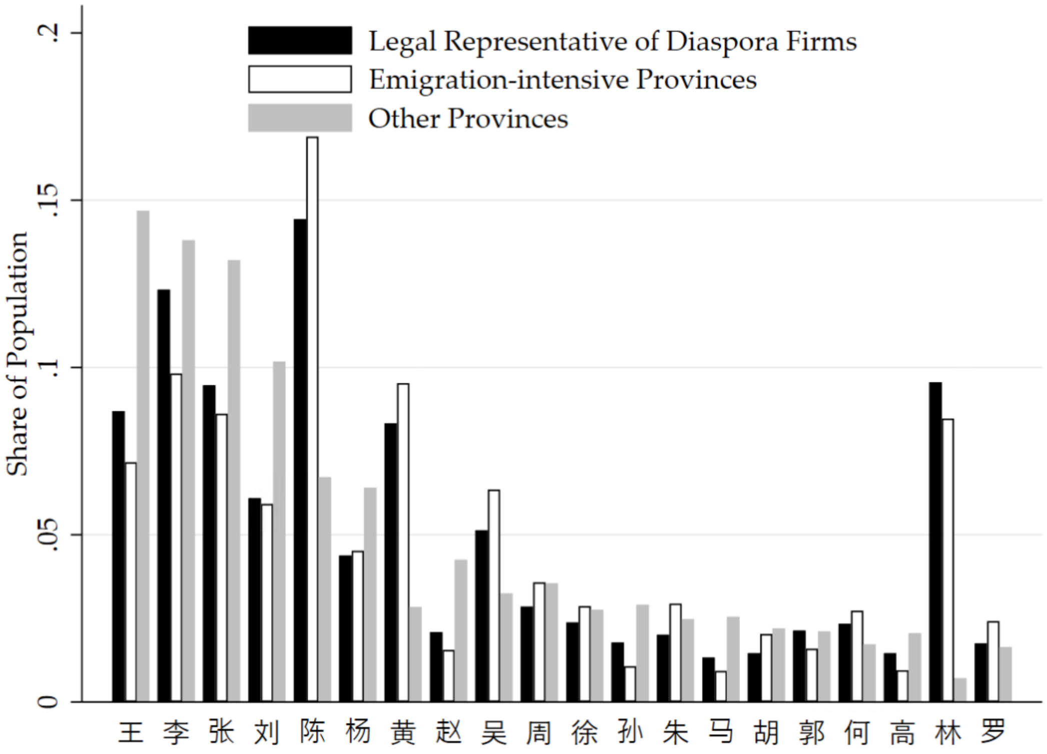
the uncertain and even hostile environment in those states. A future research topic is to explore whether DDI can serve as pioneering firms even in these countries.

Figure 1: Entry and Survival of Foreign Firms



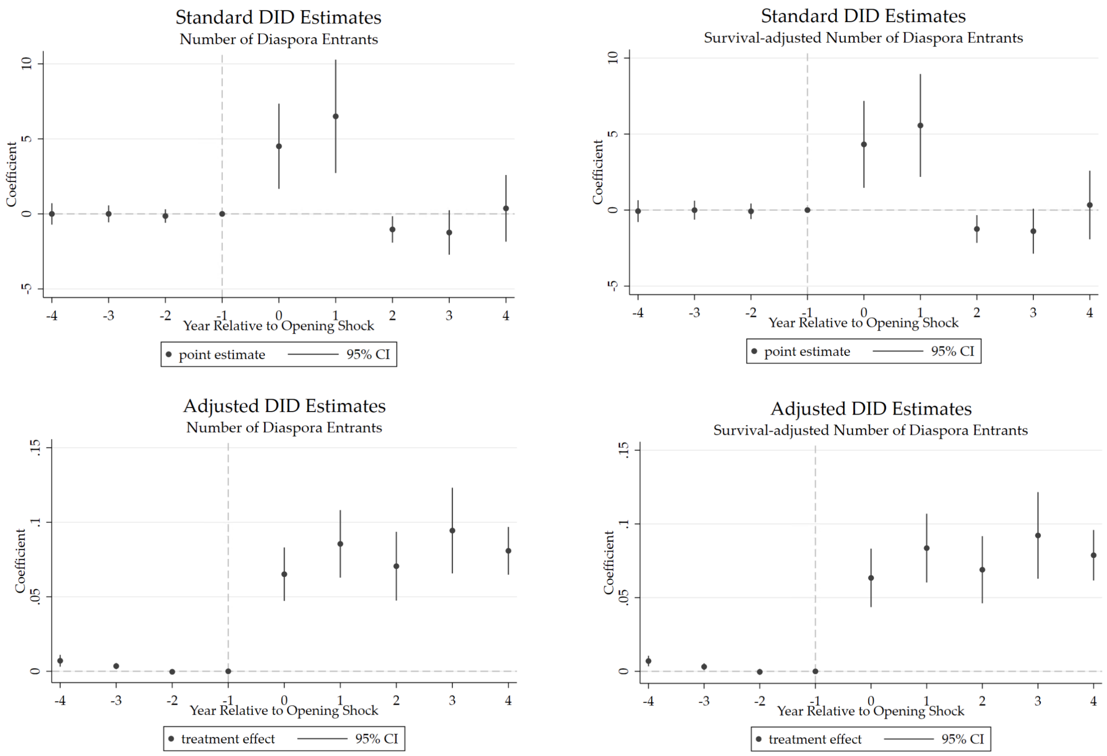
Note: This figure uses firm registration data from the SAIC database. The top panel plots the number of foreign firms (and surviving foreign firms) by year of entry over 1980-2014. The bottom panel displays the number of diaspora and non-diaspora entrants on the left axis as well as the share of diaspora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

