Posted Wage and Compensation Inequality*

Xuanli Zhu†

November 4, 2022
(Latest Version Here)

Abstract

We develop a new method to study the determinants of wage and compensation differentials using job vacancy data and machine learning algorithms. This novel method generates direct controls for heterogeneous skills and tasks required on different jobs and thus can work as an alternative to the common approach that relies on both worker and firm fixed effects and employer-employee panel data. More importantly, it sheds new insights on the determinants of labor market inequality by opening the black box of the worker effect and uncovering the missing piece of firm-provided non-wage compensations. Applying our method to the vacancy data of a Chinese IT-centered job board, we estimate the shares of different wage components to be consistent with the results found in recent studies of rich countries. During the estimation procedures, our machine learning approach discovers a data-driven skill and task structure featured by different levels of specificity. Occupation-specific skill and task variations account for an important share of wage variation in terms of both job effect and firm-job sorting especially in high skill occupations, while experience- and position-related skills and tasks are the most important in low skill occupations. In contrast, most general skills, whether cognitive, interpersonal, or noncognitive, barely matter for posted wage inequality. Lastly, we find that high wage premium firms sorted with high skill jobs also provide better compensations except for generous work-time and such amenities are not subject to compensating differential, thus aggravating labor market inequality. We suggest that a new theory that combines compensating differential with efficiency compensation and firm-worker sorting can reconcile this finding.

Keywords: wage inequality, skill and task, firm wage premium, firm-worker sorting, compensation differential

*I am very grateful to Tetsuji Okazaki for his guidance throughout the development of this project. I thank Daiji Kawaguchi, Naoki Wakamori, and participants in the seminar of University of Tokyo, the 4th Monash-Warwick-Zurich Text-as-Data Workshop, the "Labor, Firms, and Macro" Workshop, and the Tokyo Labor Economics Workshop for valuable feedback and comments. I thank Adam Oppenheimer for his help on using the Pytwoway package. This paper is previously circulated with the title "Job Task Variation, Firm Wage Premium, and Firm-Provided Compensation".

†University of Tokyo, Graduate School of Economics. Email: zhuxuanli46@gmail.com.

1 Introduction

Workers are paid differently in the labor market. The determinants of the inequality in terms of both wage and other compensations that different workers receive have long been a key research agenda for economists. One major economics and econometric problem that economists often face when studying the labor market inequality is unobserved worker and job characteristics. A common approach to resolve this problem is to use fixed effects and panel data to control for those time-invariant confounding factors that are unobserved in the data. Abowd et al. (1999) (hereafter AKM) pioneered to use two-way fixed effects, i.e. both worker fixed effects and firm fixed effects, to separate wage inequality into different components. The following literature in general find that while workers' observed and unobserved characteristics ("worker effect") account for a majority share of the wage differentials, the different levels of firm wage premium ("firm effect") and the assortative matching between worker quality and firm wage premium ("sorting") are also important determinants of wage inequality (see the review in Card et al. (2018) and other recent papers discussed in Section 2). Moreover, several recent studies (see, among others, Sorkin, 2018; Lamadon et al., 2022) suggest that the estimation results under the AKM approach and a perspective of revealed preference imply that there is potentially important roles played by other firm-provided non-wage compensations on wage determination and worker mobility.

Despite these substantial progresses, relatively less is known about the more granular factors and determinants within these board components. For example, we know that the worker fixed effects capture those potentially high-dimensional heterogeneous skills of workers beyond formal education levels, but we know little what are those skills and what structures or features do they have. Without knowing the details about what is behind the worker effect, we will also have a limited understanding on how worker ability is sorting with firm wage premium. Similarly, the values of non-wage compensations as a whole backed out from structural estimations (often as a wedge) sketch the important outlines but do not help to depict a clear and sharp picture on the ingredients.

In this paper, we develop an alternative way to study the determinants of labor market inequality in wage and non-wage compensations and show that this new method can bring important new insights on wage and compensation determination by directly exploiting those most granular wage drivers. In short, our approach applies machine learning algorithms to online job vacancy/advertisement data to distill all the wage-predictive information, including different skills, tasks, and other non-wage compensations, from the job description texts, and to generate direct controls for the captured job characteristics so that we can replace the worker effect in the AKM framework with a "job effect". The key idea behind this new approach is that while we cannot observe many important characteristics of workers in the census or survey data, in vacancy data firms actually document all the information about their jobs—the skills required, the tasks conducted, the amenities provided, etc.—so that they can attract and match with their ideal workers. Moreover, firms' posted wages (often wage ranges) in their vacancies reflect their valuations on these job characteristics, i.e. the skills, tasks, and amenities that they require or provide, and also work as the justification for their posted contents. This perspective is natural under the view of directed job search models where firms post wages and other job properties and workers direct their search on different submarkets segmented by different post

contents.[1] As a result, we can replace the real worker with the posted job or the ideal worker, of which we have the full information that matters for the firms' wage determination observed, and replace the real wage with the posted wage.[2] The difficulty, however, is to correctly capture those useful information from the high dimensional text data and to bring them back to the otherwise typical econometric estimation. We tackle this task by taking advantage of a series of machine learning algorithms to find various skills, tasks, and compensations embedded in the job vacancy text and then to generate a set of low dimensional proxy variables for them. These proxy variables allow us to estimate the job effect, firm effect, and firm-job sorting of the posted wage inequality, which correspond to the worker effect, firm effect, and firm-worker sorting in the AKM framework.[3] The distilled information on non-wage compensations also allow us to study how the provision of different compensations interacts with other components in wage determination.

There are several advantages of our new method which make it complementary to the popular AKM-style two-way fixed effect approach. The first main advantage is that through this new approach we can now open the black box of the worker effect in the previous studies and examine how important are different types of skills and tasks in accounting for the labor market wage inequalities. It can also help to improve our understanding on the firm-worker sorting by examining what are the most important part of the job or worker characteristics that contribute to the firm-worker sorting. The second main advantage is that through the information in the vacancy data, we can get a more clear picture about the missing piece of non-wage compensations in many census and survey data. In particular, we can investigate

[1]Although random search and wage bargaining have been a typical setting in the job search models, recently there are increasing evidences showing that directed search and wage posting is more realistic way to thinking about job search and wage setting in recent labor market (Banfi and Villena-Roldan, 2019; Marinescu and Wolthoff, 2020). In fact the fast development and the prevalence of online job boards in recent decades is itself the best evidence that firms and workers recognize such matching process as the efficient way to match the ideal counterparts. One potential concern may be if there could be misinformation or strategic posting by firms. However, note that in an online job portal, a job post often attracts dozens or hundreds of applicants and a typical jobseeker also applies to dozens or over hundred different jobs. Given the resulted large screening costs (and opportunity costs), it is very unlikely that firm will post wrong or misleading job information which would generate mismatches. Our assumption on wage setting is also consistent with recent empirical results that previous jobs have very limited impact on the starting wages in new jobs and that bargaining has an only moderate role on wage setting (see e.g. Di Addario et al., 2022; Lachowska et al., 2022).

[2]In other words, our implicit presumption is that the information documented in the job vacancies reflects the firms' true demand and pricing on the various worker and job characteristics, and, at least in expectation, represent the skills owned, the task conducted, and the compensation enjoyed by the workers that the firm will eventually hire. Even if there exists mismatches between firm's idea workers and actually hired workers, our approach can still represent the true labor market demand and pricing in expectation as long as the level of such mismatches are not systematically different across different types of firms and workers.

[3]Because we replace the controls for worker characteristic in the AKM framework with the controls for job characteristics, the firm effect estimated in our method also does not hold exactly the same interpretation as the firm effect obtained in the AKM framework. To be specific, while the firm effect in the AKM framework is the firms' systematic pay differences after controlling for all worker characteristics, the firm effect in our framework is the firms' pay differences after controlling for everything documented in the job vacancies (except the firm name). The difference could occur when some job characteristics are rather firm-specific. For example if a firm pays higher wage because it assigns the otherwise similar workers with some specific and productive tasks that other firms cannot imitate and such specific tasks are documented in the vacancy text, then such higher wage level will be counted as firm effect in the AKM framework but as job effect in our framework. However, in practice we do not find such firm-specific job characteristics in our distilled job characteristics.

what are the sets of non-wage compensations that different firms use to attract their potential workers and how do these compensations affect firms' wage determination. In addition to these two main advantages that help to shed new insights on the labor market inequality, there are also two other advantages in terms of data and method issues. First, while employer-employee panel data has been widely used in the recent literature with a main focus on rich countries, such data is often either not available or not accessible in many developing countries, for example the research interest of this paper, China. In comparison, vacancy data is more easily to access and also more up-to-date. Second, because in our approach there is no worker fixed effects but only firm fixed effects, the restriction of connected firm set and the assumption of exogenous mobility that are required in the AKM approach are no longer necessary here. Also, the finite sample bias, which is stemmed from high dimensional fixed effects and known as the "limited mobility bias" in the AKM framework, will be moderate as long as firms in the vacancy data do no post too few vacancies.

In this paper, we apply our new approach to the vacancy data of the largest IT-centered online job board in China, which is called Lagou.com. In total, we collected over 6 million job vacancies posted on the website between 2013 and 2020, and use a bunch of cleaning procedures to obtain our final sample of 4 million vacancies for the main analysis. Due to the nature of the job board, one third of the job vacancies in our sample belong to Computer occupations like IT engineers or programmers, but the typical firms in our data also post a large amount of vacancies in other occupations including Design & Media, Business Operations, Financial, Legal, Sales, Administrative, etc. This allows us to study the job characteristics and the wage inequality both at firm level and across occupations with different skill levels. With fully acknowledging that our data can only represent a submarket but not the entire labor market in China, we posit that our approach can be easily applied to any other job vacancy data in any other countries, and that our analysis in this specific labor market uncover many important new facts about wage and compensation inequality that we believe hold general implications for other labor markets. The key information in the job vacancy for our analysis is the raw texts of job description in which employers document their skill requirements, task descriptions, and non-wage compensations in order to attract and match with their ideal workers. We also use the systematic information like the post wages and the requirements on education and experience that firms enter or select from the website system when posting vacancies. Therefore, we are in fact using almost all the information that the potential jobseekers observe in the job posts to study firms' wage posting behavior. To better illustrating both the intermediate results generated during our analysis and the main results on posted wage inequality, we conduct all the analysis on and show the results for both the pooled sample and three subsamples of different (major) occupations. These three subsamples/occupations are Computer, Design & Media, and Administrative, which are the typical high-, medium-, and low-skilled occupations in our data.

In order to distill the useful information embedded in the job texts and to generate the proxy variables, we apply a series of machine learning algorithms to the vacancy text data. The first step is feature selection. The aim is to limit the entire vocabulary of the vacancy text into a subset of terms/tokens/features that matter for wage determination. We achieve this by running a least absolute shrinkage and selection operator (Lasso) regression of posted wage on the token indicator vector of each vacancy and then selecting those features with nonzero estimated coefficients. To avoid overfitting and reduce the randomness in selection, we use

the Bayesian information criterion (BIC) to tune the Lasso model. This procedure shrinks the entire vacancy vocabulary set of over 0.1 million tokens to a subset of only a few thousand. Although the estimation results in a high-dimensional penalized model like Lasso are in general uninterpretable and not necessarily casual due to multicollinearity and high model-flexibility in the high-dimensional context, we verify our selected features through subsampling and sanity check and find that these features are rather robust and intuitive. Our second step is feature clustering. Our aim here is to understand the structure of the wage-predictive job characteristics by clustering our Lasso-selected features into different categories. Importantly, we want to achieve this through a way that does not rely on any prior domain knowledge but let the text data speak for itself. In order to do so, we first train a natural language processing (NLP) model, the word embedding model, on our entire vacancy documents. The word embedding model learns the relationships between terms through the context of each term (i.e. the adjacent terms within a sentence) and represents each term in a latent embedding space. We then apply an unsupervised K-Mean clustering algorithm on this latent embedding space to classify our Lasso-selected features into eight clusters. In essence, the clusters gather the terms based on if the employers talk them in a similar context in the vacancy text. After inspecting the tokens within these auto-generated clusters, we now use our human knowledge to label these eight clusters as the following: a cluster of compensations and amenities; a cluster of general human capital terms on cognitive, noncognitive, and interpersonal skills; a cluster of terms about education and other relevant terms; a cluster of terms featured by experience- or position-related skills and tasks including managing, subordinating, or coordinating specific tasks; and remaining four clusters of occupation-specific skills and tasks. We interpret this structure as that our data-driven approach discovers a skill and task structure featured by different levels of specificity and confirm this claim by inspecting the occurrence frequencies of the features in different clusters across different occupations. The last step of our machine learning procedures is dimensional reduction. In particular, we split our feature indicator matrix used in Lasso regression into eight sub-matrices based on our clustering results and then use the partial least squares regression (PLS) algorithm to transform each sub-matrix into a low dimensional representation with only three proxy variables, so that we can easily add them into the standard wage differential estimation.

We recognize the proxy variables obtained through above procedures as a full set of controls for the job characteristics that affect firms' wage posting. We then embed those proxy variables of skills and tasks (all clusters except the compensation one) into a posted wage regression along with education and experience requirement dummies and firm fixed effects, and conduct the variance decomposition to distinguish the job effect, firm effect, and firm-job sorting as well as further granular components within the job effect and firm-job sorting. Our main findings on the components of the post-wage inequalities are the following. First, our estimation on the pooled sample show that the total share of the wage variance can be accounted 45.0 percent by the job effect, 13.6 percent by the firm effect, and 14.2 percent by the firm-job sorting. The levels of the firm effect and firm-job sorting is consistent with the findings in the recent literature that use the employer-employee data in the U.S. and European countries and bias-corrected AKM approach, suggesting that at least in this high-end labor market in China, the composition of wage inequality is similar to other developed countries. Second, despite the fact that we extract way more job characteristics for the high-skilled sample of the Computer occupation than the low-skilled sample of the Admin occupation, the estimation results show

substantially smaller share from job effect and larger share from firm effect and firm-job sorting in Computer occupation comparing to Admin occupation. This result suggests that the firm wage premium and firm-worker sorting observed in the labor market are potentially linked with how different firms adopt different specific skills and tasks. Third, we find that while most of the explanatory power of the education dummies are absorbed by the proxy variables that directly extract education information from the job text, the experience dummies still account for nearly half of the job effect and the sorting between job effect and firm effect and are highly correlated with our proxy variables. We suggest that this is because our machine learning approach mainly extracts the extensive margins of different skills and tasks while the experience requirement can represent the intensive margins of those occupation-specific skills and tasks and thus complements to our proxy variables. Fourth, our further decomposition on the extensive margin show that those occupation-specific skills and tasks account for an important share of the job effect and firm-job sorting in the pooled sample and the high-skilled sample of Computer occupation, but their importance declines significantly in the low-skilled sample of Admin occupation. In comparison, experience- and position-related skills and tasks, which arguably have medium levels of specificity, account for a major share of the effects of the extensive margin in low-skilled Admin sample, and also have non-negligible effects in the pooled and high- or medium-skilled sample. However, those most general skills turn out to play little roles in explaining posted wage differentials, in spite of the fact that firms do mention these cognitive and noncognitive terms in their job ads. The third and the fourth findings in combination suggest that occupational specific skills and tasks are not only a key part of the potential job or worker differences that directly account for the posted wage inequalities but also a key factor that generates the assortative matching between firms and workers or jobs. Fifth, we find that our estimated firm effects can be partially explained by firm size and location dummies, which is again consistent with the estimated firm effects in the AKM framework. Finally, we conduct several robustness checks and show that our results are affect by the finite sample bias or the compositional differences across different samples.

In the final part of this paper we investigate the patterns of non-wage compensation provision in our online vacancy data and how do those compensations affect posted wage and total compensation inequality. Further embedding the proxy variables of the compensation cluster into our wage differential estimation, we find that those compensation terms selected by the Lasso regression explain the posted wage variances not by the variances of themselves but by the positive covariance terms with job effect and firm effect. Put it differently, these non-wage compensation can predict the posted wage largely because they can indicate the job quality and firm wage premium. This suggests that different firms in different jobs also systematically provide different non-wage compensations. We thus investigate a group of different types of non-wage compensations and find that betters firms sorted with better jobs are also more likely to provide advanced insurance package, backloading wage and stock options, and high qualified coworker and flexible work-time, but less likely to provide weekend, holiday, and fixed work-time. In addition, we run a hedonic regression on these compensations with a full control of job characteristics and find that those amenities that high wage premium firms are more likely to provide are significantly and positively related with the posted wage, whereas those amenities that low wage premium firms are more likely to offer have a significantly negative correlation with the posted wage. These stylized facts cannot be explained by the classic compensating differential theory. We thus suggest that a new theory that combines the

compensating differential with two other elements, efficiency compensation and firm-worker sorting, can reconcile these puzzling findings.[4] Essentially, as long as we accept the idea that many non-wage compensations can be efficient or inefficient in production or firm operation, there will be an additional efficiency channel along with the traditional equalizing differential channel when firms decide their levels of compensation. This new channel can work in either the inverse or the same direction to the compensating differential mechanism, and to what extent it matters depends on the level of the firm-worker sorting. For those compensations that are efficient like advanced insurance or backloading wages, and in those high-pay firms sorted with high-skilled workers and jobs, the efficiency effect will be large and can even dominate the compensating differential mechanism so that the firms that provide better compensations will not reduce but increase their workers' wage. Whereas in low-pay firms with low-skilled workers and jobs, such efficiency effect is low and if the compensation is mandated and its costs cannot be fully equalized from the efficiency benefits, it will be compensated from a reduction in the workers' wage. On the other hand, inefficient compensation like generous work-time or work-load will cause a large efficiency loss in those high pay-premium firms and high quality jobs, resulting that only low pay-premium firms and low quality jobs will bear the cost and provide such amenities. Our new theory can thus generate flexible patterns of compensation provision and wage impact, and thus provide important implications that help to better understand the wage and compensation inequalities in the labor market.

The outline of the rest of our paper is following. In next section we discuss the related literature and our contributions. Section 3 introduces our data. In Section 4 we set up our econometric model and conduct a preliminary estimation. In Section 5, we apply a series of machine learning approaches to the vacancy text data to exploit wage-predictive information and generate proxy variables. Section 6 shows our main results on the posted wage inequalities. Section 7 documents our empirical findings on non-wage compensations and a new theory that can explain these findings. Finally, we provide some concluding remarks in Section 8.

2 Related Literature

This paper links to and contributes to three board literature that focus on different determinants of the wage or compensation inequalities in the labor market. While these different determinants are often studied separately, our integrated examination here shows that these drivers are in fact closely linked with each other, and thus it is important to investigate their interactions for a better understanding on the mechanisms behind wage and compensation determination.

The first literature strand is the voluminous literature that use heterogeneous workers to explain the wage inequalities in the labor market. Since Mincer (1958) and Becker (1964), human capital, whether general or specific, has long been recognized by economists as the main factor behind wage differences. However, observed worker characteristics, including education, experience, occupation, and other demographic factors, often explain only a frac-

[4]Here by "efficiency compensation", we mean the efficiency aspects of compensations that is similar to the ones proposed in the efficiency wage theory: eliciting effort, reducing labor turnover costs, etc. In fact, we argue in Section 7.2 that this is a more natural property of non-wage compensations than monetary wage.

tion (around 30%) of the total wage variation in a typical wage regression. While various types of specific human capital like industry- or occupation-specific human capital have been examined in the early literature, more recently, there is a converging consensus in the labor literature that occupation and industry categories are just serving as measurable proxies for the underlying tasks performed and skills required across different jobs and firms, and that multidimensional task-specific skills are the most natural way in thinking about human capital (see Sanders and Taber (2012) for an early survey on this literature). Following this idea, the recent empirical studies have begun to stress on the importance of multidimensional skills and tasks for wage determination and discrepancies (see Spitz-Oener, 2006; Autor and Handel, 2013; Deming and Kahn, 2018; Yamaguchi, 2012; Lise and Postel-Vinay, 2020, among others). Despite this surging popularity, in practice the entire space of the multi-dimensional tasks and skills are often classified and compressed into a very limited number of pre-determined board and abstract dimensions of cognitive, social, abstract, manual, routine, etc. And the potential specificity of skills and tasks (i.e. the necessary width of the space), though once discussed intensively in the literature, are now often completely circumvented. As a result, the examination of multi-dimensional skills and tasks in many studies are often eventually constrained in a pre-defined and fairly low dimension, which puts even distinctive occupations into very similar positions.[5] Also, many studies on multidimensional skills and tasks have limited their attention on between-occupation skill and task variations, potentially due to data limitation, even though the recent empirical studies find clear evidences of within-occupation task or skill variations and their significance in wage prediction (see e.g. Autor and Handel, 2013; Deming and Kahn, 2018). Our paper thus contributes to this literature by developing a method to investigate the indeed high dimensional skill and task space spanned both between occupations and within occupations with no priors holding on what are the most important dimensions. Indeed, we let the online vacancy text data to tell us what are the structure of the skill and task space based on how employers document the skills and tasks about their jobs. Our approach thus generates a data-driven skill and task structure, and it turns out that this structure is distinguished by different levels of specificity. Our following estimation results show that it is those most specific skills and tasks that play the most important role in accounting for the posted wage inequalities. On the other hand, general skills, whether cognitive, interpersonal, or noncognitive, matter little for the posted wage differential in our data.[6] Therefore,

[5]Although such simplification has been proven very useful in studying some labor market issues including what types of workers are or will be substituted by machines or robots or AI, it can be potentially misleading when studying worker heterogeneity and wage differences. It is because that the wage determination in the labor market are potentially based on very detailed skills and tasks which, even if similar in a low dimension, can be completely different and thus largely nontransferable at a high dimension. Such distinction is particular important when thinking about issues like job assignment, job mobility, and human capital investment, all of which could be potentially important for wage determination. For example, skill indexes calculated in those low and board dimensions often recognize that economists and biologists or electronic engineers have very similar skill compositions and skill levels. Consequently, using an analogy similar to the one in Sattinger (1993), those low dimensional skill index would indicate that this paper can be equally written by a biologist or an electronic engineer. More recently, Frank et al. (2019) suggests that studying skills and tasks with further increased specificity could provide better insights even for understanding the technological impact on labor market.

[6]Note that here we are not arguing that those cognitive, interpersonal, or noncognitive skills and their relevant dimensions, which have been extensively used in the literature, are not important or uninformative in wage determination. Actually occupation-specific skills and tasks (or skills and tasks with any levels of specificity) can be classified as cognitive or interpersonal or etc., and thus the specificity that we stress here is just another

our results suggest that the specificity is still an important dimension when considering high dimensional skill and task variations, and those highly specific skills and tasks are especially important when thinking about within-occupation skill and task variations.

The second closely related literature is a recently booming literature on estimating the firms' role in wage inequalities at both cross-sectional level and at chronological level (see Abowd et al., 1999; Card et al., 2013; Barth et al., 2016; Song et al., 2019; Bonhomme et al., 2020, among others).[7] In order to overcome the unobserved worker abilities and characteristics, these papers use linked employer-employee panel data and both worker and firm fixed effects to estimate and decompose the entire wage differential into worker effect, firm effect and sorting between firms and workers. Although the initial results of AKM show no evidences for firm-worker sorting, more recent studies equipped with better data and bias correction methods generally find that both the firm wage premiums and the assortative matching between firms and workers are important to account for wage inequality.[8] Our paper contributes to this literature by providing an alternative method to deal with the problem of unobserved worker characteristics and to estimate the firm pay differences. Instead of estimating the worker effect, we apply machine learning methods to online job vacancy text data to generate a full set of controls on the firm-documented job skills and tasks and then estimate a job effect as a replacement. The estimated wage components using our Chinese IT-centered job vacancy data are consistent with those found in the previous literature that use employer-employee panel data and AKM approach (see Bonhomme et al., 2020). Moreover, our method allows us to open black box of the worker fixed effect in the AKM framework and to examine what are the important skills and tasks that contribute to the sorting between workers and firms. Our estimation results find that those occupation-specific skills and tasks contribute for a major amount of firm-job sorting in our pooled sample and in high-skilled computer occupations, while experience- and position-related skills and tasks are the most important drivers of firm-worker sorting in low-skilled administrative occupations.[9]

Thirdly, our paper also contributes to the recently resurgent literature of compensating dif-

dimension that is orthogonal to those previously studied board dimensions. These different dimensions could have different importance when facing different economics questions about the labor market. Moreover, although those most general skills turn out to be not important in our results, these general skills can be important for workers' developing their occupational specific skills and workers' wage changes within firms (since these general skills are likely to be cheap talks and firms need time to confirm them).

[7]Before the pioneered work of AKM, labor economists had discovered strong evidence of significant and consistent wage differentials at industrial level even after controlling for all observed worker characteristics. This stylized fact called for many theories to explain, and one major explanation at that time was efficiency wage theory which generally argues that high wage can elicit workers' effort or avoid turnover costs. For detail see for example Krueger and Summers (1988) and Katz (1986). Since AKM, the main focus of the literature has turned into the differentials in firm level wage premiums.

[8]See Card et al. (2018) for the review on the findings in this literature. For recent improvements in econometric methods, see Kline et al. (2020); Bonhomme et al. (2019) and also the comparison of different methods in Bonhomme et al. (2020).

[9]Note that another feature we observe in our results is that high-skilled professional occupations have significantly more shares of wage differentials accounted by firm effects and firm-worker sorting, and thus contributing more to the aggregate results of firm effects and sorting in the pooled sample. Therefore, our results in combination suggest that those specific skills and tasks in those high-skilled professional occupations are perhaps key for understanding the firm-worker sorting in the labor market, either in terms of cross-sectional levels or trends over time.

ferential. As shown by the classic paper of Rosen (1986), firms can provide different levels of amenities or disamenities to compensate their wage cost and workers will sort into different packages of wage and compensation to maximize their utility. In spite of the theory's intuitive idea and straightforward predictions, early empirical studies that run hedonic wage regressions on different types of compensations often lead to mixed results with both supporting and counterintuitive evidences, even with worker fixed effects (see the survey in Rosen, 1986). More recently, two different types of studies start to revive this topic. First, several recent empirical studies show evidences for compensating differential by using experimental or quasi-experimental methods to identify the wage effects of certain types of compensations in specific situations (see e.g. Mas and Pallais, 2017; Wissmann, 2022, among others). Second, through a perspective of revealed preferences, a few studies begin to model the labor market by using unobserved compensation as a wage wedge to justify job moves, especially for those moves to low wage-premium firms with wage loss observed in data (e.g. Card et al., 2018; Sorkin, 2018; Taber and Vejlin, 2020; Lamadon et al., 2022). We contribute to this literature through two aspects. Firstly we provide new empirical evidences on firms' non-wage compensation provision and their impact on wage determination by taking advantage of the online vacancy data, in which firms document their most important non-wage compensations to attract potential workers. In our data, we discover a large set of pecuniary and nonpecuniary compensations including insurance, backloading wage, stock option, coworker quality, training, weekend and holiday, and flexible work-time among many others, all of which hold predictive power on the posted wage. Moreover, we find high wage premium firms and low wage premium firms have distinguished patterns in the provision of non-wage compensations. In particular, those high wage premium firms are also more likely to offer advanced insurance, larger backloading wage, stock option, and better coworker quality, and that these amenities are not compensated by the posted wage. On the other hand, low pay premium firms are more likely to provide basic insurance and less work-time or more rest days to attract potential workers, and they equalize the costs of these amenities by reducing their posted wages. These findings are consistent with several recent studies that regress the provision of non-wage compensations on the firm fixed effect obtained from the AKM approach and find evidences of high-wage premium firms also providing better non-wage compensations (Sockin, 2022; Bana et al., 2022). Such a positive relationship between wage premium and non-wage compensation provision are, however, at odds with the prediction of compensating differentiation theory.[10] The second aspect of our contribution is to build a new theory that can reconcile these stylized facts. In particular, we combine two elements observed in our findings, efficiency wage and firm-worker sorting, with the traditional compensating differential mechanism and show that this new theory can generate flexible results on firms' compensation provision and impacts on

[10]In particular, the theory of compensating differential is largely based both the heterogeneous cost functions of compensation provision across and the heterogeneous preferences on these compensations across workers. The compensations that the economists often had in mind at the time when the theory was built are job injury, job mortality, or workplace pollutions that are no longer the major concerns in today's labor market. For the compensations found in our data, many are pecuniary and thus firms have exactly the same cost function. Also, for nonpecuniary compensation like health insurance, Dey and Flinn (2005) shows that the cost function of providing the insurance is likely to be similar across different employers. In a similar vein, workers are likely to have similar preference on these amenities although to what extent is an empirical question. Therefore, the theory of compensating differential will fail to generate systematic differences in non-wage compensation provision in the recent labor markets.

the wage determination.[11]

3 Data

In this paper, we use the vacancy data from a Chinese online job board called Lagou.com, which is the first and the largest information technology (IT)-centered online job board in China. The Lagou website starts its service at 2013 and grows fast by specializing on the labor market towards the relatively less-experienced workers in the Chinese Internet industry and acquiring a large customer base of both IT-producing and IT-using firms.[12] Until the end of 2020, about 8 million vacancies have been posted on the website, and we successfully collected the information of over 6 million vacancies between 2013 and 2020.[13] For each vacancy we observe the information of the job name, the wage range, the job location and address, the education level and experience years required, if full-time or part-time or intern, the job descriptions on the tasks and skills required, the job benefits or firm amenities, the firm name, the firm industry category, the firm size category, and the posted time of this vacancy.[14] Different from many other papers in the literature that use pre-processed skill indexes or generate specific skill indexes based on a pre-specified dictionary of terms capturing certain pre-defined skill categories, we will fully utilize the raw text data of each vacancy's job descriptions by adopting a completely data-driven method to distill and classify the important skills and tasks embedded in the text. In addition, we will also exploit one previously ignored information in the

[11]The efficiency nature of alternative pay schemes have long been argued in the organizational literature, see for example Lemieux et al. (2009). And the efficiency of nonpecuniary compensation has been also argued in Dey and Flinn (2005), where the authors suggest that offering health insurance can be efficient for the employers through reductions in exogenous worker exit.

[12]The slogan of the website (https://www.lagou.com/) is "Find an Internet Job—Go to Lagou Recruitment". In 2017, 51job, a leading provider of integrated human resource services in China listed in the NASDAQ stock market and also the owner of one of the largest general online job board in China, announced that it will acquire a 60% equity interest in the parent company of Lagou for $119 million because they think the IT labor markets that Lagou specializes on will be a large complement to their general job board.

