Familiar Strangers: The Role of Diaspora Networks in Foreign Investment and Long-run Development*

Fanghao Chen, Jinan University, Guangzhou, China

Ruichi Xiong, Wuhan University, Wuhan, China

Xiaobo Zhang, Peking University & IFPRI, Beijing, China

Abstract

This paper characterizes the dominant role of diasporas in bringing in pioneering investment during the early stages of market formation, and examines whether they generate positive spillovers on subsequent local industrial development. Using a unique administrative dataset on the universe of foreign firms in China, we document that following China's opening up in 1979, initial foreign direct investment was largely driven by the Chinese diaspora, while massive non-diaspora foreign investment did not materialize until the late 1990s. Leveraging the staggered opening up of Chinese prefectures during 1981-96 and the variation in geographic distribution of surnames as an identification strategy, we show that diaspora firms tended to venture into newly opened prefectures which are more likely to be the ancestral origins of their legal representatives. Moreover, diaspora firms were more likely to be pioneering firms and the entry of diaspora firms stimulated subsequent entry of both non-diaspora foreign and domestic private enterprises.

Key Words: Diaspora Network, Market Formation, Foreign Investment, FDI Spillover, Long-run Development

JEL Codes: F21, F22, F23, O19

*Fanghao Chen: fhchen2017@nsd.pku.edu.cn, School of Economics at Jinan University. Ruichi Xiong: ruichixiong@gmail.com, Economics and Management School of Wuhan University. Xiaobo Zhang: x.zhang@gsm.pku.edu.cn, Guanghua School of Management at Peking University and the International Food Policy Research Institute. We thank Nathaniel Baum-Snow, April Franco, Kailin Gao, Ruixin Wang, Shuo Chen, Shilin Zheng, and Florian Mayneris as well as seminar participants at the University of Toronto, International School of Business and Finance at 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, European Winter Meetings of the Econometric Society, NEUDC at Yale University, Dongbei University of Finance and Economics, Lingnan College at Sun Yat-Sen University, Guangdong University of Foreign Studies, Shanghai University of Finance and Economics, Beijing Normal Univeristy at Zhuhai, and the Ninth International Symposium on Quantitative History for helpful comments and discussions. We thank Bohan Yang for excellent research assistance. All errors remain our own.

1 Introduction

Historical records and empirical analyses have documented that diasporas play an important role in international trade, especially during the early stages of market formation (Greif, 1993; Rauch, 2001; Rauch and Trindade, 2002; Pomeranz and Topik, 2017). Rigorous empirical studies examining the role of diasporas in bringing foreign direct investment (FDI) back to the countries of their origins, by comparison, are more limited. A notable exception is the study by Burchardi, Chaney and Hassan (2019) which finds a large positive effect of historical migration on contemporary FDI flows between U.S. counties and the migrants' origin countries. However, existing literature has not answered if diasporas also play a particularly important role in channeling back pioneering FDI when their origin countries are in the early stages of development with imperfect institutional environment.

Moreover, it is interesting to test whether diaspora investment fosters the long-run industrial development, particularly in developing countries, where FDI is badly needed. While it is widely acknowledged by theoretical literature that social affinity like diaspora networks could mitigate transnational frictions *statically* by enforcing contracts and providing information (Greif, 1993; Gould, 1994; Rauch and Casella, 2003), whether the positive effect of diaspora networks on economic development persists in the *long run* is debatable. On the one hand, diaspora networks enable international trade and investment to happen even in imperfect institutional environment, which may generate positive spillovers on local economy. On the other hand, relying too much on diaspora networks in doing business may inhibit the institutional transformation from personal trade to impersonal trade, which is key to maintaining greater scale of economic exchanges (North, 1990; Greif, 1993). The empirical literature has rarely examined the aforementioned questions, largely due to the lack of micro data spanning the whole development process.

To fill the gaps, we exploit the administrative business registration data maintained by the State Administration of Industry and Commerce of China (SAIC). This data covers the universe of foreign and domestic firms in China during 1980-2014, a period of rapid reform and opening up after 30 years of planned economy. 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 (Huang, 2003). A salient feature is that the initial FDI was primarily driven by the Chinese diaspora, from Hong Kong, Macao, Taiwan, and South Asian countries, and was concentrated in coastal regions that are known as the ancestral hometowns to many overseas Chinese (Vogel, 1990; Lever-Tracy, Ip and Tracy, 1996; Huang, Jin and Qian, 2013; Ye, 2014). The massive FDI from multinationals in developed countries, which were less related to the diaspora, did not come to fruition until the late 1990s.

Our empirical analyses proceed in several steps. First, we identify whether a foreign firm's legal representative is a member of the Chinese diaspora (i.e. overseas Chinese) based on their name and nationality ID, as indicated in the business registration database. We then define diaspora firms as foreign firms with overseas Chinese as their legal representatives. We find that diaspora firms, consistent with the qualitative records (Huang, 2003; Ye, 2014), accounted for a dominant share of total FDI in the early opening up period of China (see Figure 1).

Second, to quantify the effect of diaspora networks on the previously documented descriptive patterns, we measure the degree of lineage connections between diaspora firms and each prefecture as the probability of their legal representatives having the prefecture as their ancestral origin. This probability is inferred from the surnames of these legal representatives and the geographic distribution of surnames across Chinese prefectures constructed from the China Population Survey of 2005. Our construction of the lineage connection measure is motivated by the fact that the Chinese lineage sustains ties between members of a surname group and their ancestral hometowns (Szonyi, 2002), and the fact that this lineage culture is largely extended to overseas Chinese (Kuhn, 2008).

Finally, by leveraging the staggered opening up of Chinese prefectures from the 1980s to the early 1990s, we conduct an event study, showing that after a prefecture's opening up, diaspora firms were more likely to go to prefectures with stronger lineage connections. We also find, through an IV strategy, that prefectures that initially attracted more diaspora firms witnessed a greater number of foreign and domestic private firms decades later. The economic effect is sizable. The seed of a diaspora firm in 1996 at the prefecture level brought about an

additional 3.232 non-diaspora foreign firm, and about 162 domestic private firms survived as of 2014, providing 4,534 additional jobs by 2015. In conclusion, we find that diaspora networks with overseas Chinese spurred the entry of diaspora firms during China's early opening period, and the long-term benefits of these pioneering diaspora entrants in fostering industrial development outweighed the potential negative consequences of relying on informal institutions for conducting business.

We perform a series of checks to ensure that our findings are robust. First, we address concerns on our measure of lineage connections. We demonstrate that the geographic variation in surname distributions is stable across birth cohorts in the China Population Survey of 2005 and our results remain unchanged if using instead the older cohorts which were less likely to migrate prior to 2005 to construct our lineage connection measure. We then use permutation tests to show that potential errors in our lineage connection measure do not drive our results. Second, to deal with concerns on our outcome of interest, we use alternative indicator of diaspora firm entry considering the problem of zero entry, the survival of entry, and round-trip entrants. Third, we show that there are rich variation in the geographic distribution of surnames and our results are present in both the common and the rare surnames, and both the diaspora-intensive and the non-diaspora-intensive prefectures. Finally, we corroborate that the empirically identified sorting of diaspora firms into surname-prefecture pairs based on lineage connection could neither be accounted for by differences in the origin countries nor industry classifications of diaspora firms — factors commonly emphasized by the prior literature.

This paper also speaks to the literature on FDI spillovers.[1] In the context of China, this strand of literature primarily relies on the Chinese Annual Survey of Industrial Firms (the so called above-scale industrial firms), which did not start until 1998 (Lin, Liu and Zhang, 2009; Huang, Jin and Qian, 2013; Wang and Wang, 2015; Lu, Tao and Zhu, 2017; Ma, 2018). By 1998, China already opened up most cities to foreign investment. Our article complements these studies by using a new data source that covers all the firms regardless of their size and allows us to zoom in the early opening up period prior to 1998. We also discovered a new spillover channel associated with diaspora investment: diaspora firms were more likely to be pioneering firms that

[1] For thorough reviews of the literature on FDI spillovers, see Harrison and Rodríguez-Clare (2010) and Alfaro (2017).

entered new sectors and diversified the industrial composition of a prefecture. Moreover, the FDI literature focuses primarily on identifying spillovers on the productivity of incumbent domestic firms (Aitken and Harrison, 1999; Javorcik, 2004; Alfaro-Urena, Manelici and Vasquez, 2022), largely ignoring the potential benefits on the extensive firm entry margin, which is a focus of this article.[2]

Besides, our paper adds to the debate on the economic consequences of international emigration. One stream of this literature documents that how diasporas either finance the development of their home countries, or support the consumption of their family members at home through remittances (Giuliano and Ruiz-Arranz, 2009; Yang, 2011). Another line of research investigates both the negative impacts and the positive spillovers of the outflows of skilled labor, i.e. the "brain drains", on the sending countries (Agrawal et al., 2011; Docquier and Rapoport, 2012). Several recent studies also pointed out the unexpected spillovers of forced emigration, i.e. returned refugees, on the economic development of home countries (Bahar et al., 2022; Mayda et al., 2022; Bahar, Parsons and Vézina, 2022). We complement this literature by identifying the bridging role of the Chinese diaspora in directing foreign investment to their ancestral hometowns in China and the seeding effects of those early diaspora investment on the long-run local economic development.

Finally, our paper contributes to the emerging literature on the economics of names. Most studies in this field attempt to extract information embedded in names to explore issues including assimilation, discrimination, social mobility, and culture persistence, often in the context of Western Europe or North America (Bertrand and Mullainathan, 2004; Güell, Rodriguez Mora and Telmer, 2015; Clark, 2015; Abramitzky, Boustan and Eriksson, 2020; Bazzi, Fiszbein and Gerresilasse, 2020). Focusing on China, Bai and Kung (2022) quantify the importance of surname distances between Chinese prefectures in shaping the patterns of technology diffusion within China. Our paper proposes an alternative way to infer the ancestral origins of Chinese diaspora members from their surnames based on the Chinese lineage culture (Szonyi, 2002). This approach helps us overcome the common issue of insufficient data on the ancestral origins

[2]Markusen and Venables (1999) theoretically analyzes the relationship between the size of foreign and domestic sector in a country, incorporating both competition and sectoral linkage effects. Méndez and Van Patten (2022) empirically documents the long-run spillovers of the United Fruit Company as a large multinational on people's living standards in Costa Rica.

of returned diasporas in developing countries.

The findings of this paper have policy implications for other developing countries: in the absence of ideal market institutions, 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 FDI. As shown in Du, Lu and Tao (2008), U.S. multinationals have preferred to invest in regions with better quality of economic institutions since 1998, when China had launched the nationwide opening-up policy and diaspora FDI had demonstrated a success.