Figure 2: Surname Distributions in Different Samples



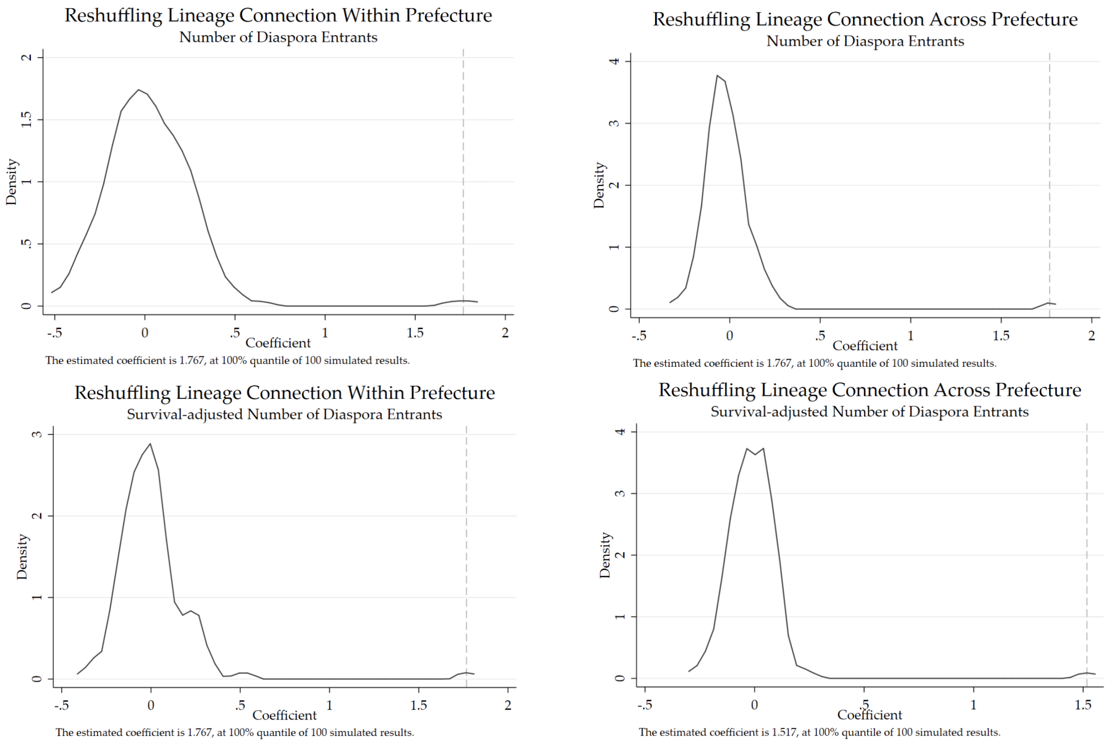
Note: This figure plots the shares of the population with the top 20 surnames based on the China Population Survey of 2005 in three samples: the legal representatives of diaspora firms in the SAIC database, the emigration-intensive provinces, and other provinces. Emigration-intensive provinces include 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 of 2005, but it sharply differs from other provinces.

Figure 3: Event Study



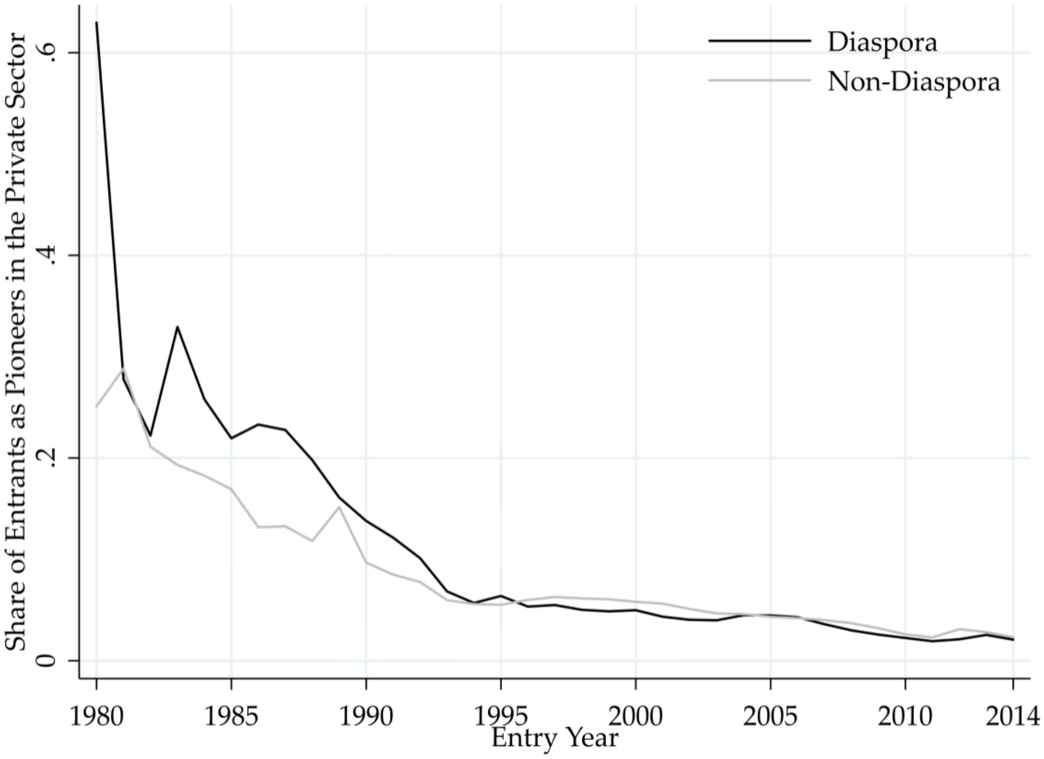
Note: This figure plots the coefficients β^{τ} obtained from a standard event study specification: $Y_{spt} = \eta_{sp} + \theta_{st} + \delta_{pt} + \sum_{\tau=-4}^{4} \beta^{\tau} \times Open_{pt}^{\tau} \times m_{sp} + \lambda \times S_{spt} + \epsilon_{spt}$ where $Open_{pt}^{\tau}$ indicates opening 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 the 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 in the left and right panels are the number of diaspora entrants and the number of survival-adjusted diaspora entrants, respectively.

