Appendix A. Variable Summary Information

Table A1: Variable Definition and Summary Information

Variable Name	Definition	Data Source	Mean	St. Dev.
Patent Intensity (ln)	One plus the logarithm of the ratio of the number of successful invention patent applications to total number of employees.	SIPO	0.014	0.327
New Innovation Sales Intensity	One plus the logarithm of the ratio of new product and process sales to total firm revenues in year *t*.	ASIF	0.035	0.298
Firm Ownership Structure	Firm classification follows Hsieh et al. (2015) where	ASIF		
- POE	information from both the firm's legal status and		0.849	0.358
- Semi-POE	the structure of paid-in capital is used to assign		0.044	0.205
- SOE	firms into one of three categories: privately-owned enterprise (POE), privatized state-owned enterprise (Semi-POE) and state-owned enterprise (SOE).		0.107	0.309
Firm Knowledge Stock	The perpetual inventory method is used to calculate patent knowledge stocks with a 15% dep. rate.	SIPO	0.241	4.234
Firm Size	Logarithm number of employees.	ASIF	4.612	1.039
Firm Age	Logarithm of the number	ASIF	2.115	0.953

	of years the firm has been in operation.			
Industry Sales Growth	Logarithm of the difference in real sales of each 3-digit industry between year *t* and year *t* – 1.	ASIF	10.121	19.120
Industry Innovation Rate	Average share of new product and process sales of each 3-digit industry in year *t*.	ASIF	0.030	0.027
Average Industry Size	Logarithm of the number of employees in each 3-digit industry in year *t*.	ASIF	15.478	1.720
City Size	Employment density calculated as the logarithm of $\frac{N_r}{A_r}$, where N_r is the size of the working population in each city, and A_r is the area of the city (km^2) in year *t*.	ASIF	5.28	1.33
Density	Proxy of inter-industry relatedness as introduced in Hidalgo et al. (2007)	ASIF	0.014	0.011
- Input-Output Linkages	Calculated as the share of industry *j's* inputs bought by industry *h*.	2002 IO Tables	0.113	0.107
- Similar Workers	Calculated as the correlation coefficient between the share of two industries' respective employment with the same educational level	2004 Economic Census	0.071	0.125
- Related Kn. Spill	Calculated as the share	SIPO	0.033	0.059

	of technologies associated with industry *h* that cites technologies associated with industry *j*			

Appendix B. Marshallian Sources of Relatedness

Following Howell (2017a), proxies are developed for the three main Marshallian sources of relatedness: input-output linkages, labor pooling, and knowledge spillovers. A description of the variable development is provided below.

Input-Output Linkages

Firms that have better access to customers and suppliers reduce transportation costs. Firms are therefore expected to enjoy higher productivity gains when they share closer input-output relationships with local leading industries. China's 2002 I-O Tables published by the National Bureau of Statistics (NBS) are used to measure supplier-buyer linkages. The I-O Tables provide input-output relationships between 122 sectors. A concordance between the I-O sectors and the CIC industries published by the NBS is used to map the I-O sectors to the 4-digit CIC industry from the ASIF. $Input_{hj}$ is defined as the share of industry $h's$ inputs that come from industry j and $Output_{hj}$ as the share of industry $h's$ outputs that are sold to industry j. The share of inputs and outputs are respectively substituted into Equation 1 above to get region-industry varying proxies for customer and buyer linkages.

Labor Pooling

Firms that enjoy better access to labor pooling of suitable workers are expected to minimize costs including risk sharing and better matching. Firms are expected to enjoy higher productivity gains when they use similar types of labor as the local predominant industries. As a proxy for labor pooling of similar workers, the 2004 Chinese Economic Census (CES) data is used to compute the industry's skill intensity and similarity between industries in terms of their education levels[1]. In the 2004 CES, there are five education categories: junior high school or below, senior high school, diploma, undergraduate/college, and postgraduate.

The capital stock (measured by fixed assets) from the 2004 CES is also used to compute capital intensity along with skill intensity to control for industrial characteristics. The industry skill intensity and capital intensity are calculated at the 4-digit CIC industry level. $Share_{ho}$ is defined as the fraction of industry $h's$ employment with education level o. Similarity of workers in industries h and j is measured through the correlation of $Share_{ho}$ and $Share_{jo}$ across education levels. Next, the correlation coefficient is substituted into Equation 1 above to get region-industry varying proxy for labor pooling of similar workers.

Knowledge Spillovers

Firms may also co-locate to benefit from knowledge spillovers – the direct or indirect transfer of information or ideas. A proxy for information exchange is developed using patent citation information on Chinese patents. Patent citation data is commonly employed employed to assess intellectual spillovers, but are obviously an imperfect measure (Jaffe et al., 1993). While Howell (2017a) relies on co-exporting information to proxy for knowledge spillovers, here a measure is developed based on the extent to which technologies associated with

[1] Note that the 2004 CES contain all-scale firms, whereas the ASIF data includes only the above-scale firms. To keep consistency over the sample, only the above-scale firms are kept in the 2004 CES data.

industry h cite technologies associated with industry j, and vice versa. Both of these measures, $PatentIn_{h \leftarrow j}$ and $PatentOut_{h \rightarrow j}$ are normalized by total citations for each industry. Due to the high correlation between the two measures, I construct a single PatentInOut $_{hj}$ measure by combining the two separate indexes together.

Chinese patent citations data is obtained from IncoPat's Technical Innovative Intelligence Platform that provides comprehensive information on Chinese patents and the relevant information related to prior art. It is expected that the other two proxies – $InputOutput_{hj}$ and $Educ_{hj}$ – better capture knowledge sharing between customers and suppliers and the exchange of workers, respectively. As a result of these two issues, the citations measure most likely serves as a better proxy for the exchange of technologies rather than capturing all forms of intellectual spillovers, and is expected to be weaker in magnitude relative to the other proxies.