[13]The amount of posted vacancies per year grows over time along with the growing popularity of the website. As a result, vacancies between 2013 and 2016 account for less one third of the data, and vacancies between 2017 and 2020 accounts for over two thirds of the data. Our scrapper successfully collected around 60 percent of the vacancies for the 2013-2016 period and over 80 percent of all vacancies posted in the 2017-2020 period. In Appendix A.1 we explain the details of the data collection and show the patterns of both collected vacancies and the missing vacancies over time.

[14]A sample of the job vacancy posted in Lagou can be found in Figure A2. The information of the wage range, the requirements on education and experience, whether full-time or not, and the job location is either selected by firms within given choices provided by the website or be filled in with certain formats when posting the vacancy. This setting ensures that almost all the vacancies in our data have the unambiguous information on the level of post wage and the required education and experience, making it straightforward to generate consistent job variables. In contrast, the format of the job name and the descriptions on job skills, tasks, and amenities is arbitrary and as a result these text contents vary in the length and structure and are often entangled together and hard to distinguish. For example, while there is a certain space for entering job or firm amenities, firms sometimes also mention amenities along with job skill or task descriptions or, in the inverse cases, mistakenly write skill or task terms in the space of amenities. This problem, which is often seen in the real-world text data, partially incentivizes our machine learning methods introduced in the later sections, i.e. we will simply combine all the descriptions on job skills, tasks, and amenities as one integrated text for each vacancy and conduct textual analysis to distinguish the different information types of different words or phrases.

vacancy data: the information about non-wage compensations and amenities that firms claim in the vacancy to attract potential applicants. These firm-provided compensation information will be also automatically extracted from the text data and then distinguished with those skills and tasks terms in the clustering step. Although these compensations claimed in the vacancy data do not constitute the full package of non-wage compensations and amenities provided by the firms, they are likely the most important ones in the labor market perceived by firms and thus allow us to study the patterns of firm compensation provision and its impact on the wage determination and earning inequality in the labor market.

One inevitable drawback of any vacancy data is that it does not constitute the whole labor market in an economy. It is well known and fully discussed in the literature that in most if not all cases the job vacancy data are biased to high skilled and high education jobs, to internet-related jobs, to jobs from large firms and in large cities, and to jobs targeting young or less-experienced workers.[15] Given that our data here is a highly professional part of the online job market, the labor market we study is thus even more biased in this way than other vacancy data. To be specific, our vacancy data is mainly composed a variety of jobs required from 0 to 10 years experience posted by Chinese IT-producing firms and IT-using firms, a majority of which locate in large cities in China.[16] One third of the vacancies in our data belongs to Computer occupations, and the other two-thirds of the jobs come from both other professional occupations like Design & Media occupations, Business Operation occupations, Financial Occupations, and Legal Occupations, and low-skilled occupations like Sales occupations and Administrative occupations.[17] Given the popularity and the low charge of the website, we think these IT-producing and IT-using firms are likely to post all types of jobs that they demand, allowing us to study the firm-level wage premium. In our main analysis and results, we will show both the result for pooled sample including all vacancies along with the results for three typical major occupations in our data: Computer occupation, Design & Media occupation, and Administrative occupation. We pick these specific occupations because they are the representative high-, middle-, and low-skill occupations in our data and thus allow us to study how the skill composition and wage determination vary across occupations with different level of skills that are normally defined in the literature. Unlike Computer occupations, the jobs in Design & Media and Administrative occupations are largely confined to those jobs in the IT-producing or IT-using companies and may not be representative for those occupations in the entire labor markets. With full awareness of the limited coverage of our vacancy data, we argue that our aim in this paper is to illustrate how our new method can be applied to vacancy data and provide new insights on wage determination and compensation provision, and we anticipate our results being examined or validated under other vacancy data in other labor markets or other countries.

[15] See the relevant discussion in Kuhn and Shen (2013) for the Chinese vacancy data and the discussion in Hershbein and Kahn (2018) for the U.S. vacancy data.

[16] IT-using firms mainly incorporate firms in a variety of industries in the tertiary sector like finance, real restate, retail, etc.

[17] Here we define these occupations by following the U.S. SOC classification 2018 and using most board two-digit or three-digit categories. For example, our "Computer occupations" refers to the 2-digit "15-0000 Computer and Mathematical Occupations" and it will incorporate all the occupations in the further 3-digit "15-1200 Computer Occupations" but only some occupations like data scientists in the 3-digit "15-2000 Mathematical Science Occupations". We explain the details of the occupation classifications below and in Appendix A.2.

We next explain our method of occupation classification, describe our sample cleaning procedures, and finally show the summary statistics of our sample used for analysis.

Occupation Classification. One empirical problem in our data is that there are no ready-for-use occupation categories for the job vacancies, as like many other online vacancy data. Although our machine learning method introduced in Section 5 does not rely on using occupation dummies, pre-classified occupations will help us to conduct our analysis on the board occupational level. Moreover, we argue that the procedure of occupation classification, under whatever methods, actually shares the same objective with our main approach, namely mapping individual jobs from a high-dimensional skill and task space to a low dimensional space. However, unlike our data-driven approach, occupation classification relies on pre-specified rules to determine a bunch of subsets in the skill and task space, and thus ignore any within-occupation skill and task variations.

Here we briefly explain some key points of our original method of occupation classification, which combines a dictionary approach and a supervised classification approach. The details of the procedures and the comparison between our approach and other alternative approaches used in the literature are described in Appendix A.2. Whether through human classifying or machine learning methods, the task of occupation classification is to learn some information about a job, in our case from the job title and job description of a vacancy, and then label it to one of the pre-determined set of occupation categories based on that information. Given that our data contains limited scope of jobs comparing to the whole labor market, we first reduce the target occupation categories to a set of 55 6-digit ("minor") occupation categories within 8 2- or 3-digit ("major") occupation categories in the U.S. Standard Occupational Classification (SOC) 2018.[18] Next we prepare a dictionary by selecting multiple keywords for each of those selected 6-digit occupations according to their occupation descriptions in the SOC. The rule here is to select specific phrases or compound phrases so that the chances that these keywords appear in the non-targeted occupations due to the multiple meanings of natural language are low.[19] Perhaps not surprisingly, phrases selected following this rule are basically specific skills and task contents that only used in that specific minor occupation. During this procedure, we further combine some 6-digit occupations when it turns out hard to find exclusive keywords to distinguish these occupations, reducing the 55 6-digit occupations to 34 minor occupation categories. With the dictionary in hand, we then check for each vacancy to see if its job title and job description contains these keywords. If a vacancy is matched with only one minor

[18] We use the U.S. SOC because there is no well-designed official occupation classification for the Chinese labor market and the Chinese IT industry closely follows the technological trend in the U.S. market. This reduction relies on some human inspection on the vacancy data and the official occupation classification and thus might be, to some degree, arbitrary in the occupation choices, but it can largely increase the accuracy of the occupation assignment under even very simple classification algorithms. Within this selected set, major occupations vary in the number of minor occupations selected. For the Computer occupation, we include all 6-digit occupations in the SOC, while for other major occupations, the selected 6-digit occupations are rather limited compared with the full lists in the SOC. Also in practice, we further add 8 more 2- or 3-digit major occupation with no detailed 6-digit minor occupations appended to form an "Other" major category which is used to help increase accuracy of the classification and the vacancies classified to this major category, which have a fairly small share in the whole data, will be removed from the final pooled sample.

[19] We use the corresponding Chinese translation of the English phrases, which sometimes requires to transform those phrases to the Chinese terms that are specific to the Chinese labor market context.

occupation, we regard it as a success of our dictionary method, label it with that matched occupation, and assign it to the "training" sample. If a vacancy has no match or multiple matches, we regard it as a failure and assign it to the "unknown" sample. In the next step we use our "training" sample to train a Naive Bayes classifier, which takes the vectorized text of job titles and job descriptions of a vacancy as input to predict the probabilities that this vacancy belongs to each of the minor occupations. We then apply the trained classifier to the "unknown" sample and assign those vacancies with the most likely occupation predicted. Finally, we also apply the trained classifier back to our "training sample" to rectify the potential misalignment under my dictionary method.

In summary, our occupation classification approach uses terms of specific skills and tasks to first identify the correct occupations for a subgroup of vacancies, and then uses this subgroup to learn the occurrence probability of all skills and tasks terms (along with some other terms) of a vacancy conditional on the vacancy belonging to that occupation. In other words, our algorithm relies on the perspective that occupation categories are different bundles of skills required and tasks conducted on the job. In some sense, this way is even more natural than strictly sticking with the guidelines in the official classification documents because it directly follows a general understanding of various occupations on the labor market, where such understanding may vary across different firms and evolve over time.[20] However, as we have mentioned earlier, these occupation categories can only represent the differences between different centroids of the subsets in the skill and task space, i.e. between-occupation skill and task variations, and thus do not contain any information about within-occupation skill and task variations. We will show later in our analysis that although the occupation dummies generated here can account for a large part of the skill and task variations across different vacancies in our data, the full-scale skill and task variables generated by our approach in Section 5 make it clear that the within-occupation skill and task variation is also an important part for the posted wage variation.

Sampling. To remove invalid vacancies and to reduce measurement errors in the vacancy data, we first drop all vacancies that are not full-time jobs, have outlier wages, or have job descriptions with less than 20 words.[21] We also drop all the vacancies posted in the website launch year 2013 from our sample due to the fact that both the sample size and the share of successfully scraped vacancies are substantially smaller than later years. We further trim our sample by dropping the vacancies from firms that have less than 10 posts and from all the locations that have less than 1000 vacancies over the observation periods. This trimming removes firms and locations with limited samples and thus both reduces the potentially invalid posts and reduces the measurement errors in our data. But it also largely reduce the proportion of small firms and small cities in our sample, resulting the majority of the firms in our sample to be middle or large size firms in large cities. Moreover, we identify the duplicated vacancies

[20] Spitz-Oener (2006) shows that the compositions and levels of the tasks indicated by the occupation actually change over time under technological or organizational changes, and there are large variations within the same occupation for different workers in different firms and positions.

[21] To be specific, we remove the vacancies with a wage lower bound larger than 100,000CNY or smaller than 1,000CNY, and the vacancies with a wage upper bound larger than 200,000CNY or smaller than 2,000CNY. The words in the job description are counted either as Chinese characters or English words. Given the large size of our dataset, our results are not sensitive to any of the thresholds selected here.

that have exactly the same job descriptions and education and experience requirements, and only keep the one with the highest wage posted.[22] Finally, we also remove a small share of vacancies with only English job descriptions that are mainly posted by multinational firms in order to focus our textual analysis on Chinese.

Table 1: Summary Statistics

	Pooled	**Major Occupation**					
	-	Computer	Design_ Media	Business_ Operations	Financial_ Legal	Sales	Admin
Vacancy #	3,999,005	1,330,001	561,236	1,162,404	214,661	452,771	277,932
- share	1.00	.33	.14	.29	.05	.11	.07
Avg # Words	108.91	104.26	103.05	115.60	110.69	120.31	95.09
Wage (1k CNY):							
- Mean	13.64	17.38	10.68	14.19	11.95	10.21	6.32
- SD	9.24	9.79	6.31	9.52	9.19	6.53	3.90
Firm:							
- #	86,330	67,369	68,092	78,244	41,285	58,847	59,016
- Avg Posts	46.32	19.74	8.24	14.86	5.20	7.69	4.71
- Median Posts	20.0	9.0	4.0	6.0	2.0	3.0	2.0
Vacancy Share (%)							
Firm Size (Employees):							
- -15	.03	.03	.05	.02	.02	.03	.03
- 15-50	.18	.17	.25	.16	.15	.19	.20
- 50-150	.23	.21	.26	.22	.22	.23	.26
- 150-500	.21	.21	.21	.22	.23	.20	.23
- 500+	.15	.16	.12	.16	.18	.15	.14
Education:							
- Vocational College	.33	.24	.38	.29	.27	.51	.52
- Bachelor	.54	.66	.47	.61	.63	.22	.24
- Master/Doctor	.01	.02	.00	.01	.03	.00	.00
- Not Specified	.12	.08	.15	.09	.07	.27	.23
Experience:							
- 0	.22	.12	.21	.16	.25	.48	.50
- 1-3	.37	.33	.48	.37	.36	.31	.38
- 3-5	.31	.41	.25	.33	.26	.16	.10
- 5-10	.11	.14	.05	.14	.13	.05	.03

Notes. From the raw data, we drop all vacancies that fit either of the following conditions: not full-time jobs, having outlier wages, having job descriptions with fewer than 20 words, posted at year 2013, posted by firms with less than 10 posts, with work locations that have less than 1000 vacancies, and non-Chinese posts. The average number of words are the number of Chinese characters or English words in the job descriptions. The posted wage is calculated as the mean of the wage lower bound and wage upper bound documented in the vacancy. Vocational school in China means a 2- or 3-years college curriculum which focuses on vocational training comparing to academic training and does not offer Bachelor degree. Not specified education can have different meanings on different cases but generally would indicate a lower bound of education level down to high school or vocational college.

[22] We use this keeping strategy to avoid the case that firms post the original vacancy with wage too low to attract any fitted workers and have to repost the same vacancy but with a raised wage which is now more close to the market level. However, this strategy will also remove the case that the firm simply repost the same job with an inflated wage.

Summary Statistics. Table 1 shows the summary statistics both for the pooled sample and for three selected major occupations. In total our final sample contains around 4 million posted vacancies from over 86 thousand firms. Under our occupation classification, this includes 33 percent vacancies in Computer occupations, 14 percent in Design & Media occupations, 29 percent in Business Operation occupations, 5 percent in Financial & Legal occupations, 11 percent in Sales occupations, and 7 percent in Administrative occupations. The numbers of firms that post vacancies in each major occupation are between 70 percent to 90 percent of the total number of firms in the pooled sample, except for Financial & Legal occupations (50 percent). In fact over 40 thousand firms in our data post vacancies in more than four major occupations, although on average firms have fewer vacancies posted in Sales and Admin occupations (5-8 vacancies) comparing to Computer and Business Operation occupations (14-17 vacancies). Hence, a majority of the firms in our sample post vacancies in multiple occupations, which allows us to study both the firm level pay differences and the potential pay differences across different occupations within the firm. Also, the average number of words in a vacancy is quite similar across different board occupations, suggesting that firms do not behave very differently on their information closure.

As we have explained earlier, the information on firm size and education and experience requirement shows that our vacancy data inclines to a young and high-end part of the labor market. Most firms in our data are middle to large sized, evenly distributed across four size categories: 15 to 50 employees, 50 to 150 employees, 150-500 employees and more than 500 employees. In comparison, firms with less than 15 employees accounts for only 3 percent, mainly due to our sample trimming strategy which cuts off all firms with less than 10 vacancy posts. The fact that firm size distributions are close across different occupations again suggests that we have the same set of firms that post jobs in different occupations. In terms of required education, among all the vacancies, 33 percent requires some college degree, 54 percent requires bachelor degree, 1 percent requires post-graduate degrees, and 12 percent has no requirement on education.[23] The high requirement on education level is due to both the nature of online job market and the large demand on cognitive-intensive jobs in the IT-producing and IT-using industries. In terms of required experience, close to 70 percent of the vacancies require 1 to 5 year experience, 22 percent do not require any experience, and 11 percent require 5 to 10 years experience.

Different from vacancy text length or firm size, education and experience requirements and posted wage vary substantially across different major occupations. In particular, Computer occupations vacancies have the highest average posted wage at CNY 17.4 thousand per month.[24] In comparison, the average posted wage of administrative vacancies is only around one-third of this number, CNY 6.3 thousand. Monthly wage in other occupations locate in between CNY 10 thousand to 14 thousand. This difference in posted wage goes hand in hand with education and experience requirements. While over 60% of the vacancies in Computer, Business Op-

[23] No specified requirement on education can have different meanings depending on different cases. But in general this indicates that the firm will have a lower requirement on the formal education level than the normal case and in most of the cases this means the lower bound can go down to high school or vocational college degree.

[24] The average wage is calculated as the mean of the lower bound and higher bound of the posted wage range. This mean wage of Computer occupations translates to 31,600 US dollars annual earning by using a currency ratio of 1USD:6.6CNY and then multiplying with 12 (months), and is three times over the Chinese GDP per capita in 2020 (10,500USD).

eration, and Financial & Legal occupations require bachelor degree or graduate degree, only around 20% of Sales and Administrative vacancies require an undergraduate or above. Those occupations requiring a higher education level also more often require higher than three year experience, while those occupations requiring lower education levels usually require 0 or 1 to 3 years work experience. This may indicate potential complementarity between college education and on-the-job training or learning by doing, and, if training or learning on the job develops within-occupation skill variations, complementarity between formal education and specific skills or tasks required on the vacancies. Given this distinction in the posted wage and education and experience requirements, we thus select Computer occupations, Design & Media occupations, and Administrative occupations as the representative high-, middle- and low-level occupations and show their results in the following analysis. However, all of our qualitative results hold if we pick say Business Operation occupations as middle-level and/or Sales occupations as low-skill occupations.

4 A First Look At The Posted Wage Inequality

In this section, we set up our baseline econometric model by following the literature of wage differential (Abowd et al., 1999; Card et al., 2013; Barth et al., 2016; Song et al., 2019) and preliminarily estimate different components of posted wage variation in our data. At this moment, we simply rely on the traditional controls on the worker characteristics, i.e. education, experience, and occupation, which arguably represent rough proxies of the skills of workers and generate biased results under unobserved skills and tasks required on the job. One of our aim here is to do a preliminary check on the estimates of labor market wage variations based on online vacancy data, by taking the results in the previous studies that use administrative data of different countries as the reference. In Section 6 we will return back to the estimation conducted here, and that time, with the full controls on all specific job characteristics documented in the job vacancy text and extracted in Section 5.

Our baseline specification of the log wage regression is

$$\ln w_{i,j,o,t} = X_i \beta_o + \psi_{j,o} + \iota_{t,o} + \epsilon_{i,o} \tag{1}$$

, where $j \equiv j(i)$ is the firm that posts vacancy i, $o \equiv o(i)$ can be either the major occupation that the vacancy is classified or the labor pooled market, and $t \equiv t(i)$ is the year that the vacancy was posted. X_{it} is a vector of job vacancy characteristics, which currently can incorporate the dummies for education and experience levels required and for the minor occupation that we assign for each job vacancy in Section 3. β_o is the coefficients for job characteristics, $\psi_{j,o}$ is the firm effects, $\iota_{t,o}$ is the year effect, and $\epsilon_{i,o}$ is the wage residual, all of which can vary across occupations for the same vacancy i.[25]

[25]The employer identifier used here is likely to be coarser than the one used in other studies using administrative or census data. In our data, while in some cases different establishment establishments or subsidies of one firm are identified differently, in other cases they are labeled as the same firm. As a result, some part of the between-establishment wage differential might be identified here as within-firm wage difference if firm have different pay policies across different branches. However, our estimation at occupation level would relieve some of this bias if firms incline to gather workers of the same occupation into one branch.

There are several deviations of this specification from the literature that are worth mentioning. First, instead of using worker characteristics (Barth et al., 2016) or worker fixed effects (Card et al., 2013; Song et al., 2019), here we replace it with job characteristics in the posted vacancies to represent the worker side effects on wage determination. By doing this we implicitly assume that firms post the education and experience level, along with all the other skills and tasks that we will study later, of the ideal job candidates that they will eventually hire and the posted wages are partly based on their pricing of these requirements. As a result, the estimated β can be seen as the average price of various skills and tasks required by firms. If there is mismatch in the labor market and such mismatch is rather random, our setting will be safe and even reduce some measurement error, but in other cases the wage and job information in the vacancy data may be deviated from the real labor market to some degree. Second, in additional to the estimation on the pooled sample, we also conduct our estimation separately on the major occupational level. This thus allows for the skill and task prices, the firm wage premiums, and the year effects to be variant across different occupations o, and the estimated results on the pooled data will be some aggregates of the results in the occupational level.[26] We will later see in our results that firms do have different levels of wage premium and of firm-worker sorting across different occupations. Another reason for occupational level estimation is that it helps use to more clear the distinction between within-occupation skill and task variations and between-occupation skill and task variations. Third, given that we do not use two-way fixed effects which rely on job movers between firms to identify the firms effect, there is no need to construct connected sample set or worry about the exogenous mobility assumption. However, we still need to assume additive separability for the job effect and the firm effect and there can still have limited mobility bias if the number of vacancies posted by some firms are quite small. We will return back to this problem in the end of this section. Finally and relatedly, lacking of a worker fixed effect, we potentially fail to control for many unobserved skills and tasks variations that are correlated with both the currently observed job characteristics and the firm fixed effects, and thus our estimation results on $\hat{\beta}$ and $\hat{\psi}$ are likely to be upward biased.[27] This motivates us to use machine learning methods to unwrap those unobserved job characteristics from the vacancy text data in Section 5.

Next we conduct variance decomposition to the results we obtained from our baseline specification so that we can divide the total dispersion in the posted wage to components of job

[26] We suggest that there is no reason to restrict that all firms in the labor market would pay exactly the same prices for one skill or task and exactly the same wage premium for all their workers. For example, economist have found that there are large discrepancies in the wage premium of college majors across industries and regions. In the case of firm wag premium, it is also an issue of definition that on which level do we interested in the unknown effect that firm pay differently for similar workers. For example, Card et al. (2016) examine the gap in firm wage premium between male and female workers and Kline et al. (2020) find different levels of firm effects and firm-worker sorting between young and old workers. Also, Card et al. (2013) find that a large and increasing part of the between-occupation variance in mean wages is due to the covariance between the mean worker effect within an occupation and the mean establishment wage premium for that occupation, indicating that firms might have different wage policy across occupations within the firm.

[27] For example, we can assume the error term in our main specification has the structure $\epsilon_i = \alpha_i + \varepsilon_i$, where α is unobserved skills and tasks, ε is the real random error. Then, if either $\text{cov}(\alpha_i, X_i) \neq 0$ or $\text{cov}(\alpha_i, \psi_j) \neq 0$, we would not have $E\left(\epsilon_i \mid X_i, \psi_j\right) = 0$ and the estimated β and ψ will be biased. In fact, it is likely that both observed worker characteristics and job characteristics are positively correlated with the unobserved job characteristics, and thus both $\hat{\beta}$ and $\hat{\psi}$ would be overestimated.

characteristics, firm pay policies, or their interaction. To simplify the notation, we will drop the subscript o throughout the rest of the paper unless when it is necessary. From Equation (13), and by ignoring the year effects ι and denoting $X_i\beta \equiv \theta_i$, we obtain

$$\text{var}(\ln w_i) = \underbrace{\text{var}(\theta_i)}_{\text{Job Effect}} + \underbrace{\text{var}\left(\psi_j\right)}_{\text{Firm Effect}} + \underbrace{2\,\text{cov}\left(\theta_i, \psi_j\right)}_{\text{Sorting}} + \text{var}(\epsilon_i) \tag{2}$$

. We denote the variance component due to job characteristics, $\text{var}(\theta_i)$, as the job effect, corresponding to the worker effect in the literature. Consequently, the variance component due to the firm fixed effects, $\text{var}\left(\psi_j\right)$, is the firm effect due to firm wage premium, and the covariance of these two variances, $2\,\text{cov}\left(\bar{\theta}_j, \psi_j\right)$, is the sorting between job quality and firm wage premium. Following Song et al. (2019), we can further rewrite the variance decomposition in (2) into

$$\text{var}(\ln w_i) \quad = \underbrace{\text{var}\left(\theta_i - \bar{\theta}_j\right) + \text{var}(\epsilon_i)}_{\text{Within-firm component}} + \underbrace{\text{var}\left(\bar{\theta}_j\right) + 2\,\text{cov}\left(\bar{\theta}_j, \psi_j\right) + \text{var}\left(\psi_j\right)}_{\text{Between-firm component}} \tag{3}$$

, so that the total wage variance is divided into within- and between-firm components. The within-firm component include the variance of the deviation of each job's characteristics from the firm average level, $\text{var}\left(\theta_i - \bar{\theta}_j\right)$, and the variance of wage residual, $\text{var}(\epsilon_i)$. The between-firm component contains the variance of firm-mean level job requirements, $\text{var}\left(\bar{\theta}_i\right)$, along with the variance of firm pay premium and covariance of sorting between job and firm effects. The results of these two types of variance decomposition under two specifications—with or without our minor (close to 6-digit) occupation dummies added in X—are shown in Table 2.

We then document several patterns about the variance decomposition results that we obtain. We first focus on the Pooled sample. The specification of only education and experience dummies in X in Panel A generates a job effect accounting for only 28 percent of the total posted wage variation, substantially smaller than the results in the AKM literature, which in most cases find a worker effect over 50 percent. Correspondingly, the variance due to the residual wage (37 percent) is abnormally high and the estimated firm effect (21 percent) are likely to be upward biased due to unobserved job characteristics. However after we add detailed occupation dummies into X in Panel B, the job effect, $\text{Var}(\theta_i)$, now accounts for 41 percent of the overall wage differential, and the firm effect, $\text{Var}\left(\psi_j\right)$ and residual wage variance, $\text{Var}(\epsilon_i)$, now reduce to 18 percent and 28 percent, respectively. Although the wage variance accounted by job effect is still lower than the level in the literature, the estimated level of firm effect and level of firm-job sorting (13 percent) are already within the range that have documented in the early studies that relies on administrative data and two-way fixed effects.[28] Moreover, the further decomposition of the job effect in our pooled data shows that the within-firm job

[28] Bonhomme et al. (2020) compare the effect of bias correction on the existing results in previous papers that use linked employer-employee data and AKM approach. They find that while in the literature the firm effect ranges from less than 10 percent to over 25 percent and the firm-worker sorting ranges from minus 10 percent to 10 percent, after using correlated random-effects (CRE) method (Bonhomme et al., 2019) or heteroskedastic fixed-effects method (Kline et al., 2020) to correct the limited mobility bias, the firm effect now ranges from 5 to 15 percent and the firm-work sorting ranges from 5 to 20 percent.

Table 2: Posted Wage Variance Decomposition

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.360	-	.279	-	.251	-	.164	-
Panel A: X={EDU, EXP}								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
Within-Firm:								
$\text{Var}(\theta_i - \bar{\theta}_j)$	.072	.199	.037	.133	.036	.144	.033	.204
$\text{Var}(\epsilon_i)$	.132	.367	.089	.318	.078	.310	.061	.371
Between-Firm:								
$\text{Var}(\bar{\theta}_j)$	.030	.084	.015	.055	.017	.068	.017	.102
$\text{Var}(\psi_j)$	.076	.212	.102	.365	.086	.342	.041	.253
$2\,\text{Cov}(\bar{\theta}_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel B: X={EDU, EXP, OCC}								
$\text{Var}(\theta_i)$	.146	.407	.065	.232	.061	.243	.052	.320
Within-Firm:								
$\text{Var}(\theta_i - \bar{\theta}_j)$	.103	.286	.049	.176	.040	.159	.035	.214
$\text{Var}(\epsilon_i)$	.101	.280	.077	.275	.074	.295	.059	.361
Between-Firm:								
$\text{Var}(\bar{\theta}_j)$	.044	.121	.016	.057	.021	.085	.017	.107
$\text{Var}(\psi_j)$	.064	.179	.096	.344	.079	.314	.040	.245
$2\,\text{Cov}(\bar{\theta}_j, \psi_j)$	.048	.134	.041	.148	.037	.148	.012	.074
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Notes. In the specification of panel A, X only contains education and experience dummies, while in the specification of panel B, X also includes minor (close to 6-digit) occupation dummies. The variance and covariance terms related to year dummies have been subtracted from the total variance of log wage, and thus the sum of all with-firm and between-firm components would be equal to the total variance of log wage. Job effect $\text{Var}(\theta_i)$ is the sum of with-firm job variations $\text{Var}\left(\theta_i - \bar{\theta}_j\right)$ and between firm job variations $\text{Var}\left(\bar{\theta}_j\right)$. For each major occupation sample estimated, we drop vacancies belong to the firms that have less than two vacancy posts in this major occupation.

characteristics differences, $\text{Var}\left(\theta_i - \bar{\theta}_j\right)$, more than doubles the between-firm differences in firm average job effect, $\text{Var}\left(\bar{\theta}_j\right)$, which is also consistent with the results in the literature.[29] These results thus give us some confidence on using vacancy data to study wage differential and also incentive on further improving the estimation by looking for better controls of job characteristics.

Now we turn to two systematic differences between the estimation results from the Pooled sample and the results from occupational-level estimations. The first difference is that, while in the pooled sample adding occupation dummies increase 12 percent points for job effect, the effect is much smaller in the occupational level, range from 1 to 4 percent points. And as a result, the share of job effect in the Pooled sample is significantly larger than the shares in all individual occupations. This result suggests that those detailed occupation categories work as a good control for between occupation skill and task variations mainly by separating very different board occupation categories and account for much less variations for the skill and task variations within each major occupation. The second difference is that the share of wage variance accounted by the firm fixed effects estimated in individual occupations are substantially greater than the one estimated in the Pooled sample. There can have several reasons contributing to this result. First, simply following the argument above, if there are more unobserved job characteristics comparing to the observed ones within the occupational level comparing to the overall level, then a stronger unobserved bias can contribute to a more upward biased firm effects within each individual occupation. Second, when we split our Pooled sample for different major occupations, in each subsample of a single major occupation we will have more firms with few posted vacancies, and thus our plug-in variance and covariance terms calculated under the results of Equation (13) for each single occupation will be subject to more finite sample biases due to the high-dimensional firm fixed effects. This bias is in nature related to the limited mobility bias in the AKM literature but in a way simpler form in our case.[30] To solve this problem, we try two bias-correction methods in the literature: the homoscedasticity estimator suggested by Andrews et al. (2008), and the heteroskedastic leave-out method suggested by Kline et al. (2020). The results in Table D1 show that for our case, both method generate very similar corrections, reducing the variance of firm effect and increase the variance of residual wage.[31] This correction is more significant when the finite sample problem is severe, like the Administrative occupation, where we have more firms with very few vacancies posted and the share of firm effect decrease over 6 percent points, and is more limited in our Pooled sample

[29] In Song et al. (2019), their estimation results on the U.S. labor market between 2007 and 2013 shows that the within-firm worker effect accounts for 38 percent of the total wage variance while the between-firm worker effect accounts for 13 percent.

[30] This econometric problem can be traced back to Krueger and Summers (1988) where they use individual characteristics and industry dummies to study industrial level wage premiums. In our vacancy data case, the bias can be also recognized as a kind of "limited mobility bias" even though we don't rely on observing worker moving between different firms. This is because in fact each vacancy can be seen as a worker movement to a new job in the labor market, whether from another firm or from unemployment or formal education, and the idea that a firm has too few job vacancies is roughly equal to the idea that we observe too less mobility in the employer-employee data for this firm.