Figure 4: Permutation Tests



Note: In this figure, we show estimates under permutations that randomly reshuffled 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 firm is defined as the first private 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 Southern 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
Surviving 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 Surviving 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 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 the surname-prefecture-year level. The survival-adjusted number of entrants is defined as the number of diaspora entrants that survived for at least 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 were obtained from the SAIC database. The remaining six variables in panel D were retrieved from official documents or the China Stock Market & Accounting Research 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*** (0.574)	1.517*** (0.527)		
Open × High Dummy			0.079*** (0.007)	0.077*** (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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, 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 four 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 a continuous connection measure mainly because the estimation method used here is computationally infeasible with 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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, 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 were surviving as of 2014 conditional on entry of diaspora firms. “Average Registered Capital in 2014 Conditional on Entry” refers to the average registered capital for the diaspora firms that were still active as of 2014.

Table 5: 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 in 2015 (thousands)
	(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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Standard errors are clustered at the province level and shown in parentheses. 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 are from the China Population Survey of 2015. Panel A presents the OLS estimates, while panel B reports the reduced-form estimates from regressing the outcomes of interest on our IV: the predicted diaspora firms in 1996 driven by the interaction of opening shocks and lineage connection in the 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, and James Kai-sing Kung. 2022. “Surname Distance and Technology Diffusion: the Case of the Adoption of Maize in Late Imperial China.” *Journal of Economic Growth*, 1–39.

Barwick, Panle, Luming Chen, Shanjun Li, and Xiaobo Zhang. 2022. “Entry Deregulation, Market Turnover, and Efficiency: China’s Business Registration Reform.” 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. 2021. “Privatization and Productivity in China.” *RAND Journal of Economics*, 52(4): 884–916.

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, D.C.

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.

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.

Szonyi, Michael. 2002. *Practicing Kinship: Lineage and Descent in Late Imperial China.* Stanford University Press.

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.* Washington D.C.: World Bank.

Xu, Chenggang, and Xiaobo Zhang. 2009. “The Evolution of Chinese Entrepreneurial Firms: Township-Village Enterprises Revisited.” International Food Policy Research Institute.

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 State Administration of Industry and Commerce of China (SAIC) database all the registered personnel working in foreign firms according to the firm's ownership code, with the following variables available for each person: name, ID type, registry address, executive position, and a dummy for the legal representative status.

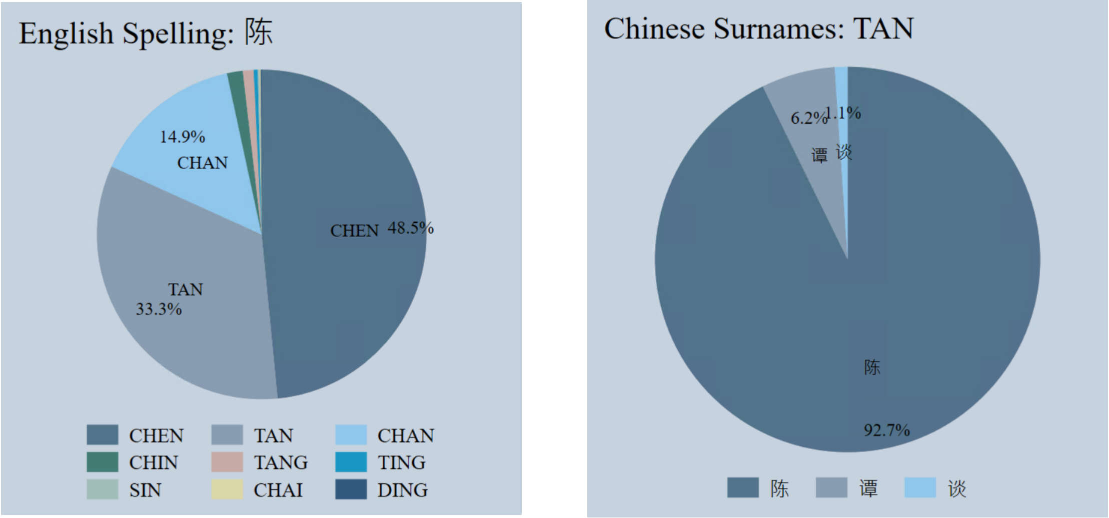
Before formally executing our algorithm, we perform a preparatory cleaning procedure to leave out all symbols (such as comma, period, or semicolon) and word content that is unrelated to names (such as titles,"Mr.," "Dr.," or "appointed by the parent company"), since many name entries are unstructured in the raw data we acquired. 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, may also register their names using Chinese characters. This further complicates the analysis. In light