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

2 Historical Background

2.1 The Gradual Opening Up of China

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). Policy uncertainty remained high until late 1990s when China set to join the WTO.

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.[3] Then Yingkou prefecture, Weihai prefecture, and Hainan province were also added to the list of opening regions in 1985, 1987, and 1988, respectively. The political crisis in 1989 put a halt to the process of opening-up. Although Shanghai Pudong district became one of the SEZs in 1990, the turbulent policy environment in China was not stabilized until Deng Xiaoping visited 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 visualizes the gradual opening process of China. It 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*.[4] The number of foreign entrants spiked in 1992 and 1993, after 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 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 stable policy environment and solid market institution for attracting foreign investment according to the conventional wisdom.

Diaspora foreign firms made the major contribution to this achievement. The bottom panel in Figure 1 decomposes foreign entrants into diaspora and non-diaspora ones from 1980 to 2014, where diaspora firms are defined in Section 3. The share of diaspora entrants was over 90%

[3]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 in the US). The administrative unit closest to the size of a city is a prefecture in China (despite great variance in administered area), including those of a higher administrative status (such as Beijing, Shanghai, Chongqing, and Tianjin). Rural counties under the administration of a city are also included in the statistics for that city.

[4]Foreign firm entrants stagnated from 1989 to 1990 after the 1989 political crisis. The inclusion of Shanghai Pudong district in the list of SEZs in 1990, on the other hand, generated a surge in number of foreign firm entrants during 1990-1991.

from 1984 to 1994, highlighting the dominant role of diaspora firms in channeling FDI 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.

After diaspora firms had demonstrated success, other foreign and domestic private firms followed suit (Ye, 2014). Therefore, diaspora firms are integral part of the post-reform growth story in China in the sense that they were "pioneers" in both the foreign sector and the private sector, propelling China's transition both from isolationism to globalism and from a planned economy to a market economy.

2.2 Chinese Diaspora's Lineage Connections with Ancestral Hometowns

China has a long history of emigration dating back to as early as the 1300s (Kuhn, 2008). According to the UN Migration Database, the stock of overseas Chinese was above ten million in 2020, ranking in fourth globally in terms of origin countries.[5] Among the vast 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.[6]

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, thanks to the persistent lineage culture. In China, lineage refers to the group of descendants of one common patrilineal ancestor across multiple generations who share the *same surname* and typically reside in *the same or neighboring villages*. Over thousands of years, lineage organizations in China serve many practical functions, including settling village disputes, sheltering members from predatory taxation, and providing within-lineage public goods (Szonyi, 2002). The Chinese lineage culture is particular in sustaining strong ties between the Chinese diaspora and their ancestral homelands. Overseas Chinese would be considered perpetual members of their lineage and included in the genealogy books even if they emigrated, and so would their descendants. Therefore, many overseas Chinese still honor their ancestral hometowns and remain

[5] https://www.un.org/development/desa/pd/content/international-migrant-stock

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

socially connected to the group members of the their lineage (Kuhn, 2008). Anecdotes suggest that, even during China's closure, overseas Chinese kept sending remittances back to and exchanging letters with their relatives in ancestral hometowns (Tan, 2006).

After China reopened its doors in 1979 and strove to attract foreign investment, the persistent yet dormant networks between overseas Chinese and their ancestral land was rekindled.[7] 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, we present summary statistics of the key variables used in the empirical analyses.

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 classification code, 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

[7] 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.

surnames of a large representative sample with more than 2.5 million observations. Since the individuals in the data are randomly sampled at the prefecture level, the surname distributions drawn from the survey should be rather accurate at the prefecture level. We leverage the geographic variation in surname distributions 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. We identify foreign firms by the administrative ownership code provided by the SAIC.[8] Following this definition, our micro dataset of foreign firms closely matches the official aggregate FDI statistics, as shown in Figure A.6.

Having pinned down foreign firms, we proceed to identify diaspora firms controlled by overseas Chinese. Specifically, we define diaspora firms as foreign firms with overseas Chinese as their legal representatives. According to the corporate law in China, legal representatives take the major legal responsibility of the registered firms and are supposed to be one of the top ranked executives such as the chairman of the board or the general manager. Legal representatives therefore have been used as a proxy for entrepreneurs in the literature (Dai et al., 2019).

Our way to identify diaspora firms has two key advantages. First, the unique legal representative of a firm is always a natural person rather than a legal person. For a natural person, it is possible to infer whether she or he is overseas Chinese or not. For a legal person, on the other hand, it is not clear whether the owner of it has a Chinese ancestral origin or not.[9] Second, the names of legal representatives are readily available in the business registration database. We

[8] A firm is officially classified as "foreign" by the SAIC if its foreign ownership exceeds 25% based on SAIC's evaluation. The other approach in the literature uses immediate shareholder information. However, a firm's immediate shareholder structure does not necessarily coincide with its structure of ultimate ownership. For example, investors can exert control over a firm through holding shells (Bai et al., 2020). The official classification by the SAIC is therefore more reliable. Moreover, our sample of analysis includes only foreign firms with a legal representative who holds an ID from somewhere other than Mainland China. These firms are unlikely to be disguised foreign firms. Adopting instead the approach identifying foreign firms by immediate shareholder information would not change our conclusions. We also demonstrate the advantages of our foreign firm dataset by comparing it with other commonly used datasets on Chinese FDI in Appendix A.2.

[9] A potential alternative approach, other than ours, is to identify diaspora firms as foreign firms with overseas Chinese as their controlling natural-person shareholders. But 92% of the shareholders of foreign firms are foreign legal persons. To find out the ultimate natural-person owners of the foreign legal persons in China, we need to penetrate the ownership structure for the universe of firms in all countries, which is literally infeasible (Bai et al., 2020).

develop an algorithm as described in Appendix A.1 to extract the surnames of the legal representatives of foreign firms. We then label a legal representative as overseas Chinese if she or he has a Chinese surname[10] and holds an ID from somewhere other than Mainland China.[11] Diaspora firms therefore 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 Table A.1, using legal representatives to identify diaspora firms has the highest sample coverage than using other top ranked executives. 97% of all foreign firms have unique legal representatives, but the list of directors, chief executive officers, and top-ranked executives are often incomplete. Second, using registered personnel other than the legal representative to identify diaspora firms would be problematic, as 52% of foreign firms registered only one person (almost all of them are legal representatives) in the database. Finally, Table A.2 reports that the chance of a legal representative being a top ranked executive within a foreign firm is 93%. This means that the legal representative is likely to be the person in charge of the business.

Our classification of diaspora firms and empirical findings are not subject to the concern on round-trip foreign firms controlled by Chinese citizens from Mainland China (Huang, 2003).[12] We screen out the round-trip foreign firms with legal representatives as Chinese citizens. By restricting our definition of diaspora firms to be those with legal representatives holding an ID from places other than Mainland China, we effectively exclude them from our dataset of diaspora firms.[13] As shown in Table A.9, our results are robust to the exclusion of foreign firms from Hong Kong, Macao, Taiwan, and tax havens which are believed to be the main origins of round-trip FDI in China (Geng, 2019).

[10] In a few countries, such as Thailand and Indonesia, some Chinese diaspora abandoned their Chinese names in exchange for local names under the pressure of national assimilation policy. In this case, our method then would underestimate the scale of diaspora firms. However, as documented in Tan (2006), the Chinese diaspora often use their original Chinese names to signal their Chinese identity when dealing with Chinese businesses. So the possibility of name assimilation would not systematically affect our classification of diaspora firms.

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

[12] Chen (2022, p. 393-98) 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." Dual citizenship is not allowed by the law of China.

[13] 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. As a robustness check, we include this group of foreign firms as diaspora firms in section 4.3. As expected, the effect becomes even stronger. We therefore adopt the more conservative definition of diaspora firms and manage to make sure that our main conclusions are not driven by the disguised round-trip foreign firms.

3.3 Measurement of Lineage Connections

To test whether the international lineage networks have an effect on diaspora firms' location choice, we need to know the diaspora's ancestral hometowns. However, such information is normally unavailable, in particular in developing countries. To overcome the challenge, we instead use the geographic distribution of surnames to infer the ancestral origins of the Chinese diaspora by making use of a defining feature of the Chinese lineage culture, i.e. the distribution of surnames are often location specific and the pattern persists over time (Szonyi, 2002).

Figure 2 visualizes the geographic distributions of the 20 most populous surnames in China, which are ranked based on the China Population Survey of 2005, from three different 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")[14] 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 that the geographic distribution of surnames could be used to reveal the bond between overseas Chinese and their ancestral hometown, i.e. the lineage connections.

Therefore for a given surname s of a diaspora firm's legal representative, we measure its lineage connection with a particular prefecture p by taking advantage of the surname's geographic distribution across prefectures on the basis of the China Population Survey of 2005 as follows:[15]

[14] 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.

[15] We chose prefecture as the geographic unit of our empirical analysis because the sample was randomly drawn at the prefecture level in the China Population Survey of 2005.

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

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 has several merits. First, it can be intuitively interpreted as the probability of one overseas Chinese with surname s having prefecture p as their ancestral origin. Second, our measure does not depend on the population size because it is normalized by the total population with the same surname for China as a whole.[16] Third, as revealed in Figure 3, our lineage connection measure at the surname-prefecture pair level has large variations, and could capture well the location preference of diaspora firms for their legal representatives' ancestral hometowns. Last, as demonstrated in Table A.6, the geographic variation in surname distributions is stable across birth cohorts in the China Population Survey of 2005. Bai and Kung (2022) show that the surname distribution across prefectures in the China Population Survey of 2005 matches well with with the surname distribution in Harvard's China Biographical Database which is computed based on data from earlier years. Therefore, although our measure of lineage connection is based on data in 2005 after the initial reform period, it can still largely capture the degree of historical lineage connection at the surname-prefecture level.

3.4 Summary Statistics

We restrict our attention to the early opening-up period from 1981 to 1996 before the Asian financial 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

[16] 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.

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.[17] Merging datasets, we arrive at our main sample consisting of 284 surnames in 296 prefectures spanning 16 years.

Table 2 presents the summary statistics for all the variables used in the empirical analyses. Panel A shows that only 2.1% of the surname-prefecture-year cells have at least one diaspora entrant. The average number of diaspora entrants is 0.06. The number becomes slightly smaller once we adjust entry by survival. The lineage connection measure varies widely, with a mean of 0.002 and a standard deviation of 0.008, as indicated in panel B. Panels C summarize the main outcomes of interest at the prefecture level.

4 Lineage Connection and Diaspora Firm Entry

4.1 Identification Strategy

We employ the following baseline specification to empirically identify 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).[18] In addition, we use the registered capital as an outcome measure of the

[17] The four autonomous regions and four mega cities host 1.2% and 9.6% of the total foreign firm entrants respectively during 1981-96. Our results are robust to including them back in the sample.

