Industries as in a Network: Micro Evidence from Job Search

By JIACHENG KANG AND KAILING SHEN

We consider the labor market linkages across industries based on the concept of cross-industry skills (CRISs) here. CRISs are skills that are productive beyond any single industry. CRISs connect industries through their impacts on workers' mobility hurdles across industries. In particular, we empirically estimate labor mobility network across industries (LMNInd) using online job board data and recently available machine learning algorithm. Based on the estimated LMNInd, vacancy-applicant skill match indices are then constructed and tested on individual job application outcomes. We further demonstrate how this estimated LMNInd can be used to predict the transmission of an exogenous shock on one industry to all the other industries – the "ripple effect". Our results show a one standard deviation increase in CRISs is associated with 0.51 percentage points increase in callback probabilities, which is equivalent to 1.16 times of the impacts of being more experienced than required. The results also suggest that the effect of CRISs is stronger for lower-paying jobs. Lastly, our aggregate level results suggest the so-called "ripple effect" can be non-linear and complicated due to the existence of CRISs and the specific configuration of LMNInd in an economy.

I. Introduction

Human capital has been attracting considerable critical attention. In the last few decades, one of the most significant discussions in human capital formation is the understanding of skills. In the human capital theory, Becker (1964) distinguishes between general-purpose and firm-specific skills. Recently, several specific human capitals and their economic significance in the labor market have been studied, such as industry-, occupation-, and task-specific skills (Neal, 1995; Gibbons & Waldman, 2004; Kambourov & Manovskii, 2009; Sullivan, 2010). However, the transferability of skills across industries is not well understood. It would be useful to understand better the significance of such transferable skills across industries as well as their effects on labor market behavior.

In this paper, we propose a concept, cross-industry skills (CRISs), to measure professional human capital applied in multiple industries. The distinction among CRISs, industry-specific skills, and general skills is intuitive and straightforward. According to Becker (1964), general skills are defined to increase workers' marginal productivity not only in firms providing general training but also in other firms. Neal (1995) documents the importance of skills that are specific to a given industry and finds that such industry-specific skills explain wage-tenure profiles better than firm-specific skills. Thus, general skills such as communication skills are useful in all industries, while industry-specific skills are useful in a specific industry such as mining skills applied in mining industry only. Yet, it is possible for some professional skills to be useful in multiple industries and we propose them as CRISs. In other words, this kind of transferable professional skills increases marginal productivity of workers in multiple industries. Therefore, CRISs play an intermediate role between industry-specific skills and general skills, and CRISs could be further regarded as a fundamental concept for our investigation of labor mobility network across industries (LMNInd). An intuitive example is that data analysis skill is one of CRISs and not an industry-specific skill because it is potentially useful in multiple industries (e.g., finance and trade industries). In the meantime, data analysis skill is also not a general skill since it is unlikely to be applied in all industries (e.g., art industry).

Empirically, we investigate CRISs by using Chinese online job board data from XMRC.com. The data contain a complete picture of firms' industry diversification, workers' industry mobility, and detailed application information. By taking advantage of internal records of the platform from January 2018 to October 2019, we find that it is pervasive for firms to do business in multiple industries, and applicants commonly have multi-industry working experience. A similar situation occurred in the United States. Kambourov and Manovskii (2008) points out that industry mobility in the United States is high and has increased dramatically over the past decades.

Motivated by the phenomenon, we first investigate LMNInd because it implicitly reflects CRISs and is essential to construct the measure of CRISs. Conceptually, industries are the categories of economic activities that produce the same nature of products or provide the same kind of services. Industries act as networks because industries are connected, interacted, and operated together through not only (intermediate) products, but also (similar) technologies (i.e., a reflection of CRISs). Therefore, characterizing and understanding LMNInd are crucial to study CRISs.

Specifically, we employ two approaches to investigate LMNInd. On one hand, by taking advantage of innovative methodology in computer science, we use machine learning introduced by Mikolov et al. (2013) to investigate LMNInd from a complete pattern of industry information. This machine learning approach essentially transforms industries into vector forms. Essentially, these meaningful industry vectors formulate LMNInd from three perspectives (i.e., either industries of firms' business, industries of workers' experience, or their combined industry information in applications). On the other hand, we use conditional probabilities to calculate the probability of any pair of industries occurred together. We find that to some extent the machine learning technique outperforms the approach using conditional probabilities as well as a method used by Neffke and Henning (2013) who studied skill relatedness between industries.

To better understand LMNInd, we employ t-SNE and k-means approaches to visualize LMNInd. We find that LMNInd performs very well in terms of reflection of CRISs. From the visualization, the “closeness” reflects the likelihood that the given industries appear near each other in the industry information. It implies that the closer the industries in LMNInd, the more intensive CRISs are applied in these industries. In addition, industry clusters indicate that industries within cluster share more intensive CRISs.

LMNInd also captures the heterogeneity of three perspectives of LMNInd such as firms’ industry business, workers’ industry experience, and their combined industry information in applications. For example, from workers’ and applications’ LMNInd, internet/e-commerce industry is most closely related to video game and computer software industries, while it is most closely related to telecom operations and video game industries from firms’ LMNInd. In contrast, internet/e-commerce industry is most far away from pharmacy, mining, and machinery industries from firms’, workers’, and applications’ LMNInd respectively. Therefore, LMNInd provides reasonable and solid foundations on construction of the measure of CRISs.

Second, we construct the measure of CRISs based on LMNInd. In light of capturing labor mobility across industries and further investigating the effect of CRISs in job search, we use workers’ LMNInd to construct the measure of CRISs although we have additional firms’ and applications’ LMNInd. Specifically, we use vacancy-applicant skill match index represented by the similarity between industries of firms’ business and industries of workers’ experience within applications based on workers’ LMNInd as the measure of CRISs.

Third, we empirically test the validation of the measure of CRISs by examining the role of CRISs in explaining the employability gap in job search. We find that workers with higher CRISs have a significantly higher chance to be interviewed by employers. Specifically, the regression results indicate that a one standard deviation increase in CRISs is associated with 0.51 percentage points increase in callback probabilities across occupations. The economic magnitude is 1.16 times than being over-experienced and 2.68 times than over-education. Furthermore, we find that this effect is more pronounced in lower-paying jobs than higher-paying jobs. It implies that the effect of CRISs is more sensitive to lower-paying jobs.

Finally, we further demonstrate how this estimated LMNInd can be used to predict the transmission of an exogenous shock on one industry to all the other industries – the "ripple effect". In practice, we apply LMNInd along with CRISs in a simulation of an industry-specific shock. Specifically, our investigation focuses on both micro- and macro-level effect. On one hand, from micro-level view, we investigate what industries for those individuals who previously applied jobs in the affected industry would apply. Also, we examine the change of employment in each industry before and after the industry shock. On the other hand, from macro-level view, we test whether there exist multiplier effects in unemployment. In other words, we test if the change of unemployment rate is more than the change of number of jobs due to the industry-specific labor demand shock.

We find that industries response differently to the industry shock due to the significance of LMNInd. In addition, we find that there are multiplier effects of an industry-specific demand shock, because the size of shock regarding total callbacks is more than size of the shock regarding total job ads. This aggregate effect highlights the spillover of an industry-specific demand shock on other industries. Therefore, our aggregate level results suggest the so-called "ripple effect" can be non-linear and complicated due to the existence of CRISs and the specific configuration of LMNInd in an economy.

The paper makes several contributions to the literature. First, we contribute to the literature on the significance of human capital. Previous literatures have investigated the importance of specific human capital. For occupation-specific human capital, Gathmann and Schönberg (2010) study transferability of skills across occupations by investigating task data; Kambourov and Manovskii (2009) examine the effect of occupation-specific skills on wage growth. Moreover, Neal (1995) finds that industry stayers with industry-specific skills earn significantly greater than industry switchers from the perspective of displaced workers. Our paper introduces CRISs to empirically measure transferable skills across industries, which supplement the scopes of industry-specific skills and general skills. We also show that CRISs play an important role in job search.

Second, we provide the first attempt to use machine learning to study LMNInd as well as construct the measure of CRISs. Our paper provides the new application of machine learning technique introduced by Mikolov et al. (2013) to study LMNInd. It also overcomes the interpretability issue of regression with a great number of dummy variables.

Third, we contribute to the literature on the study of industry shocks. Due to LMNInd, it is not reasonable to assume other industries are not affected and thus used as the control group in purpose when there is an exogenous industry-specific demand shock. However, industries response differently to the industry shock due to LMNInd and CRISs.

The reminder of the paper is structured as follows. In Section II, we describe the data. Section III introduces the methodology for the creation of LMNInd and the construction of the measure of CRISs. Section IV presents LMNInd, regression analysis to test the validation of the measure of CRISs in job search, and simulation analysis. In Section V, we conclude the paper and discuss economic implications for related research.

II. Data

In this article, we use proprietary data provided by XMRC (www.xmrc.com.cn). XMRC is a private company founded by Xiamen Talent Service Center, a public institution subordinated to Xiamen Human Resources and Social Security Bureau. XMRC provides a conventional structure like other online job boards, such as indeed.com. It allows employers to post a job ad containing job title, education requirement, experience requirement, offered wage (if specified), offered bonus (if specified), industries of the enterprise, and geographical location, and so on. On the other hand, job seekers who visit XMRC could specify some information on the top of the website to direct search jobs by selecting the desired occupation from a list, indicating a range of job creation dates, and stating some keywords in the company name. After job seekers send an application together with their resumes, employers decide whether to offer a callback via an internal messaging system for their further contacts.

Our data set consists of vacancies posted on XMRC from January 2018 to October 2019, whereas the data from XMRC used by Kuhn et al. (2020) for the study of gender-targeted job ads was extracted between May 1 and October 30, 2010. Our sample contains rich information about job postings, firms, applicants, and applications. An important advantage of the sample is that it includes detailed records for industries operated by firms as well as applicants' industry experience. In China, firms can specify one or multiple industries for their business operation in the process of registration and XMRC also allows firms to select multiple industries of business from a drop-down menu. The number of industries classified by XMRC is 55 and the industry list is shown in Table A1.

Descriptive statistics are provided in Tables 1 and 2 for job and applicants' characteristics, respectively. Table 1 presents the summary statistics of job characteristics. It is arranged in ascending order of the number of industries that job openings stated. The analysis sample consists of 120,073 job postings by 16,759 distinctive firms.

What stands out in this table is the distribution and general pattern of the number of industries of job ads. As shown in Table 1, more than half of job openings state multiple industries, while 43.84% of job postings indicate a single industry. This stylized fact highlights the pervasiveness of multi-industry business operations in the labor market.

The summary statistics further imply that job characteristics are associated with the number of industries stated. With the number of industries listed, job ads increase the education requirement which from an average year of schooling 13.92 to 14.29. In the meantime, jobs require more college or above with the number of industries stated. Similarly, experience requirement has been raised from one-industry to multi-industry job ads. With the number of industries stated, job openings are less likely to have explicit age requirements and appear to have lower age requirement if an explicit age requirement is specified.

In contrast to the age requirement, with the number of industries indicated, job ads are more likely to explicitly post wage information, and the average monthly wage increases substantially from RMB5,615 to RMB5,833 if posted wages are specified. Also, bonus information appears similar pattern as the offered wage. Further, with the number of industries listed, advertised jobs provide more vacancies. On average, each job ad posts 2.344 vacancies. In terms of firm ownership type, private firms account for a larger proportion with the number of industries stated. Therefore, it shows that jobs indicating multiple industries business are likely to require higher educational attainment, offer higher wages and bonus, have higher demand of workforce, and relax experience and age requirements.

To empirically examine the association between the number of industries in which jobs are involved and jobs' skill demands, Table A2 presents the effect of jobs' skill demands on the number of industries in which jobs are involved. From the regression results, Table A2 suggests that jobs operating in multiple industries have higher education requirements, fewer years of experience requirement, and higher offered wage. Meanwhile, private firms have more industries operated. It provides evidence that jobs involved in multi-industry business could be regarded as higher-quality jobs.

Second, Table 2 presents the summary statistics of applicants' characteristics. The framework of the table is according to the number of industries in which job seekers have worked. What is striking about the figures is that, when applicants specify previous working industries, more than half of job seekers have multi-industry working experience. Of the total number of 246,566 applicants in our analysis sample, 50.62 percent of workers state multi-industry working experience.

The summary statistics also shed light on the stylized fact that applicants' characteristics may be correlated with the number of their working industries. With the number of working industry experience, workers' years of education increases from 14.44 to 14.71 years. Specifically, the proportion of colleges increases from 34.5% to 40.2%. Similarly, when job seekers have a larger number of industry experience, they are more likely older and married. Besides, with the number of industries stated, applicants tend to have more years of experience and higher current wage and intended wage. The proportion of females also increases with the number of industries listed.

Similarly, we investigate the association between the number of industries in which applicants experienced and applicants' characteristics. Table A3 presents the effect of applicants' characteristics on the number of industries in which applicants experienced. The regression results in Table A3 implies females switch industries more frequently than males, which might contribute to explain gender wage gaps. Also, workers with higher educational attainment have higher labor mobility across industries. However, higher labor mobility across industries is associated with lower current wages, which suggests the positive effect of industry-specific human capital on wages as documented by Neal (1995).

Further, Table 3 presents matching characteristics in applications. Of the total number of 3,618,944 applications, we divide them into three terciles by vacancy-applicant skill match index which introduced in Section III. In terms of callback rates, on average, 19 percent of applications receive callbacks. Specifically, with higher vacancy-applicant skill match index, the average callback rates increase significantly, from 18.3 to 19.8 percent. By investigating other matching characteristics, there is no substantial difference among the three groups by vacancy-applicant skill match index. Taken together, Table 3 suggests the importance of CRISs on recruiting behavior in job search.

Taken together, Tables 1 and 2 document that more than half of job openings are involved in multi-industry business, and more than half of job seekers have multi-industry working experience. These facts provide fundamentals to study CRISs. Also, Tables 1—3 potentially suggest the importance of CRISs in the labor market by both the observed association between multi-industry business and job characteristics, and the underlying correlation between multi-industry experience and applicants' characteristics.

III. Methodology

In this section, we present the methodology of the paper. First, we describe how to create LMNInd from three perspectives, such as industries of firms' business, industries of workers' experience, and their combined industry information in applications. Second, we present how to construct the measure of CRISs based on LMNInd.

A. Create LMNInd

LMNInd is fundamental to the investigation of CRISs. As underlying understanding, industries operate as networks because industries are linked, collaborated, and worked together by not only (intermediate) goods, but also technologies (i.e., a reflection of CRISs).