Figure A.1: 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. In the left panel, we show the proportion of each possible English spelling of the Chinese surname "陈" among registered personnel who have name entries with both English characters and Chinese characters in the SAIC database. In the right panel, we instead present the proportions of each possible Chinese character that 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.

of these concerns, we apply the following procedures to identify ethnic Chinese and their surnames:

1. For each string of a name entry, if the length of the string is greater than four Chinese characters, we tag the string as non-ethnic Chinese, since common ethnic Han Chinese names rarely contain more than four Chinese characters. If the length of the string is exactly four 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 four 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 1,000 Japanese surnames obtained from Wikipedia. If the following three conditions are satisfied simultaneously, we tag the string as a Japanese sur-

name (otherwise a non-Japanese surname): (1) the match with the most common 1,000 Japanese surnames is successful; (2) the ID type of the personnel is foreign passport; and (3) the registry address starts with a typical location 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; and (3) the registry address starts with a typical location 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 leftmost 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 four characters. We extract the leftmost 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 four.

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 leftmost string or rightmost string. Furthermore, some irregular name entries that fail to insert a 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 a blank space in the name entry), and the other with a single string (without a blank space in the name entry).

2. For the group with multiple strings, we match the leftmost and rightmost character with the Chinese surname dictionary. If the match is successful for either the leftmost character or the rightmost 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 leftmost character as the surname by default (in our database, surnames are more likely to be identified in the leftmost 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 surnames 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 leftmost English string and rightmost 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 along the following steps. We first match sequentially

the leftmost English string and rightmost English string with the identified Chinese characters from the 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 leftmost spelling as the default.

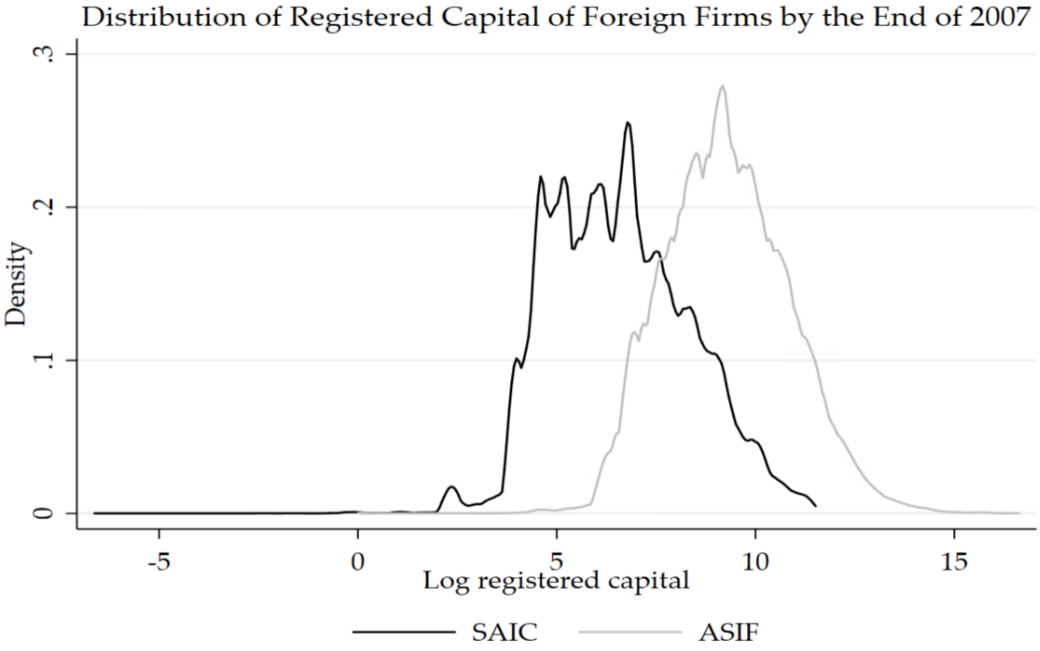
5. This procedure produces not only the surname for each ethnic Chinese, but also an English-Chinese mapping that can be used for randomly assigning English spellings into Chinese characters with empirical weights bootstrapped from the sample of overseas Chinese.

Determining overseas Chinese. The previous steps identify whether a person is ethnic Chinese or not. We further separate the mainlander and overseas Chinese according to the ID type associated with each registered personnel. An overseas Chinese is an ethnic Chinese who holds an ID from somewhere other than Mainland Chinese—including a passport of a foreign country and a residency card of Hong Kong, Macau, Taiwan.