[18] Our baseline results are insensitive to the choice of survival year thresholds as shown in Figure A.7 in the

intensive margin in Table A.3.

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 effect 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 the status of "opening city" since year t, and equals zero otherwise.[19] m_{sp} is 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. To empirically test the parallel trend assumption, we employ the following event-study framework:

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

Appendix.

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

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 all $\tau < 0$.

As shown in the top two plots in Figure 4, 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

The baseline results based on equation 2 are reported in columns (1) and (2) of 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 $\times$ Lineage Connection" variable remains highly positive and significant, suggesting that the lineage connection facilitates the entry of diaspora firms even after entry quality is taken into account.

To examine the time-varying effect, the top two panels in Figure 4 plot the event study estimates based on equation 3 for the same two outcome variables as in Table 3. Interestingly, the effect is highly statistically positive only in the first two years after opening up and then diminishes.

Following De Chaisemartin and D'Haultfoeuille (2020), we check whether our baseline results are still robust after correcting for potential biases as pointed in the literature. 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. As a second best, We divide the surname-prefecture pairs with continuous lineage connection into two discrete

groups: those above or below the median lineage connection and use a dummy variable corresponding to the two groups as a proxy for the measure of lineage connection.

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 4, there are no parallel trends prior to a prefecture's opening up, further validating our identification strategy.

As shown in the bottom two plots of Figure 4, the effect of lineage connection based on the new method lasts beyond the first two years. This is likely because the traditional staggered DID estimations fail to capture the spillovers from the early diaspora entrants on later ones, after all the prefectures opened to foreign investment.[20] 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 in operation as of 2014. The survival ratio is defined the share of entrants that were still alive as of 2014 conditional on the entry of diaspora firms. 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 granular 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

[20] The persistent effects of lineage connection do not appear if we also use the connection dummy in our standard DID estimation without correcting for the potential bias of staggered DID.

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.3 show that the effect of lineage connections still holds, even after considering the survival of entrants in four years or in 2014. On the other hand, if diaspora entrants are particularly small, the effect of lineage connection on firm entry could be less important. We deal with this problem in column (4) of Table A.3 and still find a large positive effect on the size of survived registered capital of diaspora firms as of 2014.

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 same as in columns (3) and (4) of Table 3. As displayed in Table A.4, the results of the Poisson regressions are reassuringly robust.

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.

Alternative lineage connection measure. There is a concern that our measure of lineage connection constructed from the China Population Survey of 2005 is likely subject to domestic migration prior to 2005. 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). At the time of population survey in 2005, the likelihood of migration for older cohorts was slim.

In Table A.6, we compute the pairwise correlations between our lineage connection measures calculated from different birth cohorts in the China Population Survey of 2005. It is apparent from the table that the distribution of our lineage connection measures is very stable across birth cohorts, echoing the findings of Clark (2015) and Bai and Kung (2022). 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) find that the geographic distribution of surnames across prefectures inferred from the China Population Survey of 2005 closely mirror the surname distribution in Harvard's China Biographical Database.

As a robustness check, we repeat our baseline exercise by using only older cohorts to measure lineage connection, who were less likely to migrate prior to 2005 than their younger counterparts. Columns (1) to (3) in Table A.7 report almost identical estimates with the lineage connection measure constructed from each of the three birth cohorts observed in the China Population Survey of 2005: the ones born before 1945, the ones born before 1949, and the one born before 1960.

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 5 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, some 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.8. 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 diaspora-intensive surnames. The coefficient for the key variable remains significantly positive. Column (4) excludes FDI-intensive prefectures. The results barely change. In columns (5) and (6), we turn to show that it is neither the most common surnames nor the rare surnames driving the significance of our estimates.

Heterogeneity by origin and industry. By comparing surname-prefectures based on their lineage connections, our novel setting allows us to abstract away from the issues of selection on origins, destinations and industries, which are commonly encountered in the FDI literature (Lu, Tao and Zhu, 2017; Burchardi, Chaney and Hassan, 2019). A shortcoming of our approach is that potential heterogeneity in the effects of lineage connection is masked. To remedy this concern, we show the heterogeneous effects of lineage connection by origin in Table A.9 and by industry in Table A.10. In terms of origin heterogeneity, the effects of lineage connection are concentrated in attracting diaspora foreign firms from Hong Kong, Macao, Taiwan and South East Asia. This is not a surprising finding since these places are known for hosting the largest number of the Chinese diaspora. In terms of industry heterogeneity, the effects of lineage connection are mainly in the manufacturing sectors, consistent with the manufacturing-driven growth patterns in China in the reform and opening up period.

In addition, these heterogeneity analyses indicate that our baseline results could not be accounted for by the origin or industry differences across surname-prefecture pairs since the comparisons we perform here are essentially *within* an origin group in Table A.9 or an industry group

in Table A.10.

5 Seeding Effects in the Long Run

Having 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 further investigate whether the industrial "seeds" planted by the early diaspora firms have had lasting spillover effects on the subsequent entry of non-diaspora foreign firms and domestic private firms. The fact that the extensive margin of private sector is the key engine of Chinese economic growth motivates our exercise (Wei and Zhang, 2011). 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 at the prefecture level; (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.

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.[21] Figure 6 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 number of unfilled 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

[21] 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 the government, not driven by market forces.

firms has closed.

These patterns indicate that the diaspora played an important role in setting up pioneering firms in the early opening-up period when market institutions were in the infant stage. As market institutions were developed, the special role played by the diaspora diminished.

5.2 Estimation of the Long-run Seeding Multipliers

Hausmann and Rodrik (2003) reckon that pioneering firms likely generate a positive externality for subsequent firm entrants in developing countries. Yet, given that most pioneering firms were set up by diasporas, there is a concern about the club nature of affinity group in inhibiting institutional development (North, 1990; Greif, 1993), which in turn may slow the entry of non-diaspora firms. Thereby, it is an empirical question to test whether the externality is positive or negative. In this section we estimate the multipliers of early diaspora firms on subsequent entrants.

Following the methodology in Sequeira, Nunn and Qian (2020), we first isolate the lineage-driven diaspora entrants 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 diaspora entrants predicted by the "zero stage," $\hat{\beta}$, is then used to construct an instrument for the observed diaspora firm stocks in 1996. With the estimated coefficient, we obtain the predicted value of $\hat{\beta} \times Open_{pt} \times m_{sp}$—the lineage-driven diaspora entrants 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 diaspora firms for each prefecture in 1996 as follows:

$$\text{Diaspora } \hat{\text{F}}\text{irm Stock}_p = \sum_s \sum_{t=1981}^{1996} \hat{\beta} \times Open_{pt} \times m_{sp} \tag{4}$$

Next, we use the predicted lineage-driven diaspora firm stocks as an IV for the observed diaspora firm stocks in 1996 in the following two-stage least squares (2SLS) regression:

$$\pi_p = \alpha + \gamma \times \text{Diaspora Firm Stock}_p + \lambda \times X_p + \epsilon_p \tag{5}$$

where π_p is our long-term outcome of interest, Diaspora Firm Stock$_p$ is the observed number of diaspora firms in prefecture p in 1996, and X_p denotes a set of controls at the prefecture level. Our IV strategy therefore relies on an exclusion restriction: the aggregated heterogeneous effects of opening up on diaspora firm entry across surname-location pairs during 1980-96 at the prefecture level only affects our long run outcome of interest through changing the observed prefecture-level diaspora firm stocks in 1996, conditional on the set of controls included in X_p. The potential violations of this restriction could come from the correlation between omitted prefecture characteristics and the observed diaspora firm stocks in 1996.

To test the validity of the IV, we perform a balance test as shown in Table A.11. 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 number of years since opening up, 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 highly 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, 2SLS, and first-stage regressions in three 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 as of 2014. In addition to the control variables in the balance test, we also include province fixed effects. As shown in Table A.12, 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 likely because historically less arable places pushed more people out for a living and therefore had more diaspora resources when China opened its gate to foreign firms in 1980s as shown in Table A.11.

According to the IV estimates shown in columns (1) and (2) in panel C of Table 5, one additional diaspora firm in 1996 brings about an increase of 3.232 subsequent non-diaspora foreign firms and 162 domestic private firms. The positive effect is observed not only for the number of firm stocks, but also for the total registered capital of these firms in 2014. One

more diaspora firm in 1996 leads to an increase of 0.396 percent in total registered capital for foreign enterprises and 0.252 percent for domestic enterprises that survive as of 2014, as indicated in column (3) and (4) in Table 5.[22] The emerging private enterprises led to greater local employment, consistent with the findings in the literature (Wei and Zhang, 2011; Imbert et al., 2022). Column (5) in Table 5 quantifies that one more diaspora firm in 1996 generated 4,534 more jobs by 2015 for an average prefecture.

6 Conclusions

The inflow of massive amount of diaspora investment in the early opening up period is a salient feature of China's growth story. However, the destinations of diaspora firms were not random. Our paper has shown that areas with stronger lineage connections attracted more diaspora firm entrants during the early opening up period when market institutions were far from perfect. Moreover, we found that early diaspora firm entrants, many of which were pioneering firms, facilitated the subsequent entry of non-diaspora foreign and domestic private enterprises.

China's experience in tapping diaspora resources in the early stage of development may shed light on other developing countries. Plagued with poor human capital (Noorbakhsh, Paloni and Youssef, 2001), low institutional quality (Alfaro, Kalemli-Ozcan and Volosovych, 2008), high political risk (Julio and Yook, 2012), and a combination of these factors (World Bank, 2017), developing countries often struggle to attract FDI, in particular pioneering foreign firms. If a developing country, like Peru, could manage to improve its human capital and institutional quality to the levels of Australia, it is already a developed country. However, it is a long journey to improve human capital and build sound institutions, which are the two fundamental factors. As shown in China, a more practical solution is to tap the diaspora networks to direct foreign investment into developing countries in the early stage of development despite lacking of the fundamental factors.