[31] The reason why the correction does not affect the job effect and the firm-job sorting is that given the low dimension nature of our controls for the job characteristics, the job effect will not be subject to the finite sample bias at all. And in this case, the bias on the sorting of two effects will be on a second order, and in practice be very small.

and the subsample of Computer occupation, only reducing the firm fixed effect for less than one percent point and one and half percent point respectively.[32] The third reason that for the higher share accounted by the firm effects estimated in individual occupational level is that if firm have different wage premium policies across different occupations, then our estimates on the Pooled sample would discard any variations of firm fixed effects across different occupations within the firm and produce low estimated share for firm effect variances.[33] We will explore this possibility more carefully in the later section.

The final pattern in Table 2 is the difference in shares of wage components across different types of occupations. Computer occupation has the least share of job effect and highest share of firm effect, while Administrative occupation has the inverse result. This indicates that in high-skill occupation like Computer occupations, the variations of unobserved skills and tasks is large and firms likely differ in the requirement of such job characteristics, while in low-skill occupation like Admin occupations, education and experience are already the most important variances of job qualities. Moreover, the share due to firm-worker sorting in either Computer occupation or Design & Media occupation (15 percent) doubles the share of sorting in Administrative occupation (7 percent), and the increased share due to sorting in Panel B after adding minor occupation dummies is smaller than the other occupations. This implies that sorting is probably more important in high-skill or professional occupations, and potentially links to the unobserved skill and task variations beyond education and experience differences.

In summary, our preliminary results show that the estimated composition of posted wage differential in our Chinese vacancy data is roughly consistent with the results in previous studies that use administrative employer-employee data in rich countries. However it also shows that our current results are likely to be biased due to unobserved job characteristics, which appears to be especially important for within-occupation wage differentials. In particular, the shares of wage variation accounted by job effect are likely to be underestimated, the shares accounted by firm fixed effects are likely to be overestimated, and the shares due to firm-job sorting is biased unambiguously. In next section, we try to extract all unobserved skills and tasks as well as the job compensations and amenities that are documented in the job vacancy texts, and in Section 6 we will return back to the wage differential estimation conducted here and see how the new results improve our understanding of different wage determinants.

[32] Another simple way to solve this problem is to drop the firms with too few vacancies. In practice, we find if we further drop more low vacancy firms in each major occupation sample, the firm effects would keep declining but other components changes either slightly or moderately depending on the subsample. In particular, removing all firms with less than five vacancies in Computer occupations will decrease the variance share of firm effect about 0.4 percent points while all other components barely change. Whereas doing the same dropping in Administrative occupation will reduce firm effect about 5 percent points and between-firm job component 2 percent points and increase within-firm job component three percent points. This large change in Admin occupation is because over 80 percent firms post less than five vacancies in Admin occupation and thus removing them hugely reduce the sample size.

[33] However, this is also a problem of the definition of firm premium. Strictly speaking, a firm wage premium might be defined as a fixed premium equally given to all its employees. But if a firm decides only pay half of its member a wage premium that is higher than market level, how do we decide the level for this firm's wage premium? If such cases do exit, it also raises the following question that why and how a firm decides its wage premium across different employees. And more fundamental question eventually goes to where do firm wage premiums come from.

5 Use Machine Learning To Understand Vacancy Text

In this section, we apply a series of machine learning algorithms to the job vacancy text data to extract information on detailed job skill requirements and task descriptions, as well as job benefits and compensations. Our first aim here is to select wage-predictive terms from the high-dimensional vacancy text data where both informative and meaningless information about wage determination coexist. In other words, we want the data to tell us what are the useful job characteristics embedded the vacancy text, whether skills, tasks, amenities, or any other terms, that can explain the posted wage variations. We achieve this in Section 5.1 by using regularized linear regression which reduce the effective feature dimension from the whole vocabulary in the data to a few thousand terms. Our second aim is to understand what are the job characteristics that we selected in the last step and, if possible, to classify them into board genres, again through a data-driven perspective. We achieve this in Section 5.2 by using both natural language processing (NLP) algorithm and unsupervised clustering algorithm to conduct feature clustering based on how firms talk about different things in the job vacancy. In this process, we show that our algorithms automatically separate different job skills and tasks and non-wage compensation and amenities, and generate a data-driven skill-task hierarchical structure. Our final aim in this section is to construct low dimensional proxy variables for the useful job characteristics that we have identified and classified in above steps so that we can bring these information back to our wage differential estimation and show how these previously unobserved skills, tasks, and compensations could improve our understanding on the wage determination and total earning inequality. This further dimensional reduction is achieved by using supervised dimensional reduction algorithm in Section 5.3. Throughout this section, our selections on a variety of machine learning algorithms largely follow the suggestions in Gentzkow et al. (2019), in which the authors review the applications of a wide range of machine learning techniques on text data and economics topics.

5.1 Features Selection

Our first step is to select important features from the raw text of the job descriptions on the job vacancies that incorporate a variety of skills, tasks, non-wage benefits, and perhaps other contents, and by important here we mean holding some predictive power for the posted wage, whether causal or not. Because a feature is often called a "token" in the textual analysis and means either a word or a phrase (or more generally a term), we will use these words interchangeably throughout the paper. To this end, we need first transform the raw job vacancy texts, denoted by $\mathbf{D}$, into a numerical token matrix $\mathbf{C}$ which has dimension $N \times K$. Here N is the number of vacancies in the sample data, and K is the number of tokens in the whole vocabulary set V. V is tokenized from all vacancies' texts of the data after standardization and removing words that do not convey meaningful or interpretable information.[34] Each entry of $\mathbf{C}$, indexed by c_{ik}, is an indicator of the presence of token k in vacancy i (1 if present otherwise 0). The details of this transformation are described in Appendix B.1.

[34]For example, all numerical numbers either in Arabic or in Chinese are removed because we have no idea what they interpret without the context. We also remove all the firm name from the vocabulary because it will catch the firm effect and distort the clustering of selected features.

Next, we regress the log posted wage on the token matrix $\mathbf{C}$ to estimate the explanatory power of each tokens in V. Unlike a normal regression problem, the high dimensionality of $\mathbf{C}$, in which many dimensions could be totally irrelvant, makes standard techniques like Ordinary Least Square (OLS) infeasible and unsuitable. We thus apply the penalized (also called regularized) linear models to this high-dimensional regression problem for feature selection. The penalization here add additional costs for deviations of any estimators from zero, which helps to shrink the effective dimension of explanatory variables, and the linearity retains the model to be rather intuitive and interpretable. In particular, here we choose a least absolute shrinkage and selection operator (Lasso) regression which extends the Gaussian linear regression and uses a L_1 penalization. The L_1 penalization here means that the penalization cost function is linear (zero curvature and constant shrinkage), and thus the non-differentiable spike of the additional cost at zero leads to sparse estimators, with some coefficients to be exactly zero. This strong form of penalization are particularly suitable for feature selection in text analysis because it limits nonzero estimators for prediction to a rather reasonable size and thus helps to throw out a large amount of potentially uninformative tokens in the raw text. Our lasso estimator is written as

$$\hat{\zeta} = \arg\min_{\zeta} \sum_{i=1}^{N} \left(\ln w_i - \sum_{k=1}^{K} c_{ik}\zeta_k \right)^2 + \lambda \sum_{k=1}^{K} |\zeta_k|$$

where $\lambda > 0$ is a parameter of the model that indicates the level of the "penalty". Note that the first part within the minimization is the residual sum of squares (RSS) in the normal OLS estimator, which is an unregularized objective proportional to the negative log likelihood, $-\log \mathrm{P}(\ln w_i \mid \mathbf{c}_i)$, and the second part is the penalization term.

A key difference in the estimation of Lasso comparing to the estimation of traditional econometric models like OLS is that there is now a pre-determined hyper-parameter λ, i.e. the parameter is to be set before the estimation through other procedures. This parameter controls the extent to which the model penalize non-zero estimators. The larger the λ the more sparse will be the selected non-zero estimators, and as $\lambda \to 0$ it approaches to the usual maximum likelihood estimation. The standard practice to determine this prior parameter (or "model turning" in the jargon of machine learning field) is to define a criterion to measure the performance of the estimates from different values of λ and then choose the best one from them. Although the commonly used criterions in the machine learning literature is some metrics of the model's out-of-sample prediction power, such approach has been argued to be more suitable for achieving the best predictive performance rather than for selecting features used for further analysis because it often leads to λ too small and overfitting.[35] Instead, we follow the suggestion in

[35] To be specific, this most popular tuning approach that follow the out-of-sample accuracy idea in the machine learning literature is called cross-validation. The basic step is to randomly partition a part of the data as the test sample separated from the training sample that are used to train the model, and then to apply the trained model back to test sample to calculate the results of the pre-defined measure, such as mean squared error. The one repeats this procedure for a large set of parameter values and for different random partitions to get the optimal parameter values. Although the idea of out-of-sample efficiency is exactly designed to target the problem of overfitting, empirical findings in the machine learning literature often suggest that such prediction-based approach is still very likely to overweight the predictive power to model rigidity and interpretability in many real world settings. Also note that although we also partition samples in our occupation classification approach, we do not need to

Gentzkow et al. (2019) and use the Bayesian information criterion (BIC) as the criterion to choose the optimal λ for our feature selection. Similar to the well-known Akaike's information criterion (AIC), BIC is an approximation to the Bayesian posterior marginal likelihood subject to an adjustment on degrees of freedom. In our Lasso case, the BIC is defined as

$$\text{BIC}(\lambda) = \frac{\|\ln \mathbf{w} - \mathbf{C}\hat{\zeta}_\lambda\|^2}{\sigma^2} + \widehat{df}_\lambda \log N$$

, where σ is the common variance of Gaussian noises, and $\widehat{df}_\lambda$ is the degrees of freedom of the estimation with λ.[36] In practice, we pass a grid of different λ values to the Lasso regression and find the λ^* that yields the lowest BIC score.

Table 3: BIC Tuned Lasso Models

	Pooled	Computer	Design_ Media	Admin
λ^*	332.0	190.3	238.5	155.0
MSE	.162	.149	.142	.100
R^2	.566	.494	.461	.418
BIC/N	.446	.527	.561	.613
df	3,144	1,922	929	691
K	109,123	51,602	39,306	24,896
N	3,999,005	1,330,001	561,236	277,932

Notes. For each major occupation, the hyperparameter λ^* of the Lasso model is tuned by minimizing BIC(λ) as defined in the text. The smaller the λ, the less the penalization for nonzero coefficients and the more features are picked by the Lasso model.

The tuned Lasso models and their estimation results for both Pooled sample and three selected occupations are shown in Table 3. The tuned λ^* ranges from 155 to 332 in different samples due to the different levels of tradeoff between decreased normalized RSS and increased penalty from the increased degree of freedom with a higher $lambda$. As a result, the number of tokens with nonzero estimated coefficients also varies across samples, ranging from over 3100 tokens in the Pooled sample to less than 700 tokens in Administrative occupation, a substantial reduction from the norm of the original vocabulary K. The R-squared values for the estimated models are between 57% (Pooled) to 42% (Admin), indicating that the job characteristics extracted by our Lasso model can account for around half of the entire wage

conduct such model tuning because the Naive-Bayes classifier is too simple and requires no hyper-parameters.

[36] More generally the BIC is defined as BIC $= -2\log(\hat{L}) + \log(N)\widehat{df}$, where $\hat{L}$ is the maximum likelihood under estimation. For a linear Gaussian model, the maximum log likelihood can be derived as: $\log(\hat{L}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{2\sigma^2}$, where y and $\hat{y}$ are any true and predicted targets, and σ is the "true" error variance. By bringing this term back to the definition and removing the constant terms one obtains the formula in the text. Note that there is no general way to estimate σ^2, which works as a baseline unit for the RSS so that there is a "fair" comparison between the reduction in estimation error and the increase in number of parameters. In practice, we simply estimate the σ to be Var($\ln \mathbf{w}$).

variance.

One caution that has been repeatedly raised from the literature is that the selected features and their coefficients of any high-dimensional penalized models are generally uninterpretable (see for example Belloni et al., 2014; Mullainathan and Spiess, 2017). In general the results in these models and thus in our Lasso model suffer two problems for further interpretation: multicollinearity and flexibility.[37] The first is that given the high-dimensionality and the penalization, and especially in our case with a linear regularization, the penalized regressions will likely to pick one feature at random from a highly correlated group. In other words, multicollinearity among features in such high-dimensional penalized model could cause the set of nonzero variables selected to be highly unstable. As a result, the tokens selected by our Lasso model in general do not necessarily indicate any casual relationships. The second problem is on the interpretation of the coefficient levels. Even after regularization, our models still left hundreds or thousands nonzero control variables, which makes the both the levels and the signs of the coefficients hard to be taken for any serious interpretation.[38]

Because we do want to go beyond wage prediction and to learn something about the contents of various job characteristics and their impact on wage determination from the selected features, we now check how severe are these problems of statistical uncertainty in our Lasso estimation results. To this end, we use subsampling, which is a nonparametric approach of inference and has been argued can retain robust in the cases where the estimator has non-differentiable loss function and potential model selection.[39] In practice, we randomly partition our sample into ten pieces and re-estimate the Lasso model equipped with previously tuned λ^* separately on each subsample. We repeat this procedure for ten times and gather all the results to calculate the standard deviation of our parameters of interest—the coefficients of the nonzero features selected in our Lasso estimation with the full samples.[40] The result of the

[37] On the top of these two, there could still be unobserved bias in that our captured job characteristics predict the wage not because they have direct casualty links but because they link to some other unobserved casual factors or to the firm effect. For the first possibility, we suggest that this will affect our main results because our models have extracted a fairly large amount of skills, tasks, amenities and other potential job characteristics, and eventually we will cluster these job characteristics based on their textual relationship in the text. For the second possibility, as we explained before, in our vectorization process we have already removed the terms of firm names from the vocabulary and perhaps more directly, in our selected tokens by the Lasso model we don't find specific terms are that can be used to identify any particular firms. However, if different types of firms with different levels of firm wage premiums also post certain job characteristics, being either skills, tasks, or non-wage compensations, then some features selected here will hold strong prediction power on the posted wage simply because of this indirect relationship. We show later that this hypothesis is actually true in our data.

[38] One simple but intuitive example of this problem is that if one put both age and experience variables into a high dimensional wage regression along with one hundred of other individual variables, there are chances that the coefficient of one of the age and experience variables might turn out to be negative. Although this result can still tell something informative but one need to fully acknowledge what has been conditioned on to give a reasonable interpretation.

[39] Within the inference computation algorithms that approximate the sampling distribution, the commonly-used nonparametric bootstrap uses with-replacement resampling and thus fails for the statistics models that involve non-differentiable loss functions like Lasso here. In comparison, in subsampling each subsample is a draw from the true data generating process, and thus it works for estimation algorithms even with non-differentiable losses. For more details about subsampling, we refer to Gentzkow et al. (2019,?).

[40] To calculate the statistics of interest, one needs to translate the uncertainty in the subsamples to the one in the full sample because each subsample will be smaller than the entire sample of interest. We follow the convention to assume the estimator's rate of convergence to be $\sqrt{n}$ so that the corrected standard deviation of the coefficients

subsampling is shown in Figure D2, from which we can see that the coefficients of the tokens selected in our full-sample Lasso estimation are generally robust—they don't easily flip the sign in different subsamples and their standard deviations are actually quite small in most cases. Although the robustness of our Lasso results shown in this uncertainty check does not necessarily dispel all the potential problems due to the multicollinearity and flexibility in our high-dimensional regularized feature selection, and hence any causal inference for the estimates is still largely forbidden, we think that at least it gives us some confidence that our selected features are likely to represent some stable and consistent patterns pertain to the posted wage determination.

Acknowledging the potential interpretability problems of the Lasso results, we then inspect the most important features estimated by the model to see if they make some intuitive sense and if they expose some stylized features that desire more validation and reasoning. In Table 4 and Table 5 we show the top positive and negative tokens that have the largest absolute coefficient and occurs in more than one percent of the vacancies.[41] We next document several patterns found from these top tokens and discuss their potential implications. First, not surprising, education terms appear to be the highly predictive features in our Lasso estimation. And the sign of these coefficients make intuitive sense: bachelor degree and master degree in the top positive tokens and vocational college degree in the top negative tokens. Also, several terms related to fresh graduates appear to be negatively correlated with the posted wage, which may represent the effect of working experience.[42] These intuitive features thus provide some additional sanity checks verifying that our Lasso models do find the key features that are important for posted wage determination in our data.

The second stylized fact is that a few compensation terms appear in the top tokens and these terms hold consistent pattern across occupations but distinctive pattern between positive and negative tokens. For positive top tokens, we can find backloading compensation (e.g. "14th month pay"), fringe benefits (e.g. "three meals"), advanced insurance and fund (e.g. "six insurance & one pension"), coworker quality (e.g. "guru", "maestro"), and equity compensation (e.g. "stock", "options"). For negative top tokens, we can see compensation terms of mandated insurance (e.g. "five insurance", "social insurance"), leisure time (e.g. "two-day weekend", "holiday"), and also fringe benefits (e.g. "accommodation").[43] This result thus echos the confusing

can be calculated as $\text{sd}(\hat{\zeta})\sqrt{\frac{B}{N}}$, where B is the sample size of each subsample. The robustness for our estimation results would not change significantly even if we choose a higher rate of converge.

[41] There are other ways to define the importance of coefficients of Lasso estimation, for example the absolute coefficient values scaled by the associated standard deviation or the order in which the coefficient of a covariate first turns to nonzero in a series of Lasso estimations with decreasing penalties. Here by simply showing the top positive and negative tokens with larger than 1 percent occurrence we aim to display the tokens with the highest prediction power that are not rare. We show the full list of all nonzero tokens selected by the Lasso estimations on Appendix D.2.

[42] While we have the education terms in our features, we don't have any direct work experience terms in our vocabulary V because the working experience is often documented in the vacancy as "required n year experience" where n is pure number and thus dropped from the vocabulary because they are also used in many other cases and hard to interpret.

[43] The most representative form in Chinese social insurance system are "five insurance and one fund". "Five insurance" means endowment insurance, medical insurance, employment insurance, employment injury insurance and maternity insurance and is mandated by law. "One fund" means housing provident fund, which is not compulsory by law but a large percent of formal firms, especially those large sized, will pay this fund for their workers. "Six insurance" means five basic insurance and one additional commercial supplementary medical insur-

Table 4: Top Positive Tokens (Frequency > 1%) in Lasso Regression

	Pooled			Computer			Design_Media			Admin		
	token	coef	freq	token	coef	freq	token	coef	freq	token	coef	freq
1	14薪(14th month pay)	.152	.014	15薪(15th month pay)	.181	.010	14薪(14th month pay)	.193	.011	大学本科(undergraduate)	.161	.014
2	三餐(three meals)	.143	.014	三餐(three meals)	.148	.014	带领(lead)	.155	.025	本科(undergraduate)	.157	.156
3	大平台(large platform)	.131	.019	14薪(14th month pay)	.140	.017	三餐(three meals)	.129	.015	总裁(president)	.120	.014
4	硕士(master degree)	.126	.015	硕士(master degree)	.109	.027	c++(c++)	.121	.017	ceo(ceo)	.117	.010
5	带领(lead)	.107	.041	带领(lead)	.089	.038	危机(crisis)	.113	.011	搭建(build)	.117	.016
6	c++(c++)	.092	.051	golang(golang)	.080	.017	游戏(games)	.098	.180	带领(lead)	.105	.017
7	算法(algorithm)	.082	.061	大牛(guru)	.079	.047	欧美(europe & america)	.090	.011	政府(government)	.103	.030
8	大牛(guru)	.082	.028	深度学习(deep learning)	.078	.022	引擎(engine)	.090	.046	高薪(high salary)	.089	.018
9	知名(famous)	.079	.019	知名(famous)	.070	.014	4a(4a)	.090	.014	翻译(translation)	.083	.012
10	机器学习(machine learning)	.077	.016	高薪(high salary)	.070	.018	六险一金(six insurance & one fund)	.086	.046	本科学历(bachelor degree)	.082	.018
11	组建(formation)	.076	.013	牛人(maestro)	.068	.012	财经(finance)	.084	.016	战略(strategy)	.077	.015
12	本科(undergraduate)	.074	.319	海外(overseas)	.067	.010	本科(undergraduate)	.078	.238	大型(large scale)	.076	.030
13	海外(overseas)	.072	.026	go(go)	.065	.027	上市公司(listed company)	.076	.021	落地(landing)	.070	.018
14	react(react)	.072	.020	c++(c++)	.064	.144	金融(finance)	.076	.031	项目管理(project management)	.067	.011
15	开发(development)	.071	.374	算法(algorithm)	.064	.164	外包(outsourcing)	.074	.012	海外(overseas)	.066	.021
16	大学本科(undergraduate)	.066	.029	react(react)	.064	.061	大牛(guru)	.070	.022	背景(background)	.064	.032
17	高薪(high salary)	.063	.028	机器学习(machine learning)	.061	.045	海外(overseas)	.068	.024	制定(develop)	.063	.097
18	落地(landing)	.060	.067	落地(landing)	.061	.037	记者(journalists)	.068	.011	13薪(13th month pay)	.063	.019
19	战略(strategy)	.057	.047	开发(development)	.059	.776	13薪(13th month pay)	.068	.023	统招(unified recruitment)	.058	.031
20	直播(live streaming)	.056	.014	音视频(audio & video)	.058	.012	c4d(c4d)	.066	.021	预算(budget)	.057	.021
21	上市公司(listed company)	.055	.027	统招(unified recruitment)	.054	.044	知名(famous)	.065	.023	重大(major)	.055	.019
22	大型(large scale)	.055	.072	北京(beijing)	.053	.012	unity(unity)	.065	.043	装修(decoration)	.055	.016
23	职责(responsibilities)	.055	.048	直播(live streaming)	.052	.011	高薪(high salary)	.064	.016	资源(resources)	.053	.043
24	班车(shuttle)	.054	.018	推荐(recommend)	.052	.023	管理工作(management)	.063	.010	推动(promote)	.051	.029
25	金融(finance)	.054	.070	管理工作(management)	.051	.016	3d(3d)	.063	.106	金融(finance)	.051	.036
26	六险一金(six insurance & one fund)	.053	.055	ai(ai)	.051	.015	大型(large scale)	.063	.043	英语(english)	.050	.054
27	python(python)	.052	.066	股票(stock)	.049	.025	性能(performance)	.063	.016	商务谈判(business negotiations)	.048	.010
28	总监(director)	.052	.022	本科(undergraduate)	.048	.365	统招(unified recruitment)	.059	.019	优化(optimization)	.046	.079
29	统招(unified recruitment)	.051	.042	薪资(salary)	.048	.049	大学本科(undergraduate)	.059	.023	职责(responsibilities)	.046	.035
30	hive(hive)	.051	.013	补充(supplementary)	.045	.019	ip(ip)	.057	.017	统筹(integrated planning)	.046	.028
31	技术(technology)	.049	.285	金融(finance)	.045	.057	指导(guidance)	.054	.047	上市公司(listed company)	.045	.020
32	引擎(engine)	.049	.017	建设(construction)	.045	.078	设计(design)	.054	.546	出差(business trip)	.045	.038
33	团队(team)	.048	.552	高级(advanced)	.045	.022	职责(responsibilities)	.054	.043	集团(group)	.044	.018
34	期权(options)	.047	.052	大型(large scale)	.043	.113	主导(leading)	.052	.025	指标(indicators)	.043	.033
35	收入(revenue)	.047	.019	六险一金(six insurance & one fund)	.041	.057	动效(dynamic effects)	.050	.016	整体(overall)	.042	.023
36	集团(group)	.046	.022	职责(responsibilities)	.041	.049	数值(numerical value)	.050	.012	规划(planning)	.042	.036
37	生态(ecology)	.045	.012	期权(options)	.041	.062	作品集(portfolio)	.049	.021	转化(transformation)	.042	.011
38	主导(leading)	.045	.025	指导(guidance)	.040	.076	角色(roles)	.049	.053	梳理(combing)	.041	.016
39	增长(growth)	.044	.021	架构设计(architecture design)	.040	.133	落地(landing)	.049	.041	公关(public relations)	.040	.021
40	股票(stock)	.044	.022	广告(advertisement)	.040	.015	产出(outputs)	.048	.033	管理工作(management)	.039	.110

Notes. These are the tokens selected our Lasso models that have highest (or lowest) coefficients and occurs in more than 1 percent of the sample vacancies. In the Appendix D.2 we list all the nonzero features selected by Lasso. Although the Lasso coefficients of our model means the percentage rise of the expected wage for the occurrence of the certain word in the vacancy text, these coefficients generally do not indicate any casual relationship due to the strong multicollinearity among features and the flexible structure of the Lasso model (see our discussion in the main text). The more recommended way of interpretation is that these features hold strong prediction power for the posted wage and potentially are or are correlated with some important factors of wage determination.

Table 5: Top Negative Tokens (Frequency > 1%) in Lasso Regression

	Pooled			Computer			Design_Media			Admin		
	token	coeff	feq	token	coeff	feq	token	coeff	feq	token	coeff	feq
1	应届生(freshmen)	-.155	.018	毕业生(graduates)	-.205	.013	应届生(freshmen)	-.188	.017	五险(five insurance)	-.070	.052
2	五险(five insurance)	-.136	.030	五险(five insurance)	-.197	.016	实习(internship)	-.133	.011	毕业生(graduates)	-.061	.082
3	毕业生(graduates)	-.128	.033	大专(vocational college)	-.134	.072	五险(five insurance)	-.132	.033	中专(vocational school)	-.059	.038
4	专科(vocational major)	-.100	.036	社保(social insurance)	-.121	.012	毕业生(graduates)	-.132	.030	应届生(freshmen)	-.057	.048
5	双休(two-day weekend)	-.098	.166	专科(vocational major)	-.119	.030	双休(two-day weekend)	-.090	.176	实习(internship)	-.056	.012
6	大专(vocational college)	-.094	.148	双休(two-day weekend)	-.115	.147	应届(recent graduate)	-.072	.026	实习生(interns)	-.053	.017
7	助理(assistant)	-.079	.011	应届(recent graduate)	-.106	.011	大专(vocational college)	-.070	.144	双休(two-day weekend)	-.051	.214
8	客服(customer service)	-.075	.030	测试用例(test cases)	-.067	.068	社保(social insurance)	-.068	.023	玩家(player)	-.046	.024
9	社保(social insurance)	-.073	.028	安装(installation)	-.067	.048	专科(vocational major)	-.066	.041	普通话(mandarin)	-.046	.172
10	会计(accounting)	-.071	.019	th(th)	-.066	.014	有限公司(ltd.)	-.059	.012	女性(women)	-.038	.015
11	住宿(accommodation)	-.067	.016	电脑(computer)	-.065	.011	专业不限(any major)	-.055	.011	社保(social insurance)	-.037	.060
12	行政(administration)	-.067	.027	售后(after sales)	-.061	.011	人性化(humanization)	-.055	.019	qq(qq)	-.037	.036
13	专员(commissioner)	-.063	.011	年轻(young)	-.060	.013	漫画(comics)	-.053	.014	轻松(easy)	-.035	.043
14	淘宝(taobao)	-.059	.015	五险一金(five insurance & one fund)	-.059	.273	cad(cad)	-.052	.010	网站(website)	-.033	.032
15	协助(assistance)	-.058	.164	出差(business trip)	-.051	.030	photoshop(photoshop)	-.049	.235	清洁(cleaning)	-.030	.015
16	ps(ps)	-.056	.029	记录(records)	-.048	.015	cdr(cdr)	-.047	.012	卫生(health)	-.029	.024
17	有限公司(ltd.)	-.056	.012	吃苦耐劳(hardworking)	-.048	.015	网站(website)	-.047	.180	文员(clerks)	-.029	.014
18	安装(installation)	-.055	.020	节日(holidays)	-.046	.059	协助(assistance)	-.046	.131	考勤(attendance)	-.029	.104
19	photoshop(photoshop)	-.052	.039	客户(clients)	-.046	.078	ps(ps)	-.045	.142	电子商务(e-commerce)	-.029	.031
20	细心(careful)	-.050	.032	轻松(easy)	-.043	.017	吃苦耐劳(hardworking)	-.044	.023	录入(input)	-.028	.044
21	吃苦耐劳(hardworking)	-.050	.032	软件测试(software testing)	-.043	.047	动漫(anime)	-.044	.019	轮班(shift)	-.028	.013
22	核对(verification)	-.048	.011	微信(wechat)	-.041	.042	轻松(easy)	-.044	.033	接听(answer the phone)	-.027	.101
23	人力资源(human resources)	-.047	.032	.net(.net)	-.041	.034	接触(contact)	-.042	.011	行政(administration)	-.027	.256
24	网站(website)	-.047	.090	耐心(patience)	-.040	.023	编辑(editor)	-.039	.204	全勤奖(perfect attendance award)	-.026	.032
25	专业不限(any major)	-.047	.020	网站(website)	-.039	.101	美工(artwork)	-.038	.032	应聘(apply for the job)	-.025	.018
26	人性化(humanization)	-.046	.012	专注(focused)	-.038	.011	论坛(forum)	-.038	.034	移动(mobile)	-.025	.013
27	excel(excel)	-.046	.047	网络设备(network equipment)	-.037	.016	淘宝(taobao)	-.038	.024	吃苦耐劳(hardworking)	-.025	.055
28	普通话(mandarin)	-.045	.027	bug(bug)	-.036	.053	年轻(young)	-.038	.034	加入(join)	-.024	.041
29	交代(explanation)	-.044	.013	作品(works)	-.035	.023	提成(commission)	-.037	.017	游戏(games)	-.024	.039
30	年轻(young)	-.044	.025	节假日(holiday)	-.034	.037	客户(clients)	-.037	.096	前台(front desk)	-.023	.088
31	接触(contact)	-.044	.010	分红(dividend)	-.034	.012	微信(wechat)	-.037	.172	部门经理(department manager)	-.023	.014
32	轻松(easy)	-.043	.027	故障(failure)	-.033	.055	玩家(player)	-.037	.017	资料(information)	-.023	.122
33	致力于(commitment)	-.043	.014	自主(autonomy)	-.033	.014	coreldraw(coreldraw)	-.037	.041	倒班(shift)	-.023	.015
34	应届(recent graduate)	-.043	.029	双薪(double pay)	-.033	.035	上级(higher)	-.036	.034	淘宝(taobao)	-.022	.047
35	五险一金(five insurance & one fund)	-.043	.294	培训(training)	-.033	.076	上传(upload)	-.036	.014	广阔(wide)	-.022	.024
36	编辑(editor)	-.042	.042	ssh(ssh)	-.033	.010	细心(careful)	-.033	.028	服从(obedience)	-.022	.029
37	招聘(recruitment)	-.041	.057	xcode(xcode)	-.033	.016	加入(join)	-.033	.048	客户档案(customer profile)	-.022	.016
38	seo(seo)	-.041	.010	细心(careful)	-.032	.015	耐心(patience)	-.031	.036	社会保险(social insurance)	-.022	.015
39	成立(established)	-.041	.011	专业优先(professional priority)	-.032	.024	节日(holidays)	-.031	.084	档案(archives)	-.022	.046
40	电脑(computer)	-.039	.014	测试报告(test report)	-.032	.037	文字(text)	-.031	.229	地点(location)	-.022	.045

Notes. See the note in Table 4

results in the compensation differential literature: the estimated coefficients from the hedonic regression are mixed and sometimes inconsistent with the theoretical prediction. However, one can notice that many compensations represented by the positive tokens are performance pay or fringe benefits that potentially encourage effort, long-hour or inflexible worktime, and learning, and prevent turnover cost. On the other hand the compensation in the negative tokens seem to indicate that the compensations related to work-life balance follows the classic compensation differential mechanism. Also, the case of insurance and fund provided by firm is quite interesting: the basic mandated level of insurance is negatively correlated with posted wage while the enhanced package of insurance and fund is positively correlated with posted wage, running exactly inverse to the theory of compensation differential. These features suggest that the efficiency of different non-wage compensations might be an important aspect when thinking about non-wage compensation provision in the labor market. One possible reason that these non-wage compensations hold such strong power on posted wage prediction is that they are actually correlated with firm effect. However, this argument will suggest that firms have very different strategies for compensation provision even when they are likely to have similar cost functions for providing these non-wage compensations, so that these differently provided compensations are correlated with the posted wage in a distinctive way. It thus raises the question of how and why firms decide the different packages of wage and non-wage compensations for their workers and the question that if there are some mechanisms other than compensation differential working in the labor market. We will try to examine and answer these questions in Section 7.