A.2 Comparison with Other Data Sources on Chinese Inward FDI

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

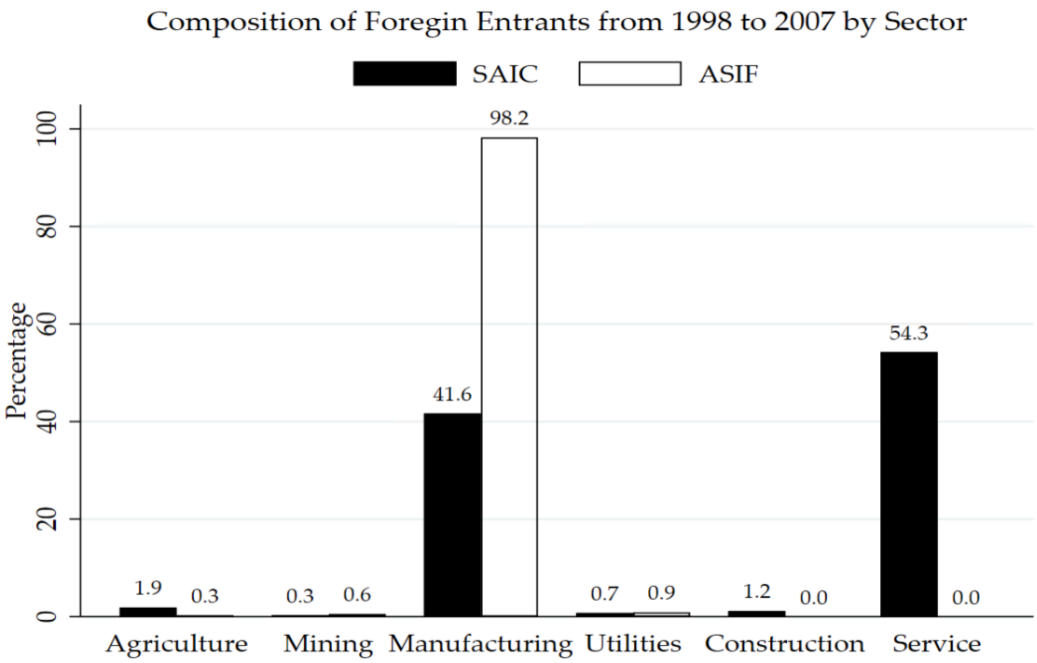
Figure A.2: SAIC versus ASIF: Firm Size



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 large number of small firms.

Foreign firms covered by the ASIF data versus those covered 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 with the ASIF data. First, the SAIC database is representative of foreign firms of all sizes, while the ASIF data only cover large firms with sales of more than 5 million CNY. Figure A.2 plots the distribution of the registered capital of foreign firms at the end of 2007, constructed from the SAIC data and the ASIF data, respectively. It clearly shows that the SAIC data cover smaller firms, while the ASIF data do not. Second, Figure A.3 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: Sectors

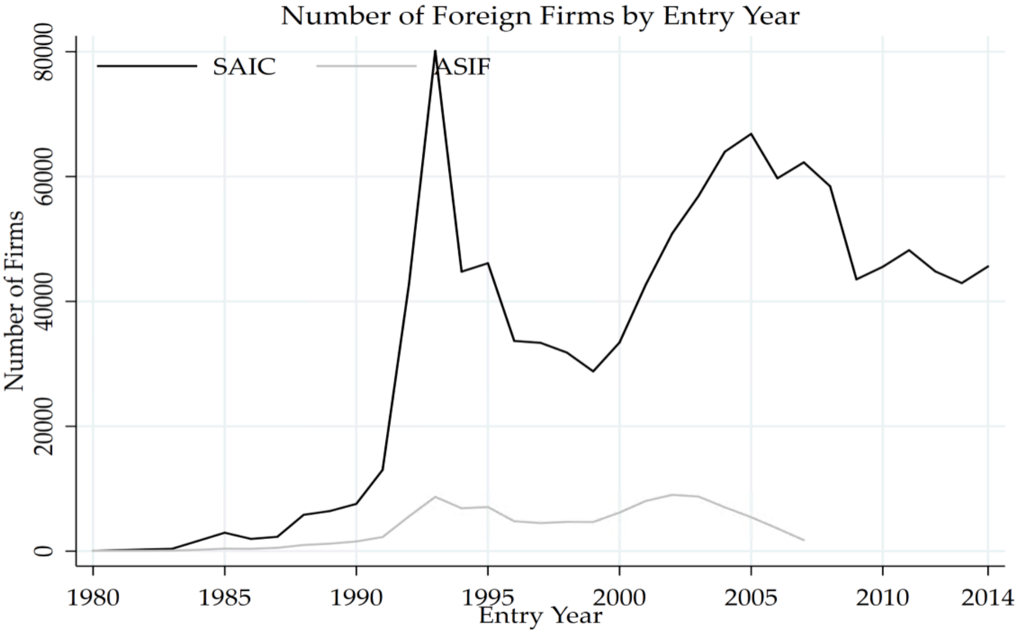


Note: In this figure, we plot the compositions of foreign firms in the SAIC database and the ASIF data. As expected, the ASIF data only cover manufacturing firms, while in the SAIC database manufacturing firms account for only about 40% of all firms by count. Therefore, using the ASIF data to study foreign firms in China would miss a large number of non-manufacturing firms that influence the country's economic development.

Diaspora firms versus Hong Kong, Macau, and Taiwan (HMT) firms. Due to data limitations, 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 an approach. First, a lot of Europe-based and United States-based multinationals invest in Mainland China through Hong Kong as a conduit, but they are not actually diaspora firms. Second, residents in HMT only make up a small fraction of the 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 widened during the process of China's gradual accession to the World Trade Organization from 1995 to 2007.