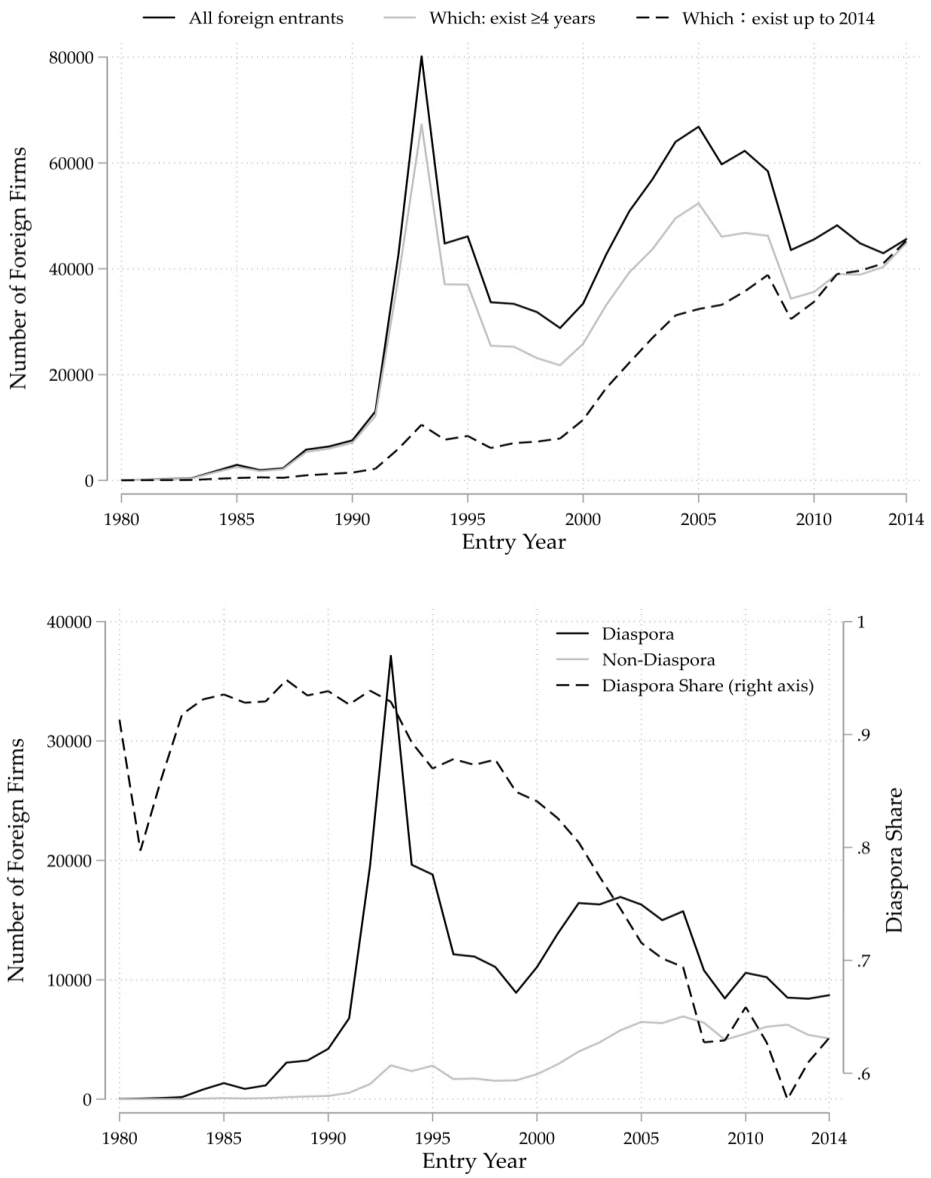
A survey in 2005 by the International Organization for Migration found that more than 90% of the countries had policies or programs targeting their diaspora.[23] Diaspora firm en-

[22] Our results are robust to using total registered capital without log as outcome variable.

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

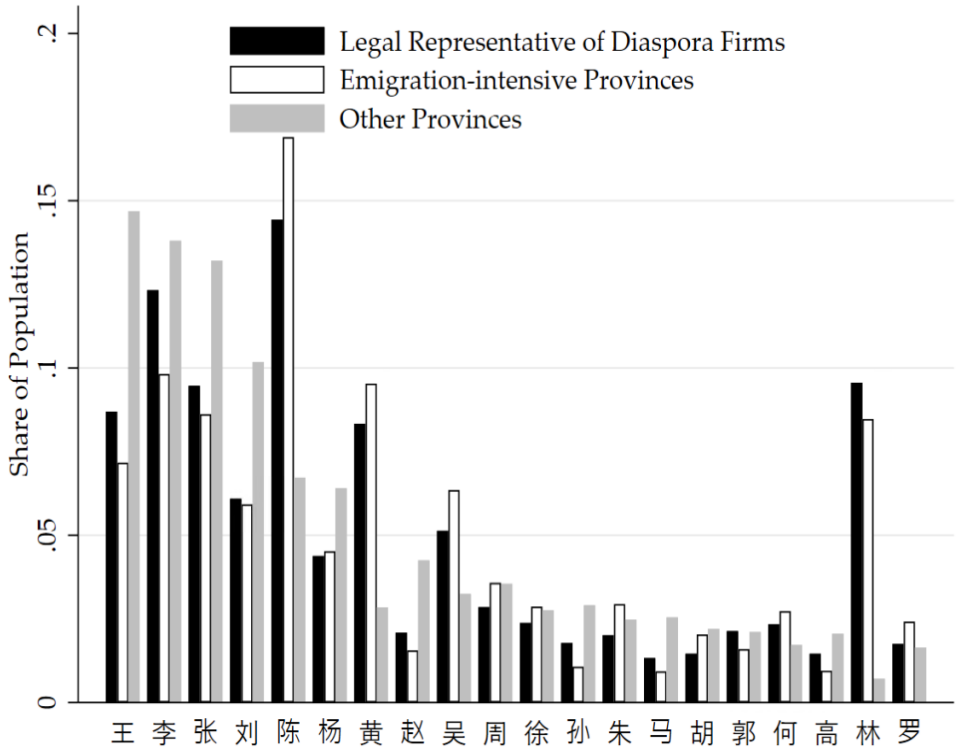
try accounts for a significant proportion of the FDI in other developing countries as well, for example 25% in Armenia during 1998-2014 (Riddle, Hrivnak and Nielsen, 2010) and 26% in India during 1991-2001 (Wei and Balasubramanyam, 2006). Collier, Gregory and Ragoussis (2019) call for more active policies 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 the uncertain and even hostile environment in those states. A promising future research topic is to explore whether diaspora firm entrants can serve as a touching stone for non-diaspora foreign and domestic private enterprises to enter even in these countries.

Figure 1: Entry and Survival of Foreign Firms in China



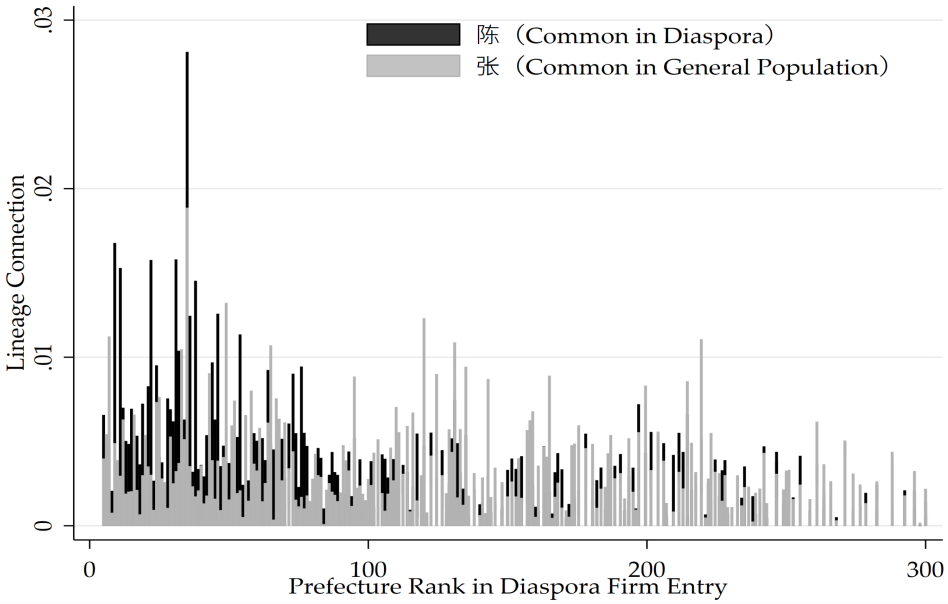
Note: This figure uses firm registration data from the SAIC database. The top panel plots the number of foreign firms (and surviving foreign firms) by year of entry over 1980-2014. The bottom panel displays the number of diaspora and non-diaspora entrants on the left axis as well as the share of diaspora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

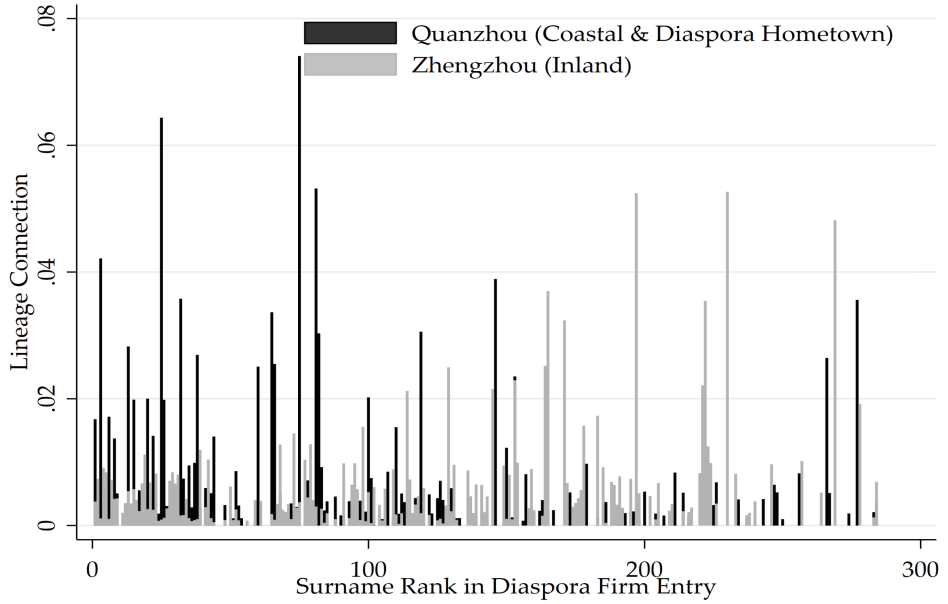
Figure 2: Surname Distributions in Different Samples



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.