The third stylized fact is that there are many occupation-specific and professional terms in the top tokens. In top positive tokens we observe for example "deep learning", "golang", and "c++" in Computer occupation, "engine", "3d", and "journalist" in Design & Media occupation, and "translation" and "business negotiation" in Admin occupation. And in top negative tokens we can observe for example "installation" and "computer" in Computer occupation, "photoshop" and "editor" Design & Media occupation, and "mandarin" and "answer the phone" in Admin occupation. Again the signs of these features take intuitive sense in that within the occupations, those of positive tokens are high level skills and those of negative tokens are relatively low level skills. This thus confirms our prior that firms describe the detailed skills and tasks that they demand in their job posts and these terms are important for the posted wage determination. This result is also consistent with the perspective of multi-dimensional skills and tasks, as we have argued earlier, in which occupations are different compositions of various skills and tasks and there could also have important within-occupation skill and task variations.

Finally, in additional to the occupation-specific skill or task terms, we also observe two groups of terms in the top tokens that consistently appears across different major occupations. The first group is a set of terms related to management and within-firm hierarchy. In the top positive tokens we observe terms like "lead", "management", and "team" across all samples. Whereas in the top negative tokens we observe terms like "assistance" and "supervisor" that represent the position of the job within the firm.[44] The second group is a set of non-cognitive

ance, which is only provided by a few well-paid firms. In some rare cases we can also observe "seven insurance" or "two fund" which basically indicates further advanced insurance or fund support.

[44]Here we check the raw data and find the term "supervisor" does not necessarily mean a supervisor job in many cases but mainly occurs in sentences like "follow the order of supervisor". This case is a good example showing the tricky part of textual analysis: all tokens should be interpreted by its meaning in the context rather

human capital terms like "hardworking", "careful", "patience", or "focused" in top negative tokens. These are general human capital that should be valued in any jobs, and one possible explanation for their negative relationship with posted wage can be that for some reasons they are more likely to be required in low-skill jobs or by firms with low wage premiums but less likely to be mentioned in high-skilled jobs or by firms with high wage premiums.

To sum up, our Lasso models select the most predictive features for the posted wage from the vacancy text data. In general, these features are constituted by various skills, tasks, and non-wage compensations or amenities. Although each coefficient is not interpretable due to multicollinearity and flexible structure embedded in our high-dimensional and penalized model, we do observe intuitive and interesting patterns within these selected features. Our next step is to classify these features into different types so that we can not only understand the underlying structure of these job characteristics but also together them into different bundles to study different questions about wage inequality.

5.2 Features Clustering

The penalized linear model in the last subsection reduces features of interest to less than 3 percent of the entire token vocabulary. However, the number of remaining tokens is still large, and it is thus hard to get a general picture about what are these features and what relationships do they hold. Although one can simply look at those selected nonzero tokens and decide for each what type it is based on some prior knowledge, here we will show how we can achieve this in a less arbitrary way by using natural language processing (NLP) model to learn the associations between terms in our vacancy text and then using unsupervised machine learning algorithm to search for potential clusters and patterns within our selected tokens.

In the last subsection we have represented our job text documents through the presence indicator matrix $\mathbf{C}$. Though simple and useful in many cases, this matrix does not tell us anything about the relationships between the tokens.[45] This motivates the development of the word embedding models in NLP, which go beyond simple counts of individual words or phrases and learn from the rich syntactical structures embedded within the human-written text to understand the "meanings" of the words. In particular what these models do is to map each word to a latent vector space in $\mathbb{R}^H$ where the dimensions of this latent vector space H correspond to some hidden aspects of meaning of which different words or phrases will hold as the endowment to fulfill their content, and where the relationships between words can be represented through some internally consistent arithmetic calculations. Among many methods to generate this mapping, we will use the most basic neural network method, the Word2Vec

than taking its superficial meanings. Another example is the positive feature "subcontracting" in Design & Media occupation. We find that it often occurs in sentence like "assigning, supervising and checking the subcontracted works" and thus means the tasks and skills of managing subcontracted workers rather than doing subcontract works.

[45] For example, consider two terms have proximate meanings. They can simultaneously occur in the same vacancy if this meaning is rephrased several times in the vacancy text. But they might also never simultaneously occur in the same vacancy if only either word would be used even though they mean very similar things. Therefore, simply through the vectors of presence or counts we cannot have enough information to tell any two words are proximate or distant.

model, for our vacancy text.[46] The key idea of the Word2Vec model is that words in similar contexts, represented by the words with close sets of adjacent words, share the similar semantic meanings in the vocabulary, and vice versa. Consequently, we can obtain such relationships by training a neural network with a single hidden layer to perform either a task of given an input word, predicting the probability distribution of the nearby words, or the mirror task of given inputs of context words, predicting the center word. The projection weights that turn the input word or context words to the hidden layer are then interpreted as the word embeddings.[47] In practice, we use the version of the Word2Vec model which predict the center word given the surrounding context words, which is also called continuous bag-of-words (CBOW) because the order of context words does not influence prediction (bag-of-words assumption).[48] The details of the CBOW word embedding algorithm is described in Appendix B.2.

The result of our word embedding model is a $K \times H$ embedding weight matrix $\mathbf{U}$, where each row of the matrix, $\mathbf{u}_k$, is the representation vector of the word or phrase k in the latent embedding space. Note that although we will only use the embedding vectors of those nonzero tokens that are selected by our Lasso estimation in Section 5.1, i.e. $\mathbf{U}' \equiv \{\mathbf{u}_k\}$ where $k \in V'$ and $V' \subset V$ is the set of selected features, each of these embedding vectors is jointed estimated with and thus defined by all the words in the entire vocabulary V. With these embedding vectors in hand, we now can apply unsupervised clustering algorithm to classify our nonzero tokens into different clusters based on their meanings in the text. Here we also use a simple and popular method, K-Means, which find the centroids for the clusters in the target space (here the embedding space) to minimize the sum of within-cluster Euclidean distances. To conduct K-Means clustering, we first need to decide a primary parameter, the number of the clusters, denoted as P. Then we look for the P-partition of the selected vocabulary V', $\{V'_1, V'_2, \ldots, , V'_P\}$,

[46] Our choice is based on a suitability and performance combined consideration. Although models with deeper neural networks like Bidirectional Encoder Representations from Transformers (BERT) are more powerful, the training of such models is significantly computation-demanding and time-consuming, making many researchers directly use already trained models based on internet text contents like Wikipedia or web news. One major strength of such more sophisticated models is that they can learn the different meanings of one token in different contexts (a trouble feature of the human nature language), while the Word2Vec model can assign only one context meaning for one token. However, given that vacancy text data is a very specific environment for language usage, such compound issue would less likely to happen in our case. More importantly, many words about job characteristics might have specific meanings deviated from the one for normal usage, and many specific terms could not exist in the vocabulary of any pre-trained models at all. Therefore, it is important to directly train any work embedding models on our specific job vacancy text data to get the best results.

[47] Note that the task here is often called synthetic or auxiliary task because we are not actually going to use that neural network for the task we trained it on—the problem of predicting surround words or center word. Rather our aim is just to learn and obtain the weights of the hidden layer. Therefore, although the Word2Vec model itself is an unsupervised machine learning task—unsupervised extraction of semantics for words from the corpus, the way it is phrased is using an auxiliary supervised machine learning task to learn the embeddings as useful representations of the words.

[48] The another version of the Word2Vec model that predict the adjacent words given a single word is called skip-gram. This architecture weighs nearby context words more heavily than more distant context words and performs well in the cases of infrequent words. We choose the CBOW architecture mainly because its generic model is more simple and nature, and its algorithm is faster.

to minimize the distance from each token to the centroid of the cluster it belongs to:

$$\underset{\{V_1',V_2',\ldots,V_P'\}}{\arg\min} \sum_{p=1}^{P} \sum_{k \in V_p'} \left\| \mathbf{u}_k - \frac{1}{\left|V_p'\right|} \sum_{j \in V_p'} \mathbf{u}_j \right\|^2$$

. The pre-determined parameter P is the only hyper parameter of the algorithm and is arbitrary unless we know the number of the "true" clusters of the data, which often does not even exist.[49] In practice, we select $P = 8$, i.e. eight clusters for each samples of analysis, in order to avoid some obvious entanglements, but our main findings hold for selecting other close numbers. To visualize the clustering result, we use t-distributed stochastic neighbor embedding (t-SNE) algorithm to first reduce the embedding matrix $\mathbf{U}$ to a two-dimensional representation and then plot all tokens in V' on this reduced two dimensions with their assigned clusters labeled in different colors. We show this for the Pooled sample and for Computer occupation in Figure 2, and same plots for other occupations can be seen in Figure D3.

We then document several consistent patterns that we find in the results of the K-Means clustering across different samples. Firstly, it is not surprising that in each sample we can find a cluster that contains all the compensation words and phrases given that they have a rather special context in the job vacancy text. We label this cluster as V_1'. In this cluster we also find many synonyms of the compensations in the top tokens that we have mentioned in the Lasso results, for example "business insurance", "five-day workday", and many similar compensations, for example "seven insurance & one fund", "two fund", "bonus", "tea time", "gym", "taxi", as well as some other typical compensations in the literature for example "flexible worktime", "overnight shift". The general picture of the compensation provided by firms are thus close to what we conclude from the top Lasso features: backloading payment and bonus, insurance & fund, worktime and leisure, fringe benefits, and learning or training environment.[50]

Secondly, we can observe a similar cluster cross all major occupations that contains a combination of words about cognitive skills, noncognitive skills, and interpersonal skills. These words include "hardworking", "patient", "responsible", "challenging", "logic", "critical thinking", "self-learning", "problem-solving", "open mind", "communication", etc. Some of these words have been used in the prior studies (e.g. Deming and Kahn (2018)) to measure the level of cognitive and social skills and found important in determining job wage. It might be a little surprising that although the terms in this cluster have slightly different stress between different

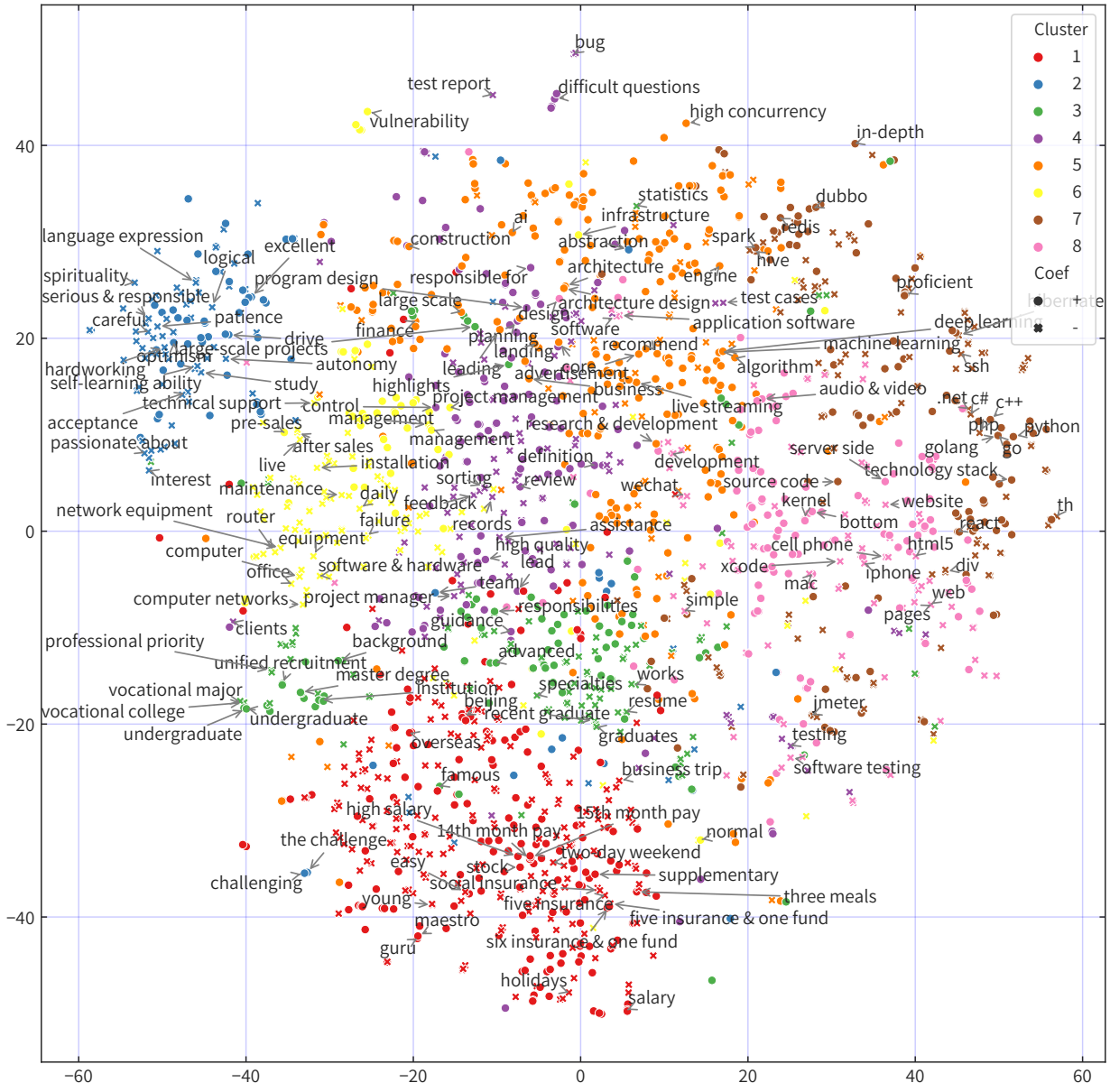
[49] This procedure is analogous to decide the hierarchy of the occupation categories by human knowledge. Both board categories and granular categories make some sense for understanding the structure of the occupational space and there is no one particular ideal number of the categories of occupations.

[50] Given the nature of the job vacancy text, we would generally not have terms of disamenities in our vocabulary because firms will not voluntarily claim the cons of their job in the vacancy. As a result, part of the compensation that have been examined in the previous literature and are for sure disamenities like work injury and safety would not be taken into account here. However, most jobs in our sample, and also a majority of the jobs in the recent labor markets of middle-income or rich countries are office jobs, and thus typically not subject to those kinds of absolute disamenities in the traditional mining or construction industry. Also, many other amenities could have priorly undetermined level of benevolence due to the varying preference of workers or firms, and thus could be found in our data. Considering the general trend of technology advance and work environment improvement, we think our data illustrates a major part of non-wage compensations and amenities that are provided by firms in recent days and are recognized by both firms and workers as important factors of hiring in the labor market.

Figure 1: Feature Clustering on the Embedding Space

(a) Pooled Sample

(b) Computation Occupation



Notes. The clustering of tokens is created by applying K-means algorithm to the latent dimensions of token context obtained from the word embedding model. Here, we visualize all the nonzero features on the 2d dimension by using TSNE algorithm to reduce the latent dimensions of word vectors from one hundred to two. This further dimensional reduction is only for illustrating the results of the clustering and is not used in any results of the main text. The features with text labels are the top 20 words in each cluster with the highest absolute value of coefficients and with frequency larger than 1%. The same plots for other occupations are shown in Figure D3.

occupations, in general the composition of the words are similar across occupations, indicating that these skills are fairly general and firms of all occupations require a similar set of these general skills whether cognitive, noncognitive or interpersonal. We index this cluster as V_2'.

Thirdly, we can find a cluster that contains the education related tokens in all occupations. It incorporates tokens about the general education levels, like high school, vocational college, college, new graduates, etc., and tokens on more specific education requirements like college majors and professional certificates. It also includes requirements on experience in certain fields and the most fundamental skills or tasks in the board occupations, probably because firms often write these terms in together with the education requirements. Therefore, this cluster can be seen as an extensive education control, which indicates relatively more specific skills and tasks than the ones in V_2', and we index it as V_3'.

The forth cluster that we are able to identify from our clustering result is less consistent and more ambiguous comparing to above three clusters. To be specific, in each major occupations we can find a cluster that incorporates words or phrases related to within-firm hierarchy and coordination like management, planning, allocation, collecting, subordination, and assistance. However, for each major occupation this cluster also incorporates occupational-specific tasks that are linked with these hierarchical or organizational terms, and for the Pooled sample it includes a variety of administrative tasks. This cluster is gathered together by the algorithm likely due to the fact that whatever the occupation is, there are always similarly stated tasks about (manager) assigning tasks, (subordinate) following manager's order to accomplish tasks, and coordinating different tasks for departments within the firm or between firm and outsider clients or suppliers, although these tasks can be specific to different occupations. We consider this cluster as an extensive or complementary control for (potentially occupational) experience, which largely indicates the job position in the firm hierarchy or the job ladder, and index it as V_4'.

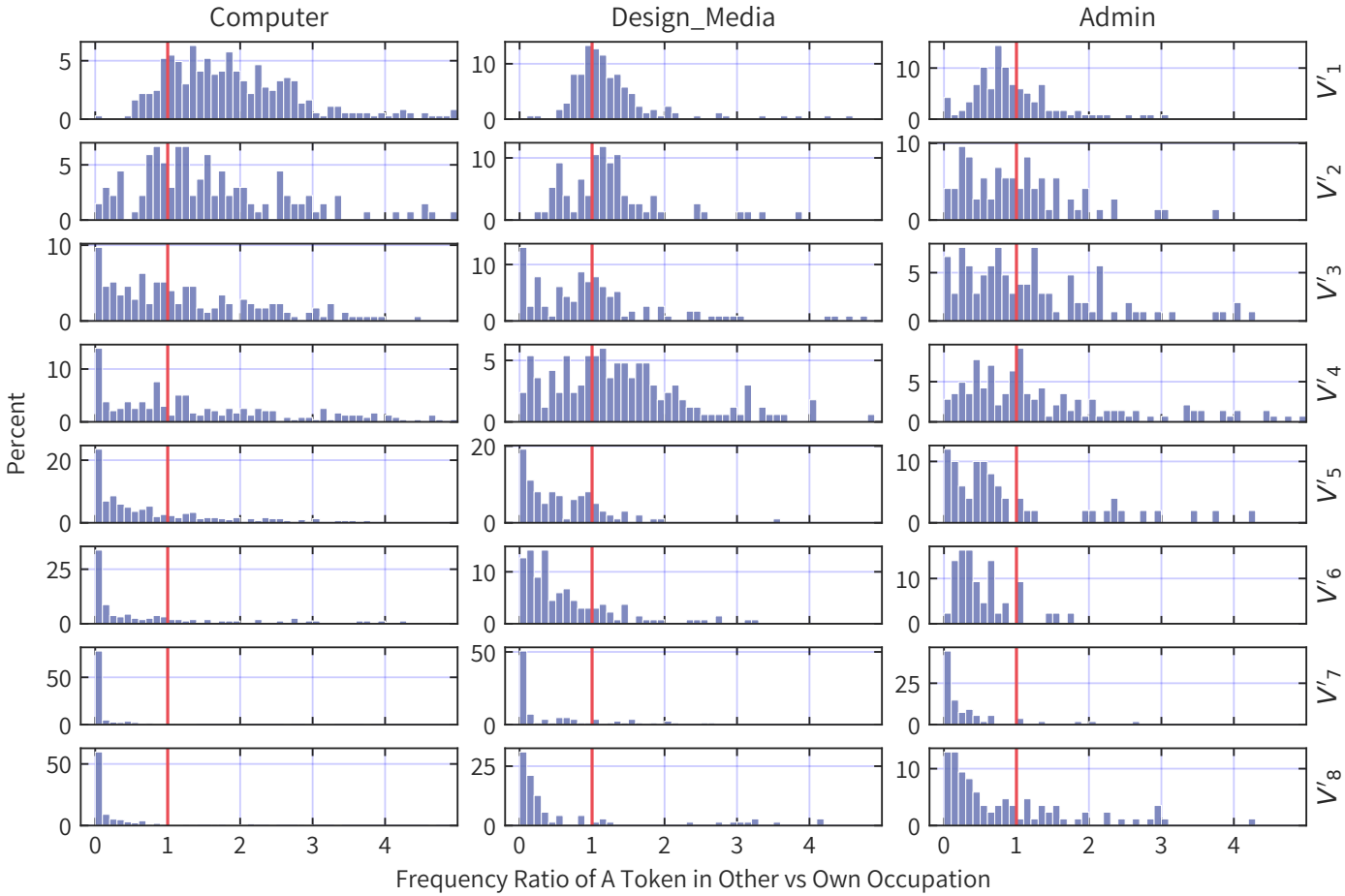
For the rest of the clusters, it becomes difficult to find the similar counterparts across different samples and occupations. In particular, for the Pooled sample, the rest four clusters (and also the V_4' to some extent) seem to be the clusters of skills and tasks stemmed from different major occupations. And for single occupation, the rest four clusters seem to be further partition of the skills and tasks in that major occupation into distinguished groups.[51] In other words it seems that our clustering algorithm conducted on the word embedding space of vacancy text mimics what the official occupation categories do: classifying jobs into different hierarchies based on skills and tasks. As a result, we recognize these rest clusters as occupation-specific skills and tasks and label them $V_5', \ldots, V_8'$ arbitrarily.

Our definition above on different clusters derived from the algorithm is based on our human learning on the terms in those clusters and one may doubt that to what extent do they make sense. To confirm, we measure the specificity of a token $k \in V'^o$ selected in occupation o by compare its occurrence rate in V'^o with the weighted mean of its occurrence rate in $V'^{o'}, \forall o' \neq o$. We plot the distributions of this other-vs-own occupation frequency ratio for all tokens in each cluster separately and for all three major occupations in Figure 2.[52] It shows

[51] This is easiest to see in Computer occupation, where these clusters contain many terms about programming languages and other IT-specific technical words. Whereas in other occupations, the skills and tasks might not be completely specific. For example, analysis and planning could be important for many occupations although for different occupations the content for analysis or planning might be very different.

[52] It's not possible to plot the same figure for the Pooled sample but given that the structure of the clustering

Figure 2: The Distribution of The Ratio of Feature Frequency in Other Occupations to in Own Occupation



Notes. We calculate this ratio by dividing each token's occurrence percentage in the vacancies out of the major occupation of this token by its occurrence percentage in this own major occupation. The cluster index is the same as the one in . In particular, the cluster 1 is the compensation cluster, the cluster 2 is the general human capital cluster, the cluster 3 is the education related cluster and the cluster 4 is the management and subordination cluster. Both cluster 3 and cluster 4 also contains some occupation specific skills and tasks. Cluster 5 to 8 are undefined occupation specific skills and tasks.

that for the compensation cluster (V_1') and the general skills cluster (V_2') the token's relative frequency ratios are concentrated in value 1 with a shape close to normal distribution, indicating that these tokens are close to equally mentioned in different major occupations. For the education-related cluster (V_3') and the experience or position-related cluster (V_4'), the distributions become more dispersed and have more concentration close to 0 in high-skill Computer occupations, suggesting that they likely contain both general and occupation specific skills and tasks. For the rest of the clusters $V_5', \ldots, V_8'$, their tokens mainly concentrated close to 0, indicating that a majority of these words are likely to be very occupational-specific as they are way more likely or sometimes only to be mentioned in their own occupations. This left-skewed distribution is again more significantly in Computer occupations than in Administrative occupation, which makes intuitive sense because while specific skills in Computer are more likely to be some specific programming languages and thus very unlikely to be mentioned in other occupations, the specifics skills in Admin occupation involve more general terms like analysis, arrangement, or report which would likely to be used in many other occupations.

To sum up, in this section we classify the features selected in Lasso models to different types without any prior (except for the number of categories) and completely based on their associations in the job vacancy context, i.e. how firms write their vacancy text. We find that the results indicate a data-driven skill and task structure that is featured by skills and tasks with different levels of specificity. This skill and task structure or space distinguish with the official occupation categories in that it add very general skills that are irrelevant to occupation, and that it fulfills the within-occupation variations with detailed skills and tasks. It also distinguishes with the skill structures used in some recent labor literature that summarize the entire skill and task space using several board abstract categories like cognitive, noncognitive and interpersonal skills by showing that such classification will lose the dimension of skill and task specificity, which could potentially be important for thinking about issues like how the workers obtain different skills and to what extent do different skills transferable across different jobs. Moreover, in this clustering process we separate a cluster of compensation along with other skill and task clusters, which allows us to study their different impacts in the posted wage determination and discrepancy. In next subsection, we further reduce the dimension of the indicator matrices of these clusters so that we could bring these clusters of different job characteristics back to our wage differential estimation.

5.3 Dimension Reduction

In order to bring the selected hundreds or thousands of features in V' back to the wage regression in Section 4, we now further reduce the dimension of the indicator matrix of the selected tokens, $\mathbf{C}' \equiv \{\mathbf{c}_k\}, k \in V'$, to a reasonable size to ease the estimation. Relying on the clustering results that we have derived in Section 5.2, we will do this dimensional reducing separately for each cluster in each sample, i.e. reducing the dimension of $\mathbf{C}'_p \equiv \{\mathbf{c}_k\}, k \in V_p'$ for all p, so that we can distinguish the effects from the different types of job characteristics. For the task of dimension reduction, unsupervised methods like principal component analysis (PCA) are often used. In fact, PCA projects the target data onto a lower dimensional space so that

in the Pooled sample is similar to those in the occupational samples, our evidence here is also suggestive for the Pooled sample.

the variance of the projected data is maximized along each axis In other words, PCA finds a low-rank representation of $\mathbf{C}'_p$ that best preserves its covariance structure, but use no information about the structure of its predictive power and thus could generate unsatisfied results for our purpose here.[53] Instead, here we follow another suggestion in Gentzkow et al. (2019) to use a supervised method, partial least squares regression (PLS), to achieve a better performed dimension reduction.

In contrast to PCA, PLS performs dimension reduction by taking account of the information in the relation between the predictive and target variables. In particular, PLS projects both predictive and target variables into a lower-dimensional subspace such that the covariance between these two projections is maximized. Here because our target variable log wage has dimension one, this boils down to simply project each $\mathbf{C}'_p$ to 1D dimension and maximizing the covariance between this projection and the log wage. This procedure is iterated with orthogonalization to reach the desired number of PLS components Q. The details of the computation procedure are described in Appendix B.3. In essence, PLS forms the components by taking all features into a small set of linear combinations where the weights are decided by the predictive power of the features. We denote the resulted matrix of each cluster as $\Xi_1, \Xi_2, \ldots, \Xi_8$, whose indices correspond to the vocabulary clusters $V'_1, V'_2, \ldots, , V'_8$. In practice, we choose Q to be three, which means each Ξ will contain three vectors that represents three most useful dimension of the cluster in wage prediction.[54] Therefore, for each sample, we can now replace the indicator matrix of hundreds or thousands features with only twenty-four synthetic continuous variables. Running an OLS regression of all twenty-four variables on the posted wage, we find that, for all major occupations, the obtained R-squared is over 95 percent of the R-squared obtained in our Lasso regressions in Section 5.1 which use the full set of tokens, indicating that our dimension reduction successfully preserves the majority of the predictive power of the tokens selected by the Lasso estimator.[55]

6 Posted Wage Inequality

In last section we exploit machine learning methods to distill all wage-predictive job characteristics from vacancy text, to classify them into different clusters of skills and tasks, and to generate the low dimensional proxy variables that preserve most of the information of these features. In this section we bring these job skill and task variables back to the econometric

[53] In particular, a predictive regression using principal components may perform poorly in data where the prediction target is strongly correlated with directions that have low variance because these directions, despite that their high predictive power, will be dropped in PCA. This problem could happen in our case due to the fact that the ill-understood features of the indicator matrix $\mathbf{C}$ that we have talked about in our motivation of using the word embedding model in the last subsection can carry over to each $\mathbf{C}'_p$.

[54] This choice of Q is again somehow arbitrary. We choose $Q = 3$ because three reduced variables under PLS are already able to account for most of the prediction power of the original token matrix of each cluster. Increasing Q further has little marginal improvement in the R-squared of the linear regression that use these reduced variables. Changing Q to two or four would not affect any of our results qualitatively.

[55] In comparison, the result from PCA or LSA (Latent Semantic Analysis, a direct singular value decomposition on the data without normalization and thus often used for sparse data in textual analysis) with three principal components ($Q = 3$) is significantly worse, achieving only around 50 percent of the R-squared in the Lasso regression.

model in Section 4 such that we can accomplish our wage differential estimation and examine how do these newly-obtained and often-observed information improve our understanding of the wage determination and inequality in the labor market. Although during our machine learning procedures we also discover a cluster of non-wage compensations and amenities, which indicates that non-wage compensation provision might also be an important potential driver of wage determination, here we focus on job skills and tasks and leave the investigation of that specific aspect to Section 7.

In order to better show our main results, in Section 6.1 we first bundle our skill and task clusters into different groups and specify the final composition of X. We then show our main results of posted wage differential in Section 6.2, where we not only illustrate the major components of posted wage variance but also further decompose the job effect to examine how different types of skills and tasks contribute to the job effect and firm-job sorting. In Section 6.3, we use the firm effects estimated in Section 6.2 to test some more features of the firm wage premium. Finally, in Section 6.4, we conduct several robustness tests on our estimation results.