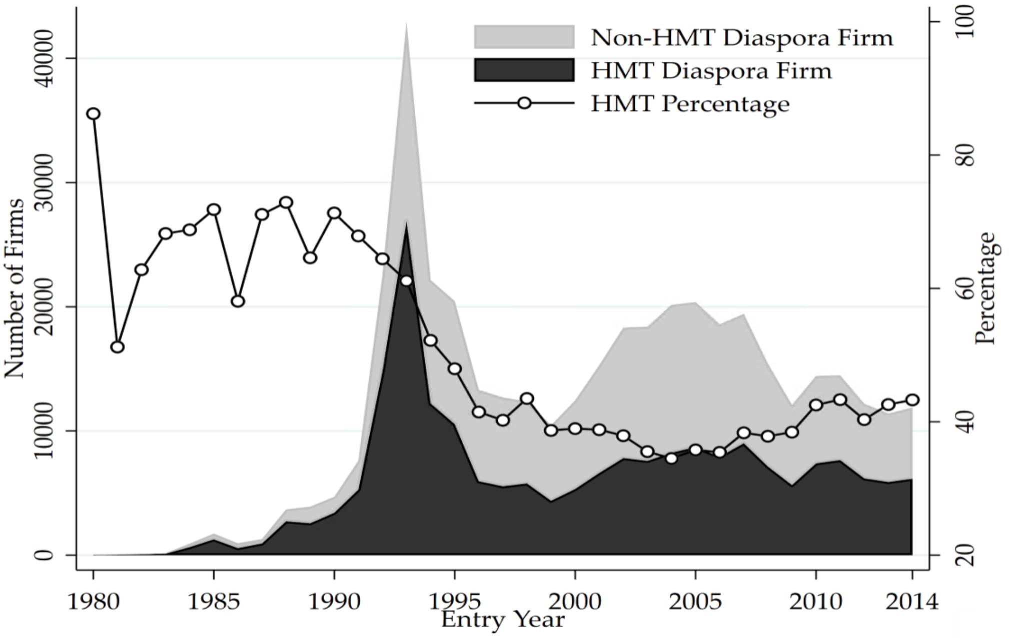
Foreign firms versus foreign direct investment (FDI). As a monetary concept, FDI includes both the initial and follow-up investment from a foreign owned entity. We cross validate our dataset of foreign firms both in counts and volumes with the official FDI data provided by the Ministry of Commerce of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

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.

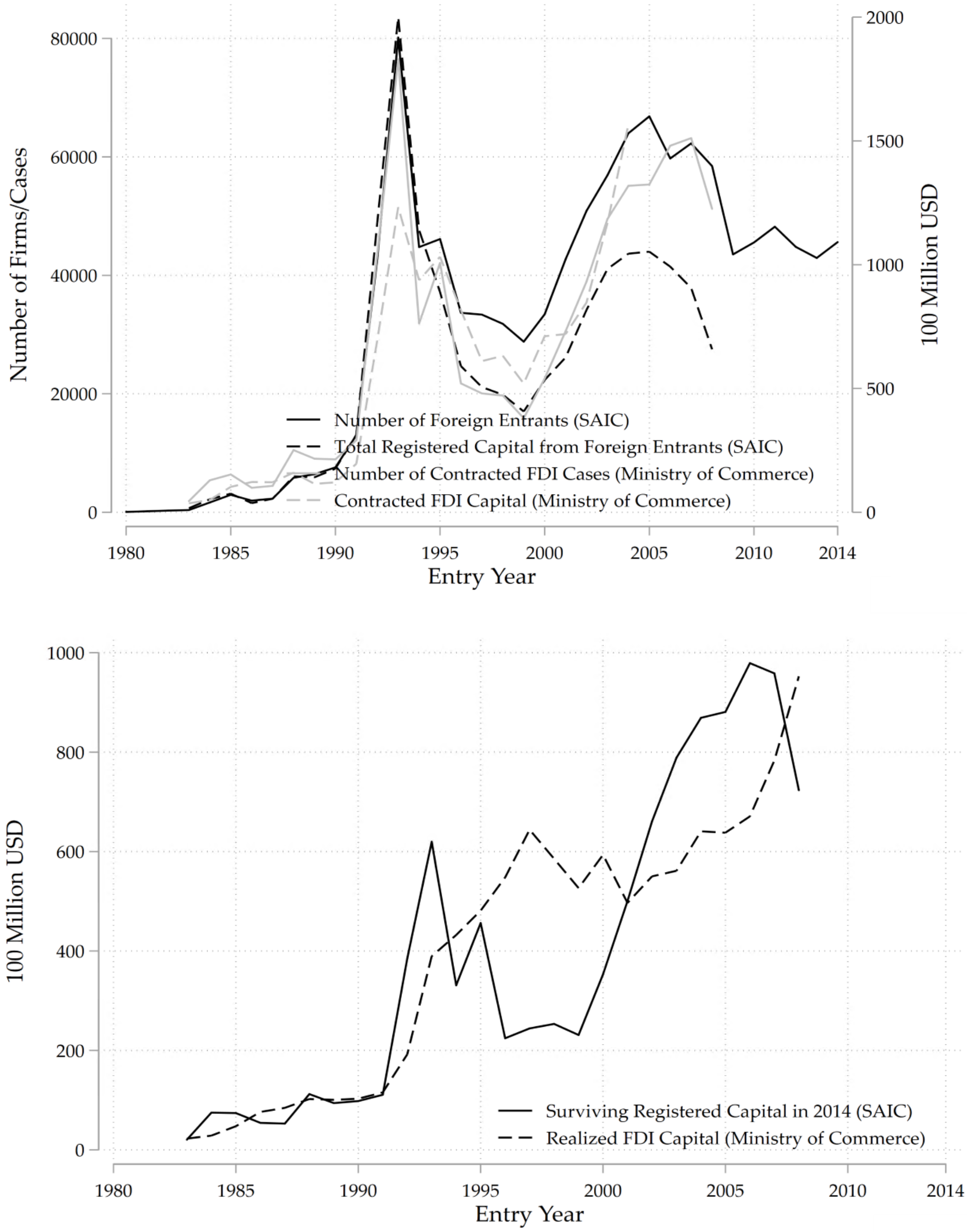
Figure A.5: Diaspora Firms versus HMT Firms