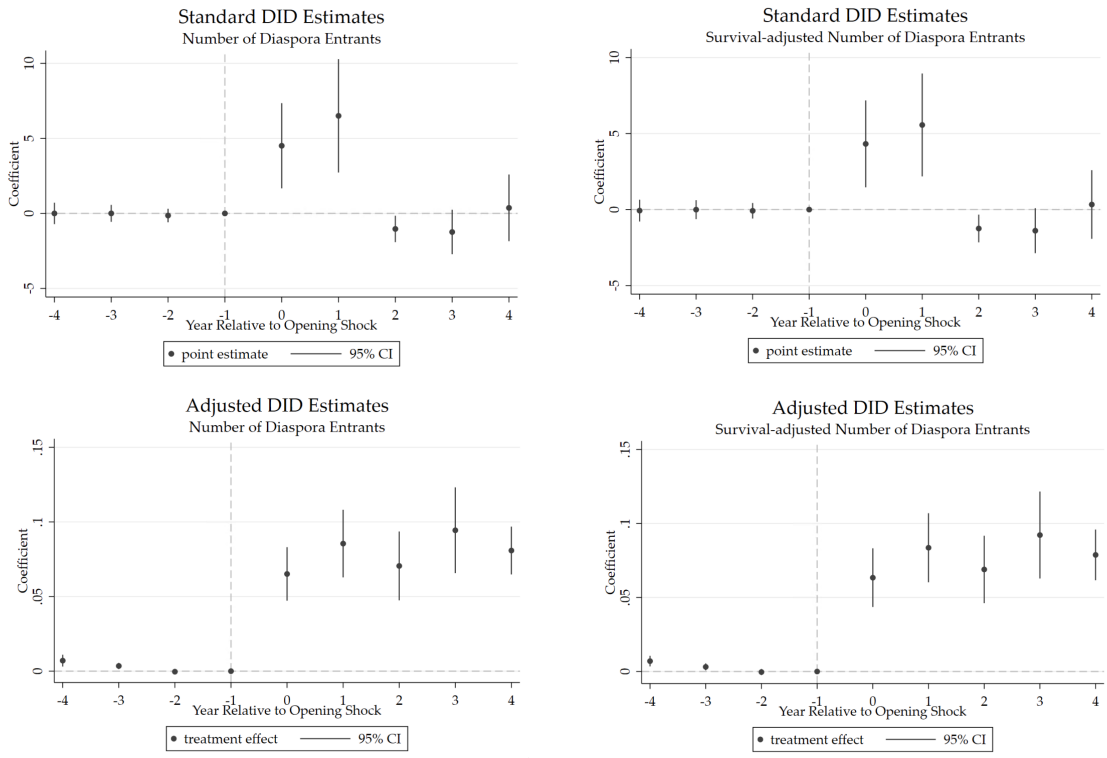
Figure 3: Lineage Connection across Surnames and Prefectures





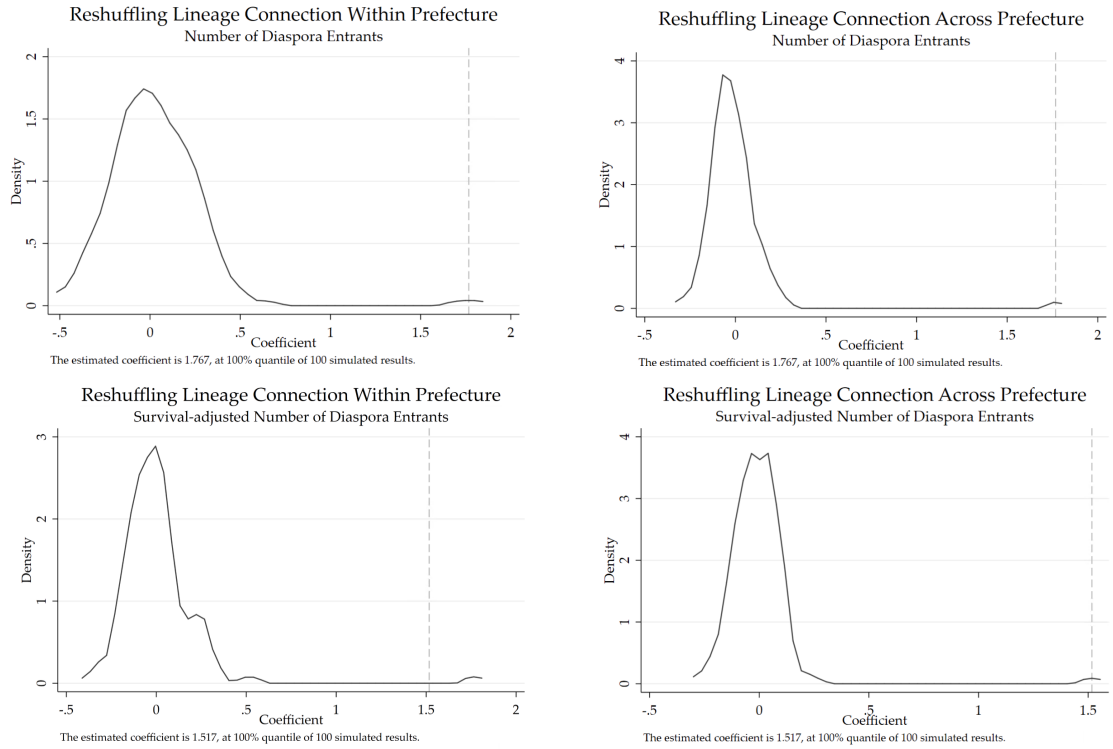
Note: This figure visualizes the variation of our measure of lineage connection using selected surnames and prefectures. The top panel compares the distribution of the measure for "陈", a common surname among Chinese diaspora, and "张", a common surname in the general population, in relation to the prefecture's rank in diaspora firm entry. The bottom panel compares the distribution of the measure for Quanzhou, a coastal city and a well known ancestral hometown to many Chinese diaspora, and Zhengzhou, a large inland city with no record of overseas connection, in relation to a surname's rank in diaspora firm entry.

Figure 4: Event Study



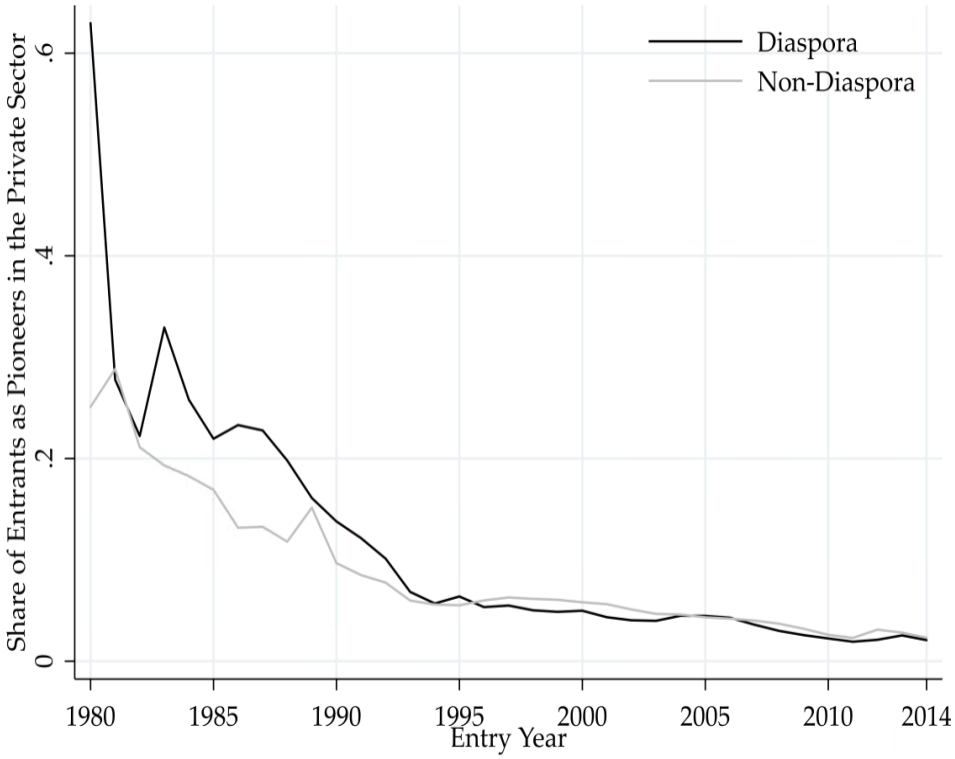
Note: This figure plots the coefficients β^τ obtained from a standard event study specification: $Y_{spt} = \eta_{sp} + \theta_{st} + \delta_{pt} + \sum_{\tau=-4}^{4} \beta^\tau \times Open_{pt}^\tau \times m_{sp} + \lambda \times S_{spt} + \epsilon_{spt}$ where $Open_{pt}^\tau$ indicates opening status 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 diaspora firm stocks 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 5: 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 6: Shares of Pioneering Firms Among Diaspora and Non-Diaspora Entrants



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 4-digit industry in a prefecture.

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 Diaspora Entrant	1,345,024	0.021	0.142
Number of Diaspora Entrants	1,345,024	0.060	1.437
Survival-Adjusted Number of Diaspora Entrants	1,345,024	0.052	1.234
Number of Surviving Diaspora Entrants in 2014	1,345,024	0.009	0.181
Surviving Registered Capital of Diaspora Entrants in 2014 (10^4 CNY)	1,345,024	27.017	940.778
Survival Ratio of Diaspora Entrants in 2014	27,846	0.168	0.338
Average Registered Capital per Surviving Diaspora Entrant in 2014(10^4 CNY)	27,846	806.588	3,967.006
Panel B: Surname-Prefecture			
Lineage Connection (All Cohort)	84,064	0.002	0.008
Lineage Connection (Prior to 1945)	84,064	0.002	0.008
Lineage Connection (Prior to 1949)	84,064	0.002	0.008
Lineage Connection (Prior to 1960)	84,064	0.002	0.008
Panel C: Prefecture			
Diaspora Firm Stocks in 1996 (10^4)	332	0.032	0.114
Non-Diaspora Foreign Firm Stocks in 2014 (10^4)	332	0.101	0.329
Domestic Private Firm Stocks in 2014 (10^4)	332	5.658	11.605
Non-Diaspora Foreign Registered Capital in 2014 (10^4 CNY)	332	226.390	744.190
Domestic Private Registered Capital in 2014 (10^4 CNY)	332	8,244.674	101,000
Employment in 2015 (10^4)	278	264.084	213.128

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 from equation 1, based on surnames of different birth cohorts in the China Population Survey of 2005 at the surname-prefecture level. Panel C reports summary statistics for the variables at the prefecture level. The first five variables were constructed from the SAIC database while the data of employment sizes in 2015 are from the China Population Survey of 2015.

Table 3: Lineage Connection and Entry of Diaspora Firms

	Standard DID		Adjusted DID	
	Number of Diaspora Entrants	Survival-Adjusted Number of Diaspora Entrants	Number of Diaspora Entrants	Survival-Adjusted Number of Diaspora Entrants
	(1)	(2)	(3)	(4)
Mean of Dep. Var.	0.060	0.052	0.060	0.052
Open × Lineage Connection	1.767***	1.517***		
	(0.574)	(0.527)		
Open × High Connection Dummy			0.079***	0.077***
			(0.007)	(0.010)
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: 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 Connection 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: Testing Sorting on Entry Quality

	Entry Quality	
	Survival Ratio of Diaspora Entrants in 2014	Log Average Registered Capital per Surviving Diaspora Entrant in 2014
	(1)	(2)
Mean of Dep. Var.	0.168	6.693
Open × Lineage Connection	-0.333	9.291
	(0.825)	(6.529)
N	20,633	20,633
Number of Incumbent Firms	Y	Y
Surname-Prefecture FE	Y	Y
Surname-Year FE	Y	Y
Prefecture-Year FE	Y	Y

Note: 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 of Diaspora Entrants in 2014” is defined as the share of entrants that were surviving as of 2014 conditional on entry of diaspora firms. “Log Average Registered Capital per Surviving Diaspora Entrant in 2014” refers to the natural logarithm of the average registered capital for the diaspora firms that were still active as of 2014. We deal with log zero by inverse hyperbolic transformation in column (2).