6.1 Specification on Job Characteristics

The set of skill and task proxy variables obtained in Section 5, denoted $\tilde{\Xi} \equiv \{\Xi_2, \ldots, \Xi_8\}$, can be recognized as a representation of the full set of skills required and tasks conducted on the job vacancies by which firms use to attract their ideal workers and justify their posted wage. These variables thus incorporate not only the between-occupation skill and tasks variations as captured by the occupation dummies but also previously unobserved within-occupation skill and task variations that we failed to control for in Section 4. In addition, these variables (in particular the cluster $V_3^{'}$) likely contains and extends the information in the education dummies because our machine learning method directly capture the terms related to both formal education level and other extensive requirements. Similarly, but to a less extent, we expect that the information in the experience dummies will be largely accounted by the specific skills and tasks that we have captured in the job description and requirements.

To confirm, we do a preliminary check by comparing the R-squared values of posted wage regressions with different specifications in Figure D6. The results show that, consistent with our expectation, after controlling for our skill and task proxy variables $\tilde{\Xi}$, adding the education dummies has very limited improvement (about 1 to 2 percent points) to the R-squared value, and adding our constructed occupation dummies now almost does not further increase the R-squared value.[56] However, somehow contrary to our prior belief, we find that the experience dummies still hold a significant additional explanatory power to our proxy variables (around 10 percent points). This result suggests that there are potentially some skill or task information embedded in the experience information but are not captured by our distilling procedure. One possibility could be that while our machine learning algorithm find specific skills and tasks, the information about the required level of the proficiency and competency on these skills and tasks is likely to be missing. As a result, we still add the education and experience dummies,

[56] One reason that the education dummies still hold some explanatory power is that in some cases the requirement of education is not documented in the vacancy text and thus cannot be captured by our machine learning procedure.

denoted as $X_e \equiv \{\text{EDU}, \text{EXP}\}$, into our final specification of the job skills and tasks X as they (mainly the experience information) could complement to our text-generated skill and task variables.

Finally, to help present and interpret our estimation results, we split $\tilde{\Xi}$ into three groups based on the level of specificity. In particular, we set the group of the most general skills as $\Xi_g \equiv \{\Xi_2\}$, the group of the medium specific skills and tasks as $\Xi_m \equiv \{\Xi_3, \Xi_4\}$, and the group of the most specific skills and tasks as $\Xi_s \equiv \{\Xi_5, \ldots, \Xi_8\}$. One potential problem of this partition is the cluster Ξ_4 which contains organizational and hierarchical skills and tasks within the firm. Because these skills and tasks are closely linked with occupational specific skills and tasks, it's somehow difficult to decide their level of the specificity despite the fact that they have a medium level of specificity in term frequency. For this concern, in Section 6.4 we also show the results for the case when Ξ_4 is included in the Ξ_s, which also helps us to the see relative importance of education-related information versus experience- or position-related information in explaining posted wage variance. A final note on the Ξs is that the clusters and groups within $\tilde{\Xi}$ are not orthogonal variables but can be correlated with each other. In fact, the sum of the R-squared value for individual cluster shown in Figure D6 is way larger than the Lasso results, indicating strong complementarity between different clusters of skills and tasks. We will also check such complementarities or sorting between different groups of skills and tasks in our results.

6.2 Main Results

The main results of our posted wage variance decomposition with full controls on job characteristics are shown in Table 6. Panel A shows that in our Pooled sample, the total share of wage variance is accounted by 45 percent job effect, 14 percent firm effect, and 14 percent sorting. Comparing to our preliminary results in Table 2, the large increase in the share attributable to job effect and the significant decline in the share attributable to firm effect indicate that firms pay differently partly because they require different skills and tasks that cannot be captured by education, experience, and even the most granular occupation categories, but are captured by our machine learning methods.[57] The levels of firm effect and firm-job sorting is consistent with the results in the recent literature that use employer-employee panel data in rich countries and bias-corrected AKM approach (see Bonhomme et al., 2020), suggesting that at least in this high-end labor submarket in China, the wage inequality composition is similar to the labor markets in other developed countries. In contrast to this consistency, our estimation at major occupational level shows that the determinants of posted wage differential vary significantly across board occupations. In particular, in spite of the fact that we obtain more features

[57] While the additional increase in the job effect for the Pooled sample is limited to about one third of the increase due to adding occupation dummies in Panel B of Table 2, one should be careful to interpret the results as that when targeting a board labor market, the between-occupation skill and task variations are three-fold more important than within-occupation skill and task variations. This is first because our occupation dummies are constructed directly based on skill and task contents and might perform well in distinguishing skill and task variations comparing to the occupation information in other dataset. More importantly adding highly granular occupation dummies will mechanically increase the explanatory power of job effect but does not necessarily represent meaningful distinctions about skill and task variations. In fact if we check the results of each major occupation sample the increase in job effect due to adding granular occupation dummies are less than the increase due to adding detail skill and task controls, $\tilde{\Xi}$.

Table 6: Wage Variance Decomposition

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.330	.067	.407
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.352
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.028	.168
$2\,\text{Cov}(\theta_j, \psi_j)$	.051	.142	.054	.194	.043	.171	.011	.070
Panel B: Decompose θ **Terms**								
$\text{Var}(X_e)$	.047	.130	.031	.110	.032	.126	.020	.124
$\text{Var}(\tilde{\Xi})$	.063	.173	.029	.103	.028	.109	.023	.138
$2\,\text{Cov}(X_e, \tilde{\Xi})$	.053	.147	.022	.079	.024	.096	.024	.146
$2\,\text{Cov}(X_e, \psi_j)$	.022	.060	.022	.079	.021	.082	.006	.035
$2\,\text{Cov}(\tilde{\Xi}, \psi_j)$	.030	.082	.032	.114	.022	.088	.006	.035
Panel C: Further Decompose $\tilde{\Xi}$ **Terms**								
$\text{Var}(\Xi_g)$	.001	.002	.000	.001	.000	.002	.000	.001
$\text{Var}(\Xi_m)$	.013	.035	.004	.015	.005	.019	.012	.073
$\text{Var}(\Xi_s)$	.022	.062	.014	.051	.012	.046	.003	.017
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.003	.008	.001	.002	.001	.003	.002	.010
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.005	.013	.001	.005	.001	.004	.001	.003
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.019	.053	.008	.028	.009	.035	.006	.034
$2\,\text{Cov}(\Xi_g, X_e)$	.004	.012	.001	.005	.001	.005	.001	.008
$2\,\text{Cov}(\Xi_m, X_e)$	.019	.053	.006	.023	.011	.042	.018	.110
$2\,\text{Cov}(\Xi_s, X_e)$	.030	.082	.014	.051	.012	.049	.004	.027
$2\,\text{Cov}(\Xi_g, \psi_j)$	.002	.007	.002	.007	.001	.004	.000	.002
$2\,\text{Cov}(\Xi_m, \psi_j)$	.010	.028	.009	.033	.010	.038	.005	.029
$2\,\text{Cov}(\Xi_s, \psi_j)$	.017	.048	.021	.075	.012	.046	.001	.004
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Notes. The covariates X for the job effect θ_i now include education and experience dummies, X_e, and the proxy variables for different skill and task clusters, $\tilde{\Xi} \equiv \{\Xi_2, \ldots, \Xi_8\}$, that are derived from the machine learning algorithms in Section 5. In Panel B we decompose the variance and covariance terms that involve θ_i for X_e and $\tilde{\Xi}$ separately. And in Panel C we further decompose the terms that involve $\tilde{\Xi}$ by splitting it into three groups, $\{\Xi_g, \Xi_m, \Xi_s\}$,based on the level of skills and tasks specificity. All results here have been corrected for finite sample bias by using KSS (leave-out) correction method.

in high-skilled occupations like Computer occupation than low-skilled occupations like Admin occupation, the results in Panel A show that the job effect is smallest in Computer occupation and highest in Admin occupation. On the other hand, the levels of firm effect and firm-job sorting are higher in high-skilled occupations than low-skilled occupations. Given that the absolute value of wage variance is also larger in high-skilled occupations, this result suggests that firm wage premium and firm-job sorting are potentially correlated with those skills and tasks that we discover from the vacancy text and thus could be important drivers behind in the labor market inequality due to forces like skill-biased technological change. In other words, the high level of wage inequality in high-skilled occupations are due to not solely that these occupations have a larger amount of skills and tasks, but more importantly that such skills and tasks are used by different firms in a systematically different way along with significantly different levels of firm wage premium.

The merit of our machine learning and vacancy data approach is that, different from the worker fixed effect, we can unmask our job effect and investigate the effects of different types of skills and tasks on both job effect and firm-job sorting. In panel B we first distinguish the education and experience controls X_e and our constructed skill and task controls $\tilde{\Xi}$. The results show that X_e and $\tilde{\Xi}$ holds roughly similar explanatory power on wage inequality in both the channel of job effect and the channel of sorting with firm effect, although in the pooled sample X_e can explain some more wage variance than $\tilde{\Xi}$ in both channels. The results also illustrate that X_e and $\tilde{\Xi}$ are highly correlated by themselves. As we have explained earlier in Section 6.1, X_e contains mainly the information about experience level on different types of skills and tasks and can be regarded as an intensive margin of skill and task usage, and thus potentially complements to the information of mentioning different skills and tasks in $\tilde{\Xi}$, which can be regarded as an extensive margin of skill and task use. As a result, the main implication of our estimation results in Panel B is that both margins are important for job effect and firm-job sorting, and that there is also sorting between these two margins, i.e. firms which require more on specific skills and tasks and thus offer higher pay are also more likely to require a higher level of experience on these skills and tasks.

Next, we further decompose all terms related to $\tilde{\Xi}$ in Panel C to examine the roles played by different types of skills and tasks. The decomposition results make it clear that at both pooled level and occupational level, general skills, as represented by Ξ_g, almost does not explain for any job effect and account for a very small fraction of sorting with firm effect or with X_e. In contrast, the most specific group of skills and tasks, Ξ_s, account for a majority of both job effect and sorting with firm effect and with external margin X_e, except in the Admin occupation. In fact, while the share of posted wage variance attributable to the most specific group Ξ_s more than double the share attributable to medium specific group Ξ_m in Computer occupation and are 70 percent higher in the pooled sample, the relative importance of Ξ_s with respect to Ξ_m declines in the middle skilled Design & Media occupation and is completely inverted in the Admin occupation. These results indicate that the most important driver for the large wage variance in those high-skilled occupations are those most specific skills and tasks, potentially linking with firm technologies and productivities, and that for those low-skilled occupations, wage levels differ mainly due to different requirements on education- and experience- or position-related skills and tasks, and variations in detailed skill and task information have little impact on posted wages. We think our findings here provide some new and intuitive evidences on the microfoundation of the popular argument of skilled biased technological change (SBTC)

in the labor literature. For high skilled occupation like Computer occupation, firms' different usage of new skills and tasks like machine learning or AI largely derive the differences in wage. And because such different requirements are highly correlated with firms requirements on education- and experience-related skills, they could potentially help to generate results like college premium. Whereas for the low skilled occupation like Admin occupation, firms are likely to ask for a bundle of skills and tasks which are relatively less specific and less linked with new technologies and thus firms select workers mainly and directly on the education or experience level. Moreover, our results also suggest that below the hood of worker effect, it is those specific skills held by workers contributing to the sorting with firm effect, and the more skilled the job is, the more importance have those most specific skills and tasks hold.

To sum up our main findings, equipped with the full controls for job skills and tasks derived from the job text, the estimation on our entire sample now generates a job effect and a firm effect consistent with the results in the previous studies that use AKM approach and employer-employee panel data in rich countries. However, we also find that in the board occupational level, more high skilled occupations with more job features selected have less share of job effect but more shares from firm effect and sorting. By decomposing the job effect, we find that the extensive job skill margins in our constructed skill and task variables account for equally or more shares of the job effect and firm-job sorting comparing to the intensive job skill margin captured by the experience requirement dummies. Further decomposition on the proxy skills and tasks make it clear that it is not the variations in those most general skills but the variations in those specific skills and tasks that explain the posted wage differentials, and in general the most specific skills and tasks account for a majority of all types of effects except for the low-skilled occupation in our data, where those medium specific education and experience or position skills and tasks play the major roles. Along with our early results on the data-driven skill and task structure, the strong correlations between extensive and intensive skill and task margin and between medium specific and the most specific skills and tasks indicate that there can have complex and correlated dimensions behind the observed skill premium in the labor market.[58] Overall, our results offer a detailed picture on how different types of skills and tasks contribute to the posted wage inequality and provide new insights and hints for understanding the deep nature of firm wage premium and firm-worker sorting.[59]

[58]The strong correlation between different types of skills and tasks may suggest that there are important complementarities between skills and tasks of different level of specificity. And the potential correlation between different skills and tasks in the vacancy text that we find also raises the caution for any studies that try to identify the wage effect of any single skills or tasks from the job vacancy data without considering the entire skill and task space.

[59]Although our main finding is that occupation specific skills and tasks are perhaps the most important part to explain posted wage differential, it needs to be clear that we are not suggesting for a simple backlash to the early framework of occupation specific or industry specific human capital, which lack any within-group skill and task variations. Also, general skills, which are mentioned by firms but unsubstantiated for wage variance under our estimation, might be still important for workers sorting into different occupations in the first place. A more serious treatment might be a mapping from a multidimensional skill space to a multidimensional task space as well as a further mapping from the task space to a job space, which then helps to clarify how substitutable are many types of skills, especially those cognitive ones, across different jobs and how important is human capital investment or on-the-job learning on wage determination.

6.3 Firm Posted Wage Premium

The results in last subsection show that even after controlling for almost all the information in the job vacancy text we can still observe a substantial part of posted wage variation attributable to firm fixed effects, i.e. firms have different pay policies even if they document exactly the same requirements in the job vacancies. Moreover, high-skilled occupations have substantially larger firm effects than low-skilled occupations, whether in absolute value or in relative level. In this subsection we study two additional questions on the firms' premiums in the posted wages: first if firms pay similar wage premium across different major occupations and second what firm characteristics can explain the difference in firm wage premium.

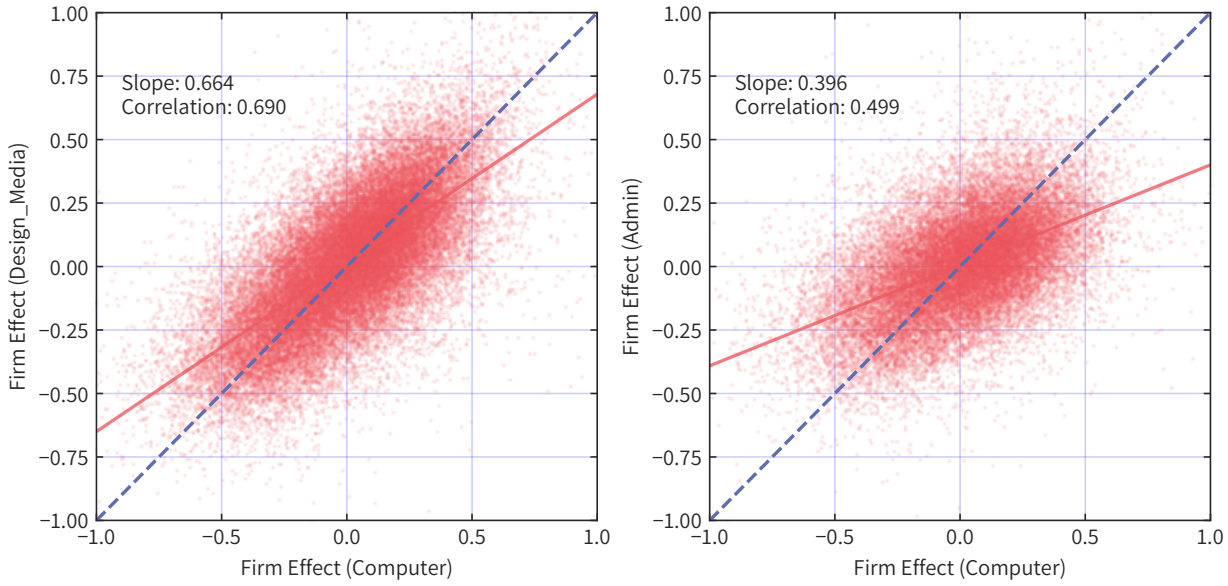
We answer the first question by using firms' fixed effects estimated in the Computer occupation as a benchmark to compare with their fixed effects estimated in the other two major occupations. All fixed effects are demeaned, and we plot the results in Figure 3, where the red lines illustrate the linear regression slopes. While both two pairs show a positive relationship indicating that firms' posted wage premiums are in general consistent across occupations, the slopes are less than one which suggests that firms incline to have a smaller variation of wage premium in Computer occupation comparing the other two occupation. Moreover, the level of the deviation from the firm effects in Computer occupation is more significant in the low-skilled Admin occupation than the medium-skilled Design & Media occupation. In particular, while for the pair between Design & Media occupation and Computer occupation the slop and the correlation is both close to 0.7, for the pair between Admin occupation and Computer occupation the slop is 0.4 and the correlation is 0.5. In other words, the firm wage premium for an Administrative vacancy is more likely to be irrelevant to the same firm's wage premium paid to a Computer vacancy. This fact again speaks to the possibility that firm wage premium varies across different occupations and is potentially linked with the specific skills and tasks used in different occupations. For example, a reasonable explanation could be that there are strong complementarities across high skilled job positions and thus firm would pay a high wage premium for these core jobs with high productivity. On the other hand, those auxiliary jobs are less complementary to those core jobs and thus firm would not pay an equally high wage premium for more auxiliary jobs.[60] However, it is also possible that our results are driven by the heavier measurement errors in the low-skilled occupations in our data. In Section 6.4 we check for the finite sample problems by looking at only firms with more than ten vacancies in both sides of the occupation pair and the results here maintain.[61]

We then move the second question and examine what firm characteristics in our data could explain the significantly different firm pay policies. In particular, we regress the firm fixed effects on firm size and firm location dummies. Because from our wage variance decomposition we already know that firm fixed effect is positively correlated with job quality, we also try to include our estimated firm-average level of job characteristics, $\bar{\theta}_j$, into the regression in case

[60] By arguing this we assume that some types of rent sharing would exist. Such empirical evidences have been documented in the literature, see for example Bloesch et al. (2021). Of course there could have other internal forces like firm norms or culture to equalizing the pay premium within a firm, and firms facing such forces would have incentive to outsource their non-core jobs.

[61] One might still worry the problem of measurement error in the case that for some reasons the firms in our data are less accurately post the wage for their vacancies in low-skilled occupations. However, note that the absolute value of posted variations in low-skilled occupations are substantially less than the absolute value in high-skilled occupations.

Figure 3: Variation of Firm Effects Across Occupations



Notes. Firm effects are estimated using the specification of X in Section 6.1. We then find the set of firms that have both estimated firm effects in each two pairs of major occupations. A linear regression is then estimated on the demeaned firm effects for each pair. The red line shows the slop of the regression and the blue dash line is the 45 degree line.

that the variables of interests are correlated with both firm wage premium and job quality. The results of our regression are shown in Table 7. Firm employee size is significantly and positively correlated with firm fixed effects in all types of specifications, and coefficients for firm size dummies only decrease by a limited part after adding job effect $\bar{\theta}_j$. This positive correlation is again consistent to the results in the literature that use employer-employee data and AKM framework (see e.g. Kline et al., 2020). Despite the statistical significance, the R-squared in the specification with only firm size categories is less than 2 percent, indicating that the part of firm wage premium that can be explained by firm size categories is very limited. In contrast, we find that the dummies of work location can still explain a large part of the firm wage premium even after controlling for average job quality and firm size categories, increasing the R-squared for slightly less than 30 percent points in most samples except for Admin occupation. This suggests that the firm wage premium may be partly due to different bargaining power under outside option differences across different regions, or due to different levels of productivities and rents under different levels of geographical agglomeration, along with other geographical reasons. More data on the firm-side is required to identify the exact sources of firm wage premium.

In summary our results in both this subsection and last subsection suggest the possibility of different levels of firm wage premium across different occupations and the different roles that different skills and tasks in different occupations play as (at least one of) the potential drivers. As a result, assuming a same firm's wage premium for all jobs of a firm could potentially underestimate the importance of firm effect in explaining for wage inequality in the labor

Table 7: Firm Fixed Effect and Firm Characteristics

	Pooled			Computer			Design_Media			Admin		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
fsize.15-50	.019**	.018**	.023**	.011+	.013*	.019**	.022**	.013**	.020**	.006	.005	.005
	(.004)	(.003)	(.003)	(.006)	(.005)	(.004)	(.005)	(.005)	(.004)	(.006)	(.006)	(.006)
fsize.50-150	.042**	.037**	.050**	.037**	.032**	.038**	.050**	.033**	.045**	.020**	.018**	.021**
	(.004)	(.003)	(.003)	(.006)	(.005)	(.004)	(.005)	(.005)	(.004)	(.006)	(.006)	(.005)
fsize.150-500	.067**	.057**	.067**	.072**	.054**	.051**	.086**	.058**	.063**	.035**	.031**	.030**
	(.004)	(.004)	(.003)	(.006)	(.005)	(.005)	(.005)	(.005)	(.004)	(.006)	(.006)	(.006)
fsize.500-2000	.095**	.078**	.085**	.108**	.074**	.066**	.127**	.087**	.086**	.050**	.043**	.040**
	(.005)	(.004)	(.004)	(.007)	(.006)	(.005)	(.006)	(.006)	(.005)	(.007)	(.007)	(.006)
fsize.2000+	.121**	.102**	.120**	.140**	.084**	.082**	.161**	.107**	.108**	.064**	.055**	.058**
	(.005)	(.005)	(.004)	(.008)	(.007)	(.006)	(.007)	(.007)	(.006)	(.008)	(.008)	(.007)
Job Effect ($\tilde{\theta}$)		.287**	.201**		.643**	.498**		.391**	.292**		.118**	.063**
		(.004)	(.003)		(.007)	(.006)		(.006)	(.005)		(.008)	(.008)
const	.146**	-1.115**	-.633**	.222**	-2.684**	-1.905**	-.030**	-1.759**	-1.208**	.024**	-.478**	-.166**
	(.003)	(.016)	(.015)	(.005)	(.030)	(.027)	(.004)	(.028)	(.024)	(.006)	(.036)	(.033)
Location FE			✓			✓			✓			✓
Adj. R^2	.016	.096	.377	.016	.168	.436	.022	.100	.390	.006	.014	.229
No. Obs	86165	86165	86165	62628	62628	62628	55664	55664	55664	41448	41448	41448

Notes. The baseline group for firm size fixed effect is the group of firms with less than 15 employees. In the case that a firm has multiple entries of size or location information we use the categories that are most recorded in the vacancies of the firm.

market. Also similar to other studies that use administrative data, our estimated firm effects are also correlated with firm size and can be partly explained by the location dummies. A more careful examination on the source or nature of the difference in firm pay policies require an integral model to incorporate all these empirical facts and we leave it for future study.

6.4 Robustness

In this section we provide several robustness tests on our results in Section 6.2 and Section 6.3.

Finite Sample Bias. As we have mentioned in Section 4, given the high-dimensional firm fixed effects in our regression model, the variance and covariance terms of firm fixed effect could be biased especially when the observed vacancies of a given firm are limited. In fact this finite sample bias can be also called "limited mobility bias" following the AKM literature due to the fact that the deep nature of limited mobility bias in the AKM approach is the finite sample bias for identifying firm-level wage differences and that limited vacancies observed can be also regarded as one type of limited job mobility for a certain firm. As a result, we can use the methods that are developed in the AKM literature to resolve the finite sample bias to correct the finite sample bias here. In particular, we use both the homoscedasticity correction approach suggested by Andrews et al. (2008) and the heteroscedasticity leave-out correction approach suggested by Kline et al. (2020), and the main results shown in Section 6.2 is under the heteroscedasticity leave-out correction. The comparison of two different types of corrections with the plug-in results are shown in Table D2. The results show that both corrections have very similar results in which the firm effects are reduced and the part accounted by the error terms are correspondingly increased. The corrections are not substantial in the pooled sample, where the changes are about 0.5 percent point, but more significant in the Admin occupation, where the changes are around 5 percent points. This difference is simply due to the fact that in the pooled sample we have less firms with a very small number of vacancies posted, and thus the finite sample bias is rather limited. Our estimated job effects are not subject to any corrections because our controls on job characteristics are either sparse categories or continuous variables. The correction also has very insignificant impact on our estimated firm-job covariance because empirically the finite sample bias in our case will only have second-order effect and thus only appear when the finite sample bias is very large.

In addition to our results in Section 6.2, the finite sample bias also matters for the results in the firm wage premium regression in Section 6.3. Kline et al. (2020) shows that under strong finite sample bias, inference of the regression will be biased and lead to potentially wrong conclusions. However given that the finite sample bias is rather limited in our case, it is likely that this bias would also be small. To confirm, we remove firms with less than ten vacancies in each sample and then redo our tests in Section 6.3, as what we have done for our pooled sample in the sample cleaning. The results are in Figure D1 and Table D5, which show that our results in Section 6.3 largely maintains under these limited samples.

Compositional Differences. One potential concern on our results in Section 6.2 is that the different importance of different types of skills and tasks across different occupations might be driven by the compositional differences across different occupations. In particular, because

in our data high-skilled occupations contain more vacancies with requirements of higher education and experience levels than low-skilled occupations, our results could be misleading if those specific skills and tasks are only important for wage differential in high education and high experience vacancies jobs and if our data does not correctly represent the true composition in the labor market. To resolve this concern, we slice our sample given certain education or experience level and redo our estimations on these sliced samples, and we find our main findings remain. For example, Table D3 shows the estimation results when the experience is conditional to be 0 (no requirement), where we can still observe that for the pooled sample and Computer occupation, those most specific skills and tasks account for the most share of total wage variance, while for the Admin occupation, it's those medium specific skills and tasks account for the major shares.

Specification of Skill and Task Groups. As we have stated in Section 6.1, the inclusion of skill and task cluster Ξ_4 into medium specific group might be problematic because it's difficult to decide how specific are those tasks of management, supervision, coordination, and subordination in different occupations. On the one hand, these organizational or positional tasks could require very occupational specific skills, and on the other hand, there might also have some generality in these tasks even across different occupations. To test how important are the classification of this cluster for our results, in Table D4 we show the variance decomposition results when Ξ_4 is classified to the most specific group Ξ_s. It turns out that Ξ_4 is a key part in Ξ_m for explaining posted wage differential. With only the education-related cluster Ξ_3 in Ξ_m, the posted wage variations accounted by all Ξ_m related terms decline about or less than half of the original level, along with corresponding increase in Ξ_s related terms. This change is especially substantial in the Admin occupation, where now it is the most specific group Ξ_s accounts for the major shares of job effect and firm-job sorting. Given that there is less specific skills and tasks in the Admin occupation, we think it is still safe to conclude that different from high-skilled occupations, medium specific skills and tasks are the most important drivers for wage differential. And the perhaps more important message in this result is that for the posted wage inequality, it seems that experience- and position-related skills and tasks are more important than education-related skills, which again implies that those specific skills and tasks learned on the job are the key for labor market inequalities.

7 Posted Compensation Inequality

In this section we take advantage of the information about non-wage compensation provision that are collected and distilled in Section 5 to investigate the inequality in non-wage compensations and its interaction with posted wage inequality. First in Section 7.1 we show some empirical facts about the provision patterns of the non-wage compensations observed in our data as well as their impact on the posted wage. We then suggest that the classical theory of compensating differential cannot explain our empirical findings. Consequently, in Section 7.2 we construct a new theory which extends the canonical compensating differential mechanism with two new elements, namely efficiency compensations and firm-worker sorting, both of which have been observed in the data and long discussed in the literature. We show that this

new theory can generate flexible provision patterns and wage impact of compensations and reconcile all the stylized facts that we find in our data.

7.1 Empirical Results

In features selected the Lasso regression in Section 5.1 we discover a bunch of non-wage compensations and through our feature clustering procedure in Section 5.2 we gather these compensations into a set V_1'. We find that this set covers a large amount of different types of non-wage compensations, ranging from pecuniary compensations like backloading payment, bonus, and stock options to nonpecuniary compensations like insurance, worktime flexibility, fringe benefits, etc. Although for each vacancy or firm the set of non-wage compensations observed in our data is not necessarily a full list of the compensations that the job or the firm offer, these non-wage compensations mentioned in the job vacancies are arguably the most important ones because they are used by firms to attract potential workers or to justify their posted wage through compensating differential.[62] To study why these non-wage compensations are selected by the Lasso model and how do they affect the posted wage differentials, we first try taking all different types of non-wage compensation as a whole and embedding the proxy variable of the compensation cluster also into our log wage regression. In particular, the specification in Section 4 now becomes

$$\ln w_i = X_i\beta + \psi_j + \delta_i + \iota_t + \epsilon_i \tag{4}$$

, where $\delta_i \equiv \Xi_{1,i} b$ is the product of the dimensional reduced proxy variables for the compensation cluster Ξ_1 and its corresponding coefficients b at vacancy level. Under the logic of compensation differential, and by assuming similar preference on any specific non-wage compensation across different workers, the value of δ_i should represent the part of wage that is differential equalized in each job vacancy due to the compensation provision. More specifically, because in our case firms are generally more likely to document amenities rather than disamenities, the value of δ_i would indicate to what extent the posted wage of a vacancy is discounted due to the non-wage compensations provided by this job. Therefore, a low value of δ_i estimated in our case means that the amenities provided by the employer of this job are highly valued by the potential jobseekers and thus justify a large discount in the posted wage of this job. On the other hand, a high value of estimated δ_i means that the job amenities offered can bring only limited utility for the potential workers and thus cannot act as much compensation for the posted wage, or even that the set of compensations are in net disamenities so that wage should rise to equalize the potential worker's loss in utility.

[62] By construction, the non-wage compensations found in our data are those that firms recognize as attractable for the workers they are looking for. As a result, they are less likely to contain those compensations that the utilities vary hugely and can be either positive or negative across different workers due to personal preference. Example of such compensations include commuting time or some other personal preferences on the workplace which are rather random across workers. Although in some recent search models like Card et al. (2018) these idiosyncratic preference could be important for job moves and wage inequalities, to what extent that different workers vary their preferences on non-wage compensations in general is an empirical question worth future investigation. To simplify the analysis, for the empirical investigation here and for the theoretical model introduced later we will either implicitly or explicitly assume that workers have the similar preference on the non-wage compensations.