We employ a popular machine learning technique in natural language processing (NLP)—Word2Vec—to construct LMNInd from the complete pattern of industry information. Word2Vec is introduced by Mikolov et al. (2013) and is designed originally to provide a way for machines or computers to interpret and even understand human language. Essentially, it takes texts which can be in any language as inputs where the context of those words is assumed in relation to each other. Then, a vector representation of words created by Word2Vec is as outputs and captures inherent relations among words. Next, the relation of words can be characterized by these word vectors. Therefore, this feature of Word2Vec is consistent with our attempt to create LMNInd from industry information.

The fundamental assumption behind Word2Vec is that the meaning of a word is affected by the words around it. Skip-gram is one of the architectures of Word2Vec which uses the center word to predict the surrounding window of context words. Practically, the skip-gram neural network algorithm uses a center word to predict the probability of each word in the vocabulary being a context word within a chosen window size. To further train the neural network, the initial setup is to construct the input layer that represents each word in the vocabulary in the form of one-hot vectors or the dummy variables. The next step is to adjust the weight of the neural network (i.e., update the dummy variables in the input layer) to get an output prediction as close as possible to the actual data. Finally, the weight updated and optimized by the gradient descent method would be the word vectors. More details are presented in Mikolov et al. (2013).

The philosophy of Word2Vev is consistent with our study of LMNInd. In this paper, we regard a set of 55 industries as vocabulary and assume industries requiring the same CRISs are likely to occur together more frequently in the same document (i.e., industries of firms' business, industries of workers' experience, and industries of applications, respectively). In other words, we assume that industries of firms or workers or applications relate to each other, respectively.

Conceptually, firms are assumed more likely to expand their business in some new industries in the sense that their current industries and those intended industries require relatively intensive CRISs than other industries. Similarly, when workers choose to switch their working industries, they prefer to work in some new industries requiring relatively comprehensive CRISs that they had obtained from their previous industry experience. Because by doing so, workers maximize their utilities as their skill sets are not depreciating much and even possibly gain more from CRISs in working on new industries. Additionally, applicants are assumed to compare their previous industry experience and industries of firms' business. Thus, application behaviors implicitly imply the potential matching of CRISs between industries of workers' experience and firms' business. Therefore, LMNInd created by machine learning can help to capture the relation of industries and understand LMNInd from different perspectives.

Except the feature of Word2Vec facilitates our investigation of LMNInd, it further helps to construct the measure of CRISs and essentially avoid the non-interpretable issue in regression with dummies representing (combinations of) industries. In regression analysis, one common way of incorporating categorical variables is to represent them as dummy variables or indicator variables. In the study of CRISs, because the number of industries classified by XMRC is 55, there are a potentially great number of combinations and matchings of industries of firms' business, or industries of workers' experience, or industries of applications. Such a considerable number of dummy variables representing industry combinations imposes severe interpretability issues in regression results. Therefore, regression analysis with dummy variables is not applicable to investigate CRISs in the circumstance. We will describe how to construct the measure of CRISs based on LMNInd in Section III.B.

Besides, the machine learning technique could be applied to create LMNInd from different perspectives. Practically, there are three types of industry information, such as industries of firms' business, industries of workers' experience, and industries of applications. Therefore, the methodology provides a systematic way to construct and compare three LMNInd for further analysis.

We then compare our approach to the method employed by Neffke and Henning (2013). Recently, Neffke and Henning (2013) investigate skill relatedness between two industries based on labor flows across industries. They count the number of labor flows from one industry to another industry and then regress it on industry characteristics by a zero-inflated negative binomial model. By utilizing regression results, they construct predicted interindustry labor flows as well as skill relatedness.

In contrast, we apply machine learning to construct LMNInd and further construct the measure of CRISs. The critical difference is that we consider the integrality of industry information when we construct LMNInd. Specifically, given complete and detailed records of individuals' industry labor flows, the integrality of industry information is important for understanding LMNInd and would avoid potential biases in constructing LMNInd and CRISs. Neffke and Henning (2013) study skill relatedness from the basic unit of labor flows (i.e., one original industry to one destination industry or equivalently pairs of industries), while our approach considers integrality of workers' industry choice (i.e., potentially multiple original industries to multiple destination industries). By doing so, we can examine LMNInd more precisely and would have better insights.

Specifically, in light of the integrality of industry information, one more important implication is the effectiveness of CRISs. As illustrated in Figure 1, we consider four industries named A–D and they are shown in circles with different colors. As mentioned, there are three distinguished skills, such as general, industry-specific, and CRISs. First, the area interacting with all industries is intuitively regarded as general skills such that they are applicable in all industries (labeled as ① in Figure 1). Second, the area not interacting with any other industries refers to industry-specific skills that are working in one single industry (labeled as ② in Figure 1). Third, the area interacting two or three or four industries then refers to CRISs which are applied in multiple industries.

The illustration of CRISs in Figure 1 highlights the difference between the method used by Neffke and Henning (2013) and our application of machine learning. Neffke and Henning (2013) measures skill relatedness between pairs of industries only (i.e., the area interacting two industries and labeled as ③ in Figure 1), while our approach emphasizes not only CRISs between pairs of industries but also CRISs among multiple industries (labeled as ③ and ④ in Figure 1, respectively). Therefore, our methodology somehow better captures LMNInd from the perspective of CRISs.

Further, we consider an alternative statistics approach as robustness check to create LMNInd and construct the measure of CRISs. We find that skill match index created by machine learning is more evenly distributed and implies that it is better to represent CRISs to some extent. Further, skill match index from machine learning is better in terms of distribution and for ease of interpretation purpose (i.e., using raw skill match index rather than adjusted ones). Additionally, vacancy-applicant skill match index from machine learning outperforms that from statistics not only in the regression analysis but also from better reflection of CRISs. The details are presented on Appendix 4.

In addition, we explore the robustness of machine learning approach. The implementation of Word2Vec considers several important parameters in the model. We examine the robustness in selecting these parameters in Appendix 5.

Taken together, Word2Vec is conceptually an economic toolbox to potentially provide a doable, interpretable, and reliable way to construct LMNInd. LMNInd captures the relationships among different combinations of industries and provide a fundamental to the measure of CRISs.

B. Construct the Measure of CRISs

In Section III.A., we introduce Word2Vec as an economic tool to create vector representation of industries. This technique transforming industry information to industry vectors internalizes the industrial relations and helps to create LMNInd based on that. As mentioned, LMNInd could be further applied to construct the measure of CRISs to address the interpretability of a large number of dummies representing industry combinations in regression analysis. This section then presents how to measure the measure of CRISs.

Before introducing the method of constructing the measure of CRISs from LMNInd, we restate our key assumption that the higher the similarity between industries of firms' business and industries of workers' experience, the higher the CRISs would be observed. This kind of similarity implicitly captures workers' application behavior such that after observing industries of firms' business workers consider their past industry experience to apply it. Intuitively, the similarity presents overall relevancy between industries of firms and workers, and it further implies the magnitude of CRISs. Therefore, we make use of such similarities to construct the measure of CRISs.

In light of capturing labor mobility across industries and further investigating the effect of CRISs in job search, we use workers' LMNInd to construct the measure of CRISs although we have additional firms' and applications' LMNInd. Specifically, we use vacancy-applicant skill match index represented by the similarity between industries of firms' business and industries of workers' experience within applications based on workers' LMNInd as the measure of CRISs.

Mathematically, our approach is first to aggregate firms' and workers' industry vectors within an application to calculate their average industry vectors, respectively. Then, vacancy-applicant skill match index could be measured by the cosine similarity between firms' and workers' average industry vectors:

(1) $$\textit{vacancy-applicant skill match index from machine learning}$$
$$= \cos(\theta) = \frac{\mathbf{F} \cdot \mathbf{W}}{\|\mathbf{F}\| \|\mathbf{W}\|} = \frac{\sum_{i=1}^{n} F_i \cdot W_i}{\sqrt{\sum_{i=1}^{f} F_i^2} \sqrt{\sum_{i=1}^{w} W_i^2}}$$

where **F** and **W** denote the firms' and workers' average industry vectors, respectively. The subscript i indexes the dimension of industry vectors.

IV. Empirical Analysis

In this section, we present several results. First, we present LMNInd to capture the relationships among industries from three perspectives. Second, in terms of job matching behavior, we examine how employers respond workers with different levels of CRISs. Third, we apply LMNInd along with CRISs in a simulation analysis to study the effect of an industry-specific labor demand shock.

A. LMNInd

Figure 2 presents the LMNInd from three perspectives. Panel A, B, and C present the LMNInd from industries of firms' business, industries of workers' experience, and industries of applications, respectively. In terms of the design of the figure, the size of the circle represents the weighted sample size of industries in their own perspective sample. It is clear to examine which industries are most observed in the three samples, respectively. To illustrate Figure 2, we take panel B as an example. From industries of workers' experience, the top 5 industries involved are electronic technology (8.02%), other (6.99%), internet/e-commerce (5.58%), construction (5.66%), education (3.55%). These statistics are presented in Table 4.

Second, the color depicts industry clustering information. In three panels, we create eight clusters by k-means methodology to possibly capture the group of industries. As shown in panel B, the light-blue color presents one industry cluster (i.e., legal service, bank, finance, insurance, consulting/HR, intermediary service, and government industries). Table 4 presents the details of industry clusters and other related characteristics.

Third, we employ t-SNE method to do dimension reduction of industry vectors from 20 dimensions to 2 dimensions for better visualization. The distance between industries implies the relatedness of industries as well as the magnitude of CRISs. Intuitively, the closer the two industries are in LMNInd, the more related the two industries are, and the higher the CRISs between the two industries. As a result, the industries within a cluster present relatively higher CRISs. For example, in panel B, workers' LMNInd implies that the relatively higher CRISs among industries in the light-blue colored cluster (i.e., legal service, bank, finance, insurance, consulting/HR, intermediary service, and government industries). In more detail, panel A of Figure A1 presents the heat plot of 55×55 skill match index matrix to capture pairwise CRISs from industries of workers.

From Figure 2, we compare three panels to have important insights. In panel A, it is intuitive that firms do business in relatively tightly related industries and the clusters show clearer industry groups than other panels. In terms of application behavior, industry clusters in panel C are also dense, which indicates that workers are likely to apply the job that they have higher CRISs. This is consistent with directed search theory.

The sparsest plot is LMNInd from industries of workers. As shown in panel B, the distance between industries is relatively large. It shows that workers' industry experience presents a more random pattern, partly because they may choose low-related industries in their careers due to various reasons. More importantly, panel B captures labor mobilities across industries in the whole labor market.

In terms of methodology, the machine learning tools, such as Word2Vec, t-SNE, and k-means, help to take a closer look at the relationship between industries as well as provide a systematic potential way to measure CRISs. By doing so, Figure 2 provides a comprehensive overview of LMNInd. It qualitatively well captures LMNInd as well as CRISs.

B. Effects of CRISs on Callback Probability

In this section, we further examine the role of CRISs in job matching. As mentioned before, the internal records of XMRC contain information about whether employers contact applicants. We then create an indicator, callback, as the measure of successful job matchings. By doing so, we can test whether applicants with higher CRISs receive higher callback probability.

We estimate the following linear probability model to check whether callback rates increase in CRISs during the application process:

$$\text{(3)} \qquad Callback_i = \alpha + \beta CRIS_i + \gamma J_i + \delta W_i + \theta M_i + \eta C_i + \phi \mathrm{O}_i + \epsilon_i$$

where i indexes applications. J_i is the set of job requirement controls. W_i is the set of detailed CV controls. M_i is the set of controls for matching outcomes between job requirements and applicant characteristics. C_i is the set of competition controls. O_i is the occupation fixed effect. ϵ_i is the error term. In this specification, our main coefficient of interest is β which captures the matching behavior in terms of the effect of CRISs on callback probabilities. In addition, the parameter θ gives the differential between those non-matching and matching outcomes.

Table 5 reports the estimates of equation (3) to capture the relationship between CRISs and callback probabilities. Without controls in column 1, the coefficient is statistically significant which indicates that a one standard deviation increase in CRISs is associated with a 0.64 percentage points increase in the probability of callback. In column 3, by adding job requirement controls and detailed CV controls, the effect becomes even larger and remains robust. Meanwhile, all coefficients of these matching outcomes are also statistically significant and give the signs as expected. The results show callback penalties for those under-educated and under-experienced, while firms give a premium on callback probability for those who are over-educated and over-experienced.

Column 5 in Table 5 presents estimated results when further including vacancy-applicant matching controls and competition controls. The result highlights that the effect of CRISs on callback probability is robust and significant in magnitude across occupation cells. We find that, across occupations, a one standard deviation increase in CRISs is associated with a 0.51 percentage points increase in probabilities of callbacks. The magnitude is statistically substantial which is 1.16 times than being over-experienced (0.44 percentage points) and 2.68 times than over-education (0.19 percentage points). In addition, workers with over qualification in education gain callback premium which is different from the results in Kuhn and Shen (2013).

As shown in columns 6, the effect on callback probabilities is still statistically significant and economically large when absorbing occupation fixed effects. We find that, within occupations, a one standard deviation increase in CRISs is associated with a 0.52 percentage points increase in probabilities of callbacks. The magnitude is 1.3 times than being over-experienced (0.40 percentage points) although the point estimate of over-education is not statistically significant.

As Marinescu and Wolthoff (2020) indicates the power of words in job titles in the matching process, we further absorb job title fixed effects in column 7. We find that, within occupations, a one standard deviation increase in CRISs is associated with a 0.65 percentage points increase in callback probabilities. This most saturated specification implies that the magnitude of the effect of CRISs is even much substantial which is 2.24 times than being over-experienced (0.29 percentage points) and 16.25 times than over-education (0.04 percentage points). Similarly, the point estimate of over-education is not statistically significant. On the other hand, the positive effect of CRISs could offset 78 percent of the effect of being under-experienced (-0.84 percentage points) and offset 60 percent of the under-education (-1.08 percentage points).

The hypothesis internalized is that the higher CRISs between jobs and workers, the higher probability the worker receives a callback. For job postings, recruiting workers who have higher CRISs is likely to reduce operation or human resource costs and potentially better for firms to develop competitivity. In other words, multi-industry firms might be interested in recruiting employees who have highly related industry experience on firms' industries even related industries, because by doing so, it could potentially reduce costs and improve efficiency.

Summing up, the main takeaway in Table 5 is that CRISs play a critical role in job matching since it well explains the employability gap in the labor market.

So far, we have already known that applicants with higher CRISs have higher callback probabilities. Next, we continue to test in the segmented labor market whether CRISs have heterogeneities in the effect on employability. In other words, we examine the effect of CRISs in different wage markets. In the literature, jobs requiring higher skills are likely to offer higher wages. Therefore, we essentially investigate whether an increase in CRISs has different effects on callback probabilities between jobs demanding higher and lower skills. To check it, we estimate the following linear probability model:

$$(4) \qquad Callback_i = \alpha + \beta CRIS_i + \psi \log(Wage_i) + \tau CRIS_i \times \log(Wage_i) + \gamma J_i + \delta W_i + \theta M_i + \eta C_i + \phi O_i + \epsilon_i$$

where i indexes applications. $\log(Wage_i)$ is the log of offered wages. Other variables in this specification are defined as equation (3).