Table 5: Seeding Effects of Diaspora Firms

	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	Local Employment in 2015 (10^4)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
1996 Diaspora Firm Stocks	1.644***	64.651***	0.046%***	0.043%***	0.104***
	(0.298)	(9.616)	(0.000)	(0.000)	(0.018)
Panel B: 2SLS					
1996 Diaspora Firm Stocks	3.232***	161.574***	0.396%***	0.252%***	0.453***
	(1.154)	(61.344)	(0.000)	(0.001)	(0.1670)
Panel C: First Stage					
	Dependent Variable: 1996 Diaspora Firm Stocks				
Predicted Diaspora Firm Stocks			46.265***		
			(16.419)		
N	231	231	231	231	231
F statistics	58.940	58.940	58.940	58.940	58.940
Controls	Y	Y	Y	Y	Y
Province Fixed Effects	Y	Y	Y	Y	Y

Note: 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 shows the 2SLS estimates using the predicted diaspora firm stocks in 1996 as IV for the observed diaspora firm stocks in 1996. Panel C reports the first stage of the 2SLS estimation. Controls include province fixed effects, number of years since opening up, 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. We deal with log zero by inverse hyperbolic transformation.

References

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2020. "Do Immigrants Assimilate More Slowly Today than in the Past?" *American Economic Review: Insights*, 2(1): 125–141.

Agrawal, Ajay, Devesh Kapur, John McHale, and Alexander Oettl. 2011. "Brain Drain or Brain Bank? The Impact of Skilled Emigration on Poor-Country Innovation." *Journal of Urban Economics*, 69: 43–55.

Aitken, Brian J., and Ann E. Harrison. 1999. "Do Domestic Firms Benefit from Direct Foreign Investment? Evidence from Venezuela." *American Economic Review*, 89(3): 605–618.

Alfaro, Laura. 2017. "Gains from Foreign Direct Investment: Macro and Micro Approaches." *World Economic Review*, 30: S2–S15.

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.

Alfaro-Urena, Alonso, Isabela Manelici, and Jose P. Vasquez. 2022. "The Effects of Joining Multinational Supply Chains: New Evidence from Firm-to-Firm Linkages." *Quarterly Journal of Economics*, 137(3): 1495–1552.

Bahar, Dany, Andreas Hauptmann, Cem Özgüzel, and Hillel Rapoport. 2022. "Migration and Knowledge Diffusion: The Effect of Returning Refugees on Export Performance in the Former Yugoslavia." *Review of Economics and Statistics*, Forthcoming.

Bahar, Dany, Christopher Parsons, and Pierre-Louis Vézina. 2022. "Refugees, Trade, and FDI." *Oxford Review of Economic Policy*, 38(3): 487–513.

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.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gerresilasse. 2020. "Frontier Culture: The Root and Persistence of "Rugged Individualism" in the United States." *Econometrica*, 88(6): 2329–2368.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

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.

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, 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.

Docquier, Frédéric, and Hillel Rapoport. 2012. "Globalization, Brain Drain, and Development." *Journal of Economic Perspectives*, 50(3): 681–730.

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.

Giuliano, Paola, and Marta Ruiz-Arranz. 2009. "Remittances, Financial Development, and Growth." *Journal of Development Economics*, 90: 144–152.

Gould, David M. 1994. "Immigrant Links to the Home Country: Empirical Implications for U.S. Bilateral Trade Flows." *Review of Economics and Statistics*, 76(2): 302–316.

Greif, Avner. 1993. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *American Economic Review*, 83(3): 525–548.

Güell, Maia, Jose V. Rodriguez Mora, and Christopher I. Telmer. 2015. "The Informational Content of Surnames, the Evolution of Intergenerational Mobility, and Assortative Mating." *Review of Economic Studies*, 82: 693–735.

Harrison, Ann, and Andrés Rodríguez-Clare. 2010. "Trade, Foreign Investment, and Industrial Policy for Developing Countries." *Handbook of Development Economic*, 5: 4039–4214.

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.

Imbert, Clement, Marlon Seror, Yifan Zhang, and Yanos Zylberberg. 2022. "Migrants and Firms: Evidence from China." *American Economic Review*, 112(6): 1885–1914.

Javorcik, Beata S. 2004. "Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers through Backward Linkages." *American Economic Review*, 94(3): 605–627.

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.

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.

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

Markusen, James R., and Anthony J. Venables. 1999. "Foreign Direct Investment as a Catalyst for Industrial Development." *European Economic Review*, 43: 335–356.

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.

Mayda, Anna-Maria, Christopher Parsons, Han Pham, and Pierre-Louis Vézina. 2022. "Refugees and Foreign Direct Investment: Quasi-Experimental Evidence from US Resettlements." *Journal of Development Economics*, 156: 102818.

Méndez, Esteban, and Diana Van Patten. 2022. "Multinationals, Monopsony, and Local Development: Evidence from the United Fruit Company." *Econometrica*, 90(6): 2685–2721.

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.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance.* Cambridge University Press.

Pomeranz, Kenneth, and Steven Topik. 2017. *The World that Trade Created: Society, Culture, and the World Economy, 1400 to the Present.* Routledge.

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

Rauch, James E, and Alessandra Casella. 2003. "Overcoming Informational Barriers to International Resource Allocation: Prices and Ties." *Economic Journal*, 113(484): 21–42.

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.

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.

Wang, Jian, and Xiao Wang. 2015. "Benefits of Foreign Ownership: Evidence from Foreign Direct Investment in China." *Journal of International Economics*, 97: 325–338.

Wei, Shang-Jin, and Xiaobo Zhang. 2011. "Sex Ratios, Entrepreneurship, and Economic Growth in the People's Republic of China." NBER Working Paper.

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.

Yang, Dean. 2011. "Migrant Remittances." *Journal of Economic Perspectives*, 25(3): 129–152.

Ye, Min. 2014. *Diasporas and Foreign Direct Investment in China and India.* Cambridge University Press.

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

Appendix for Online Publication

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.

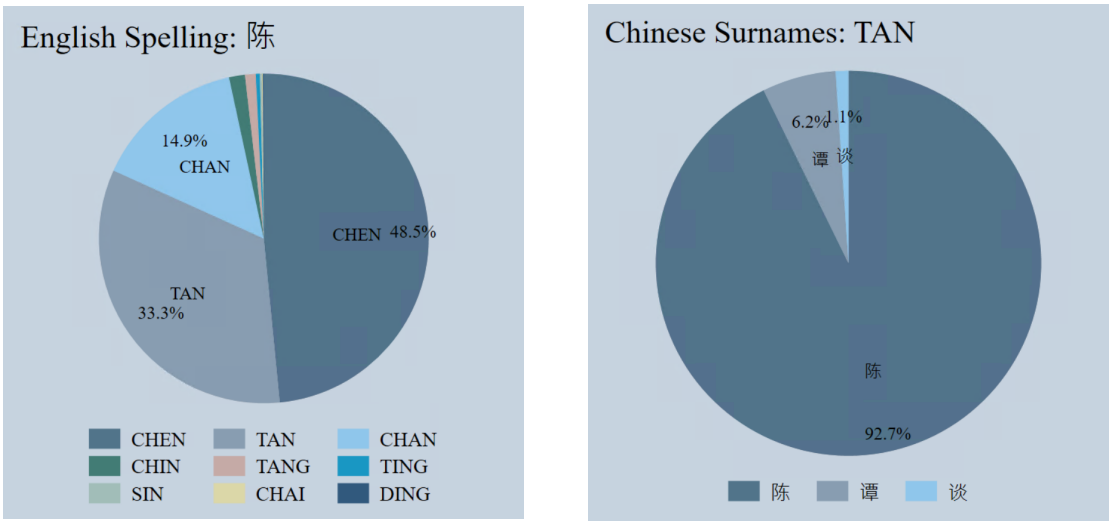
Before formally executing our algorithm, we perform a preparatory cleaning procedure to leave out all symbols (such as comma, period, or semicolon) and word content that is unrelated to names (such as titles,"Mr.," "Dr.," or "appointed by the parent company"), since many name entries are unstructured 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.

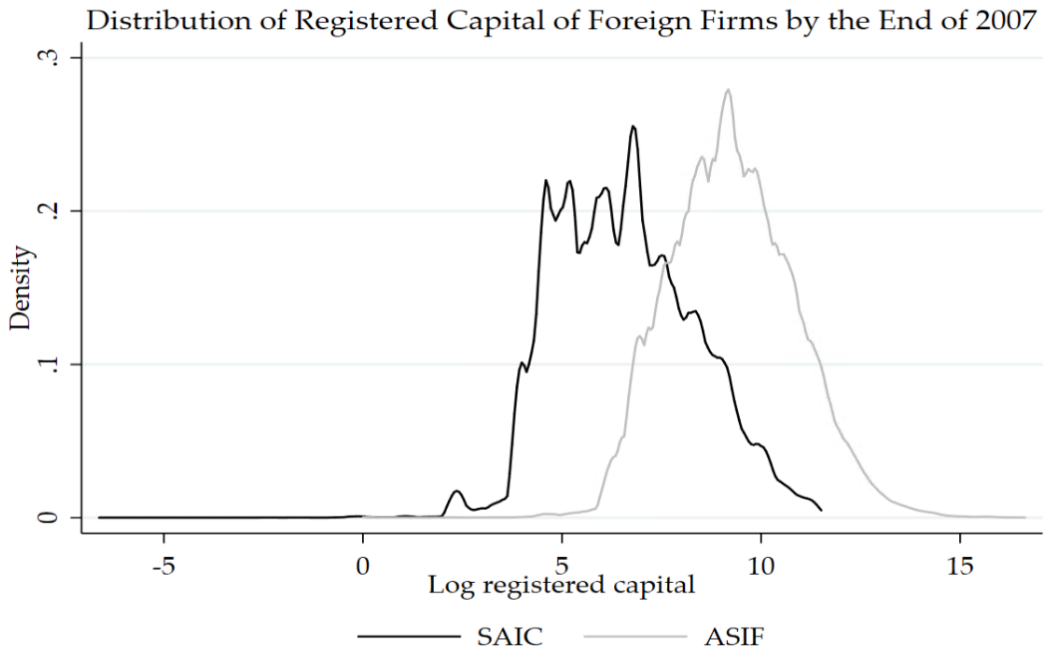
5. This procedure produces not only the surname for each ethnic Chinese, but also an English-Chinese mapping that can be used for randomly assigning English spellings into Chinese characters with empirical weights bootstrapped from the sample of overseas Chinese.