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

our dataset. Furthermore, the realized FDI measured in U.S. dollars is also highly correlated with the registered capital of the surviving foreign firms in 2014 calculated from our dataset. Our dataset therefore is consistent with official aggregate FDI statistics both in counts and volumes.

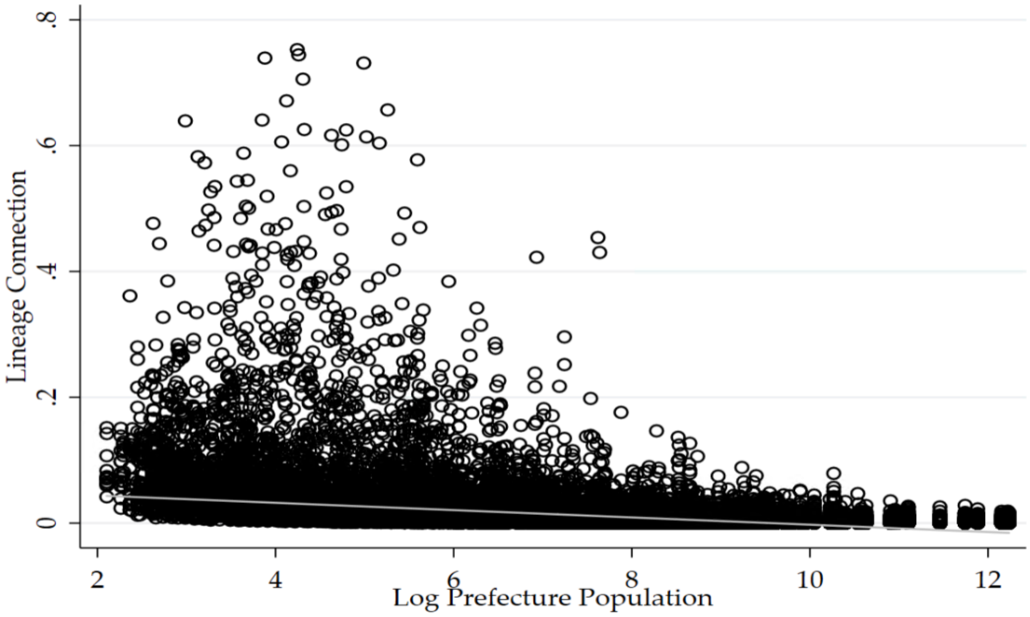
Figure A.6: Cross Validation with Official FDI Statistics in Aggregate



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 Supporting the Measurement of Lineage Connections

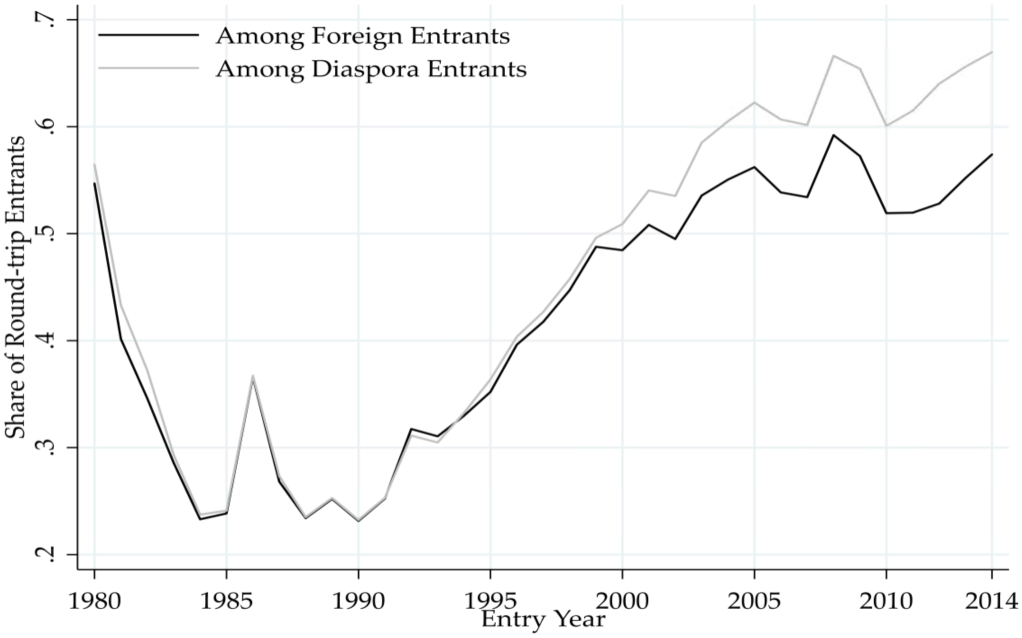
Figure A.7: Lineage Connection Measure versus Population



*Coef=-.0058, Std=.00006, R square=.14283. Observations are at surname-prefecutre level.