Table 8: Wage Variance Decomposition With Compensation

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.254	-	.164	-
Panel A: $\delta_i \equiv \Xi_{1,i}\beta^c$								
$\text{Var}(\theta_i)$	.158	.437	.079	.282	.082	.324	.063	.385
$\text{Var}(\delta_i)$	.002	.004	.001	.003	.001	.002	.001	.006
$\text{Var}(\epsilon_i)$	.097	.269	.074	.262	.070	.277	.057	.349
$\text{Var}(\psi_j)$	.046	.128	.066	.234	.052	.207	.026	.161
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.137	.051	.181	.041	.160	.011	.066
$2\,\text{Cov}(\delta_i, \theta_i)$	.006	.017	.005	.018	.004	.015	.004	.027
$2\,\text{Cov}(\delta_i, \psi_j)$	.003	.008	.006	.021	.004	.014	.001	.006
Panel B: Decompose $2\,\text{Cov}(\delta_i, \theta_i)$								
$2\,\text{Cov}(\delta_i, X_e)$	.002	.006	.002	.007	.002	.007	.002	.011
$2\,\text{Cov}(\delta_i, \tilde{\Xi})$	.004	.011	.003	.011	.002	.009	.003	.016
$2\,\text{Cov}(\delta_i, \Xi_g)$	.000	.001	.000	.001	.000	.001	.000	.001
$2\,\text{Cov}(\delta_i, \Xi_m)$	.002	.004	.001	.003	.001	.004	.002	.012
$2\,\text{Cov}(\delta_i, \Xi_s)$	.002	.006	.002	.007	.001	.005	.001	.003
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

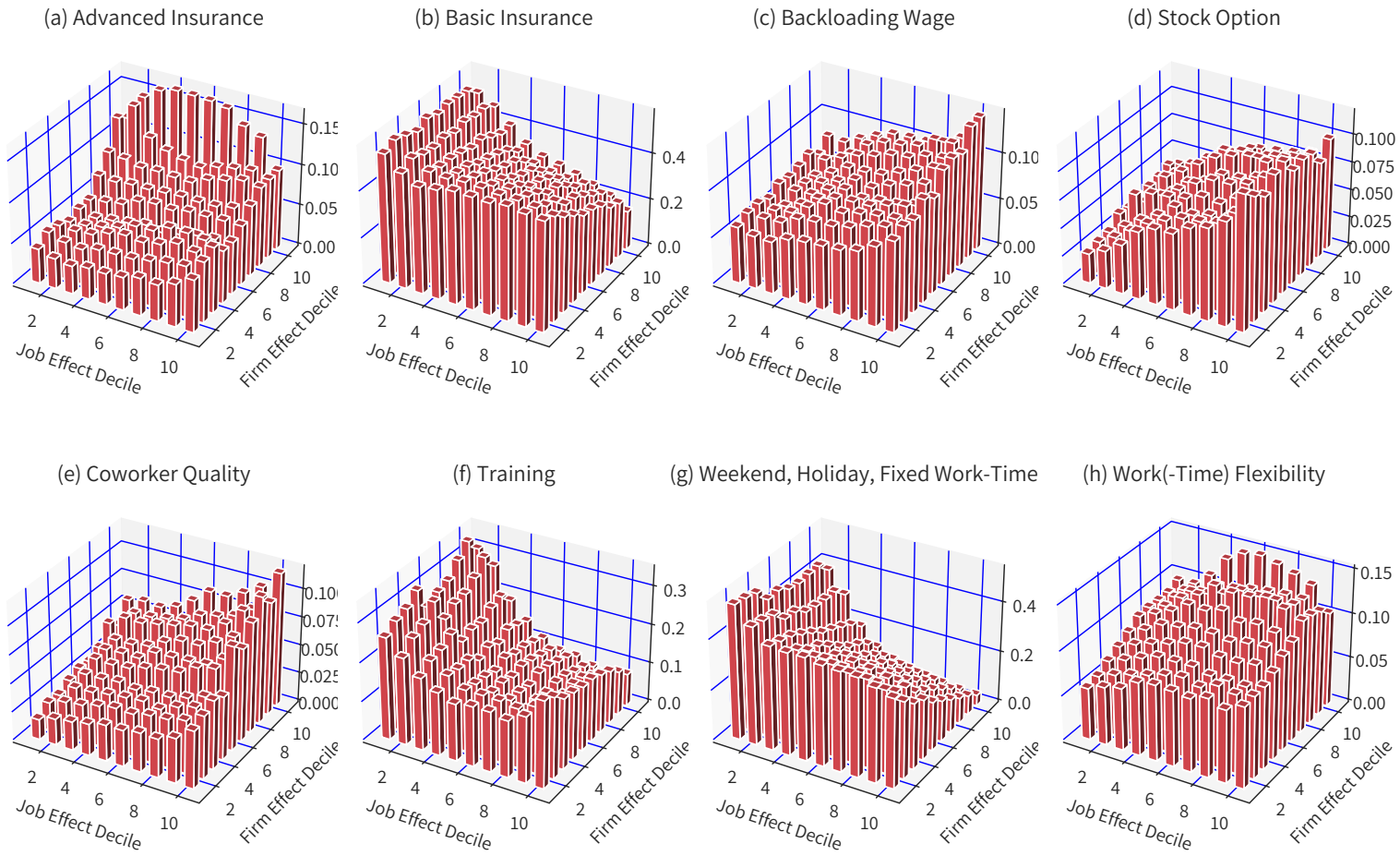
Notes. This variance decomposition is similar to the one in Equation (2) except that we now add the variance and covariance terms of additional component δ.

The estimation results of Equation (4) are shown in Table 8, where panel A documents the components of variance decomposition similar to Equation (2) but with the additional term δ, and panel B further decomposes the covariance between the job effect and the compensation effect, $2\,\mathrm{Cov}(\delta_i, \theta_i)$ into the covariance terms of different types of skills and tasks. The results in panel A make it clear that the variation of non-wage compensation provision itself, i.e. $\mathrm{Var}(\delta_i)$, holds very limited explanatory power for posted wage, accounting for only 0.2 to 0.6 percent of the for total wage variances. However, there are significant and positive relationship between compensation provision with both job effect and firm effect, as shown by the positive covariance terms $2\,\mathrm{Cov}(\delta_i, \theta_i)$ and $2\,\mathrm{Cov}\left(\delta_i, \psi_j\right)$ with shares ranging from 1 percent to 3 percent. These results indicate that our Lasso regression picks those compensation terms largely not because themselves have important impact on posted wage determination, but because that these compensation features can somehow indicate high quality jobs and high wage premium firms. In other words, high premium firms in high skill jobs somehow provide systematically different non-wage compensations. Also, if we again follow the logic of the compensation differential theory, these two positive correlations imply that high wage premium firms and high skill jobs are accompanied by amenities that have low values and thus are less compensated, while low wage premium firms with low skill jobs are more likely to provide amenities that worth more and thus are compensated more from posted wage.[63]

To investigate how different types of firms in different types of job provide systematically different set of non-wage compensations, and to inspect if the arguments following the compensating differential theory make intuitive sense in our data, we next select a bunch of important compensation topics from our Lasso results and examine their occurrence ratios across different types of firms and jobs. In particular, we select eight genres of non-wage compensations that have terms show up with large absolute Lasso coefficients and/or are considered as important topics in the literature. These eight types of non-wage compensations are basic insurance, advanced insurance, backloading wage, stock and options, coworker quality, training, weekend, holiday and fixed work-time, and work-time flexibility. Then we find out all the synonymous in our vocabulary that indicate those genres by checking the group of terms with a small Gaussian distance in the embedding space constructed in Section 5.2. Finally, we partition all the vacancies into the 10×10 job-firm joint decile cells and calculate the occurrence ratio of each compensation type for each cell by checking if its vacancy text contains any of the relevant terms. Because the patterns depicted in the results of all samples are largely similar, here we only show the plot of the occurrence distribution for the pooled sample in Figure 4 and leave the same plots for other samples to Figure D7 in appendix. For all eight types, we see compensation occurrence rate systematically changes along with either or both two axes. In particular, for advanced insurance, backloading wage, stock and option, coworker quality, and work-time flexibility, we observe that the occurrence increases in both the level of job effect and the level of the firm effect, although the extent to which effect matters more varies across compensation

[63] If we allow for large variations in idiosyncratic preference on non-wage compensations, we might also interpret the results as that high skill workers sorted with high wage premium firms value firm-provided amenities less whereas those low skill workers sorted with low premium firms value amenities more. Or if we allow for variations in firms' cost functions of non-wage compensation, we could interpret as that high premium firms are somehow more costly in providing high value amenities whereas low premium firms have a lower cost in providing amenities that are valued by their workers. In general, we can have both cases but if such substantial variations do exist is an empirical question with currently no strong evidences in the literature.

Figure 4: Compensation Occurrence in Pooled Sample



Notes. Job effects and firm effects here are the ones estimated using the specification in Section 6.1. The occurrence ratio is calculated as the percentage of vacancies in each job-firm cell of which the vacancy text contains any of the terms related with a certain type of compensation. Basic insurance means five insurance and one fund, which is the most common compensation package in Chinese labor market. Advanced insurance means any other advanced package of insurance and fund which usually have additional business insurance or fund. Work flexibility relates to the work-time flexibility in most cases. See Figure D7 for the results of major occupation samples.

types. Conversely, for basic insurance and rest day and fixed work-time, their occurrence in job vacancy decrease significantly in both firm effect and job effect, and for training, the occurrence reduce strongly with job effect with ambiguous impact of firm effect.[64] In other words, our results suggest that high-pay firms with high-skill jobs are more likely to provide also better insurance and fund package, non-wage pecuniary compensations like backloading wage and stock option, and also nonpecuniary work place amenities of better coworkers and flexible worktime, whereas low-pay firms with low-skill jobs more often mention training and weekend, holiday, and fixed work-time as the amenities.[65] Our finding here thus largely contradicts our early interpretation of the positive relationship between compensating level and the levels of firm and job effects based on the theory of compensating differential. High wage premium firms also provide better non-wage compensation or amenities in many aspects, although they would less likely to offer training and leisure.

To further examine the idea of compensating differential, we next follow the empirical literature of compensating differential and run a hedonic regression on the occurrence of those eight selected compensations. Similar to the specification in Section 6.1, we use both the full set of proxy variables on heterogeneous skills and tasks and the firm fixed effects such that our hedonic regressions control for almost all the information documented in the job vacancy. The estimated coefficients for compensations in Table 9 show mixed evidences on compensating differential. For the compensations that are positively correlated with job effect and firm effect, i.e. advanced insurance, backloading wage, stock and option, coworker quality, and work-time flexibility, the coefficients are significantly positive in almost all cases. Whereas for the compensations of basic insurance and work-time that are negatively correlated with job and firm effects, their coefficients are significantly negative.[66] Taking at the face value, these results indicate that those amenities provided by high wage premium firms in high skill jobs are not compensated at all but actually increase wage, while those amenities provided by low pay premium firms in low skill jobs are compensated from their posted wage. Therefore, our hedonic regression with detailed job controls produce results consistent with the findings in previous empirical studies: the mechanism of compensating different works in some cases, but in other cases we see exactly the inverse results that generates puzzles for the theory.

To sum up our empirical results, firms that have different levels of wage premium and post

[64]The non-monotone relationship between job effect and training occurrence is because our method cannot distinguish that if the training terms mentioned in job text indicate receiving training or offering training. Actually after checking the raw data we find that the increase in training occurrence in the top deciles of job effect is completely due to these high-skill jobs require tasks of offering training to other workers in the firm. Although we can resolve this problem by applying more advanced NLP model to our text data, we argue that such case is relatively rare in our vacancy text data, and thus we stick with simpler method. The special pattern of training occurrence can be more clearly observed in the Computer occupation sample in Figure D7, where only the top two and bottom two deciles of job effect see a large increase while the middle deciles are generally flat.

[65]One note here is that our result does not necessarily mean that better firms with better jobs are less likely to provide basic insurance and fund package. This is because first obviously that such firms are more likely to offer advanced insurance package and thus correspondingly will not mention the basic package, and second that given that the basic insurance is compulsory for formal firms, high wage firms will generally not think it as an attractive compensation for their potential workers and thus not mention it even when they are actually providing it. We don't think a similar argument will go to the work-life balance because there are a large amounts of anecdotes on long working hours in many big and well-paid firms, and because income effect will make higher income workers prefer at least not less, if not more, leisure and so high pay premium firms can use it to attract workers if possible.

[66]The coefficients for training are misleading due to the reason that we talked earlier.

Table 9: Hedonic Regression on Selected Compensations with Full Controls

	Pooled	Computer	Design_ Media	Admin
	(1)	(2)	(3)	(4)
Advanced Insurance	.017**	.016**	.011**	.004
	(.001)	(.001)	(.002)	(.003)
Basic Insurance	-.026**	-.024**	-.018**	-.014**
	(.000)	(.001)	(.001)	(.001)
Backloading Wage	.009**	.012**	.022**	.011**
	(.001)	(.001)	(.002)	(.002)
Stock Option	.089**	.071**	.064**	.042**
	(.001)	(.001)	(.002)	(.004)
Commission	.029**	-.001	.003*	.032**
	(.001)	(.001)	(.002)	(.002)
Coworker Quality	.024**	.017**	.005*	.008*
	(.001)	(.001)	(.002)	(.004)
Training	-.001*	-.018**	-.002	.014**
	(.001)	(.001)	(.002)	(.002)
Work-Time	-.020**	-.019**	-.021**	-.022**
	(.000)	(.001)	(.001)	(.001)
Work Flexibility	.015**	.010**	.013**	.008**
	(.001)	(.001)	(.002)	(.002)
const	8.872**	9.155**	8.747**	8.336**
	(.002)	(.002)	(.004)	(.006)
Education FE	✓	✓	✓	✓
Experience FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
$\Xi_2, \ldots, \Xi_8$	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓
R^2	.743	.760	.757	.711
Adj. R^2	.738	.748	.730	.656
No. Obs	3998840	1325260	548808	260364

jobs with different levels of skills also provide non-wage compensations differently. In particular, those high wage premium firms sorted with high skilled jobs also provide many other pecuniary or nonpecuniary amenities including advanced insurance, additional payment, and high qualified work place, and these amenities are not compensated from posted wage but actually positively correlated with posted wage. In contrast, low wage premium firms sorted with low skilled jobs will provide basic insurance, training, and generous work-time as the job amenities, which are significantly compensated form posted wage. These empirical evidence provide hard challenges for the compensating differential theory, which claims that firms vary their provision on nonpecuniary compensation due to different cost functions of provision, and that workers select firms with different compensation through their heterogeneous preferences and allow their wage to be partially compensated for those compensations. For many pecuniary and nonpecuniary compensations provided by firms in our data, their cost functions are arguably similar for different firms, making it impossible to generate the strong linkages between their provision and firm and job effects. Similarly, it's a difficult empirical question that to what extent workers with different skills have different preference on these compensations. Actually if we believe that there is strong income effect on leisure, then it's hard to explain that why high wage premium firms sorted with high skill workers are substantially less likely to provide amenities of weekend, holiday, and less overtime, whereas low wage firms are more likely to provide such leisure to low income workers. And why for those amenities that they do generously provide unlike those low-pay firms, why do they not discount from their posted wages? In the next subsection, we suggest that compensation differential might not be the only force in the labor market for the provision of non-wage compensations and a new theory that takes efficient compensation and firm-worker sorting also into account can reconcile for all the empirical facts that we find here.

7.2 A New Theory

In this subsection, we suggest that the puzzle in the compensating differential literature, which is also occurred in our results, is not a problem of identification but a problem of incomplete theory. In particular, we argue that as long as we combine two additional elements, which are also observed in the literature and in our data, with the canonical mechanism of compensating differential, we can then generate patterns of compensation provision and different levels of compensating differential that are consistent with our empirical findings. The first new element is efficiency compensation, i.e. non-wage compensation can be efficient in production or in firm-operation.[67] The second new element is firm and worker sorting, or the firm and job sorting in our case. While the existence of the second element, sorting, have been confirmed by the recent literature on wage inequality and by the results here, the first element, the efficiency function of compensation, is often dismissed when empirically testing the impact of compensations on wage. Here we argue that the level of efficiency (or inefficiency) is a gen-

[67] We call it "efficiency compensation" because it is analogous to the idea of efficiency wage theory, which suggests firms pay wages higher than market clearing level for various efficiency reasons such that it is optimal for the production or profit maximization. Actually we think efficiency compensation is even a more nature idea because one key critique on the efficiency wage theory is that firms should be able to take advantage of other non-wage compensations to achieve the same efficiency aim (see Katz, 1986).

eral and important feature of non-wage compensations.[68] First, it is not difficult to see the efficiency nature of those monetary compensations like backloading wage and stock option. In fact the literature have been long argued that alternative payment structure can help firm to improve efficiency through effort inducing, turnover reduction, and so on (see e.g. Lemieux et al., 2009). Similarly, it has been argued in the literature that health insurance and other insurance can reduce exogenous worker turnover (see e.g. Dey and Flinn, 2005), and that better coworker quality improves both production productivity and on-the-job learning efficiency in a complementary production setting (see e.g. Jarosch et al., 2021). In contrast, weekend, holiday, and less overtime or limited work duty are straightforward inefficient because they allow less work-time and effort. Other amenities like training or work-time flexibility are perhaps more unambiguous and if they are efficient or inefficient likely depends on the detailed cases.

A formal model setting and derivation of our new theory, which combines a simple framework of worker sorting with efficiency compensation, are documented in Appendix E.. For the rest of this subsection we briefly introduce the key ideas, intuitions, and implications of our new theory.[69] The key idea is that when an amenity is allowed to be efficient in production, then in additional to the wage saving benefit, firm will also take the marginal product of efficiency improvement into account when considering the provision of a certain compensation. With firm-work sorting in the labor market, the level of this marginal production benefits from offering efficiency compensation will be larger in high wage premium firms that are sorted with high productivity workers or jobs. In other words, the better the firm or the job, more efficient will be the compensations. As a result, higher wage premium firms and higher wage jobs are more likely to provide those efficiency compensations, and because increase in productivity will often at least partially translate into increase in wage, this efficiency gain act in counter to the classical compensating differential mechanism. And if the level of the compensation has a large span and the marginal product does not decline too fast, it is also possible that the efficiency channel dominates the compensating differential above some threshold of firm and worker level, generating positive wage effect in net, i.e. firms providing better compensations now cause wage increase rather than wage decrease. In contrast, firms with lower

[68] To be clear, in the canonical compensation differential the provision of a compensation can be also efficient or inefficient in production. However, the theory of compensation differential assumes that the sign of the impact of the compensation on the production be must inverse to the sign of its impact on the workers' utility. In other words, an amenity for the workers must cause a reduction in production productivity or a direct cost in production. Here we relax this restriction and allow an amenity to be either efficient or inefficient or having no impact on production at all.

[69] In fact, the setting in Appendix E. is one of the simplest way, but not the only way, to generate the desired results, and there are many potential or further extensions that can be added to the basic framework. To distinguish with the traditional compensating differential model and to clarify our new mechanisms, in our model we assume workers are homogenous in their preference on all non-wage compensations and firms have the same direct cost functions on providing all compensations. However, both firms and workers are heterogeneous in their productivity, and they form pairs endogenously, and the joint production function is assumed to be supermodular—a necessary condition to generate positively assortative matching between firms and workers in the economy. Compensations provided by firms are assumed to be either efficient or inefficient, i.e. they affect an efficiency terms of the firms' production which acts as another complementary input in the production function. We show that this simple and parsimonious setting that contains efficiency compensation and sorting is enough to generate rich features of compensations provision and different levels of compensating differential. More realistic models can be constructed by adding heterogeneous worker preference or search frictions so that the sorting becomes no longer monotone or perfect.

wage premiums and sorted with low productive workers or jobs are less likely to provide efficient compensations because the marginal production benefits are small. And when firms do provide such compensations in some cases, say basic insurance that is mandated by the government, their net loss between the provision cost and the efficiency effect, if any, will be equalized through reduction in wage, and the lower is the rank and productivity of the firm and the work, the severe is the level of compensating differential. Therefore, our theory can generate the feature that while an amenity is significantly compensated from wage by low pay firms in low pay jobs, the same or even a superior amenity is not compensated from but actually positively correlated with wage in high wage firms and high skill jobs. The similar logic can be applied to compensations that are inefficient, say generous work-time or work-life balance.[70] Under complementarity, the higher the productivity and rank of the firm and worker, the larger the efficiency loss coming from the provision of such compensation.[71] Consequently, as long as the income effect on leisure is not too strong, high wage premium firms and jobs will not provide such compensations, but rather compensate workers for their utility loss with higher wage. On the other hand, such efficiency cost is small when the firm and the job have low rank and low productivity, and thus low wage premium firms with low skill jobs are more likely to document such inefficient compensations for attracting workers. In other words, now the efficiency channel is in the same direction as the compensating differential channel, and the impact of firm-worker sorting on the efficiency channel in fact act as an amplifier for compensating differential. Finally, when a compensation is neither efficient nor inefficient, the efficiency channel shuts down, and the model returns back to the traditional compensating differential model.

Our new theory have three implications that are important for understanding the labor market inequality in wage and non-wage compensation. First, the efficiency aspect of different pecuniary or nonpecuniary compensations could be the key to dissolve the puzzle that is brought by the mixed results found in the empirical tests for the theory of compensating differential. As our new theory shows, the efficiency effect can totally offset the effects of equalizing differential and generate results inverse to the predictions by the compensating differential theory. Our theory thus predicts that while it might be not difficult to find the clear evidences for compensating differential in the submarket with low-pay firms and low-skill workers, the similar evidences will be hard to find when targeting to the high-end labor market or the entire labor market. Also, directly adopting the estimation results and conclusions found from a particular compensation in a particular labor market to other compensations and other labor markets could be dangerous and misleading. Second, with firm-worker sorting and efficiency compensation, the labor market inequality could be underestimated by just looking at wage or monetary payments. The high-skill workers employed in high wage premium firms are likely to also enjoy the best non-wage compensations in many aspects, including both additional earnings from bonus and stock and nonpecuniary amenities like better insurance or fringe benefits, though at the expense of high effort. Perhaps more surprisingly, our theory suggest that the

[70] It is arguable that in some cases generous rests like paid leave or maternity leave can be actually efficient if they help to retain workers and the turnover cost is very high. In fact Bana et al. (2022) find the in the U.S. high wage premium firms are more likely to participate in Paid Family Leave programs and have lower turnover rates. However, it could be a difficult empirical question to answer ex-ante that if an amenity like this is efficient or not.

[71] Note that in additional to the linkage with the firm-worker match productivity, such inefficient compensation also offsets the effect of other efficient compensations.

provision of compensations can not only generate inequality in non-wage compensation itself but also further enlarge the wage inequality. This is because efficiency compensations can simultaneously increase the workers' direct utility on non-wage compensations and increase workers' wage through a boost in their productivity. In other words, efficiency compensations work as an amplifier for the labor market inequality at both observed wage level and observed utility level. Third, our theory suggest that the set of the unobserved non-wage compensations that drive the large amount workers' moving to low-wage premium firms will be rather limited (see Sorkin, 2018; Bonhomme et al., 2019). In fact, our theory suggest that these compensations must be inefficient ones like less work-time because high-wage premium firms will also provide better efficient compensations. Moreover, a worker that goes down the firm ladder due to some changes in preferences for certain amenities like leisure will suffer not only a worse matching but also a downgrading on many other efficient compensations, both of which will negatively affect the wage that the worker receive.[72]

8 Conclusion

In this paper we develop a new method to study the wage and compensation inequality in the labor market. This method relies on vacancy data and machine learning algorithms and can work as an alternative to the popular method in the literature which uses two-way fixed effects and employer-employee panel data. Applying the method to the vacancy data of a Chinese job board, we find that at least in this high-end labor submarket in China, the compositions of posted wage inequality is consistent with other findings in the labor markets of the U.S. and European countries. More importantly during the analysis process, we unmask the most granular details of job characteristics and find a data-driven skill and task structure featured by different levels of specificity. We find that those occupational specific skills and tasks are the most important part of job heterogeneities that can account for the posted wage inequalities and especially the sorting between firms and jobs or workers. In addition, our new approach also allows us to bring new insights on the labor market inequalities in non-wage compensations. In particular, we find that high wage premium firms sorted with high skilled jobs are in general also more likely to provide better compensations, except for work-time, and these amenities seem not to be subject to compensating differential. We suggest that it is important to take compensating differential, efficiency compensation, and firm-worker sorting all into account to explain these findings.

There are two caveats on our approach and results that are worth mentioning. First, as

[72] In fact in the section 5.4 of Rosen (1986), Rosen suggests an application of the compensation differential theory as "hours of work (or work schedules more generally) may be formally treated as nonpecuniary aspects of jobs. Then the market transaction must be viewed as a tie-in in which a firm offers a fixed wage-hours package to workers, take it or leave it, with these package deals varying from firm to firm". He then suggests two sources for the equilibrium distribution of different packages generated in the labor market: coordination in production or set up costs. Our idea of labor market sorting as the source for heterogeneous provision of working hour and wage packages is close to the idea of coordinating production, but different from the classic compensation differential framework that Rosen suggest, in our argument the interpersonal differences in productivity affect the equilibrium allocation not only through the resulted heterogeneity in preference but also through firms' opportunity cost of offering such "inefficient" compensations. In addition, the nonpecuniary aspect of job we consider here can be more general and contains not only hours of work but also latent effort.

we have argued earlier, online vacancy data does not cover the entire labor market. A typical online vacancy data is inclined to those young, educated, and internet-related jobs and workers. Firms may not post all their jobs on the internet and the vacancy posting frequency could be potentially different from real job compositions within the firm. Also, the posted wages are always the entry wage and lack the information of within-firm wage changes, wage bargaining and other firm-level wage determinants. To what extent do these issues matter is an empirical question worth future investigation. The second caution is that throughout our analysis we examine the wage inequality in monthly pay rather than in an efficient unit level of hourly wage. In fact in most cases precise information about working hour is not available in the online vacancy data and thus such examination is prohibited. One might suggest that this would result overestimated labor market inequality if both higher wage and better non-wage compensations are in fact fully compensated by the difference in different working hours. Here we argue three points that could potentially alleviate this concern. First, there will often be additional wage for overtime work that are not accounted in the posted wage. Second, the variations in working hours are rather limited comparing to the variations in posted wages. Finally, labor market inequality is often more reasonable to be considered on the total compensation level because firms are likely to provide wage and working-time as an indivisible package.

In terms of the future work, one important task is to validate to what extent are the results of our new approach be consistent or different from the results of using administrative employer-employee data and AKM approach. One straightforward way to test for this is to find a country with both types of data to be available and then conduct both analysis and compare the results. Also given the fact that our online vacancy data used in this paper is limited to a labor submarket rife with IT-related firms, we expect to see if the similar results on the compositions of posted wage inequality can be obtained when applying to the vacancy data of other labor markets, though in those cases some adaptions and adjustments in the practical details of machine learning algorithms might be necessary. Finally, we think that our new theory on non-wage compensation provides new perspective for thinking about wage inequality, compensation provision, job switch on the labor market and thus further development on this idea could be promising.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008). High wage workers and low wage firms: negative assortative matching or limited mobility bias? *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(3), 673–697.

Autor, D. H. and M. J. Handel (2013). Putting tasks to the test: Human capital, job tasks, and wages. *Journal of labor Economics 31*(S1), S59–S96.

Bana, S., K. Bedard, M. Rossin-Slater, and J. Stearns (2022). Unequal use of social insurance benefits: The role of employers. *Journal of Econometrics*.

Banfi, S. and B. Villena-Roldan (2019). Do high-wage jobs attract more applicants? directed search evidence from the online labor market. *Journal of Labor Economics 37*(3), 715–746.

Barth, E., A. Bryson, J. C. Davis, and R. Freeman (2016). It's where you work: Increases in the dispersion of earnings across establishments and individuals in the united states. *Journal of Labor Economics 34*(S2), S67–S97.

Becker, G. S. (1964). *Human Capital*. University of Chicago Press.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Bloesch, J., B. Larsen, and B. Taska (2021). Which workers earn more at productive firms? position specific skills and individual worker hold-up power.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2020). How much should we trust estimates of firm effects and worker sorting? Technical report, National Bureau of Economic Research.

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A distributional framework for matched employer employee data. *Econometrica 87*(3), 699–739.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics 36*(S1), S13–S70.

Card, D., A. R. Cardoso, and P. Kline (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly journal of economics 131*(2), 633–686.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics 128*(3), 967–1015.

Deming, D. and L. B. Kahn (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics 36*(S1), S337–S369.

Dey, M. S. and C. J. Flinn (2005). An equilibrium model of health insurance provision and wage determination. *Econometrica 73*(2), 571–627.

Di Addario, S., P. Kline, R. Saggio, and M. Sølvsten (2022). It ain't where you're from, it's where you're at: hiring origins, firm heterogeneity, and wages. *Journal of Econometrics*.

Frank, M. R., D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Feldman, M. Groh, J. Lobo, E. Moro, et al. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences 116*(14), 6531–6539.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-

dimensional choices: method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Hershbein, B. and L. B. Kahn (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review 108*(7), 1737–72.

Jarosch, G., E. Oberfield, and E. Rossi-Hansberg (2021). Learning from coworkers. *Econometrica 89*(2), 647–676.

Katz, L. F. (1986). Efficiency wage theories: A partial evaluation. *NBER macroeconomics annual 1*, 235–276.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Krueger, A. B. and L. H. Summers (1988). Efficiency wages and the inter-industry wage structure. *Econometrica: Journal of the Econometric Society*, 259–293.

Kuhn, P. and K. Shen (2013). Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics 128*(1), 287–336.

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (2022). Wage posting or wage bargaining? a test using dual jobholders. *Journal of Labor Economics 40*(S1), S469–S493.

Lamadon, T., M. Mogstad, and B. Setzler (2022). Imperfect competition, compensating differentials, and rent sharing in the us labor market. *American Economic Review 112*(1), 169–212.

Lemieux, T., W. B. MacLeod, and D. Parent (2009). Performance pay and wage inequality. *The Quarterly Journal of Economics 124*(1), 1–49.

Lise, J. and F. Postel-Vinay (2020). Multidimensional skills, sorting, and human capital accumulation. *American Economic Review 110*(8), 2328–76.

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

Mas, A. and A. Pallais (2017). Valuing alternative work arrangements. *American Economic Review 107*(12), 3722–59.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy 66*(4), 281–302.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Rosen, S. (1986). The theory of equalizing differences. *Handbook of labor economics 1*, 641–692.

Sanders, C. and C. Taber (2012). Life-cycle wage growth and heterogeneous human capital. *Annu. Rev. Econ. 4*(1), 399–425.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of economic literature 31*(2), 831–880.