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

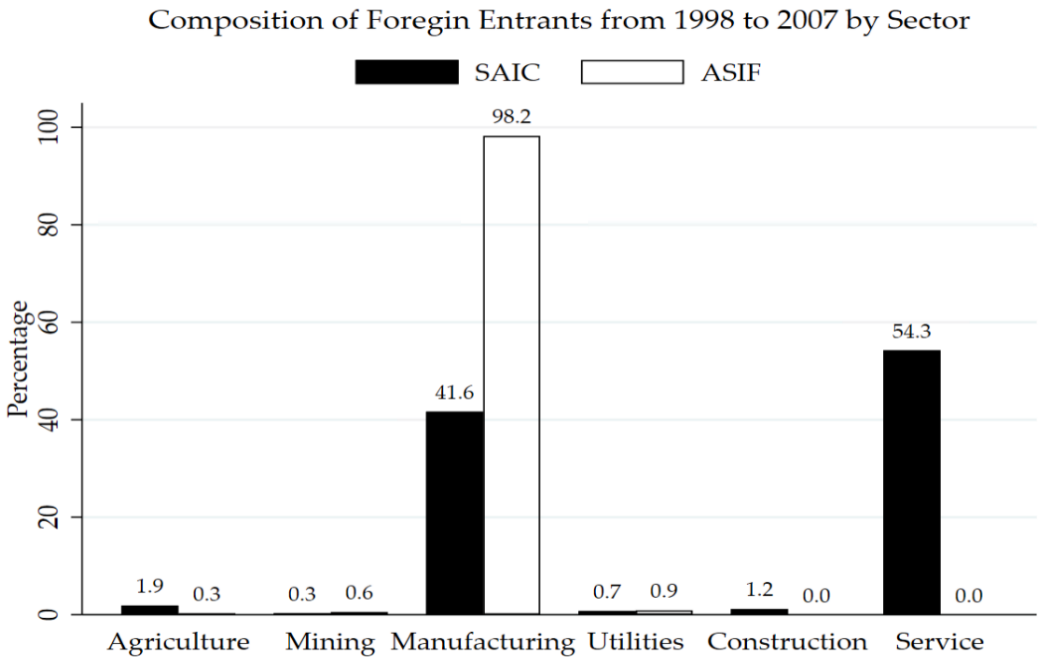


Note: In this figure, we plot the distribution of the registered capital of foreign firms observed in the SAIC database and ASIF respectively. We can immediately see that the ASIF firms are much larger than the firms in the SAIC database in terms of registered capital. Therefore using ASIF firms to study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

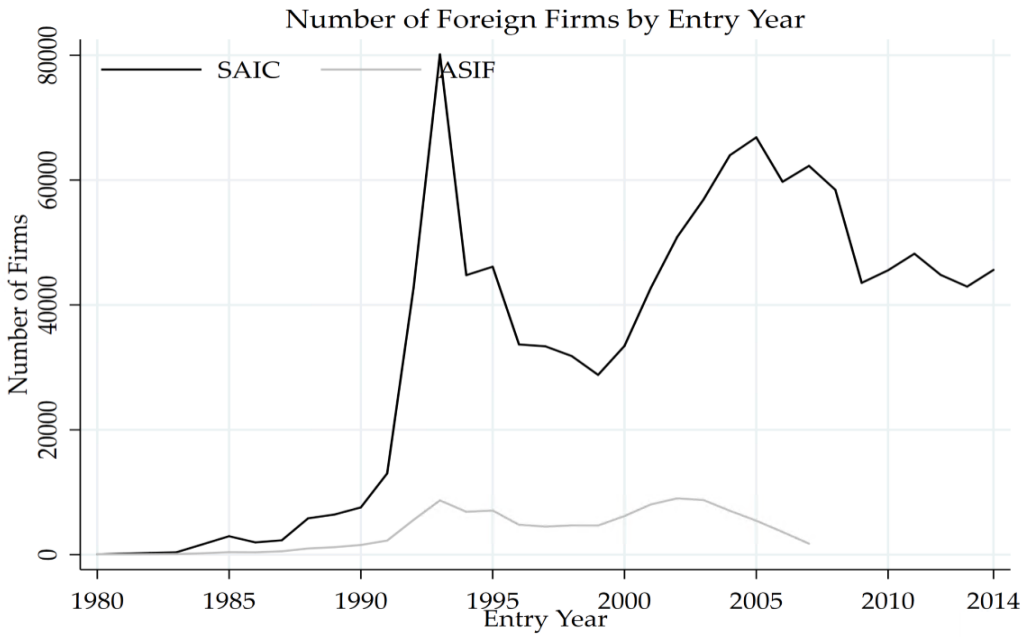


Note: In this figure, we plot the compositions of foreign firms in the SAIC database and the ASIF data. As expected, the ASIF data only cover manufacturing firms, while in the SAIC database manufacturing firms account for only about 40% of all firms by count. Therefore, using the ASIF data to study foreign 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.

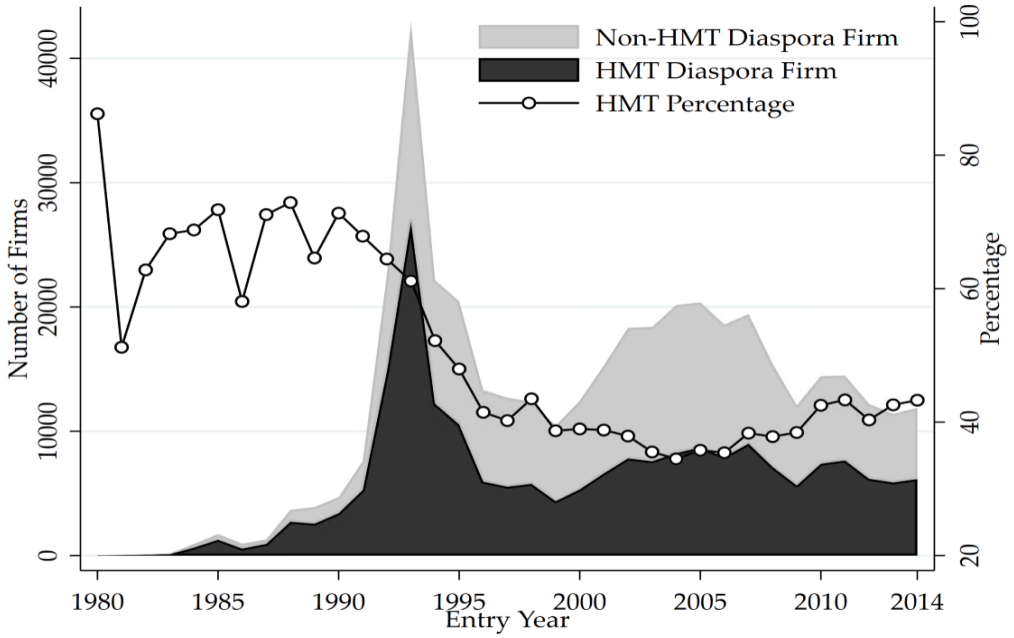
Foreign firms versus foreign direct investment (FDI). As a monetary concept, FDI includes both the initial and follow-up investment from a foreign owned entity. We cross validate our dataset of foreign firms both in counts and volumes with the official FDI data provided by the Ministry of Commerce of 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.

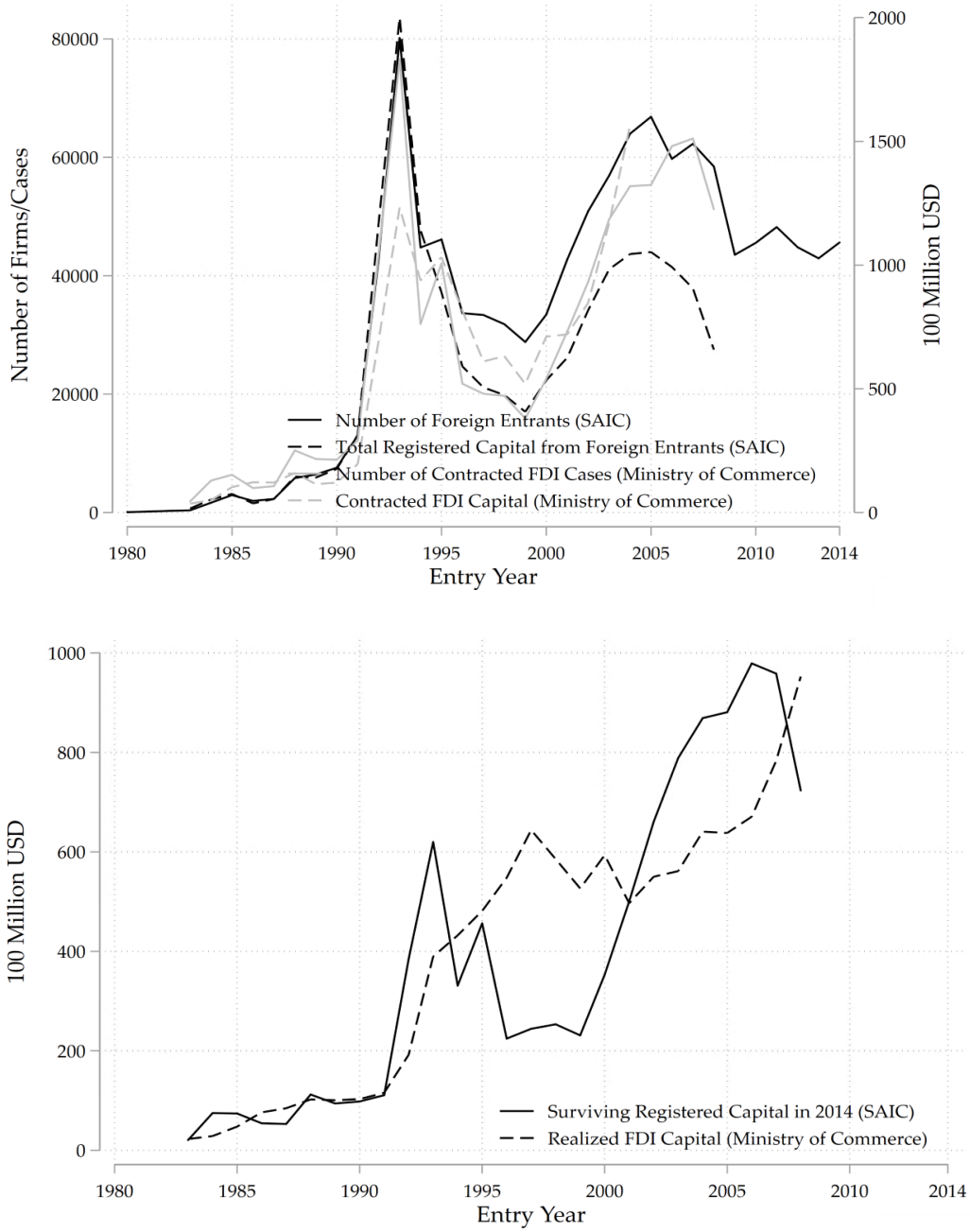
Figure A.5: Diaspora Firms versus HMT Firms



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 around 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.