Note: In this figure, we plot our lineage connection measures between surnames and prefectures over log prefecture population. 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.

Figure A.8: The Share of Round-Trip FDI



Note: In this figure, we plot the shares of round-trip foreign entrants among all foreign entrants and among just diaspora entrants. Here a round-trip foreign entrant is defined as a foreign entrant that has legal representative who hold a Chinese national ID.

Table A.1: Legal Representative versus Top Executive Position in a Foreign Firm

	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 that 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 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 chairperson on the board	23.90
Has a CEO	20.57
Has more than one legal representative	1.35
Has more than one chairperson	4.78
Has more than one CEO	1.24

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

Table A.3: Registered Capital versus 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 the Annual Survey of Industrial Firms (ASIF) during 1998-2007. Industry is at the 2-digit level. ***, **, and* denote significance at the 1%, 5%, and 10% levels, respectively. Standard errors are clustered at the firm level and shown in parentheses. 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 Surviving Diaspora Entrant after Four Years	At Least One Surviving Diaspora Entrant in 2014	Log Surviving 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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Standard errors are clustered at the surname-prefecture level. Open indicates whether the prefecture has been opened to foreign capital, and 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 Round-Trip Diaspora Entrants
	(1)	(2)	(3)
Number of Diaspora Entrants	2.874*** (0.946)	1.767*** (0.493)	3.821*** (0.820)
Survival-Adjusted Number	2.511*** (0.903)	1.517*** (0.456)	3.253*** (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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, 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 four years (included). Open indicates whether the prefecture has been opened to foreign capital, and lineage connection is between surname and prefecture calculated from equation 1. In column (1), we use an 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 the prefecture level instead of the surname-prefecture level. In column (3), we include the potential round-trip diaspora entrants which are defined as foreign entrants represented by citizens of the 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*** (0.233)	6.001*** (2.476)	1.733*** (0.332)	1.534*** (0.440)
Survival-Adjusted Number	0.482*** (0.167)	5.152*** (2.186)	1.480*** (0.292)	1.308*** (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

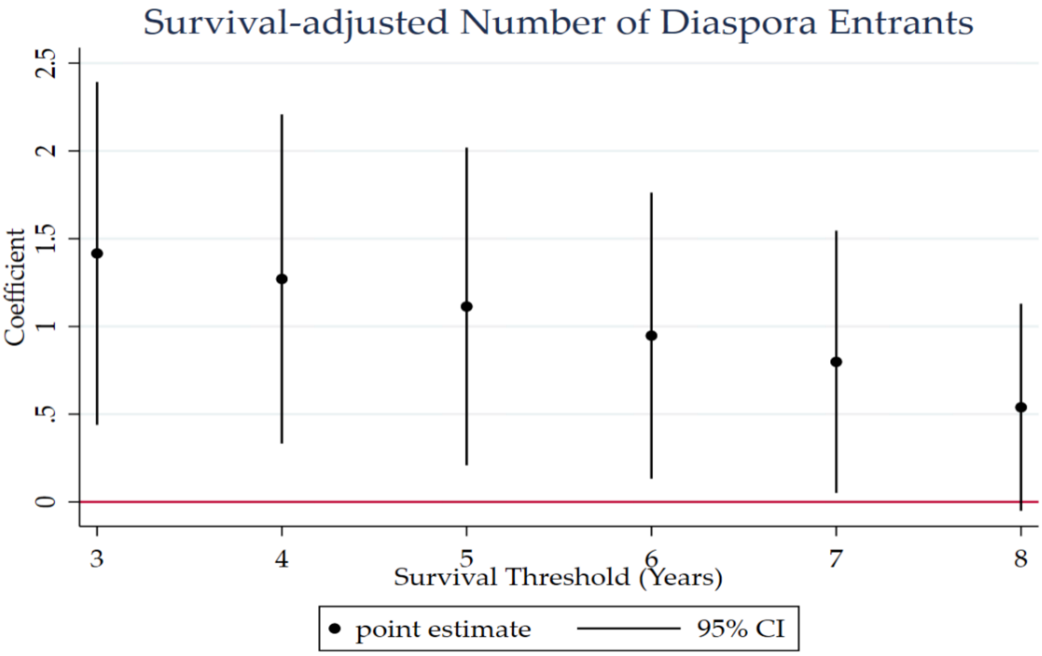
Note: ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Standard errors are clustered at the surname-prefecture level. Survival-adjusted Number of Diaspora Entrants is the number of entrants that survive for more than four years (included). Open indicates whether the prefecture has been opened to foreign capital, and lineage connection is between surname and prefecture, calculated from equation 1. Emigration-intensive provinces include Guangdong, Fujian, and 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 ranked in the top 20 among all prefectures during 1981 to 2014.

Table A.7: Poisson Regression

	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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, 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 four years (included). Open indicates whether the prefecture has been opened to foreign capital, and 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: Balance Test

	Predicted Number of Diaspora Firms in 1996	Number of Observed 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: ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. We regress the observed number of diaspora firms surviving 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 that our instrument is orthogonal to these prefecture characteristics. In column (2), the observed number of diaspora firms in 1996, shows stronger correlation with these prefecture characteristics in terms of more significant and larger coefficients and an R squared that is doubly larger.

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