Table 6 presents estimates from equation (4). In column 1, we find that there is a positive effect of CRISs and a negative impact of offered wage on callback probabilities. In column 2, by adding the interaction term between CRISs and log of offered wage, we find that all coefficients are statistically significant. Also, the coefficient of CRISs becomes much larger and the coefficient of the interaction term is negative. It implies that CRISs have a larger impact on employability in lower-wage or low-skilled jobs.

Column 7 is our most saturated specification where we include all sets of controls and occupation fixed effects. Within occupations, the estimates suggest that CRISs have a larger effect on employability in lower-wage or low-skilled jobs. It may indicate that CRISs are somehow more important to secure low-wage or low-skilled jobs than high-wage or high-skilled jobs. Meanwhile, compared to Table 7, the effect of a one standard deviation increase in CRISs is much larger than the return on over-education and over-experience when we control offered wages.

Summing up, Table 6 highlights the importance of CRISs in job matching and the effect of CRISs is more sensitive to low-wage or low-skilled jobs.

C. Effects of An Industry-Specific Labor Demand Shock

So far, we have investigated the effect of CRISs on callback probabilities and find the important role of CRISs in job search. In this section, we further apply

LMNInd along with CRISs in a simulation analysis to study the effect of an industry-specific labor demand shock.

Specifically, our investigation focuses on two scopes, such as micro-level and macro-level effect. On one hand, from micro-level view, we want to investigate what industries for those individuals who previously applied jobs in the affected industry will apply. Also, in terms of employment, we want to find the change of employment in each industry before and after the industry-specific labor demand shock. On the other hand, from macro-level view, we test whether there exist multiplier effects in unemployment. In other words, we test if the change of unemployment rate is more than the change of number of jobs due to the industry-specific labor demand shock.

To study the effect of an industry-specific labor demand shock, we simulate how the labor flow across industries affected by an exogenous industry-specific shock based on LMNInd and CRISs. We start the simulation analysis by supposing that there is an industry-specific labor demand shock, say trade industry has been affected. Then, we follow five steps among three periods to simulate the effect of an industry shock.

First, in the pre-shock period, step 1 is to find the distribution of applications across industries. It means that among all 3,618,944 applications before the industry shock, we treat each application as one unit. We then assign average weights to the job's industries within each application. In other words, we assign $1/n$ to the industry if the job's industry information in the application state n industries. Therefore, we are able to aggerate the distribution of applications across industries.

Step 2 is to find the distribution of the predicted callbacks across industries. To do so, we use the same specification as column 7 in Table 5 to predict callback probabilities. Then, we similarly assign average weights to the job's industries within each application. This allows to calculate the aggregate share of the predicted callbacks across industries.

Second, in the being-shocked period, step 3 is to randomly select one industry as being affected (e.g., trade industry) and then find the distribution of applications across industries. Specifically, we randomly remove 10/30/50 percent of job ads containing the affected industry (treated job ads). This way is plausible because more than half of job ads have multiple industries, so we are not restricted to eliminate job ads containing only the affected industry. Then, we apply similar approach as step 1 to find the distribution of applications across industries in the simulation of the industry shock.

Third, in the post-shock period, step 4 is to simulate new applications for applicants whose applications have been removed due to the industry shock and then find the distribution of applications across industries. Specifically, after removing treated job ads, we keep jobs for each affected applicant based on matching characteristics in applications as Table 3 illustrated, such as proper education and age. Then, we randomly keep twice of jobs than the number of removed jobs of each applicant. In other words, if the number of removed jobs of an applicant is n, then we randomly keep $2n$ jobs for the applicant and we call these jobs as candidate jobs. Further, we select top n jobs for each applicant based on vacancy-applicant skill match index between the industry in which jobs the applicant applied before the industry shock and the industry of candidate jobs. This process allows to maximize the possibility of simulating affected applicants' application behavior, because affected applicants would apply jobs in which CRISs are significant.

Step 5 is to find the distribution of the predicted callbacks across industries. After simulating new jobs for affected applicants, we apply similar way of step 2 to calculate the aggregate share of the predicted callbacks across industries.

After these five steps among three periods, we then investigate the effect of an industry-specific labor demand shock. We present the simulation results in Figure 3. Panel A of Figure 3 plots the difference of share of callbacks against the difference of share of applications across industries. The correlation is 99 percent which shows that the difference of share of callbacks is strongly correlated with the difference of share of applications across industries. Therefore, the reduction (increment) of callbacks in each industry is strongly associated with the decrease (increase) of applications from that industry.

Second, from the legend of Panel A of Figure 3, the number in the bracket is the rank of skill match index to the affected industry. For example, trade industry is ranked as first highest industry because it is the affect industry itself as an example; and apparel industry is ranked as fourth highest industry, and so on. There are two main takeaways. On one hand, we find that industries which have high skill match index to trade industry are negatively affected the most, such as apparel (ranked as 4th), wholesale/retail (ranked as 7th), and furniture/appliance industries (ranked as 13th). Due to their high skill match index to the affected industry, their share of application decreases and therefore their employability indicator, callback rates, are negatively affected. On the other hand, we find that industries which have higher skill match index to trade industry are not necessarily affected in a higher degree. The wholesale/retail industry are affected more than apparel industry, although wholesale/retail industry have lower skill match index to trade industry than apparel industry. This is striking and motivates to investigate how the share of applications changes due to the industry shock and the simulation of new applications in Panel B of Figure 3.

From Panel B of Figure 3, we find that due to the industry shock, those industries with high skill match index to the affected industry face significant job loss, such as apparel (ranked as 4th), wholesale/retail (ranked as 7th), and furniture/appliance industries (ranked as 13th). So, the relative labor demand of these industries is declining. In contrast, for industries with low skill match index to the affected industry, such as machinery (ranked as 25th), computer software (ranked as 39th), and electronic technology industries (ranked as 28th), their relative labor demand is conversely increasing.

On the other hand, we find that from simulation of new applications, these industries with higher skill match index to the affected industry absorb the job loss due to the industry shock. Therefore, their relative labor demand increases due to simulated applications. In other words, those labor forces who initially aiming to work in affected industry are very likely to apply and find jobs in these industries with higher skill match index to the affected industry.

More importantly, we find that industries response differently to the industry shock and simulated applications because this simulation exercise incorporates the nature of LMNInd. Figure 3 highlights the potential importance of LMNInd. Specifically, from labor economics' point of view, skills in which industries require are not distributed equally in universe skill space. Thus, some groups of industries have more common skill requirements than other groups of industries. Also, industries are not isolated islands but have inherent linkages through CRISs, although this important feature is not considered in general, specific, and task-based skills (which usually summarized by jobs). Therefore, LMNInd captures this important idea, and the simulation analysis also provides evidence on it.

In our simulation analysis, the labor supply is a decision based on expectations of the demand side. For example, when trade industry is affected, workers who worked in trade industry are more likely to apply job in apparel industry and less likely to apply in media industry now. Although workers decide on their own judgement and opinions, the research makes workers' implicit factors (i.e., judgement) to be incorporated and explained explicitly by LMNInd and cross-industry skills.

Next, we investigate if there is a multiplier effect of an industry-specific demand shock. Again, we first use trade industry as an example to illustrate in Table 7. Row 1 presents the number of total job ads is 120,073. Row 2 indicates that the number of job ads containing the affected industry is 18,308. Row 3 shows the size of shock regarding affected job ads is 50 percent, and therefore, 9,154 job ads (= 120,073 * 50 percent) containing the affected industry are removed in the simulation analysis shown in row 4. Then, row 5 calculates the size of the shock regarding total job ads, which is 7.62 percent (= 9,154/120,073). Row 6 indicates that the ratio of post-shock callbacks to pre-shock callbacks is 91.69 percent, and thus, the size of shock regarding total callbacks is 8.31 percent shown in row 7 (= 1 - 91.69 percent). To the end, row 8 presents there is a multiplier effect of an industry-specific demand shock, which is 108.96 percent (= 8.31 percent / 7.62 percent). It shows that the size of shock regarding total callbacks is more than the size of the shock regarding total job ads.

Table 8 summarizes multiplier effect of an industry-specific demand shock for three industries respectively (such as construction, trade, and internet/e-commerce industries) with respect to different sizes of shock regarding affected job ads. It provides evidence that in responding to different sizes of shock regarding affected job ads, there are different magnitudes of the multiplier effect. In other words, with different size of shock, different industries would lead to different size of multiplier effect. In all cases, an industry-specific labor demand shock leads multiplier effect larger than 100 percent, except two scenarios, such as 10 and 50 percent shock regarding affected job ads in internet/e-commerce industry.

Therefore, we find that there exists multiplier effect of an industry-specific demand shock, because the size of shock regarding total callbacks is more than size of the shock regarding total job ads. Also, the multiplier effect is mostly larger than 100 percent, and thus, in light of LMNInd, the aggregate effect highlights the spillover of an industry-specific demand shock on other industries.

Taken together, what this simulation analysis shows is that when there is an exogenous industry-specific demand shock, it is not reasonable to assume other industries are not affected and thus used as the control group in purpose. However, industries response differently to the industry shock due to LMNInd and CRISs. It might be of interest to take advantage of these large variations in the magnitude of the effect and treats them as instrument variables in further research. Moreover, there are multiplier effects of an industry-specific demand shock.

It also sheds light on the importance of LMNInd. Because with LMNInd, we now are able to investigate, if there is an industry-specific demand shock, then whether other industries are affected as well. Also, with LMNInd, we can examine the magnitude of the effect on all industries and test whether the effect of all industries is the same.

V. Discussion

In this paper, we propose a concept, cross-industry skills (CRISs), to measure professional human capital applied in multiple industries. The distinction among CRISs, industry-specific skills, and general skills is intuitive and straightforward. General skills such as communication skills are useful in all industries, while industry-specific skills are useful in a specific industry such as mining skills applied in mining industry only. Yet, it is possible for some professional skills to be useful in multiple industries and we propose them as CRISs. In other words, this kind of transferable professional skills increases marginal productivity of workers in multiple industries. Therefore, CRISs play an intermediate role between industry-specific skills and general skills.

CRISs have a critical implication on labor mobility. Under the context of CRISs, workers do not necessarily need the same industry-specific skills as the firms' industries. In other words, workers acquire CRISs even when they have completely different industry skills comparing with firms' industries. Alternatively, when a worker works in a given industry, she not only accumulates professional skills in this industry but also in other industries. Because of LMNInd, some professional skills are also applied in different industries. In our example, job seekers accumulate data analysis skills when they worked in either finance or trade industries. Since finance and trade firms require data analysis skills, they will be likely to recruit hybrid employees with such CRISs.

Workers' industry experience plays an important role in job search and matching. To some extent, whenever they have unrelated/related industry experience to the expected firms, they will have penalties/premium on employability or earnings. In other words, as LMNInd highlighted, there might be positive/negative spillover effects in individuals' career decisions in terms of industries. It has an important implication that workers should carefully consider the future career when choose to switch industries. Indeed, with an increasing number of firms engaged in multi-industry operations to pursue industrial diversity and rising mobility of workers across industries, CRISs play an important role in labor market.

This paper provides a new understanding of unemployment and employment across industries. CRISs enable workers to move away from downside industries and into emerging industries. During economic transformation, some industries have a downturn in the economy, and it often occurs with substantial unemployment. If there is LMNInd and CRISs between such declining industry with other emerging industries, unemployed workers are likely to find jobs in the developing industry.

Moreover, LMNInd from firms' perspectives also shed light on the structural transformation of the economy. Although our data in short term captures relative static LMNInd, the dynamic change in the long term would be used to investigate the structural or industrial development and it is the importance of labor mobility.

References

Becker, G. S. (2009). *Human capital: A theoretical and empirical analysis, with special reference to education.* University of Chicago press.

Gathmann, C., & Schönberg, U. (2010). How general is human capital? A task-based approach. *Journal of Labor Economics*, 28(1), 1-49.

Gibbons, R., & Waldman, M. (2004). Task-specific human capital. *American Economic Review*, 94(2), 203-207.

Kambourov, G., & Manovskii, I. (2008). Rising occupational and industry mobility in the United States: 1968–97. *International Economic Review*, 49(1), 41-79.

Kambourov, G., & Manovskii, I. (2009). Occupational specificity of human capital. *International Economic Review*, 50(1), 63-115.

Kuhn, P., Shen, K., & Zhang, S. (2020). Gender-targeted job ads in the recruitment process: Facts from a Chinese job board. *Journal of Development Economics*, 147, 102531.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Marinescu, I., & Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, 38(2), 535-568.

Neal, D. (1995). Industry-specific human capital: Evidence from displaced workers. *Journal of labor Economics*, 13(4), 653-677.

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3), 297-316.

Robinson, C. (2018). Occupational mobility, occupation distance, and specific human capital. *Journal of Human Resources*, 53(2), 513-551.

Sullivan, P. (2010). Empirical evidence on occupation and industry specific human capital. *Labour economics*, 17(3), 567-580.

Tables and Figures

Table 1: Summary Statistics of Job Characteristics

	Number of industries			Total
	1	2	3 or more	
Panel A. Job requirements				
Education				
Education required?	0.940	0.941	0.938	0.940
Education requirement (years)	13.92	14.09	14.29	14.09
Education not required	0.060	0.059	0.062	0.060
High school or below required	0.169	0.149	0.113	0.146
Tech school required	0.144	0.130	0.124	0.134
College required	0.486	0.515	0.539	0.510
Undergraduate or above required	0.142	0.148	0.161	0.150
Age				
Age required?	0.491	0.478	0.450	0.474
Age requirement	29.99	29.81	29.41	29.77
Experience				
Experience required?	0.509	0.524	0.513	0.514
Experience requirement (years)	2.363	2.345	2.282	2.332
New graduates required?	0.020	0.022	0.020	0.021
Wage				
Explicit offered wage?	0.803	0.805	0.851	0.819
Wage offered (Yuan/month)	5,615	5,732	5,833	5,717
Explicit offered bonus?	0.175	0.210	0.251	0.208
Bonus offered (Yuan/month)	6,895	6,941	7,253	7,049
Explicit offered wage and bonus?	0.144	0.173	0.216	0.175
Wage and bonus offered (Yuan/month)	13,016	13,181	13,241	13,146
Panel B. Job and firm characteristics				
Explicit vacancy numbers?	0.962	0.960	0.960	0.961
Number of vacancies	2.309	2.276	2.438	2.344
Firm size (number of workers)	549.8	340.9	354.9	437.1
Firm ownership type				
Private, domestic	0.888	0.919	0.941	0.913
Foreign	0.029	0.018	0.015	0.022
State-owned enterprise	0.084	0.062	0.043	0.065
Number of job ads	52,638	27,395	40,040	120,073
Share of job ads	43.84	22.82	33.35	100.00

Notes: This table presents descriptive statistics of job characteristics. The layout of columns follows number of industries in which job involved: "1" indicates that one industry is involved in the jobs; "2" indicates that two industries is involved in the jobs; "3 or more" indicates that three to seven industries are involved in the jobs.