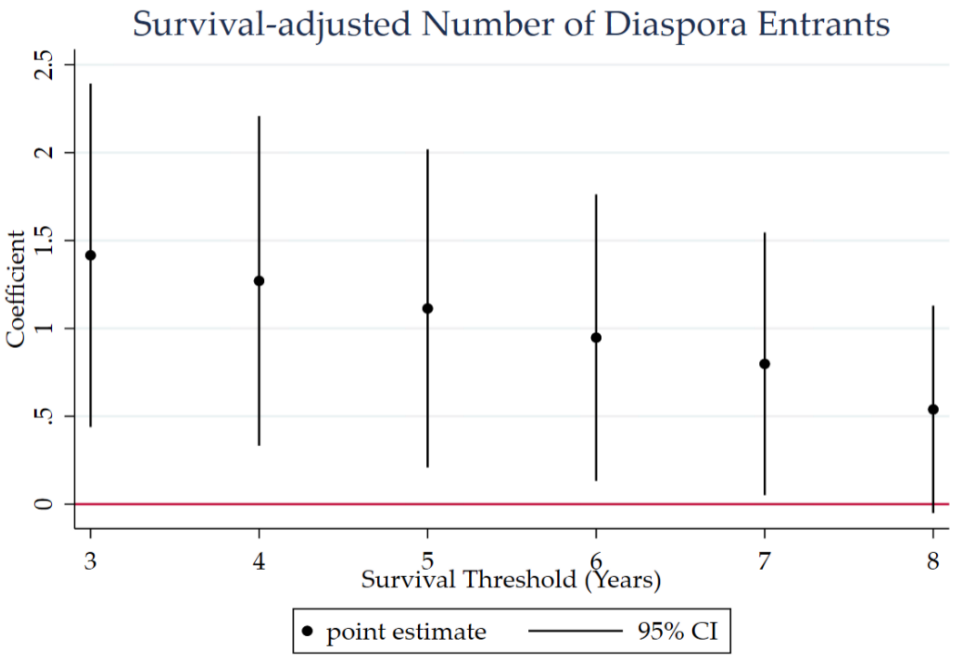
Figure A.6: Cross Validation with Official FDI Statistics in Aggregate



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

Figure A.7: 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.

Table A.1: 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.2: 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.

A.4 Additional Robustness Checks

Table A.3: Robustness Checks: 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 Registered Capital of Diaspora Entrants 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: 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.4: Robustness Checks: 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: 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.

Table A.5: Robustness Checks: Alternative Setups

	Open × Lineage Connection	
	Alternative Standard Error	Include Round-Trip Entrants
	(1)	(2)
Number of Diaspora Entrants	1.767***	3.821***
	(0.493)	(0.820)
Survival-Adjusted Number	1.517***	3.253***
	(0.456)	(0.743)
N	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y
Surname-Prefecture FE	Y	Y
Surname-Year FE	Y	Y
Prefecture-Year FE	Y	Y

Note: Standard errors are clustered at the surname-prefecture level except in column (1). 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 cluster the standard errors at the prefecture level instead of the surname-prefecture level. In column (2), 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: Pairwise Correlations: Lineage Connection Measures for Different Birth Cohorts

Birth Cohort	(1)	(2)	(3)	(4)	(5)	(6)
(1) 1940-1949	1.000					
	(–)					
(2) 1950-1959	0.603***	1.000				
	(0.000)	(–)				
(3) 1960-1969	0.548***	0.602***	1.000			
	(0.000)	(0.000)	(–)			
(4) 1970-1979	0.528***	0.569***	0.610***	1.000		
	(0.000)	(0.000)	(0.000)	(–)		
(5) 1980-1989	0.554***	0.625***	0.611***	0.568***	1.000	
	(0.000)	(0.000)	(0.000)	(0.000)	(–)	
(6) 1990-1999	0.591***	0.579***	0.625***	0.569***	0.631***	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(–)

Note: This table shows pairwise correlations for lineage connection measures constructed based on different birth cohorts in the China Population Survey of 2005. Significance levels are in the brackets.

Table A.7: Robustness Checks: Alternative Lineage Connection Measures

	Open × Lineage Connection		
	Birth Cohorts		
	Before 1945	Before 1949	Before 1960
	(1)	(2)	(3)
Number of Diaspora Entrants	2.112***	2.100***	2.019***
	(0.742)	(0.742)	(0.731)
Survival-Adjusted Number	1.809***	1.800***	1.739***
	(0.663)	(0.663)	(0.656)
N	1,344,421	1,344,421	1,344,421
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: 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. From column (1) to column (3), we use an alternative lineage connection measure calculated based on individuals born before 1945, 1949, and 1960 respectively from China Population Survey of 2005.

Table A.8: Robustness Checks: Subsample Regressions

	Open × Lineage Connection					
	Excluding Emigration-Intensive Provinces	Within Emigration-Intensive Provinces	Excluding Diaspora-Intensive Surnames	Excluding FDI-Intensive Prefectures	Only Common Surnames	Excluding Common Surnames
	(1)	(2)	(3)	(4)	(5)	(6)
Number of Diaspora Entrants	0.642***	6.001***	1.733***	1.534***	25.675***	1.303***
	(0.233)	(2.476)	(0.332)	(0.440)	(9.615)	(0.275)
Survival-Adjusted Number	0.482***	5.152**	1.480***	1.308***	22.377***	1.097***
	(0.167)	(2.186)	(0.292)	(0.380)	(9.154)	(0.239)
N	1,162,898	181,447	1,250,024	1,272,091	236,341	1,108,046
Number of Incumbent Firms	Y	Y	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y	Y	Y

Note: 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. Diaspora-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. Common surnames refer to the 50 most populous surnames in the population according to the China Population Survey of 2005.

Table A.9: Heterogeneity: Origin

	Open × Lineage Connection						
	Hong Kong, Macao, Taiwan	South East Asia	United States	Japan, South Korea	Europe	Rest of World	Excluding HMT & Tax Havens
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share of All Foreign Entrants 1981-96	0.592	0.042	0.051	0.048	0.021	0.069	0.358
Number of Diaspora Entrants	1.848***	1.151**	0.099**	0.013	0.029	0.091	0.865**
	(0.504)	(0.069)	(0.049)	(0.066)	(0.019)	(0.089)	(0.362)
Survival-Adjusted Number	1.759***	0.148**	0.068	0.024	0.025	0.092	0.622**
	(0.491)	(0.069)	(0.043)	(0.056)	(0.018)	(0.080)	(0.290)
N	1,344,421	1,344,421	1,344,421	1,344,421	1,344,421	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y	Y	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y	Y	Y	Y

Note: This table reports heterogeneous effects by origin. 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. HMT refers to Hong Kong, Macau, and Taiwan. Tax havens include Caymen Islands, American Virgin Islands, British Virgin Islands, Bahamas, Panama, American Samoa, Bermuda, Marshall Islands, Solomon Islands, and Cook Islands.

Table A.10: Heterogeneity: Industry

	Open × Lineage Connection				
	Agriculture	Mining	Manufacturing	Commerce	Service
	(1)	(2)	(3)	(4)	(5)
Share of All Foreign Entrants 1981-96	0.039	0.007	0.649	0.052	0.254
Number of Diaspora Entrants	0.119**	0.010	1.306**	-0.034	0.366**
	(0.059)	(0.009)	(0.528)	(0.064)	(0.145)
Survival-Adjusted Number	0.094**	0.010	1.175**	-0.035	0.274**
	(0.044)	(0.008)	(0.491)	(0.056)	(0.136)
N	1,344,421	1,344,421	1,344,421	1,344,421	1,344,421
Number of Incumbent Firms	Y	Y	Y	Y	Y
Surname-Prefecture FE	Y	Y	Y	Y	Y
Surname-Year FE	Y	Y	Y	Y	Y
Prefecture-Year FE	Y	Y	Y	Y	Y

Note: This table reports heterogeneous effects by industry. 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.

A.5 Additional IV Results

Table A.11: Balance Test

	Predicted Number of Diaspora Firms in 1996	Number of Observed Diaspora Firms in 1996
	(1)	(2)
Years Since Opening Up	1.477***	172.718***
	(0.414)	(55.146)
Distance to Sea	0.002	-0.004
	(0.004)	(0.211)
Log Slope	-3.371*	-363.187
	(1.925)	(270.317)
Log Elevation	1.702*	119.294
	(0.974)	(128.701)
Log Cultivated Land per capita 1996	-0.699	-421.615*
	(0.699)	(241.391)
Log Average Wage 1996	0.098	941.342**
	(1.733)	(398.290)
R^2	0.528	0.745
N	231	231

Note: 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 larger R squared.

Table A.12: Seeding Effects of Diaspora Firms: 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	Local Employment in 2015 (10^3)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
1996 Diaspora Firm Stocks	2.374***	72.352***	0.062%***	0.044%***	0.870***
	(0.614)	(12.126)	(0.001)	(0.000)	(0.114)
Panel B: 2SLS					
1996 Diaspora Firm Stocks	3.274***	162.255***	0.317%***	0.226%***	4.476***
	(1.205)	(64.060)	(0.001)	(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: Standard errors are clustered at the province level and shown in parentheses. The data on non-diaspora foreign firm stocks and domestic private firm stocks in 2014 are from the SAIC database. The data on employment in 2015 are from the 2015 Population Survey of China. Panel A presents the OLS estimates, while panel B presents the 2SLS estimates using the predicted diaspora firm Stock in 1996 as an IV for the observed diaspora firm stock in 1996. The Cragg-Donald Wald F statistic is reported for the IV regressions. We deal with log zero by inverse hyperbolic transformation.

Disclosure Statement

Ruichi Xiong and Fanghao Chen acknowledge financial support from the China Research Initiative at the Rotman School of Management, University of Toronto. Fanghao Chen acknowledges financial support from the Fundamental Research Funds for the Central Universities (project number 23JNQN16). Xiaobo Zhang is grateful for the financial support from China Natural Science Foundation (project numbers 71874008, 71441008, 71873121, and 72192844). We declare that there is no conflict of interest associated with our paper.