Sockin, J. (2022). Show me the amenity: Are higher-paying firms better all around?

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. Von Wachter (2019). Firming up inequality. *The Quarterly journal of economics 134*(1), 1–50.

Sorkin, I. (2018). Ranking firms using revealed preference. *The quarterly journal of economics 133*(3), 1331–1393.

Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of labor economics 24*(2), 235–270.

Taber, C. and R. Vejlin (2020). Estimation of a roy/search/compensating differential model of the labor market. *Econometrica 88*(3), 1031–1069.

Wissmann, D. (2022). Finally a smoking gun? compensating differentials and the introduction of smoking bans. *American Economic Journal: Applied Economics 14*(1), 75–106.

Yamaguchi, S. (2012). Tasks and heterogeneous human capital. *Journal of Labor Economics 30*(1), 1–53.

Appendices

Appendix A. Data Collection And Processing

A.1 Data Collection

We set up a scraper which scraped all the vacancy data from the website of lagou.com in 2020. Because each vacancy that has been posted in the lagou.com website is attached with a unique ID, we were able to access to the information of the historical vacancies. Given the fact that at the end of the year 2020 new vacancy posts are typically assigned with an ID slightly larger than 8,000,000, we set up our scraper to try scraping all the vacancies with an ID between 0 and 8,000,000.

Despite a part of the vacancies that had been deleted from the website at the time our scraper accessed, we successfully collect a majority (about 75%) of all the historical vacancies that were still observable. Figure A1b plots the share of successfully collected vacancies for each 10,000 chunks of the total 8 million vacancies ordered by the ID. It shows that in general we scraped a consistent share of vacancies across all the IDs. In particular, for the vacancies ID between 0 and 3,000,000, we collect over 60% of all the vacancies and for the vacancies ID between 3,000,000 and 8,000,000 we collect over 80% of the vacancies, and within the unsuccessfully corrected vacancies over 20% are invalid vacancies that we have removed when scraping the data. Although we have no information on the deleted vacancies, we think those are more likely to be invalid or repeated vacancies and does not systematically bias any main results in our paper.

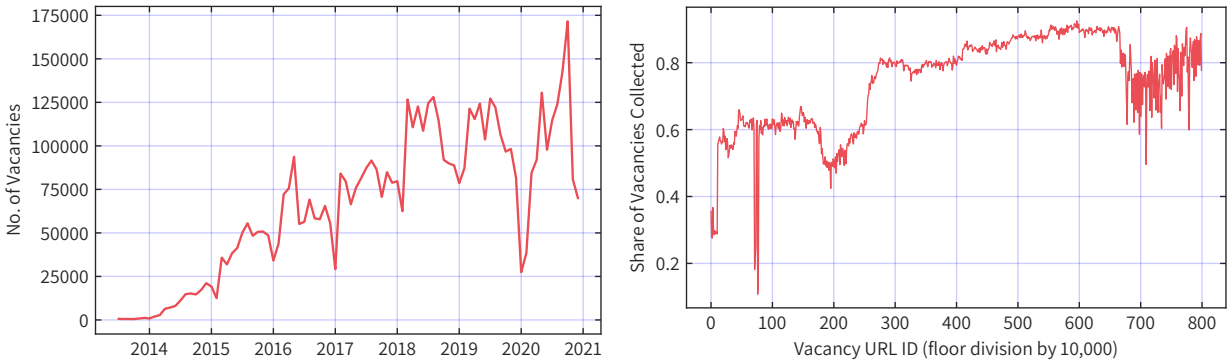
While the vacancy ID is only roughly correlated with time of posting, we can directly observe the posted time for each vacancy along with other information. Figure A1a plots the time trend of the monthly number of posted vacancies that we successfully collect. The monthly amount of vacancies increase over time which represents the growing popularity of the website. In particular, the average monthly amount of vacancies collected in year 2014 is about 12,000, and it grows to around 70,000 in 2016 and 2017, and around 100,000 between 2018 and 2020. Within a typical year, the number of posted vacancies is higher in the first half of the year and plummets in the end of the year, and this trend is consistent with other Chinese job vacancy data that target more general labor market (see e.g. He et al. (2021)).

A.2 Occupation Classification

In this section we explain the choices and the methods we use to assign the major and minor occupation for all the vacancies in our data.

There are two major steps of classifying the occupations for any vacancy data. The first step is that we need to decide that to which occupation code and in which level do we match our vacancies. Here, we decide to match our vacancy data to the official Standard Occupational Classification (SOC) 2018 designed by U.S. Bureau of Labor Statistics for two reasons. First, the U.S. SOC category has been widely used and studied in the labor literature and equipped with

Figure A1: Trends on Collected Vacancies



(a) # of Vacancies by Posted Month **(b)** % of Vacancies by Post ID Chunks

Notes. The sample here is the collected vacancies removing about 15% invalid posts that have either signals of being test posts, abnormal wages, lack of key information, or too less content in job descriptions. This sample is further trimmed to obtain the sample used in the analysis conducted in the main text.

detailed task and skill descriptions and variables that can be used to compare with our own measure (after taking average on the occupational level). Second, because our data mainly contains IT jobs and other jobs in IT firms in recent years, it requires a recently updated occupation classification to obtain a good match. Due to the fact that Chinese IT market has been advanced fast in recent years and largely followed the technological and organizational innovation in the global leading US IT market, we think the SOC 2018 would be a good fit to our data here.[73]

The second choice in the first step is selecting the occupation classification level. Ideally we want to match with the finest occupation level in SOC, which is the 6-digit occupations, so that we can use it to form the most accurate control of the heterogeneous skills and tasks between different jobs. However, Turrell et al. (2019) documents a potential tradeoff between the accuracy and the granularity in applying machine learning algorithms to assign job vacancies to the occupations codes. In particular, they argue that if matching with too granular occupation classification, the machine learning algorithm that based on job information from job title and job description text would find it difficult to accurately assign vacancies to the correct occupation. As a result, we decide to classify to the 3-digit level of the U.K. SOC.[74] We suggest that this result is mainly due to two reasons. First, adding more granular occupations as matching targets adds the possibility of the repetition of keywords across occupations which represents different meanings, and thus increase the difficulty of classifying occupations based on the job texts by any machine learning algorithms that only consider the occurrences of the keywords. Second, at the most granular level, i.e. the 6-digit SOC, some occupation cate-

[73] In comparison, the official occupation classification in China is not open to public access and largely outdated comparing to the fast development in the Chinese labor market, especially for ICT industries.

[74] The commonly used U.S. vacancy data from Burning Glass Technology has their vacancies data equipped with an occupation classification at 6-digit level of U.S. SOC. However, they do not make their machine learning algorithms public and thus one cannot tell the accuracy of their occupation assignment.

Figure A2: A Sample Vacancy From ByteDance

Notes. The style of the web page changes over time and this is a screenshot token in 2020 December. Some contents of vacancies (the part of job tasks, requirements, and benefits in left white space) are not always tidy as we have shown in this sample.

gories might not be well-defined and easily distinguished from other occupations even from a theoretical perspective— it might not easier even for the worker themselves to distinguish the similar occupation categories. This conceptual problem in occupation classification design is easy to understand in a multi-dimensional task framework, where occupations are defined as the different compositions of the multi-dimensional tasks. In such a framework, the most granular occupation is often defined as working on one specific task, or on an easy-to-recognized specific composition of tasks. However in many real world cases, the typical job that one works on can range within a set of composition of these specific tasks, and those who work on close shares of tasks would find it difficult to classify into each single one. One example is that while in the U.S. SOC 2010, the "15-1130 Software Developers and Programmers" is further divided by "15-1131 Computer Programmers", "15-1132 Software Developers, Applications", "15-1133 Software Developers, Systems Software", and "15-1134 Web Developers", the two items "15-1132 Software Developers, Applications" and "15-1133 Software Developers, Systems Software" are combined as "15-1252 Software Developers" in U.S. SOC 2018, probably due to the fact that these two occupations share very similar tasks. Considering these two problems and the feature of our data, in this paper we choose the occupation matching targets to be a limited set of SOC 6-digit occupations with some rearrangements to combining not well-defined occupations. In particular, rather than mapping to a whole set of all occupations in the SOC, we limit our target occupations to be six major occupations (Computer, Art & Design & Media, Business Operations, Financial & Legal & Educational, Sales, and Administrative Occupations) that constitute the bulk of our vacancy data and one other occupations that we use to classify any other occupations.[75] This limitation requires some preliminary check on what kind of the job the data contains, but it can significantly simplify the classification. For each major occupation, we select the relevant SOC 6-digit occupations to be the minor occupations and in some cases combining several SOC 6-digit occupations into one to make classification easy. Again the selection on the 6-digit occupations and the bundles requires the understanding of the data and is subject to potential bias. However, our machine learning algorithm introduced later would automatically refine any of these problems because in nature it will be a task-based classification. Finally, we add one new minor occupation, "product manager" into the major occupation "Business Operations", which appears to be a new occupation in our data but has no corresponding category in the 2018 SOC.[76] The set of all minor occupations are shown in

[75]By selecting these 6 major occupations, we do not include any management occupations (11-0000 Management Occupations in the SOC) although manager occupations can indicate skills and tasks important for wage determination. This is because the management occupations usually contain both some occupation-specific tasks and some general management tasks, which would usually dampen the accuracy of machine learning algorithm. This is also because often it's hard to tell the distinction between a management job and non-management job as there is no strict threshold of the share of management tasks beyond which a job will be recognized as a management job. Finally, the word manager or manage translated in Chinese is often used in non-management occupations and thus would likely to mislead the occupation classification. Note although we do not assign any vacancies to management occupations, we partly control its explanatory power on wage through our measure on experience. And eventually in our textual analysis on the job description, we would explicitly examine the importance of the management tasks and skills.

[76]This "product manager" is likely a new occupation that have been updated in the 2018 SOC. Actually the updating of SOC designs is lagged behind the real labor market, especially for the sectors with the rapid technological changes. For example, in US SOC 2018, "15-1253 Software Quality Assurance Analysts and Testers" and "15-1255 Web and Digital Interface Designers" have been newly added into "15-1250 Software and Web Developers, Programmers, and Testers", although these two occupations have been commonly recognized in the labor

Table A1.

After deciding the target for matching, the second task is to use the information in the job vacancies to match the most suitable occupations codes for each job vacancy. Prior literature (e.g. Turrell et al. (2019) and Atalay et al. (2020)) measure the similarity between each pair of a vacancy and an official occupation category and select the most similar pair as the assignment. To be specific, one typically first represent the texts of job title and job description in each vacancy and official classification documents as a numerical array and then calculate the cosine similarity between the arrays.[77] While this method is relatively simple and can be easily conducted for any vacancy data, the disadvantage of the method is that the texts of SOC occupation descriptions and sample job titles often is very limited and thus sometimes not contain enough information to distinguish different occupations. Also, these official descriptions are written by official analyst but not replacing the real words that will be used in the real job vacancies. These problems are especially severe in our case because the English description and job titles after translation is often not the similar Chinese words used in the Chinese labor market and thus does little help to distinguish the occupations of vacancies.

To overcome this problem, in this paper we use a simple dictionary method to select a learning sample and then do supervised machine learning on this sample so that we can both classify the occupations for the remaining sample and refine the result from our simple dictionary matching. In particular, we construct a dictionary that for each minor occupation I prepare several exclusive words or phrases that are either job titles or specific skills or tasks in Chinese that correspond to the terms in the SOC documents. Then for each vacancy we check if its job title or job description contains these keywords or not. If there is only one match, we directly assign the matched minor occupation to this vacancy and classify it as the learning sample. If there is no match or multiple matches, we classify it as the unknown sample. Because our man-made dictionary is likely not perfect, we would likely to have wrongly assigned vacancies in our learning sample, but by restricting the keywords to be highly specific, we ensure that the majority of the learning sample is correctly assigned. Next we use bag-of-words (BoW) method to transform the job text of vacancies $\mathbf{D}$ to a matrix of token counts $\mathbf{C}$ and apply a naive Bayes (NB) classifier to our learning sample. Each vacancy is represented by a row in the token matrix, $\mathbf{c}_i$, and each entry in this row, $c_{ik}, k \in K$, means the counts of the occurrence of token k from the entire token vocabulary K in the vacancy i. The details of this construction of $\mathbf{C}$ can be found in Appendix B.1.

The NB classifier is the most common and simple supervised classification algorithm and works quite well in many real-world situations in spite of its over-simplified assumptions. It is a generic model that assumes hypothetical distributions that generates the data and thus following the Bayes' theorem the possibility of a vacancy belonging to a minor occupation o

market years before 2018.

[77] The methods of transforming raw text to a numerical array usually includes bag-of-words (BoW), term frequency-inverse document frequency (tf-idf) and n-grams. For details of these methods one can refer to Gentzkow et al. (2019). Atalay et al. (2020) first runs a word embedding model, Word2Vec, to represent each words as a vector in a hidden feature space, and then add the vectors of all words in a vacancy to construct the vacancy-level array in the same latent feature space.

given its token vector $\mathbf{c}_i$ is

$$P(o \mid \mathbf{c}_i) = \frac{P(\mathbf{c}_i \mid o) P(o)}{P(\mathbf{c}_i)} = \frac{\prod_j P(c_{ik} \mid o) P(o)}{P(\mathbf{c}_i)}$$

, where the second equation is from the naive conditional independence assumption across tokens. The different naive Bayes classifiers differ by the assumptions on the distribution of $P(c_{ik} \mid o)$, and we follow the custom to use a multinomial version of NB classifier which is the typical one used in text classification. The probability $P(c_{ik} \mid o)$ can then be easily estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting

$$P(c_{ik} \mid o) = \frac{\sum_i c_{ik} + \alpha}{\sum_i \sum_k c_{ik} + \alpha K}$$

, where smoothing parameter α is often set to 1.

The estimated multinomial NB classifier is then used to classify the occupations for the unknown sample and also reapplied to classify the occupations for the learning sample. The latter process is done because in nature our classifier assigns the occupations by looking at how likely the tokens, which are mainly tasks and skills, occur given that it belongs to this occupation, and thus by applying the classifier back to our learning sample we can rectify the potential misassignment by the dictionary approach. This reassignment is shown in A3 from where we can see that most of the reassignments occur across the minor occupations within the major occupations. This means the confusing mainly exist across minor occupations because they share similar tasks and skills and indicate that our classifier works quit well.

We need to note that one might find our method not easily to be generalized to the whole labor market. This is because one needs to select the keywords for the dictionary used in the first step to pick the learning sample and thus the whole procedure is not fully automated but involve applying human knowledge. Hence, our method is currently more suitable for vacancies data with limited amount of occupations so that the researchers can easily construct the dictionary. However, in general we think our strategy have the potentiality to be improved and applied to the more general labor market. In particular one might find a way to automate the procedure of finding the unique keywords in the dictionary from the official descriptions. Or one might find an alternative way to obtain the sample vacancies for each occupation. The core advantage of our strategy is that in the second step, we can use simple supervised machine learning algorithms to learn from our data in hand and then classify the rest of the vacancies. The learning sample need not be 100% correct because we can apply the supervised machine learning back to itself to rectify mistakes. By using the supervised machine learning algorithms, we think the accuracy of our strategy will be largely better than the alternative methods that use cosine similarity. And this method would be more solid for the case of matching non-English vacancy data to English occupation classifications, like our paper.

Table A1: Occupations And Keywords Selected

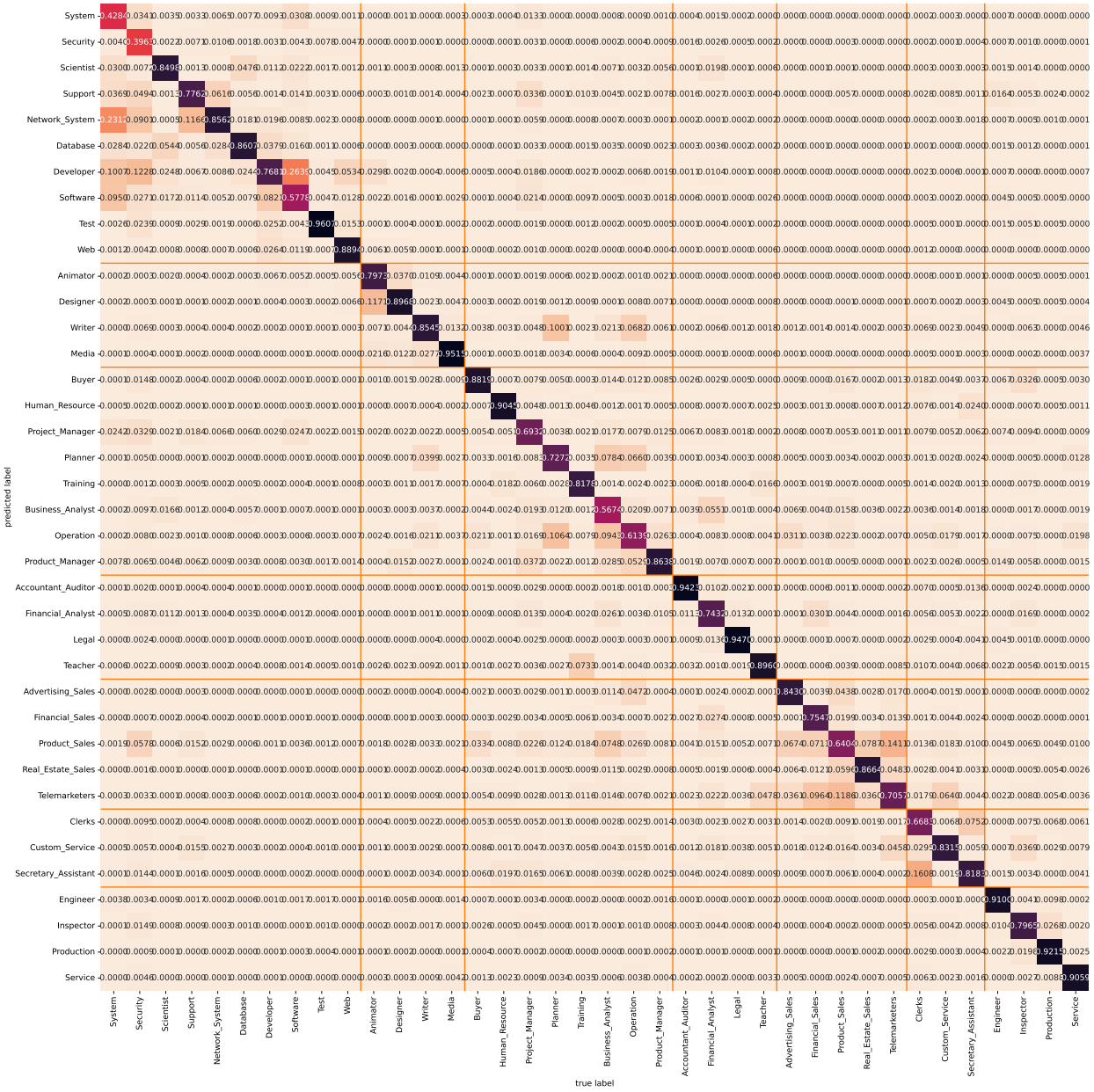
SOC Major	SOC Minor (6-digit)	Keywords Used For Assignment (Translation from Chinese)
	15-1211 Computer Systems Analysts	"Systems Analysis", "Systems Architect", "Systems Engineer"
	15-1212 Information Security Analysts	"Information Security", "Network Security", "System Security"

	15-1221 Computer and Information Research Scientists 15-2051 Data Scientists	"Data Mining", "Algorithm", "Machine Learning", "Deep Learning", "Image Processing", "Image Recognition", "Voice Recognition", "Computer Vision", "Natural Language Processing"
	15-1231 Computer Network Support Specialists 15-1232 Computer User Support Specialists	"IT Support", "Support Engineer", "Network Technician", "Network Support", "Pre-Sales Engineer", "After Sales Engineer"
	15-1241 Computer Network Architects 15-1244 Network and Computer Systems Administrators	"Network Engineering", "Network Architecture", "Network Management", "System Administration", "System Operations and Maintenance", "Operations and Maintenance Engineer"
	15-1242 Database Administrators 15-1243 Database Architects	"Data Engineer", "Data Architecture", "Database Engineering", "Database Architecture", "Database Administration", "Database Development"
	15-1251 Computer Programmers	"Development Engineer", "Programmer", "IT Engineer"
	15-1252 Software Developers	"Software Engineer", "Software Development", "Software Architect", "Application Development"
	15-1253 Software Quality Assurance Analysts and Testers	"Test Engineer"
	15-1254 Web Developers	"Frontend", "Web"
27-0000 Arts, Design, Entertainment, Sports, and Media Occupations	27-1013 Fine Artists, Including Painters, Sculptors, and Illustrators 27-1014 Special Effects Artists and Animators	"3D", "2D", "Original Painting", "Animation", "Painter", "Artwork", "Fine Art"
	27-1021 Commercial and Industrial Designers 27-1024 Graphic Designers	"Designer", "Graphic Design", "UI", "Drafting"
	27-3041 Editors 27-3043 Writers and Authors	"Editor", "Copywriter", "Editor", "Writer", "Lead Writer", "Screenwriter"
	27-4011 Audio and Video Technicians 27-4021 Photographers	"Photography", "Videography", "Editing", "Video Production"
13-1000 Business Operations	13-1022 Wholesale and Retail Buyers, Except Farm Products 13-1023 Purchasing Agents, Except Wholesale, Retail, and Farm Products	"Trade", "Import/Export", "Foreign Trade", "Purchasing", "Buyer"
	13-1071 Human Resources Specialists	"Personnel", "Human Resources", "HR"
	13-1081 Logisticians 13-1082 Project Management Specialists	"Project Management", "Process Management", "Logistics Management", "Logistics Planning"
	13-1121 Meeting, Convention, and Event Planners	"Event Planning", "Meeting Planning", "Event Operations"
	13-1151 Training and Development Specialists	"Trainer", "Training Instructor"
	13-1161 Market Research Analysts and Marketing Specialists	"Business Analysis", "Business Analysis", "Strategic Analysis", "Marketing Strategy", "Market Analysis"
	13-1190 Miscellaneous Business Operations Specialists 13-1??? Advertising, Promotions, Marketing Specialists	"Product Operation", "User Operation", "Promotion Operation", "Advertising and Marketing"
	13-1??? Product Manager	"Product Manager", "Product Design", "Product Planning"
13-2000 Financial Specialists; 23-0000 Legal Occupations; 25-0000 Educational Instruction Occupations	13-2011 Accountants and Auditors	"Accounting", "Audit", "Finance", "Tax"
	13-2041 Credit Analysts 13-2051 Financial and Investment Analysts 13-2054 Financial Risk Specialists	"Credit Analysis", "Credit Assessment", "Risk Control", "Risk Management", "Investment Manager", "Investment Analysis", "Industry Research", "Industry Analysis", "Securities Analysis"
	23-1011 Lawyers 23-2011 Paralegals and Legal Assistants	"Lawyer", "Legal", "Law"
	25-2011 Preschool Teachers, Except Special Education 25-3011 Adult Basic Education, Adult Secondary Education, and English as a Second Language Instructors	"Teacher", "Assistant Teacher", "Teacher", "Kindergarten Teacher"
41-0000 Sales and Related Occupations	41-3011 Advertising Sales Agents	"Advertising Sales"
	41-3021 Insurance Sales Agents 41-3031 Securities, Commodities, and Financial Services Sales Agents 13-2052 Personal Financial Advisors	"Investment Advisor", "Financial Advisor", "Financial Manager", "Financial Planning", "Financial Sales", "Insurance Sales"
	41-4011 Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products 41-4012 Sales Representatives, Wholesale and Manufacturing, Except Technical and Scientific Products	"Sales Representative", "Account Representative", "Sales Specialist", "Commercial Specialist", "Channel Sales"
	41-9021 Real Estate Brokers 41-9022 Real Estate Sales Agents	"Real Estate Consultant", "Real Estate Agent", "Real Estate Agent", "Real Estate Sales", "Real Estate Sales"

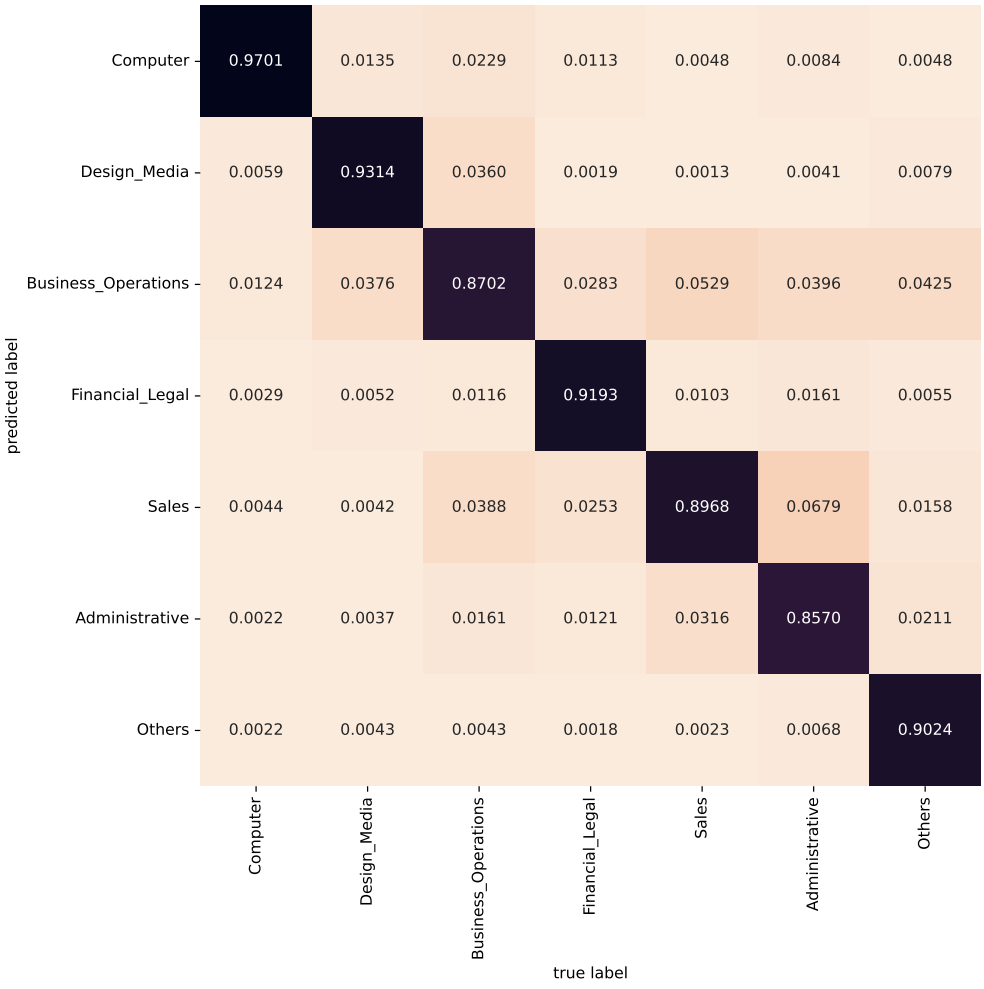
	41-9041 Telemarketers Solicit donations or orders for goods or services over the telephone	"Telemarketing"
43-0000 Office and Administrative	43-4171 Receptionists and Information Clerks 43-9061 Office Clerks, General	"Clerk", "Receptionist"
	43-4051 Customer Service Representatives	"Customer Service"
	43-6011 Executive Secretaries and Executive Administrative Assistants 43-6014 Secretaries and Administrative Assistants, Except Legal, Medical, and Executive	"Secretarial", "Administrative", "Clerical"
Others (Droped in Analysis)	17-2000 Engineers 17-3000 Drafters, Engineering Technicians, and Mapping Technicians	"Mechanical Engineer", "Process Engineer", "Equipment Engineer"
	19-4000 Life, Physical, and Social Science Technicians	"Quality Inspection", "Quality Testing", "Environmental Testing", "Equipment Testing", "Food Testing", "Communication Testing", "Chemical Testing", "Non-Destructive Testing"
	51-0000 Production Occupations	"General Laborer", "Operator", "Welder"
	35-0000 Food Preparation and Serving Related Occupations 39-0000 Personal Care and Service Occupations 41-2000 Retail Sales Workers 53-0000 Transportation and Material Moving Occupations	"Receptionist", "Delivery Person", "Courier", "Rider", "Beautician", "Driver", "Cook", "Sales Clerk", "Salesman", "Swimmer", "Taster", "Anchor", "Florist"

Figure A3: Confusion Matrix of Occupation Assignment

(a) Minor Occupations



(b) Major Occupations



Appendix B. Vectorization, Word Embedding, And Dimensional Reducing

B.1 Vectorization

In this section we explain our procedure of transforming the raw text of our vacancies **D** into the numerical token matrix **C** that are used in the machine learning algorithms. For all three machine learning methods that use **C**, namely the naive Bayes classifier for occupation classification, the lasso regression for feature selection, and the word embedding (Word2Vec) for feature clustering, there are several differences in the detailed choices but the general steps are exactly the same.

The first step is to select the individual documents $\{\mathbf{D}_i\}$ which is used to construct the individual numerical vector $\{\mathbf{c}_i\}$ (the rows of **C**). In the occupation classification and the lasso feature selection, an individual document is simply a vacancy. For the occupation classification, the $\mathbf{D}_i$ is the combined text of job title and job description. For the lass regression, the $\mathbf{D}_i$ is the combined text of job description and job benefits. In the word embedding model Word2Vec, the documents are not vacancies but the sentences in the job description and job benefits.

The second step is tokenization, i.e. breaking up raw text data **D** into short strings, and constructing the vocabulary set V, i.e. selecting the K standardized tokens (or in general features) that form the columns of **C**. In the textual analysis with English text the tokens are usually words obtained by splitting on spaces. But in Chinese a sentence is usually formed by multiple words which are present in a single sequence of characters without any spaces. To tokenizing the Chinese words, we use an open sourced Chinese tokenizer package, jieba.[78] The major advantage of jieba is that it is able to recognize Chinese compound words as well as to automatically tokenize both Chinese words and English words contained in one sentence.[79] We also add a list of IT words, education words, compensation words and etc. to the jieba tokenizer to enhance its performance. To reduce the dimension of the token/feature space, a lower bound of the occurrence of the tokens is often set to remove the words are too rare and do not convey much meaning. We set a lower bound of 10, so we only collect the tokens occurs over than 10 times. Also, after tokenizing the words from the text, a "stop words" list is often used to remove the words are very common and/or meaningless in the text. To do this, we use a commonly-used Chinese stop words list and in addition we use regular expression to remove all the tokens that are pure Chinese or English numbers or just one Chinese characters.[80] Finally, we remove all firms name from the segmented tokens because in the textual analysis firm names will be able to predict the posted wage through firm effects and thus disturb our examination on the skills, tasks and compensations. The remaining tokens then form the vocabulary V and

[78] Jieba is one of the most popular Chinese tokenizers that are open sourced. See the detailed information of its Chinese text segmentation functions in https://github.com/fxsjy/jieba.