Table 2: Summary Statistics of Applicants' Characteristics

	Number of industries			Total
	1	2	3 or more	
Education				
Education stated?	0.998	0.998	0.998	0.998
Education (years)	14.44	14.71	14.71	14.58
Education not stated	0.002	0.002	0.002	0.002
High school or below stated	0.161	0.114	0.103	0.135
Tech school stated	0.116	0.107	0.119	0.114
College stated	0.345	0.384	0.402	0.369
Undergraduate or above stated	0.376	0.394	0.374	0.381
Age	28.26	29.60	31.75	29.39
Experience (years)	7.565	8.918	11.29	8.740
New graduate?	0.001	0.001	0.000	0.001
Wage				
Current wage stated?	0.793	0.800	0.802	0.797
Current wage (Yuan/month)	5,937	6,326	6,676	6,207
Intended wage stated?	0.640	0.662	0.663	0.651
Intended wage (Yuan/month)	6,375	6,750	7,129	6,648
Female	0.427	0.474	0.488	0.454
Married	0.398	0.485	0.589	0.464
Myopic	0.321	0.335	0.336	0.328
Height stated?	0.929	0.945	0.961	0.940
Height (cm)	167.4	166.8	166.5	167.0
Photo flags available?	0.334	0.422	0.488	0.392
Number of applicants	121,742	73,796	51,028	246,566
Share of applicants	49.38	29.93	20.70	100.00

Notes: This table presents descriptive statistics of applicants' characteristics. The layout of columns follows number of industries in which job involved: "1" indicates that one industry is involved in the jobs; "2" indicates that two industries is involved in the jobs; "3 or more" indicates that three to seven industries are involved in the jobs.

Table 3: Summary Statistics of Matching Characteristics in Applications

	Vacancy-Applicant skill match index			Total
	Low	Mid	High	
Called back?	0.183	0.188	0.198	0.190
Education match required?				
Related data missing	0.055	0.051	0.052	0.053
Less educated than required	0.142	0.151	0.149	0.147
Education proper	0.418	0.413	0.418	0.416
More educated than required	0.385	0.385	0.381	0.384
Experience match required?				
Related data missing	0.469	0.433	0.426	0.443
Less experienced than required	0.025	0.019	0.022	0.022
Experience proper	0.030	0.024	0.029	0.028
More experience than required	0.475	0.524	0.522	0.507
Age match required?				
Younger than required	0.044	0.037	0.036	0.039
Age proper	0.880	0.879	0.895	0.884
Older than required	0.076	0.084	0.070	0.077
Graduation status match required?				
Fresh grad & fresh grad required	0.000	0.000	0.000	0.000
Former grad & fresh grad required	0.019	0.016	0.018	0.018
Fresh grad & former grad required	0.000	0.000	0.000	0.000
Former grad & former grad required	0.980	0.983	0.982	0.982
Number of applications	1,206,315	1,206,334	1,206,295	3,618,944
Share of applications	33.33	33.33	33.33	100.00

Notes: This table presents descriptive statistics of matching characteristics in applications. Vacancy-Applicant skill match index has been divided into three groups (i.e., low, medium, and high) based on three terciles.

Table 4: Industry Clusters

Cluster	Total share	Share of top 5 within cluster	Top 1	Share	Top 2	Share	Top 3	Share	Top 4	Share	Top 5	Share
Panel A: Firms												
1	21.82%	90.75%	Trade	8.68%	Wholesale/Retail	4.45%	Apparel	2.80%	Furniture/Appliance	1.96%	Materials Processing	1.92%
2	19.72%	94.34%	Internet/E-Commerce	10.00%	Computer Software	3.98%	Computer Service	2.67%	Telecom Equipment	1.14%	Computer Hardware	0.80%
3	18.81%	84.76%	Construction	5.29%	Other	5.26%	Transportation Service	2.16%	Room Decoration	1.94%	Environmental Protection	1.30%
4	10.93%	100.00%	Machinery	4.88%	Electronic Technology	3.76%	Industrial Automation	1.71%	Electric/Water	0.57%		
5	10.19%	95.08%	Education	3.58%	Finance	2.86%	Consulting/Human Resource	1.62%	Accounting	1.00%	Intermediary Service	0.63%
6	8.10%	88.55%	Fast-Moving Consumer Goods	2.80%	Diversified Business Group	1.33%	Real Estate	1.17%	Catering	1.08%	Domestic Service	0.80%
7	6.06%	89.61%	Hotels/Tourism	1.48%	Medical Care/Nursing	1.12%	Property Management	1.10%	Medical Equipment	0.95%	Pharmacy	0.77%
8	4.37%	100.00%	Media/Art	1.23%	Advertising	1.20%	Marketing	0.83%	Sports/Recreation	0.82%	Publishing	0.29%
Panel B: Workers												
1	25.55%	84.47%	Electronic Technology	8.02%	Machinery	5.93%	Automobile	3.14%	Furniture/Appliance	2.31%	Materials Processing	2.18%
2	24.32%	89.72%	Other	6.99%	Trade	4.96%	Apparel	4.27%	Wholesale/Retail	3.13%	Transportation Service	2.48%
3	12.72%	90.91%	Internet/E-Commerce	5.58%	Computer Software	3.12%	Telecom Equipment	1.38%	Computer Service	0.87%	Telecom Operation	0.61%
4	12.28%	100.00%	Construction	5.66%	Fast-Moving Consumer Goods	3.52%	Real Estate	1.66%	Diversified Business Group	0.90%	Agriculture/Forestry/Fishing	0.54%
5	10.22%	98.45%	Education	3.55%	Hotels/Tourism	2.56%	Catering	1.92%	Property Management	1.29%	Sports/Recreation	0.73%
6	6.07%	84.77%	Finance	2.47%	Government	0.72%	Bank	0.66%	Insurance	0.65%	Intermediary Service	0.65%
7	4.62%	93.89%	Medical Care/Nursing	1.73%	Oil/Chemical/Mineral	1.00%	Pharmacy	0.66%	Environmental Protection	0.57%	Detection/Certification	0.38%
8	4.22%	100.00%	Room Decoration	1.82%	Media/Art	1.05%	Advertising	0.69%	Marketing	0.46%	Publishing	0.20%
Panel C: Applications												
1	29.09%	91.14%	Electronic Technology	8.14%	Internet/E-Commerce	7.72%	Machinery	5.65%	Furniture/Appliance	2.53%	Automobile	2.47%
2	28.54%	78.21%	Trade	6.48%	Other	6.33%	Wholesale/Retail	3.91%	Fast-Moving Consumer Goods	3.55%	Transportation Service	2.06%
3	10.41%	98.87%	Construction	4.32%	Hotels/Tourism	1.87%	Room Decoration	1.54%	Real Estate	1.31%	Property Management	1.25%
4	8.57%	88.69%	Education	2.49%	Finance	2.19%	Accounting	1.39%	Consulting/Human Resource	1.00%	Intermediary Service	0.52%
5	8.40%	93.42%	Computer Software	3.42%	Computer Service	1.78%	Telecom Equipment	1.39%	Computer Hardware	0.70%	Telecom Operation	0.56%
6	7.73%	94.89%	Apparel	3.82%	Medical Care/Nursing	1.26%	Medical Equipment	1.04%	Pharmacy	0.79%	Beauty/Health Care	0.42%
7	4.97%	96.08%	Industrial Automation	1.78%	Environmental Protection	1.02%	Oil/Chemical/Mineral	0.94%	Electric/Water	0.69%	Detection/Certification	0.34%
8	2.31%	100.00%	Media/Art	0.86%	Advertising	0.75%	Marketing	0.52%	Publishing	0.17%		

Notes: This table presents industry clusters in three panels from firms', workers', and applications' perspective, respectively. The column "Total share" shows the share of each cluster within its corresponding panel. The column "Share of top 5 within cluster" shows the share of top 5 industries within its corresponding cluster. Within each cluster, top 5 industries and their share are presented.

Table 5: Effects of Vacancy-Applicant Skill Match Index on Callback Probabilities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Vacancy-Applicant skill match index from ML	0.0064***	0.0075***	0.0077***	0.0076***	0.0051***	0.0052***	0.0065***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0004)
Education less than requested				-0.0087***	-0.0077***	-0.0068***	-0.0108***
				(0.0010)	(0.0010)	(0.0010)	(0.0008)
Education more than requested				0.0041***	0.0019**	-0.0001	0.0004
				(0.0009)	(0.0009)	(0.0008)	(0.0007)
Experience less than requested				-0.0041*	-0.0067***	-0.0055**	-0.0084***
				(0.0022)	(0.0022)	(0.0022)	(0.0020)
Experience more than requested				0.0039**	0.0044**	0.0040**	0.0029*
				(0.0018)	(0.0017)	(0.0017)	(0.0015)
Job Requirement Controls		Y	Y	Y	Y	Y	Y
Detailed CV Controls			Y	Y	Y	Y	Y
Vacancy-Applicant Matching Controls				Y	Y	Y	Y
Competition Controls					Y	Y	Y
Occupation Fixed Effects						Y	Y
Job Title Fixed Effects							Y
N (Applications)	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944
R2	0.0003	0.0043	0.0056	0.0065	0.0153	0.0176	0.1211

Standard errors (in parentheses) are clustered at the job level. ***p < 0.01, **p < 0.05, *p < 0.10.

Notes: In addition to the covariates shown, columns 2—7 include "Job Requirement Controls": education requirement (5 categories), experience requirement (quadratic), age requirement (quadratic), an indicator for missing experience requirement, and an indicator for missing age requirement. Columns 3—7 include "Detailed CV Controls": a dummy for whether the worker married, a dummy for whether the worker is myopic, a dummy for whether the worker has photo flags, the workers' height, and an indicator for missing height. Columns 4—7 include "Vacancy-Applicant Matching Controls": matching status of age (4 categories: three dummies for whether the worker's age is less than/match with/more than the requested age, respectively, and an indication for missing age-related information), four dummies for whether the worker's new graduate status matches the requested status (new/non-fresh graduate interact with requested new/non-fresh graduate), an indicator for missing education information, and an indicator for missing experience information. Columns 5—7 include "Competition Controls": the number of applications received by the job, the number of positions advertised, and an indicator for missing the number of positions advertised. Omitted or reference groups are as follows: for the matching status of education, education matches requested; for the matching status of experience, experience matches requested. Occupation fixed effects control for the 70 categories used on the XMRC website. Job title fixed effects control for categories of job titles.

Table 6: Effects of Vacancy-Applicant Skill Match Index and Offered Wage on Callback Probabilities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vacancy-Applicant skill match index from ML	0.0085***	0.0368***	0.0466***	0.0471***	0.0478***	0.0631***	0.0457***	0.0336***
	(0.0006)	(0.0136)	(0.0134)	(0.0134)	(0.0134)	(0.0129)	(0.0128)	(0.0105)
Log (offered wage)	-0.0443***	-0.0441***	-0.0147***	-0.0101***	-0.0109***	-0.0055*	-0.0092***	-0.0106**
	(0.0027)	(0.0027)	(0.0033)	(0.0033)	(0.0033)	(0.0032)	(0.0034)	(0.0053)
Vacancy-Applicant skill match index from ML * Log (offered wage)		-0.0033**	-0.0044***	-0.0045***	-0.0046***	-0.0067***	-0.0047***	-0.0032***
		(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0015)	(0.0015)	(0.0012)
Education less than requested					-0.0092***	-0.0078***	-0.0073***	-0.0111***
					(0.0012)	(0.0012)	(0.0011)	(0.0009)
Education more than requested					0.0039***	0.0012	-0.0006	-0.0005
					(0.0010)	(0.0009)	(0.0009)	(0.0008)
Experience less than requested					-0.0043*	-0.0067***	-0.0056**	-0.0085***
					(0.0025)	(0.0024)	(0.0024)	(0.0022)
Experience more than requested					0.0047**	0.0054***	0.0051***	0.0042**
					(0.0019)	(0.0019)	(0.0019)	(0.0017)
Job Requirement Controls			Y	Y	Y	Y	Y	Y
Detailed CV Controls				Y	Y	Y	Y	Y
Vacancy-Applicant Matching Controls					Y	Y	Y	Y
Competition Controls						Y	Y	Y
Occupation Fixed Effects							Y	Y
Job Title Fixed Effects								Y
N (Applications)	2,965,566	2,965,566	2,965,566	2,965,566	2,965,566	2,965,566	2,965,566	2,965,566
R2	0.0022	0.0022	0.0047	0.0059	0.0069	0.0155	0.0181	0.1235

Standard errors (in parentheses) are clustered at the job level. ***p < 0.01, **p < 0.05, *p < 0.10.

Notes: In addition to the covariates shown, columns 3—8 include “Job Requirement Controls”: education requirement (5 categories), experience requirement (quadratic), age requirement (quadratic), an indicator for missing experience requirement, and an indicator for missing age requirement. Columns 4—8 include “Detailed CV Controls”: a dummy for whether the worker married, a dummy for whether the worker is myopic, a dummy for whether the worker has photo flags, the workers’ height, and an indicator for missing height. Columns 5—8 include “Vacancy-Applicant Matching Controls”: matching status of age (4 categories: three dummies for whether the worker’s age is less than/match with/more than the requested age, respectively, and an indication for missing age-related information), four dummies for whether the worker's new graduate status matches the requested status (new/non-fresh graduate interact with requested new/non-fresh graduate), an indicator for missing education information, and an indicator for missing experience information. Columns 6—8 include "Competition Controls": the number of applications received by the job, the number of positions advertised, and an indicator for missing the number of positions advertised. Omitted or reference groups are as follows: for the matching status of education, education matches requested; for the matching status of experience, experience matches requested. Occupation fixed effects control for the 70 categories used on the XMRC website. Job title fixed effects control for categories of job titles.

Table 7: Multiplier Effect of An Industry-Specific Demand Shock

(1)	Total job ads		120,073
(2)	Job ads containing the affected industry		18,308
(3)	Size of shock regarding affected job ads		50%
(4)	Job ads removed due to the industry shock	= (2) * (3)	9,154
(5)	Size of the shock regarding total job ads	= (4) / (1)	7.62%
(6)	Ratio of post-shock callbacks to pre-shock callbacks		91.69%
(7)	Size of shock regarding total callbacks	= 1 - (6)	8.31%
(8)	Multiplier effect of the one-industry shock	= (7) / (5)	108.96%

Notes: This table presents how to measure multiplier effect of an industry-specific demand shock based on vacancy-applicant skill match index. In this table, the shock to trade industry is as an example of an industry-specific demand shock. The multiplier effect in row 8 is larger than one, indicating that the size of shock regarding total callbacks is more than size of the shock regarding total job ads. This aggregate effect highlights the spillover of an industry-specific demand shock on other industries.

Table 8: Summary of Multiplier Effect of An Industry-Specific Demand Shock

Affected industry	(1) Size of shock regarding affected job ads	(2) Size of shock regarding total job ads	(3) Size of shock regarding total callbacks	(4) Multiplier effect
construction	10%	0.77%	0.95%	123.38%
	30%	2.31%	2.53%	109.74%
	50%	3.84%	4.25%	110.58%
trade	10%	1.52%	1.70%	111.78%
	30%	4.57%	4.99%	109.00%
	50%	7.62%	8.31%	108.96%
internet/e-commerce	10%	2.38%	2.32%	97.49%
	30%	7.13%	7.16%	100.41%
	50%	11.89%	11.82%	99.47%

Notes: This table presents the summary of the multiplier effect of an industry-specific demand shock based on vacancy-applicant skill match index. In this table, the shock to construction, trade, and internet/e-commerce industry, respectively, is as an example of an industry-specific demand shock. If the multiplier effect in column 4 is larger than one, then the size of shock regarding total callbacks is more than size of the shock regarding total job ads. This aggregate effect highlights the spillover of an industry-specific demand shock on other industries.

Figure 1: Illustration of General, Industry-Specific, and Cross-Industry Skills

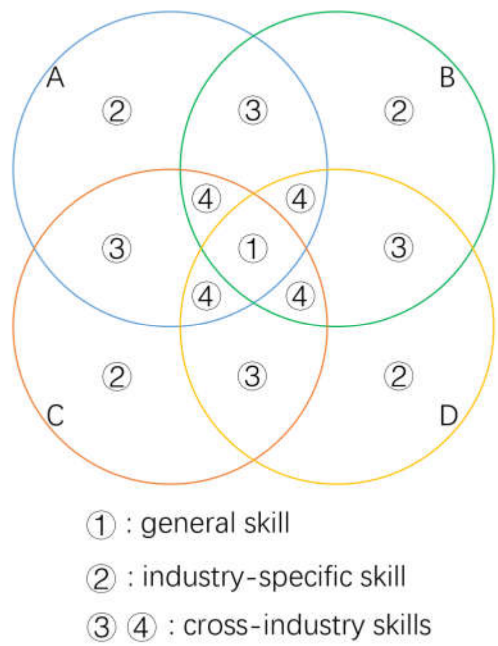
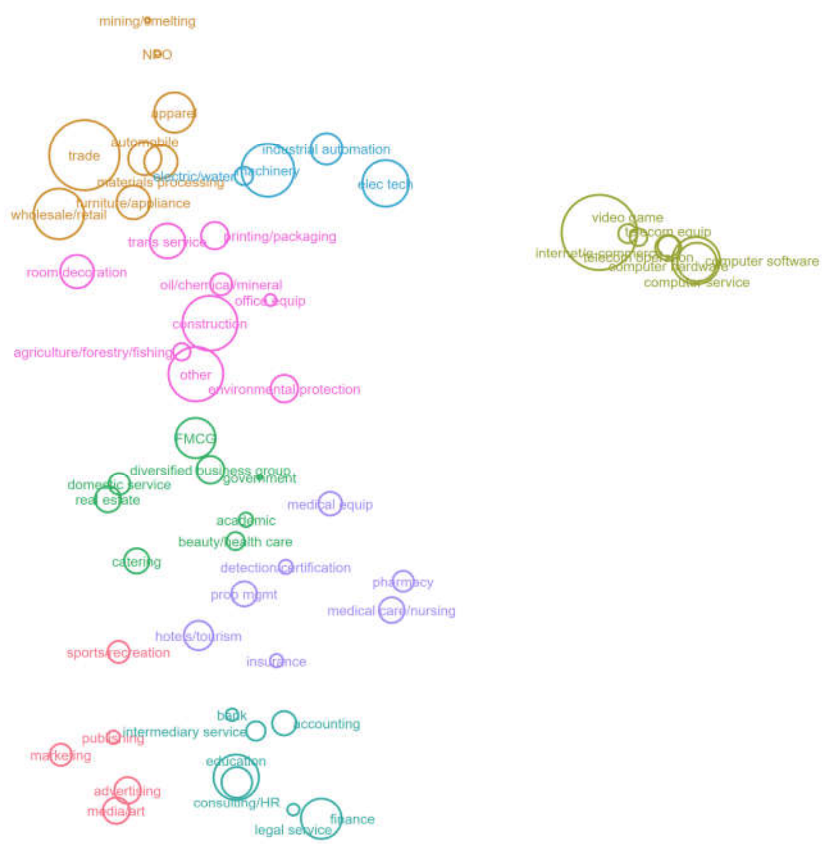
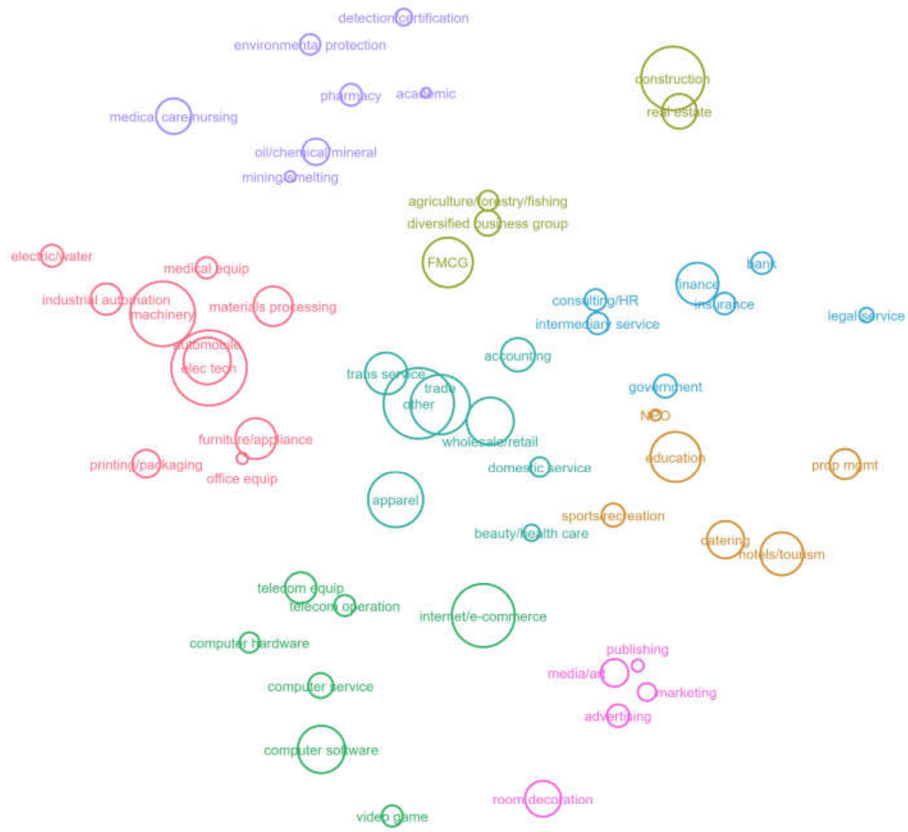


Figure 2: LMNInd

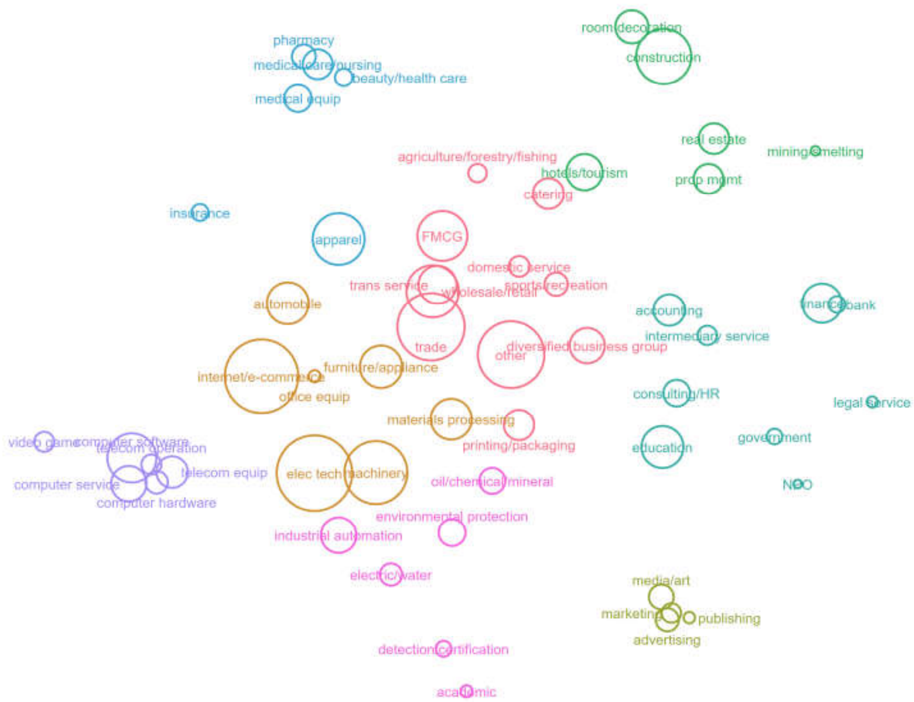
Panel A: LMNInd from Industries of Firms



Panel B: LMNInd from Industries of Workers



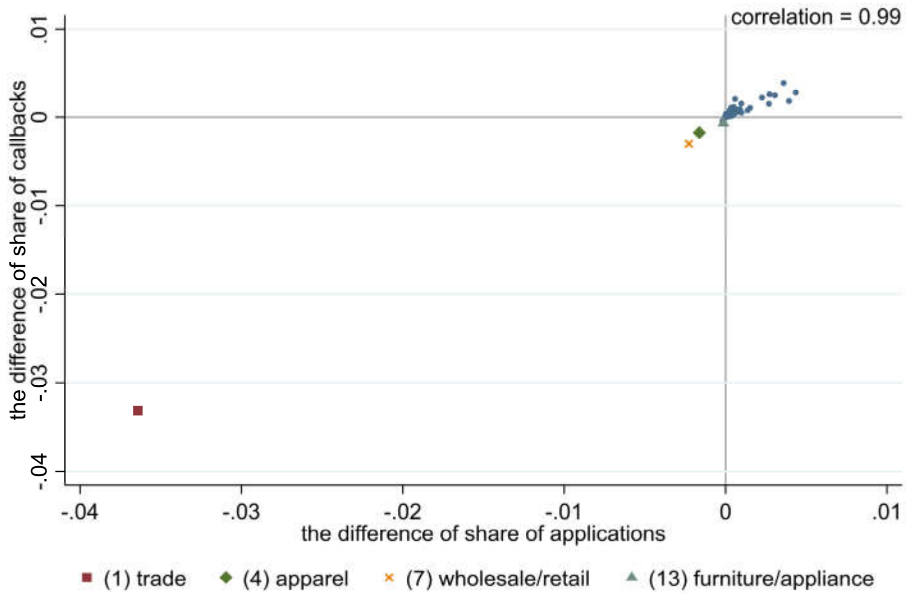
Panel C: LMNInd from Industries of Applications



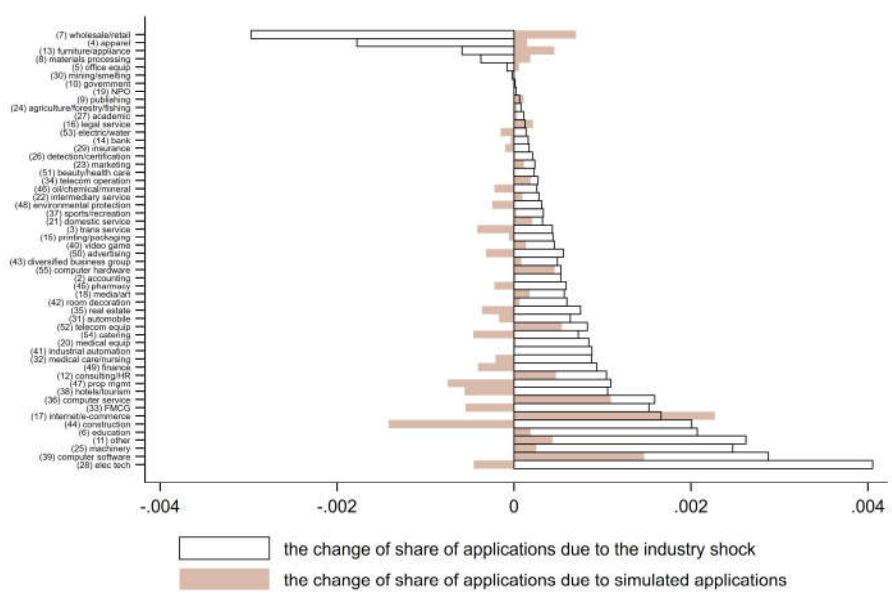
Notes: In terms of the design of the figure, the size of the circle represents the weighted sample size of industries in their own perspective sample. It is clear to examine which industries are most observed in the three samples, respectively. The color depicts industry clustering information. In three panels, we create eight clusters by k-means methodology to possibly capture the group of industries. The distance between industries implies the relatedness of industries as well as the magnitude of CRISs. Intuitively, the closer the two industries are in LMNInd, the more related the two industries are, and the higher the CRISs between the two industries. As a result, the industries within a cluster present relatively higher CRISs.

Figure 3: Effects of An Industry-Specific Demand Shock

Panel A: Effects of An Industry-Specific Demand Shock



Panel B: Industries Response Differently



Notes: Panel A of Figure 3 plots the difference of share of callbacks against the difference of share of applications across industries. The correlation is 99 percent which shows that the difference of share of callbacks is strongly correlated with the difference of share of applications across industries. In the legend, the number in the bracket is the rank of skill match index to the affected industry. For example, trade industry is ranked as first highest industry because it is the affect industry itself as an example; and apparel industry is ranked as fourth highest industry, and so on. From Panel B of Figure 3, we find that due to the industry shock, those industries with high skill match index to the affected industry face significant job loss, such as apparel (ranked as 4th), wholesale/retail (ranked as 7th), and furniture/appliance industries (ranked as 13th). So, the relative labor demand of these industries is declining. In contrast, for industries with low skill match index to the affected industry, such as machinery (ranked as 25th), computer software (ranked as 39th), and electronic technology industries (ranked as 28th), their relative labor demand is conversely increasing.

Appendix 1: Industry List

In this section, we present industry list and the share of industries from the weighted workers' industry experience in Table A1. In terms of the share of industries, we treat each applicant's industry experience record as one unit and assign average weights to each industry experience within the worker's industry experience record. In other words, we assign $1/n$ to the industry if the worker's industry experience record lists n industries.

From Table A1, 55 distinct industries are presented. The share of industries captures the distribution of workers' industry experience. Specifically, the top 5 workers' industry experience are electronic technology, other, machinery, construction, and internet/e-commerce industries. The bottom 5 workers' industry experience are publishing, office equipment, NPO, mining/smelting, and academic industries.

Table A1: Industry List

No.	Industry	Share	No.	Industry	Share
1	Electronic Technology	8.02%	29	Diversified Business Group	0.90%
2	Other	6.99%	30	Computer Service	0.87%
3	Machinery	5.93%	31	Sports/Recreation	0.73%
4	Construction	5.66%	32	Government	0.72%
5	Internet/E-Commerce	5.58%	33	Electric/Water	0.71%
6	Trade	4.96%	34	Advertising	0.69%
7	Apparel	4.27%	35	Bank	0.66%
8	Education	3.55%	36	Pharmacy	0.66%
9	Fast-Moving Consumer Goods	3.52%	37	Insurance	0.65%
10	Automobile	3.14%	38	Intermediary Service	0.65%
11	Wholesale/Retail	3.13%	39	Consulting/Human Resource	0.64%
12	Computer Software	3.12%	40	Medical Equipment	0.62%
13	Hotels/Tourism	2.56%	41	Telecom Operation	0.61%
14	Transportation Service	2.48%	42	Video Game	0.60%
15	Finance	2.47%	43	Environmental Protection	0.57%
16	Furniture/Appliance	2.31%	44	Computer Hardware	0.55%
17	Materials Processing	2.18%	45	Agriculture/Forestry/Fishing	0.54%
18	Catering	1.92%	46	Domestic Service	0.54%
19	Room Decoration	1.82%	47	Marketing	0.46%
20	Medical Care/Nursing	1.73%	48	Detection/Certification	0.38%
21	Real Estate	1.66%	49	Beauty/Health Care	0.36%
22	Accounting	1.60%	50	Legal Service	0.29%
23	Telecom Equipment	1.38%	51	Publishing	0.20%
24	Industrial Automation	1.37%	52	Office Equipment	0.17%
25	Property Management	1.29%	53	Non-Profit Organization	0.16%
26	Printing/Packaging	1.09%	54	Mining/Smelting	0.16%
27	Media/Art	1.05%	55	Academic	0.13%
28	Oil/Chemical/Mineral	1.00%			

Notes: The table presents the industry list and the share of industries from the weighted workers' industry experience. Specifically, we treat each applicant's industry experience record as one unit and assign average weights to each industry experience within the worker's industry experience record. In other words, we assign $1/n$ to the industry if the worker's industry experience record lists n industries.

Appendix 2: Effect of Jobs' Skill Demands on the Number of Industries in which Jobs are Involved

In this section, we examine the association between the number of industries in which jobs are involved and jobs' skill demands. Table A2 presents the effect of jobs' skill demands on the number of industries in which jobs are involved.

In Table A2, column 1 estimates the effect of education requirements on the number of industries in which jobs are involved. Column 2 adds the control for the experience requirement. Column 3 further controls for firm ownership type. Column 4 is the same as column 3 except using an explicitly offered wage sample. Column 5 controls for offered wage and adds occupation fixed effects.

Column 3 in Table A2 highlights the effect of education requirements on the number of industries in which jobs are involved. The estimated results indicate that higher educational attainment required, the more industries in which jobs are involved. Specifically, jobs requiring undergraduate or above are associated with 0.13 more industries which jobs are involved than jobs with missing education requirements. Further, the magnitude is striking that it is more than 5 times on jobs requiring tech school.

Column 5 implies jobs requiring undergraduate or above are associated with 0.12 more industries which jobs are involved. Similarly, the economic magnitude is more than 4 times on jobs requiring tech school.

In addition, a one percent increase in offered wage is associated with a 5.05 percent increase in the number of industries in which jobs are involved. Furthermore, private firms have 0.19 and 0.25 more industries which jobs are involved compared with foreign firms and State-owned enterprises (SOEs), respectively.

In sum, Table A2 suggests that jobs operating in multiple industries have higher education requirements, fewer years of experience requirement, and higher offered wage. Meanwhile, private firms have more industries operated.

It provides evidence that jobs involved in multi-industry business could be regarded as higher-quality jobs.

Table A2: Effects of Jobs' Skill Demands on the Number of Industries in which Jobs are Involved

	(1)	(2)	(3)	(4)	(5)
Education requirement					
Tech school	0.0260*	0.0250*	0.0244*	0.0285*	0.0279*
	(0.0132)	(0.0132)	(0.0131)	(0.0158)	(0.0156)
College	0.0972***	0.0929***	0.0919***	0.0931***	0.0865***
	(0.0147)	(0.0152)	(0.0150)	(0.0163)	(0.0155)
Undergraduate or above	0.1345***	0.1277***	0.1299***	0.1345***	0.1221***
	(0.0196)	(0.0201)	(0.0198)	(0.0207)	(0.0186)
Experience requirement (years)		-0.0017**	-0.0016**	-0.0018**	-0.0024***
		(0.0007)	(0.0007)	(0.0007)	(0.0007)
Firm ownership type					
Foreign ownership			-0.1676***	-0.1937***	-0.1907***
			(0.0375)	(0.0442)	(0.0444)
State-owned enterprise			-0.2109***	-0.2547***	-0.2540***
			(0.0275)	(0.0272)	(0.0272)
Log (offered wage)					0.0505***
					(0.0186)
Occupation Fixed Effects	Y	Y	Y	Y	Y
N (Jobs)	120,073	120,073	120,073	98,399	98,399
R2	0.0521	0.0522	0.0542	0.0496	0.0498

Standard errors (in parentheses) are clustered at the occupation level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

Notes: In all of the specifications, we control for an indicator for missing education requirements, log of firm size, the number of positions advertised, and an indicator for missing the number of positions advertised. Omitted or reference groups are as follows: for education requirement, high school or below required; for firm ownership type, private and domestic. Occupation fixed effects control for the 70 categories used on the XMRC website.

Appendix 3: Effect of Applicants' Characteristics on the Number of Industries in which Applicants Experienced

In this section, we investigate the association between the number of industries in which applicants experienced and applicants' characteristics. Table A3 presents the effect of applicants' characteristics on the number of industries in which applicants experienced.

In Table A3, column 1 estimates the gender differential in the number of industry experience. Column 2 adds the control for educational attainment. Column 3 further controls for years of experience. Column 4 is the same as column 3 but regressing with explicit current wage sample. Column 5 additionally controls for the current wage. Column 6 includes detailed CV controls.

Column 3 in Table A3 highlights the significant gender differential in the number of industry experience. Specifically, females have 0.13 more industry experience than males on average. Further, it presents the effect of educational attainment on the number of industries in which applicants experienced. The estimated results indicate that higher educational attainment, the more industries in which worker experienced. Specifically, workers attained undergraduate or above are associated with 0.40 more industry experience than workers with missing educational attainment. Moreover, the magnitude is about 2 times on applicants with tech school attainment.

Column 6 implies that applicants with higher educational attainment have a greater number of industry experience. Similarly, undergraduates or above have 0.34 more industry experience than applicants with missing educational attainment, which is about 1.7 times than applicants who attained tech school. In contrast with Table A2, a one percent increase in workers' current wage is associated with a 1.63 percent decrease in their number of industry experience. It suggests that workers' current wage is lower when they have frequent job transition across industries. In addition, females have 0.16 more industry experience than males on average.

Summing up, Table A3 implies females switch industries more frequently than males, which might contribute to explain gender wage gaps. Also, workers with higher educational attainment have higher labor mobility across industries. However, higher labor mobility across industries is associated with lower current wages, which suggests the positive effect of industry-specific human capital on wages as documented by Neal (1995).

Table A3: Effects of Applicants' Characteristics on the Number of Industries Experienced

	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.0862***	0.0689***	0.1286***	0.1375***	0.1374***	0.1556***
	(0.0040)	(0.0041)	(0.0039)	(0.0044)	(0.0045)	(0.0067)
Educational attainment						
Tech school		0.1614***	0.2035***	0.2063***	0.2063***	0.2023***
		(0.0079)	(0.0078)	(0.0083)	(0.0083)	(0.0082)
College		0.2105***	0.3638***	0.3681***	0.3682***	0.3380***
		(0.0062)	(0.0063)	(0.0067)	(0.0068)	(0.0068)
Undergraduate or above		0.1524***	0.3977***	0.4064***	0.4066***	0.3391***
		(0.0061)	(0.0064)	(0.0069)	(0.0072)	(0.0074)
Experience (years)			0.0464***	0.0442***	0.0442***	0.0440***
			(0.0004)	(0.0004)	(0.0005)	(0.0005)
Log (current wage)					-0.0005	-0.0163***
					(0.0055)	(0.0055)
Detailed CV Controls						Y
N (Workers)	246,566	246,566	246,566	196,487	196,487	196,487
R2	0.0019	0.0062	0.0819	0.0735	0.0735	0.0886

Standard errors (in parentheses) are clustered at the worker level. ***p < 0.01, **p < 0.05, *p < 0.10.
Notes: In addition to the covariates shown, columns 2—6 control for an indicator for missing educational attainment. Column 6 also includes "Detailed CV Controls": a dummy for whether the worker married, a dummy for whether the worker is myopic, a dummy for whether the worker has photo flags, height, and an indicator for missing height. Omitted or reference groups are as follows: for gender, male; for education requirement, high school or below required. Occupation fixed effects control for the 70 categories used on the XMRC website.

Appendix 4: Compare Skill Match Index by ML and Statistics

In this section, we explore the robustness of skill match index. First, we present alternative statistics method of creating skill match index. We use conditional probabilities to calculate the probability of any pair of industries occurred together from three perspectives (i.e., either industries of firms' business, industries of workers' experience, or their combined industry information in applications). In light of capturing labor mobility across industries and further investigating the effect of CRISs in job search, we use workers' LMNInd to construct the measure of CRISs although we have additional firms' and applications' LMNInd.

Specifically, we first construct a cooccurrence matrix from industries of workers' experience. By investigating what industries occur together in a same document (i.e., from each worker's industry experience), we construct a 55×55 symmetrical cooccurrence matrix. Mathematically, the cooccurrence matrix followed by

$$(5) \qquad cooccurrence\ matrix = c_{ij} = \begin{cases} \sum_d 1\,(i \in d | j \in d) & \forall\ i \neq j \\ 0 & \forall\ i = j \end{cases}$$

where c_{ij} denote the frequency of industries i and j occurring together in a same document d. In other words, c_{ij} reflects how many documents d containing both industries i and j.

Then, we calculate skill match index by conditional probability, formally by

$$(6) \qquad skill\ match\ index\ from\ statistics = p_{ij} = \begin{cases} p_{i|j} \cdot p_i & \forall\ i \neq j \\ 1 & \forall\ i = j \end{cases}$$

where p_{ij} denote the probability of industries i and j occurring together in all documents d based on the cooccurrence matrix and thus p_{ij} represents skill match index between industries i and j.

Similarly, skill match index by machine learning is formally expressed as

(7) $\quad skill\ match\ index\ from\ machine\ learning = \cos(\gamma) = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|}$

where V_1 and V_2 denote the vectors of industry 1 and 2, respectively.

Then, we calculate vacancy-applicant skill match index from statistics based on (6)

(8) $\quad vacancy\text{-}applicant\, skill\ match\ index\ from\ stats = \frac{1}{k} \sum p_{ij}\ \forall\ i \in W, \forall\ j \in J$

where W and J denote the set of industries of workers' experience and firms' business, respectively. In addition, k represents the number of vacancy-applicant pairs of industries by interacting W and J.

Panel A and B of Figure A1 present the heat plot of skill match index from machine learning and statistics approach, respectively. The industries are labelled in the horizontal and vertical symmetrically. The color in the interaction cell from rows to columns (or columns to rows) depicts the magnitude of skill match index between the two industries. The darker the interaction cell, the higher skill match index between the two industries. Since the heat plot is symmetric, it yields the highest skill match index between the same industry (darkest blue diagonal interaction cell).

From the legend of Figure A1, we find that the order of magnitude has significant differences. The skill match index from machine learning is ranged from -0.1411 to 1, while that from statistics is ranged from 0.000025 to 1. Especially, most of skill match index from statistics is smaller than 0.000939. It implies that skill match index from machine learning might be able to generate better variations. Also, by comparing these two heat plots visually, we find that skill match index created by machine learning is more evenly distributed and implies it is better to represent CRISs to some extent.

Figure A2 compares the distribution of skill match index by machine learning and statistics approaches. The upper left and middle panels show the distribution of skill match index from machine learning and statistics, respectively. We find that the distribution of skill match index from machine learning is more normally distributed and the distribution of skill match index from statistics is more right-skewed distributed with long tails. It implies that skill match index from machine learning helps investigate larger variations in skill match index. The upper right panel shows the correlation between skill match index from machine learning and that from statistics is 0.23.

Moreover, the lower left and middle panels show the distribution of standardized skill match index from machine learning and standardized logarithm of skill match index from statistics, respectively. The lower right panel shows the correlation between the adjusted skill match index from machine learning and that from statistics is 0.80. In sum, skill match index from machine learning is better in terms of distribution and for ease of interpretation purpose (i.e., using raw skill match index rather than adjusted ones).

Further, to explore the robustness of skill match index, we use similar specifications of Table 5 and add an additional regressor of vacancy-applicant skill match index from statistics in Table A4. Columns 1 and 2 include vacancy-applicant skill match index from machine learning and statistics, respectively. Column 3—8 add both vacancy-applicant skill match index from machine learning and statistics to test which is the most proper vacancy-applicant skill match index. Column 3—8 also includes additional controls which are similar to columns 2—7 in Table 5.

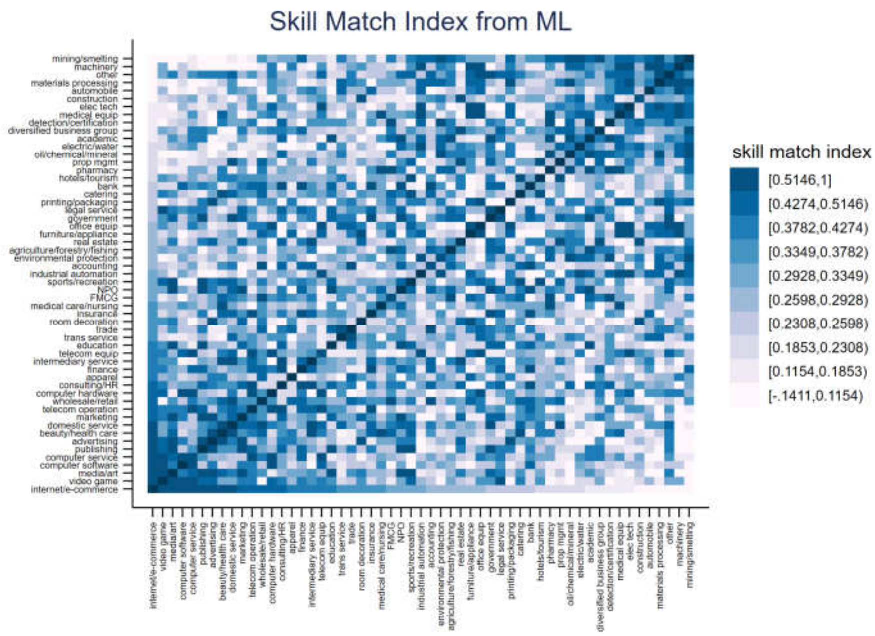
The results are shown in Table A4. Columns 1 and 2 indicate that without controls, both vacancy-applicant skill match index has significant positive effect on callback probabilities, respectively. Turning to column 3, without controls, we find that the estimate of vacancy-applicant skill match index from machine learning is statistically significant at 1 percent level, while that from statistics is statistically significant at 10 percent level. Also, the magnitude of the effect on callback probabilities is more pronounced in vacancy-applicant skill match index from machine learning (0.0053) than that from statistics (0.0014). With additional controls, the results from columns 4—8 confirm that the estimate of vacancy-applicant skill match index from machine learning is always statistically significant at 1 percent level and the magnitude of the effect on callback probabilities is more pronounced in vacancy-applicant skill match index from machine learning. However, the estimate of vacancy-applicant skill match index from statistics is either statistically significant at 10 percent level or statistically insignificant, and its magnitude turns to negative which is counter intuitive. Therefore, vacancy-applicant skill match index from machine learning outperforms that from statistics.

On the other hand, from methodological view, vacancy-applicant skill match index from machine learning is better than that from statistics in capturing the concept of CRISs. Since statistics method is consistent with the idea of Neffke and Henning (2013) where the calculation is based on weighted sum of the similarity between pairs of industries only (labeled as ③ in Figure 1). The machine learning approach is more flexible and realistic due to the vector form of industries. Specifically, the aggregate industry vectors somehow are better to capture the integrality of CRISs which applied not only in pairs of industries but also in more than two industries (again, labeled as ③ and ④ in Figure 1, respectively).

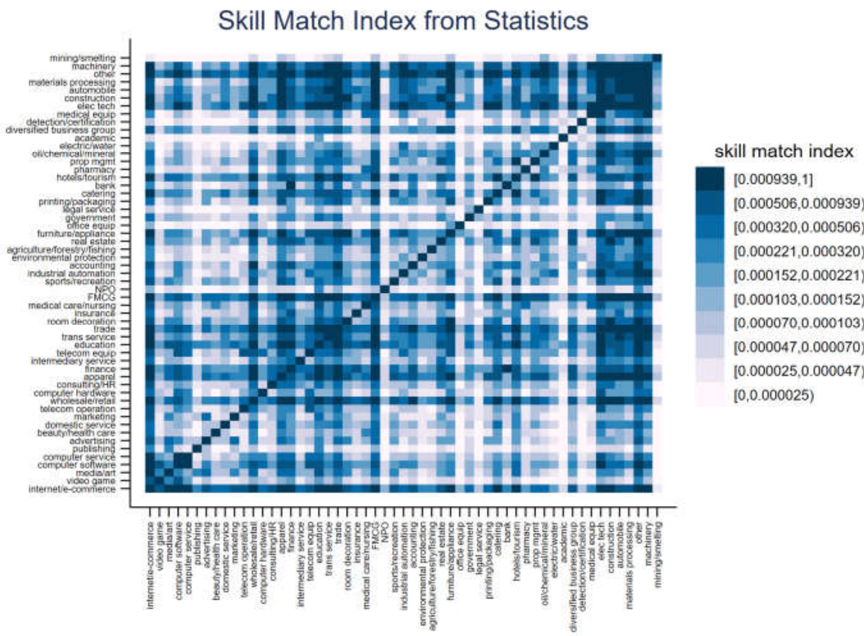
Summing up, we compare alternative statistics approach of construction of vacancy-applicant skill match index and machine learning approach. We find that skill match index created by machine learning is more evenly distributed and implies it is better to represent CRISs to some extent. Further, skill match index from machine learning is better in terms of distribution and for ease of interpretation purpose (i.e., using raw skill match index rather than adjusted ones). Additionally, vacancy-applicant skill match index from machine learning outperforms that from statistics not only in the regression analysis but also from better reflection of CRISs.

Figure A1: Compare Heat Plot of Skill Match Index by ML and Statistics

Panel A: Heat Plot of Skill Match Index from ML

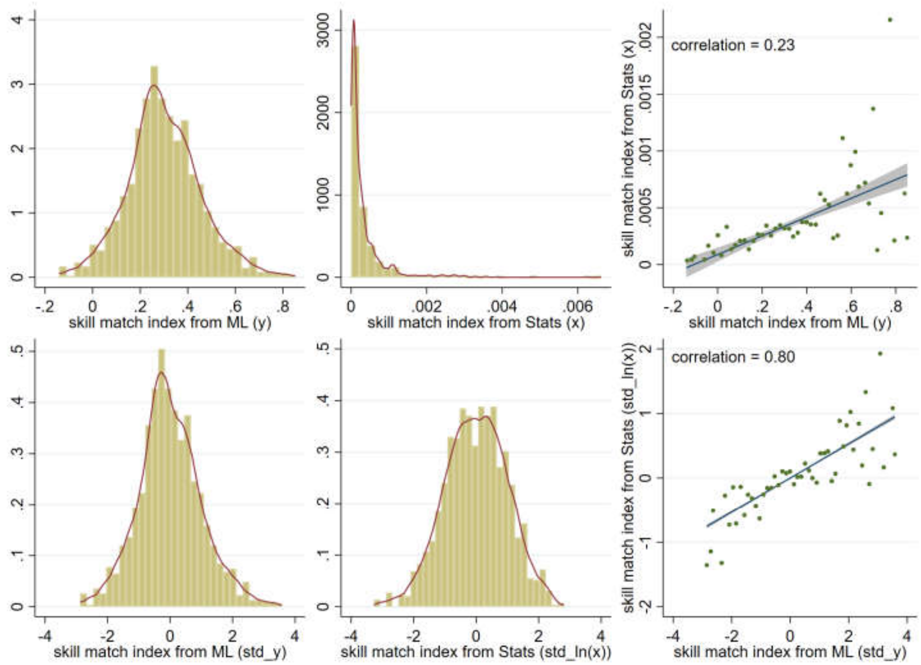


Panel B: Heat Plot of Skill Match Index from Statistics



Notes: Panel A and B of Figure A1 present the heat plot of skill match index from machine learning and statistics approach, respectively. The industries are labelled in the horizontal and vertical symmetrically. The color in the interaction cell from rows to columns (or columns to rows) depicts the magnitude of skill match index between the two industries. The darker the interaction cell, the higher skill match index between the two industries. Since the heat plot is symmetric, it yields the highest skill match index between the same industry (darkest blue diagonal interaction cell). Comparing these two heat plots visually and mathematically highlights that skill match index created by machine learning is more evenly distributed and implies it is better to represent CRISs to some extent.

Figure A2: Compare Distribution of Skill Match Index by ML and Statistics



Notes: Figure A2 compares the distribution of skill match index by machine learning and statistics approaches. The upper left and middle panels show the distribution of skill match index from machine learning and statistics, respectively. We find that the distribution of skill match index from machine learning is more normally distributed and the distribution of skill match index from statistics is more right skewed distributed with long tails. It implies that skill match index from machine learning helps investigate larger variations in skill match index. The upper right panel shows the correlation between skill match index from machine learning and that from statistics is 0.23. Further, the lower left and middle panels show the distribution of standardized skill match index from machine learning and standardized logarithm of skill match index from statistics, respectively. The lower right panel shows the correlation between the adjusted skill match index from machine learning and that from statistics is 0.80. In sum, skill match index from machine learning is better in terms of distribution and for ease of interpretation purpose (i.e., using raw skill match index rather than adjusted ones).

Table A4: Compare Explanation Power of Vacancy-Applicant Skill Match Index from ML and from Statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vacancy-Applicant skill match index from ML	0.0064***		0.0053***	0.0074***	0.0084***	0.0085***	0.0062***	0.0063***
	(0.0005)		(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
Vacancy-Applicant skill match index from Stats		0.0054***	0.0014*	0.0001	-0.0009	-0.0011	-0.0014*	-0.0015**
		(0.0005)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Job Requirement Controls				Y	Y	Y	Y	Y
Detailed CV Controls					Y	Y	Y	Y
Vacancy-Applicant Matching Controls						Y	Y	Y
Competition Controls							Y	Y
Occupation Fixed Effects								Y
N (Applications)	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944	3,618,944
R2	0.0003	0.0002	0.0003	0.0043	0.0056	0.0066	0.0153	0.0176

Standard errors (in parentheses) are clustered at the job level. ***p < 0.01, **p < 0.05, *p < 0.10.
Notes: In addition to the covariates shown, columns 4—8 include “Job Requirement Controls”: education requirement (5 categories), experience requirement (quadratic), age requirement (quadratic), an indicator for missing experience requirement, and an indicator for missing age requirement. Columns 5—8 include “Detailed CV Controls”: a dummy for whether the worker married, a dummy for whether the worker is myopic, a dummy for whether the worker has photo flags, the workers’ height, and an indicator for missing height. Columns 6—8 include “Vacancy-Applicant Matching Controls”: matching status of age (4 categories: three dummies for whether the worker’s age is less than/match with/more than the requested age, respectively, and an indication for missing age-related information), four dummies for whether the worker’s new graduate status matches the requested status (new/non-fresh graduate interact with requested new/non-fresh graduate), an indicator for missing education information, and an indicator for missing experience information. Columns 7—8 include “Competition Controls”: the number of applications received by the job, the number of positions advertised, and an indicator for missing the number of positions advertised. Omitted or reference groups are as follows: for the matching status of education, education matches requested; for the matching status of experience, experience matches requested. Occupation fixed effects control for the 70 categories used on the XMRC website.

Appendix 5: Robustness of Machine Learning Approach

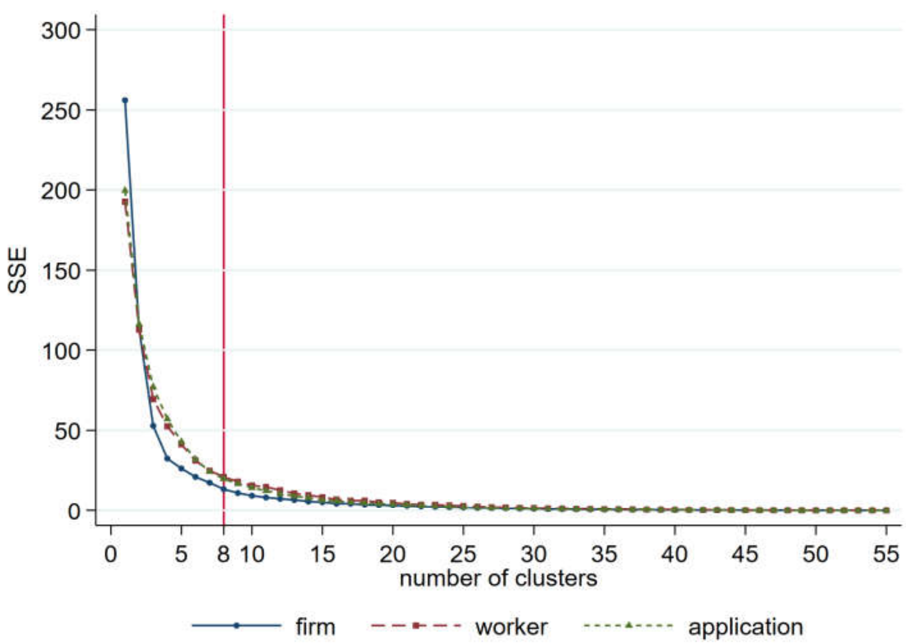
In this section, we present various robustness checks for machine learning approaches. Figure A3 presents the robustness check from number of clusters shown in Figure 2. The horizontal axis and vertical axis depict number of industry clusters and within-cluster sum of squared errors (SSE), respectively. Intuitively, the larger the number of clusters, the lower the value of SSE. For example, in an extreme case, when the number of clusters equals the number of industries (i.e., 55), SSE is equal to zero. To check how many clusters applied is optimized in the three LMNInd, we use elbow method to select 8 as the optimal number of clusters. Because at 8 clusters, all of three lines representing SSE start seeing diminishing returns by increasing number of clusters.

Figure A4 explores the robustness of dimensionality by skill match index. Panel A, B, and C of Figure A4 present robustness of dimensionality from industries of firms' business, industries of workers' experience, and industries of applications, respectively. The horizontal and vertical axis indicates the number of dimensions of industry vectors and statistics of skill match index, respectively. From the three panels, we select the number of dimensions as 20, because statistics of skill match index diverge to reasonably stable status and due to our size of corpus is small (i.e., the number of industries is 55).

Figure A5 explores the robustness of dimensionality by skill match index. Panel A, B, and C of Figure A5 present robustness of dimensionality from industries of firms' business, industries of workers' experience, and industries of applications, respectively. It shows kernel density of skill match index at different dimensions of industry vectors. From the three panels, we select the number of dimensions as 20, because the kernel density implies that larger variations can be exploited as decreasing number of dimensions and due to our size of corpus is small (i.e., the number of industries is 55).

Figure A6 explores the robustness of sample size. The horizontal and vertical axis indicates the sampling percentage and statistics of skill match index, respectively. We use sample from industries of applications because it has largest sample size (i.e., 3,618,944 applications) compared with industries of firms' business (i.e., 120,073 firms) and industries of workers' experience (i.e., 246,566 workers). From the figure, we find that all statistics of skill match index is reasonably similar by varying sampling percentage, except minimum appears some differences.

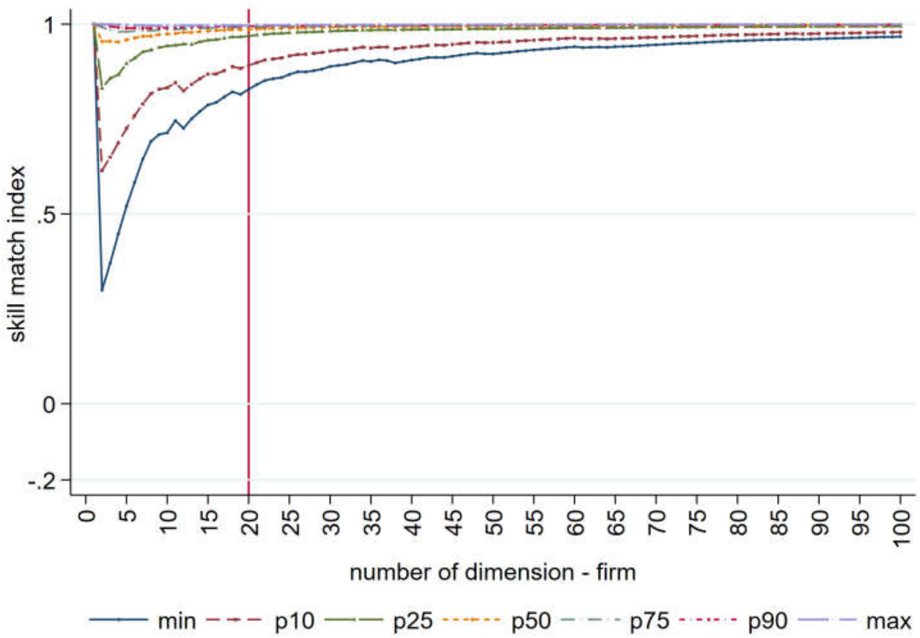
Figure A3: Robustness Check for Number of Clusters



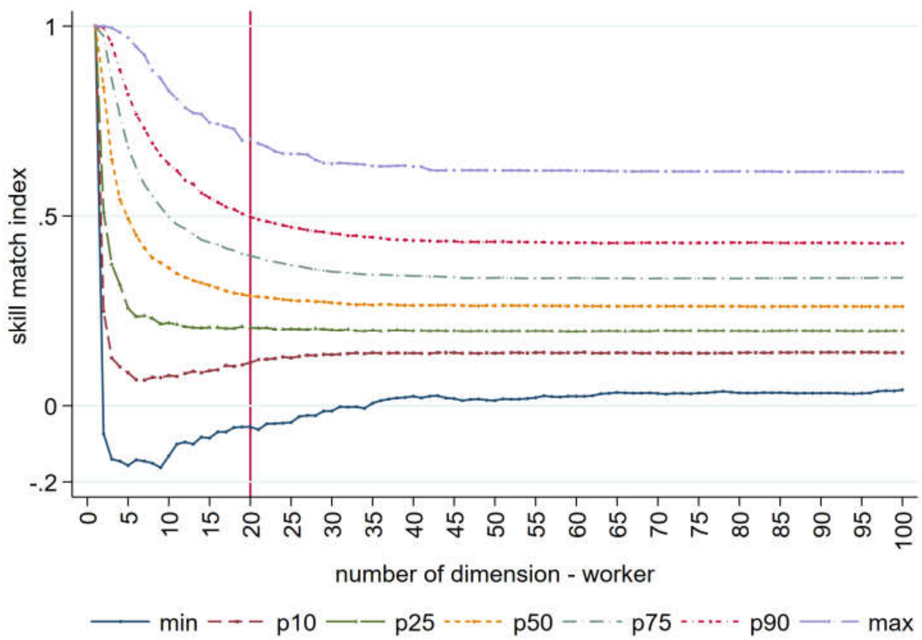
Notes: Figure A3 presents the robustness check from number of clusters shown in Figure 2. The horizontal axis and vertical axis depict number of industry clusters and within-cluster sum of squared errors (SSE), respectively. Intuitively, the larger the number of clusters, the lower the value of SSE. For example, in an extreme case, when the number of clusters equals the number of industries (i.e., 55), SSE is equal to zero. To check how many clusters applied is optimized in the three LMNInd, we use elbow method to select 8 as the optimal number of clusters. Because at 8 clusters, all of three lines representing SSE start seeing diminishing returns by increasing number of clusters.

Figure A4: Robustness Check for Dimensionality by Skill Match Index

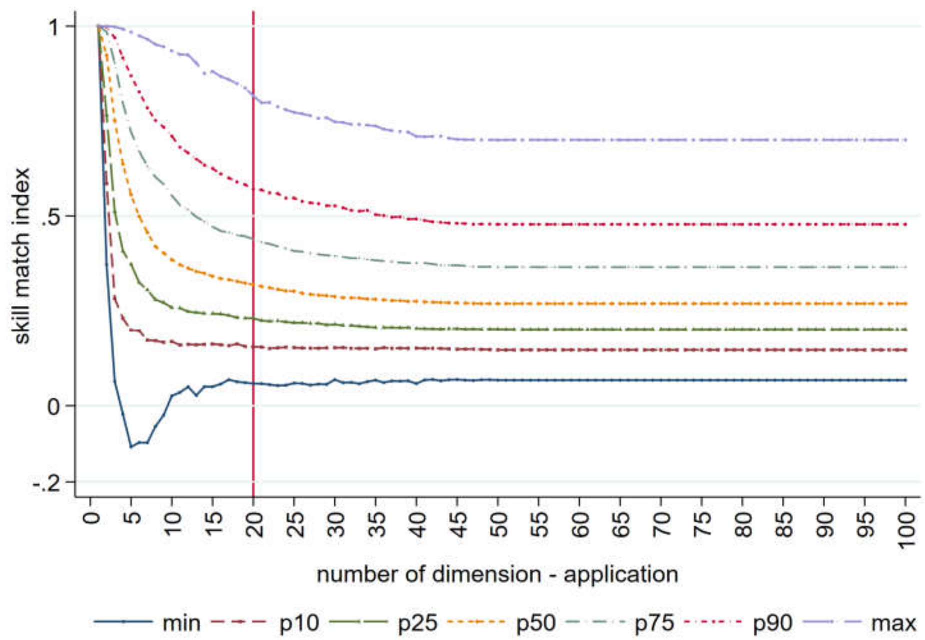
Panel A: Robustness Check for Dimensionality of Industry Vectors of Firms by Skill Match Index



Panel B: Robustness Check for Dimensionality of Industry Vectors of Workers by Skill Match Index



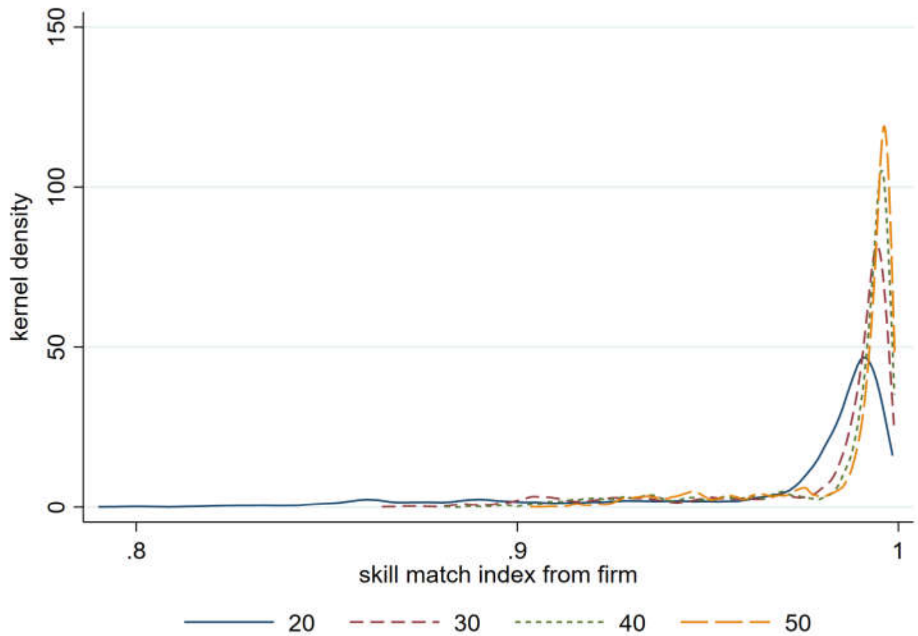
Panel C: Robustness Check for Dimensionality of Industry Vectors of Applications by Skill Match Index



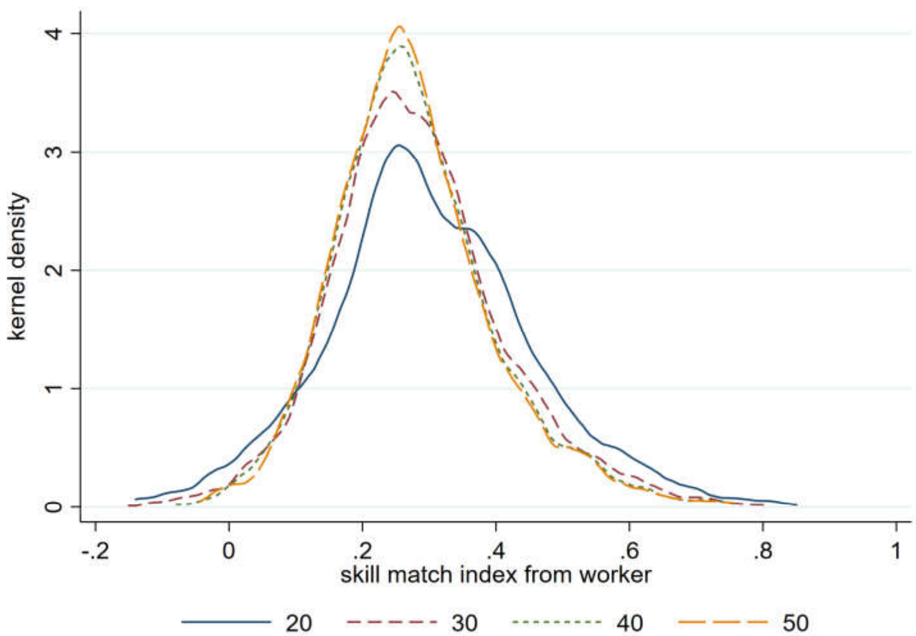
Notes: Figure A4 explores the robustness of dimensionality by skill match index. Panel A, B, and C of Figure A4 present robustness of dimensionality from industries of firms' business, industries of workers' experience, and industries of applications, respectively. The horizontal and vertical axis indicates the number of dimensions of industry vectors and statistics of skill match index, respectively. From the three panels, we select the number of dimensions as 20, because statistics of skill match index diverge to reasonably stable status and due to our size of corpus is small (i.e., the number of industries is 55).

Figure A5: Robustness Check for Dimensionality by Kernel Density

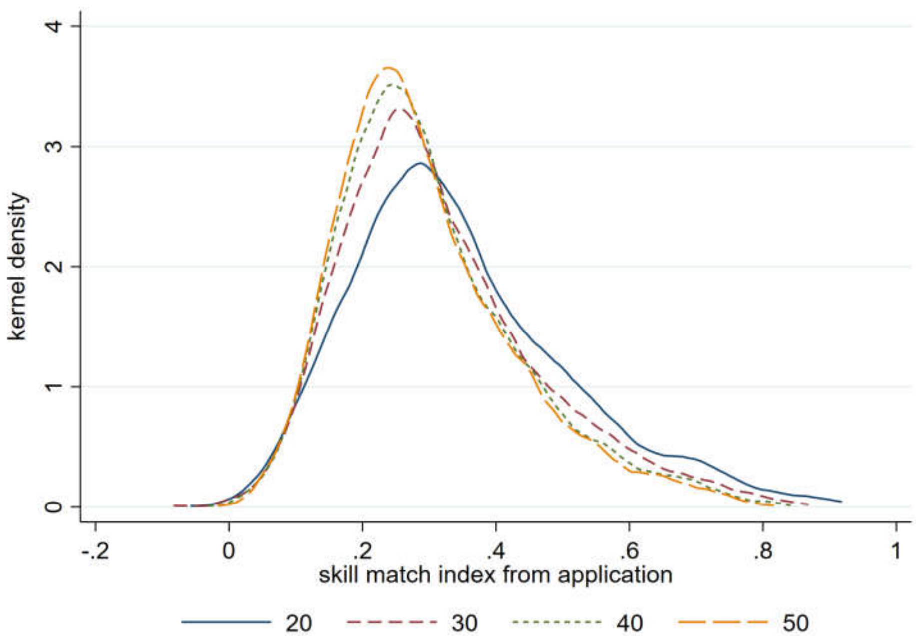
Panel A: Robustness Check for Dimensionality of Industry Vectors of Firms by Kernel Density



Panel B: Robustness Check for Dimensionality of Industry Vectors of Workers by Kernel Density

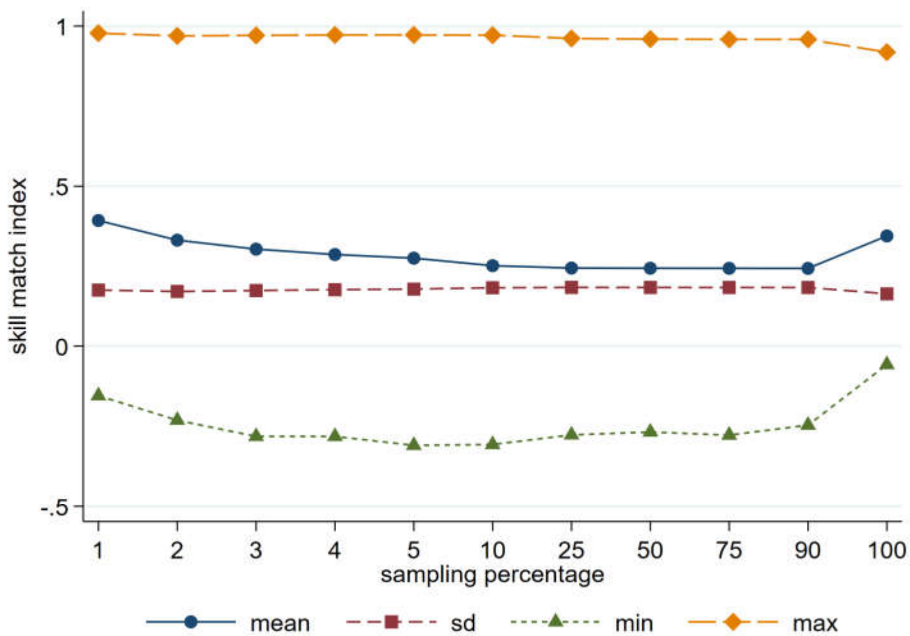


Panel C: Robustness Check for Dimensionality of Industry Vectors of Applications by Kernel Density



Notes: Figure A5 explores the robustness of dimensionality by skill match index. Panel A, B, and C of Figure A5 present robustness of dimensionality from industries of firms' business, industries of workers' experience, and industries of applications, respectively. It shows kernel density of skill match index at different dimensions of industry vectors. From the three panels, we select the number of dimensions as 20, because the kernel density implies that larger variations can be exploited as decreasing number of dimensions and due to our size of corpus is small (i.e., the number of industries is 55).

Figure A6: Robustness Check for Sample Size



Notes: Figure A6 explores the robustness of sample size. The horizontal and vertical axis indicates the sampling percentage and statistics of skill match index, respectively. We use sample from industries of applications because it has largest sample size (i.e., 3,618,944 applications) compared with industries of firms' business (i.e., 120,073 firms) and industries of workers' experience (i.e., 246,566 workers). From the figure, we find that all statistics of skill match index is reasonably similar by varying sampling percentage, except minimum appears some differences.