[79] Because jieba does not automatically standardize the English words, we first lowercase all the English words before feeding our text to jieba tokenizer. We do not do any stemming or lemmatization for the English words because they are mostly technical words. There is no need to do stemming or lemmatization for the Chinese words because there is also no concept of a stem in Chinese at all.

[80] The common Chinese stop words list is taken from https://github.com/Alir3z4/python-stop-words. The numbers and single characters are removed because they often contain ambigious information.

thus all the features in **C**.

The final step is to select method of encoding for each entry $c_{i,k}$ in **C**. For the occupation classification, we use the most common way of encoding, bag-of-word, i.e. c_{ik} are the number of times token k occurs in document i, which is classical when using the multinomial naive Bayes classifiers. For the feature selection, we encode c_{ik} as an indicator of the presence of token k in document i, which is the simplest way to interpret the lasso regression. Using alternative methods like "term frequency-inverse document frequency" (tf-idf) would not affect our results qualitatively. Although these encoding methods are extremely simple and totally ignore the order of words that represents high-dimensional structure of the text, we find these simple methods are powerful enough to study the information embedded in the job texts.

B.2 Word Embedding

The model of word embedding with continuous bag-of-words (CBOW) architecture assumes the following process.

First, following exactly the same procedure as Appendix B.1, we construct the vocabulary set V from our raw vacancy documents **D** which contains K unique words or phrases.[81] Then we partition the entire corpus **D** into sentences, and each sentence is represented by a sequence of words denoted by $\{w_1, w_2, \ldots\}$, where each w_i is a word in V. Accordingly, we define a context of a word w_i in a certain sequence as a set of its adjacent words, $O = \{w_{i-m}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+m}\}$, i.e. a subset of $2m$ words in the same sequence that locate within a m-word window of w_i.

The basic idea of the estimation is to find two mapping U and W. The first function U maps any word w_i into a real vector in the hidden embedding space with pre-determined size H. The second function W maps the transferred vectors of a context, $U(O)$, to a conditional probability distribution: $\hat{P}\left(w_j \mid O\right) = W\left(U\left(w_{i-m}\right), \ldots, U(w_{i-1}), U(w_{i+1}), \ldots\right)$. And these two mapping are chosen by matching the estimated conditional probability with the conditional probabilities observed in the corpus through maximum likelihood procedure.

In practice, each word w_i is represented as a one-hot encoded vector, $\mathbf{x}_i \in \mathbb{R}^{|V|}$, i.e. an indicator vector of length K. Accordingly, a context is then denoted as $\{\mathbf{x}_{i-m}, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+m}\}$. Also, the two mapping is created as two matrices, input word matrix $\mathbf{U} \in \mathbb{R}^{H \times K}$ and output word matrix $\mathbf{W} \in \mathbb{R}^{K \times H}$. Although the **U** is still the mapping from word to the hidden embedding space, the **W** matrix here does not directly map $\mathbf{U}(O)$ to the conditional probability. In particular, we first calculate an averaged vector that represents the context in the latent layer, i.e.

$$\hat{\mathbf{u}} = \frac{\mathbf{U}(\mathbf{x}_{i-m}) + \ldots + \mathbf{U}(\mathbf{x}_{i-1}) + \mathbf{U}(\mathbf{x}_{i+1}) + \ldots + \mathbf{U}(\mathbf{x}_{i+m})}{2m} \in \mathbb{R}^H$$

. Then we use **W** to transfer this vector into a score vector $\hat{\mathbf{v}} = \mathbf{W}\hat{\mathbf{u}} \in \mathbb{R}^K$. Finally, we pass this score vector to a softmax operator to obtain the output vector $\hat{\mathbf{y}} \in \mathbb{R}^K$, with each element

[81] In general this vocabulary V need not be exactly the same as the one used for occupation classification or feature selection. But our selection and cleaning of the gathered tokens potentially also helps for the CBOW word embedding model so we use the same vocabulary here.

calculated by

$$\hat{y_k} = \frac{\exp(\hat{v}_k)}{\sum_{j=1}^{K} \exp(\hat{v}_j)}$$

.[82] This output vector $\hat{\mathbf{y}}$ is our estimation of the conditional probability distribution $\hat{P}(w \mid O)$, and $\mathbf{U}$ and $\mathbf{W}$ are found by maximizing the objective function $\sum_{k=1}^{K} \log(\hat{y}_k)$.[83]

In our computation we follow the literature to choose the primary parameters of our word embedding model. In particular, we set the window size of preceding and succeeding context words to be five ($m = 5$), and the dimension size of the hidden embedding space to be 100 ($H = 100$).

B.3 Dimension Reduction

Here we explain the procedure of the PLS dimension reduction. To easy notation, we denote our target variable log wage as $\mathbf{Y} \in \mathbb{R}^{N \times 1}$ and our predictive token matrix simply as $\mathbf{C} \in \mathbb{R}^{N \times K}$ (note in practice we go through each $\mathbf{C}'_{\mathbf{p}} \in \mathbb{R}^{N \times |V_p|}$). Our aim is to seek a representation of $\mathbf{C}$ in the lower dimensional space, $\boldsymbol{\Xi} \in \mathbb{R}^{N \times Q}$, where Q is the predetermined number of the components.

To obtain the first component, we first find a weight vector $\omega_1 \in \mathbb{R}^K$ that maximize the covariance between projected $\mathbf{C}$ and the target log wage $\mathbf{Y}$, $\mathrm{Cov}(\mathbf{C}\omega_1, \mathbf{Y})$. This can be achieved by finding the first left singular vectors of the cross-covariance matrix $\mathbf{C}^T\mathbf{Y}$, i.e. computing the singular value decomposition of $\mathbf{C}^T\mathbf{Y}$ and retain the singular vector with the biggest singular values. Then the first component is simply obtained as the projection $\xi_1 = \mathbf{C}\omega_1$. To calculate the second and following components, we take orthogonalization for both $\mathbf{C}$ and $\mathbf{Y}$ with respect to ξ_1, i.e. finding a loading vector $\gamma_1 \in \mathbb{R}^K$ and a loading value $\delta_1 \in \mathbb{R}$ that minimize the norm between $\xi_1\gamma_1^T$ and $\mathbf{C}$ and the norm between $\xi_1\delta_1$ and $\mathbf{Y}$ respectively, and replacing the original $\mathbf{C}$ and $\mathbf{Y}$ by the errors of their approximation respectively. We then take the orthogonalized value back to above procedure and iterate the whole process to obtain all remaining components $\xi_2, \ldots, \xi_Q$. In the end we gather all the components $\xi_1, \ldots, \xi_Q$ to form $\boldsymbol{\Xi}$ which is the demanded projection matrix of $\mathbf{C}$, and $\mathbf{C} = \boldsymbol{\Xi}\boldsymbol{\Gamma}^T + \mathbf{E}$ where $\boldsymbol{\Gamma}$ consists of the loading vectors $\gamma_1, \ldots, \gamma_Q$, and $\mathbf{E}$ are the error terms.

[82] This softmax function is equivalent to the multinomial logit model in discrete choice problems.

[83] In practice updating the two matrix and calculating the objective function is computational expensive due to the large size of the latent layer and thus a technique called negative-sampling is often used as a more efficient way of deriving word embeddings.

Appendix C. Connections To Deming & Kahn (2018)

Following the method in Deming and Kahn (2018), here we also construct the indicator variables for cognitive skill and social skill respectively and study their predictive power in posted wage regression. To be specific, we prepare two keyword lists about cognitive and social skills similar to the one in Deming and Kahn (2018) and generate the indicator variables by checking if a vacancy contains any words in the keyword list or not. The results of regressing posted log wage on these two variables along with a bunch of other controls are shown in Table C1. The significant and positive relationship between wage and these indicator variables that found in Deming and Kahn (2018) is also observed in our results, even after controlling for education, experience and occupation dummies. However we find that these variables barely increase the R-squared value, and thus explain not much wage variation in the data. Moreover, after further controlling for our skill and task variables $\Xi_2, \ldots, \Xi_8$, the coefficients of these two variables decrease substantially and the coefficient of cognitive variable even turns negative in some cases. Although this is surely not surprising given that both two types of proxy variables are constructed from the same original text and in similar way, what interesting is that the large decrease in coefficients occurs no matter which subset of the skill and task variables $\Xi_2, \ldots, \Xi_8$ we control for. This indicates that these cognitive and social skills are likely to be some high level combination of various general or specific skills and tasks. Therefore in order to understand for example why cognitive and social skills are becoming more complementary or that why firms require different levels on these skills, we need to carefully examine what low level skills and tasks do these high level index represent for.

Table C1: Wage Regression With Cognitive Skill and Social Skill Indicators

	Pooled					Computer					Design_Media					Admin				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
cognitive	.045	.029	.042	-.000	.002	.041	.023	.018	-.001	.000	.082	.066	.031	.005	.001	.013	.002	-.012	.018	-.002
	(.000)	(.000)	(.000)	(.000)	(.000)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)
social	.035	.019	.041	.018	.023	.026	.002	.019	.009	.010	.036	.015	.020	.010	.010	.049	.030	.016	.044	.024
	(.001)	(.001)	(.001)	(.000)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.002)	(.001)	(.001)	(.001)	(.002)	(.002)	(.002)	(.002)	(.002)
const	9.064	9.068	9.080	8.968	9.000	9.202	9.209	9.202	9.174	9.191	8.753	8.786	8.833	8.716	8.775	8.117	8.167	8.269	8.198	8.275
	(.003)	(.003)	(.003)	(.003)	(.003)	(.004)	(.004)	(.003)	(.003)	(.003)	(.004)	(.004)	(.004)	(.004)	(.004)	(.005)	(.006)	(.005)	(.005)	(.005)
Ξ_2		✓			✓		✓			✓		✓			✓		✓			✓
Ξ_3, Ξ_4			✓		✓			✓		✓			✓		✓			✓		✓
$\Xi_5, \ldots, \Xi_8$				✓	✓				✓	✓				✓	✓				✓	✓
Education FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Experience FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	.582	.594	.608	.630	.641	.465	.484	.517	.544	.562	.465	.480	.531	.534	.557	.429	.446	.501	.470	.512
Adj. R^2	.582	.594	.608	.630	.641	.465	.484	.517	.544	.562	.465	.480	.531	.534	.557	.429	.446	.501	.470	.512

Notes. The cognitive skill and social skill variables are constructed following Deming and Kahn (2018).

Appendix D. Additional Tables And Figures

D.1 Robustness Checks on The Results in Section 6

Table D1: Bias Correction on Posted Wage Variance ($X = \{\text{EDU, EXP}\}$)

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.360	-	.279	-	.251	-	.164	-
Panel A: Plug-In								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.132	.367	.089	.318	.078	.310	.061	.371
$\text{Var}(\psi_j)$	.076	.212	.102	.365	.086	.342	.041	.253
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel B: Homoscedasticity Correction								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.135	.376	.093	.334	.087	.345	.072	.441
$\text{Var}(\psi_j)$	.073	.204	.097	.349	.077	.307	.030	.183
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel C: KSS (Leave-Out) Correction								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.135	.374	.093	.332	.085	.339	.071	.431
$\text{Var}(\psi_j)$	.074	.205	.098	.350	.079	.314	.032	.193
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.138	.036	.130	.034	.136	.011	.069
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Table D2: Bias Correction on Posted Wage Variance ($X = \{\text{EDU}, \text{EXP}, \tilde{\Xi}\}$)

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: Plug-In								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.408
$\text{Var}(\epsilon_i)$	.096	.267	.071	.252	.065	.255	.050	.304
$\text{Var}(\psi_j)$	.051	.141	.074	.263	.062	.243	.035	.216
$2\,\text{Cov}(\theta_j, \psi_j)$	.051	.142	.054	.193	.043	.171	.012	.072
Panel B: Homoscedasticity Correction								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.330	.067	.408
$\text{Var}(\epsilon_i)$	.099	.273	.074	.264	.072	.284	.059	.361
$\text{Var}(\psi_j)$	.049	.135	.070	.251	.054	.214	.026	.159
$2\,\text{Cov}(\theta_j, \psi_j)$	.051	.142	.055	.194	.044	.172	.012	.072
Panel C: KSS (Leave-Out) Correction								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.330	.067	.407
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.352
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.028	.168
$2\,\text{Cov}(\theta_j, \psi_j)$	.051	.142	.054	.194	.043	.171	.011	.070
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Figure D1: Variation of Firm Effects Across Occupations

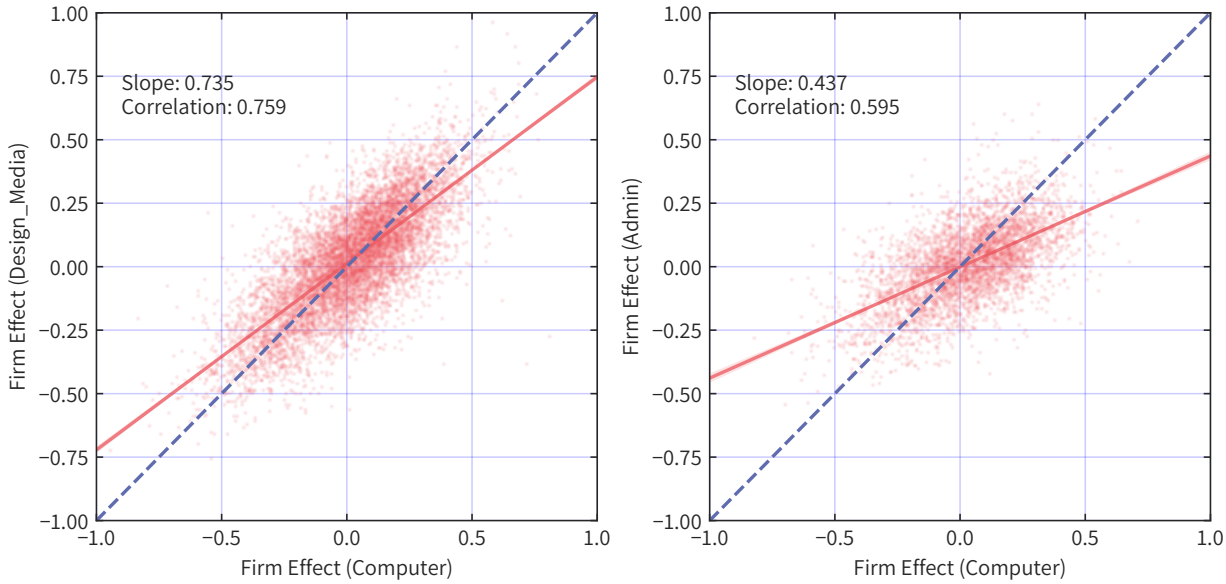


Table D3: Variance Decomposition Conditional on EXP=0

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.305	-	.407	-	.226	-	.097	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.079	.259	.068	.167	.038	.168	.014	.149
$\text{Var}(\epsilon_i)$	.115	.378	.111	.273	.084	.372	.049	.512
$\text{Var}(\psi_j)$	.068	.222	.138	.339	.075	.333	.029	.298
$2\,\text{Cov}(\theta_j, \psi_j)$	.044	.143	.090	.222	.030	.133	.004	.046
Panel B: Decompose θ Terms								
$\text{Var}(X_e)$	.001	.003	.002	.006	.001	.003	.001	.009
$\text{Var}(\tilde{\Xi})$	.073	.238	.058	.142	.036	.160	.011	.116
$2\,\text{Cov}(X_e, \tilde{\Xi})$	.005	.017	.008	.020	.001	.006	.002	.022
$2\,\text{Cov}(X_e, \psi_j)$	.002	.007	.008	.019	.002	.007	.001	.009
$2\,\text{Cov}(\tilde{\Xi}, \psi_j)$	.042	.136	.083	.203	.028	.126	.004	.038
Panel C: Further Decompose $\tilde{\Xi}$ Terms								
$\text{Var}(\Xi_g)$	.001	.004	.001	.002	.001	.004	.000	.003
$\text{Var}(\Xi_m)$	.010	.033	.009	.021	.007	.031	.005	.053
$\text{Var}(\Xi_s)$	.033	.110	.026	.064	.015	.065	.002	.023
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.002	.007	.002	.004	.001	.005	.001	.008
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.005	.018	.003	.008	.001	.006	.000	.001
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.020	.067	.018	.043	.011	.050	.003	.030
$2\,\text{Cov}(\Xi_g, X_e)$	.000	.001	.000	.001	.000	.000	.000	.001
$2\,\text{Cov}(\Xi_m, X_e)$	.002	.006	.003	.007	.001	.003	.002	.016
$2\,\text{Cov}(\Xi_s, X_e)$	.003	.010	.005	.012	.000	.002	.000	.004
$2\,\text{Cov}(\Xi_g, \psi_j)$	.003	.010	.005	.012	.002	.007	.000	.003
$2\,\text{Cov}(\Xi_m, \psi_j)$	.011	.036	.025	.063	.012	.053	.003	.030
$2\,\text{Cov}(\Xi_s, \psi_j)$	.027	.090	.052	.128	.015	.065	.000	.004
Obs	858147		144122		104960		120241	
Firm	66010		20060		19946		24807	

Table D4: Variance Decomposition If $\Xi_m \equiv \{\Xi_3\}$

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.408
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.353
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.027	.168
$2\,\text{Cov}(\theta_j, \psi_j)$	.051	.142	.054	.193	.044	.172	.012	.071
Panel B: Decompose θ Terms								
$\text{Var}(X_e)$	.047	.130	.031	.110	.032	.126	.020	.124
$\text{Var}(\tilde{\Xi})$	.063	.173	.029	.103	.028	.110	.022	.137
$2\,\text{Cov}(X_e, \tilde{\Xi})$	.053	.147	.022	.079	.024	.096	.024	.145
$2\,\text{Cov}(X_e, \psi_j)$	.022	.060	.022	.079	.021	.083	.006	.035
$2\,\text{Cov}(\tilde{\Xi}, \psi_j)$	.030	.082	.032	.114	.023	.090	.006	.036
Panel C: Further Decompose $\tilde{\Xi}$ Terms								
$\text{Var}(\Xi_g)$	.001	.003	.000	.001	.000	.001	.000	.001
$\text{Var}(\Xi_m)$	.003	.007	.001	.005	.001	.005	.002	.009
$\text{Var}(\Xi_s)$	.039	.108	.021	.074	.020	.078	.014	.083
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.001	.004	.000	.001	.000	.001	.001	.003
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.006	.017	.002	.006	.002	.007	.002	.009
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.012	.034	.004	.016	.004	.016	.005	.031
$2\,\text{Cov}(\Xi_g, X_e)$	.004	.012	.001	.005	.001	.005	.001	.008
$2\,\text{Cov}(\Xi_m, X_e)$	.009	.025	.004	.013	.004	.017	.006	.035
$2\,\text{Cov}(\Xi_s, X_e)$	.040	.109	.017	.061	.019	.074	.017	.102
$2\,\text{Cov}(\Xi_g, \psi_j)$	.002	.006	.002	.007	.001	.005	.000	.001
$2\,\text{Cov}(\Xi_m, \psi_j)$	.005	.014	.005	.019	.004	.017	.002	.014
$2\,\text{Cov}(\Xi_s, \psi_j)$	.022	.062	.025	.088	.017	.068	.003	.021
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Table D5: Firm Fixed Effect and Firm Characteristics

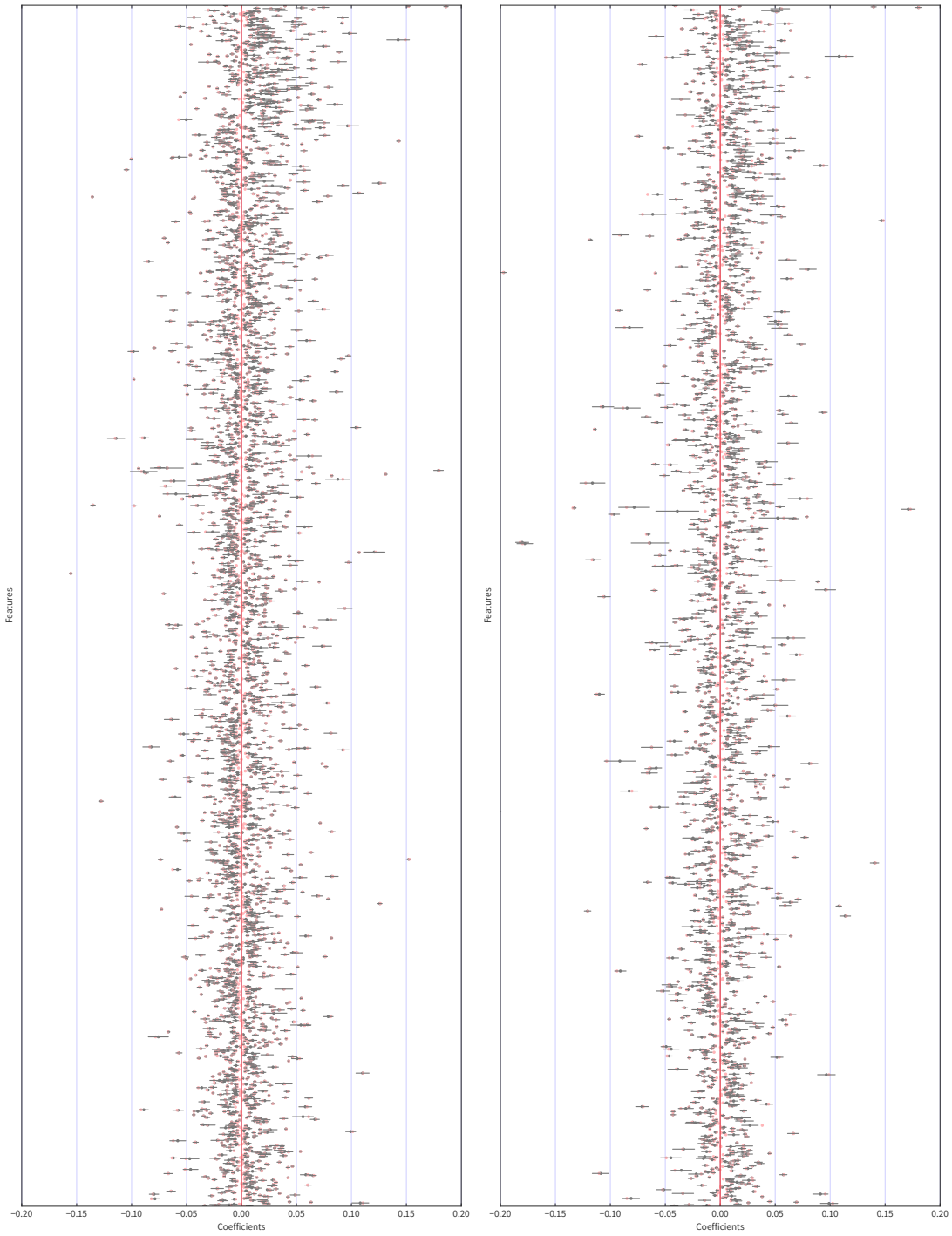
	Pooled			Computer			Design_Media			Admin		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
fsize.15-50	.019**	.018**	.023**	.012	.011	.014+	.049**	.035**	.045**	-.032	-.039	-.034
	(.004)	(.004)	(.003)	(.010)	(.009)	(.008)	(.011)	(.010)	(.008)	(.038)	(.034)	(.033)
fsize.50-150	.044**	.038**	.050**	.043**	.034**	.032**	.083**	.058**	.073**	-.023	-.038	-.035
	(.004)	(.004)	(.003)	(.010)	(.009)	(.007)	(.010)	(.010)	(.008)	(.038)	(.034)	(.033)
fsize.150-500	.069**	.059**	.068**	.079**	.053**	.043**	.127**	.087**	.094**	-.009	-.032	-.032
	(.004)	(.004)	(.003)	(.010)	(.009)	(.008)	(.011)	(.010)	(.009)	(.038)	(.034)	(.033)
fsize.500-2000	.099**	.081**	.086**	.119**	.070**	.053**	.176**	.121**	.120**	.015	-.014	-.019
	(.005)	(.004)	(.004)	(.011)	(.009)	(.008)	(.012)	(.011)	(.009)	(.038)	(.035)	(.033)
fsize.2000+	.125**	.105**	.121**	.154**	.077**	.065**	.213**	.140**	.134**	.028	-.005	-.006
	(.005)	(.005)	(.004)	(.011)	(.010)	(.008)	(.013)	(.012)	(.010)	(.038)	(.035)	(.034)
Job Effect ($\tilde{\theta}$)		.284**	.200**		.793**	.622**		.479**	.395**		.262**	.171**
		(.004)	(.003)		(.009)	(.008)		(.010)	(.009)		(.020)	(.018)
const	.148**	-1.101**	-.630**	-.176**	-3.946**	-3.018**	.157**	-1.931**	-1.488**	.175**	-.919**	-.468**
	(.003)	(.016)	(.015)	(.010)	(.042)	(.037)	(.010)	(.046)	(.040)	(.038)	(.079)	(.073)
Location FE			✓			✓			✓			✓
Adj. R^2	.017	.096	.381	.025	.243	.515	.053	.190	.473	.014	.062	.292
No. Obs	84023	84023	84023	30658	30658	30658	13871	13871	13871	5592	5592	5592

D.2 Lasso Features and Inference

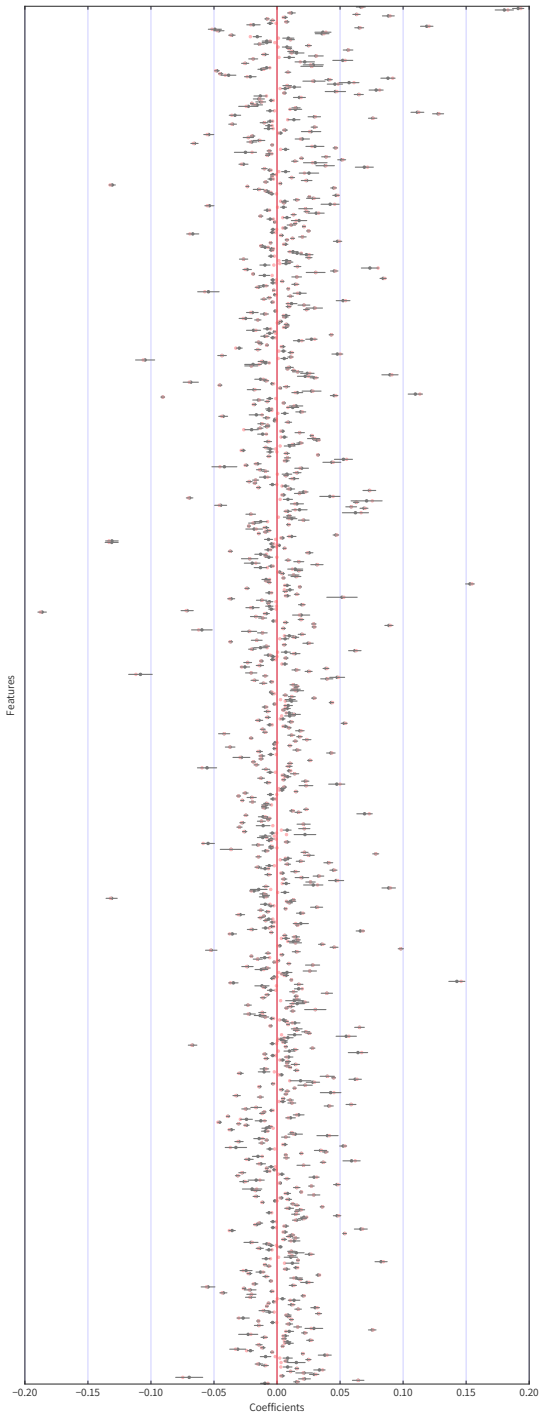
The full list of all the nonzero features (tokens) selected by Lasso estimation is very long. To save place, we are going to publish the list of tokens along with their classified clusters on web.

Figure D2: Subsampling on Lasso Non-Zero Coefficients

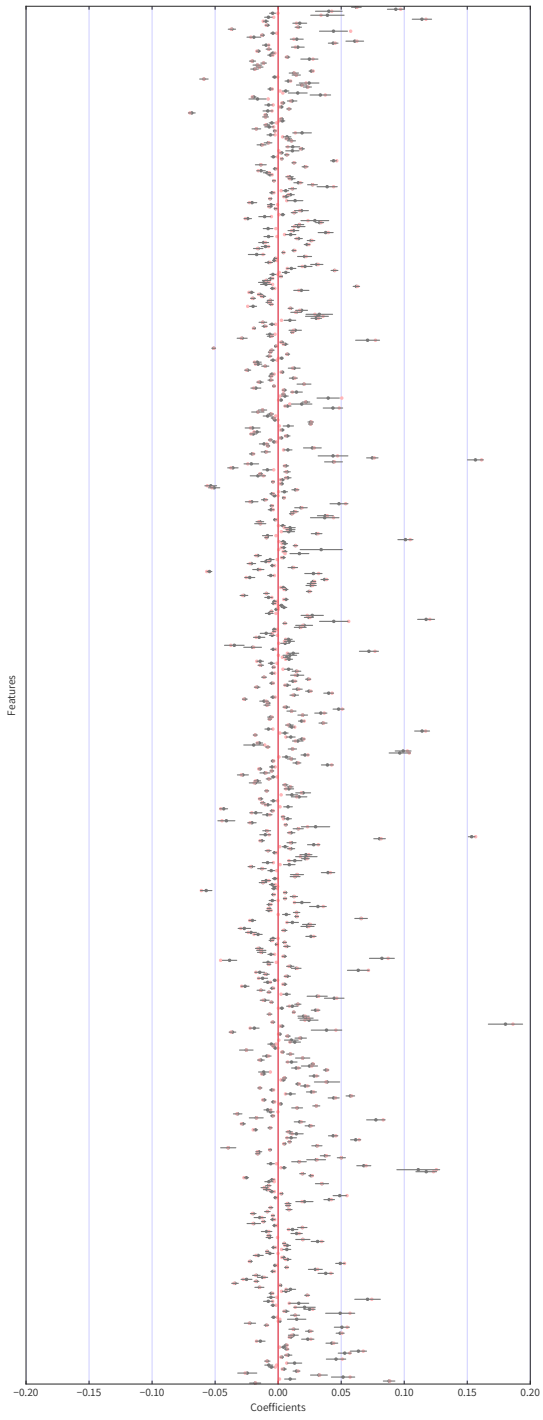
(a) Pooled **(b)** Computer



(c) Design_Media



(d) Administrative

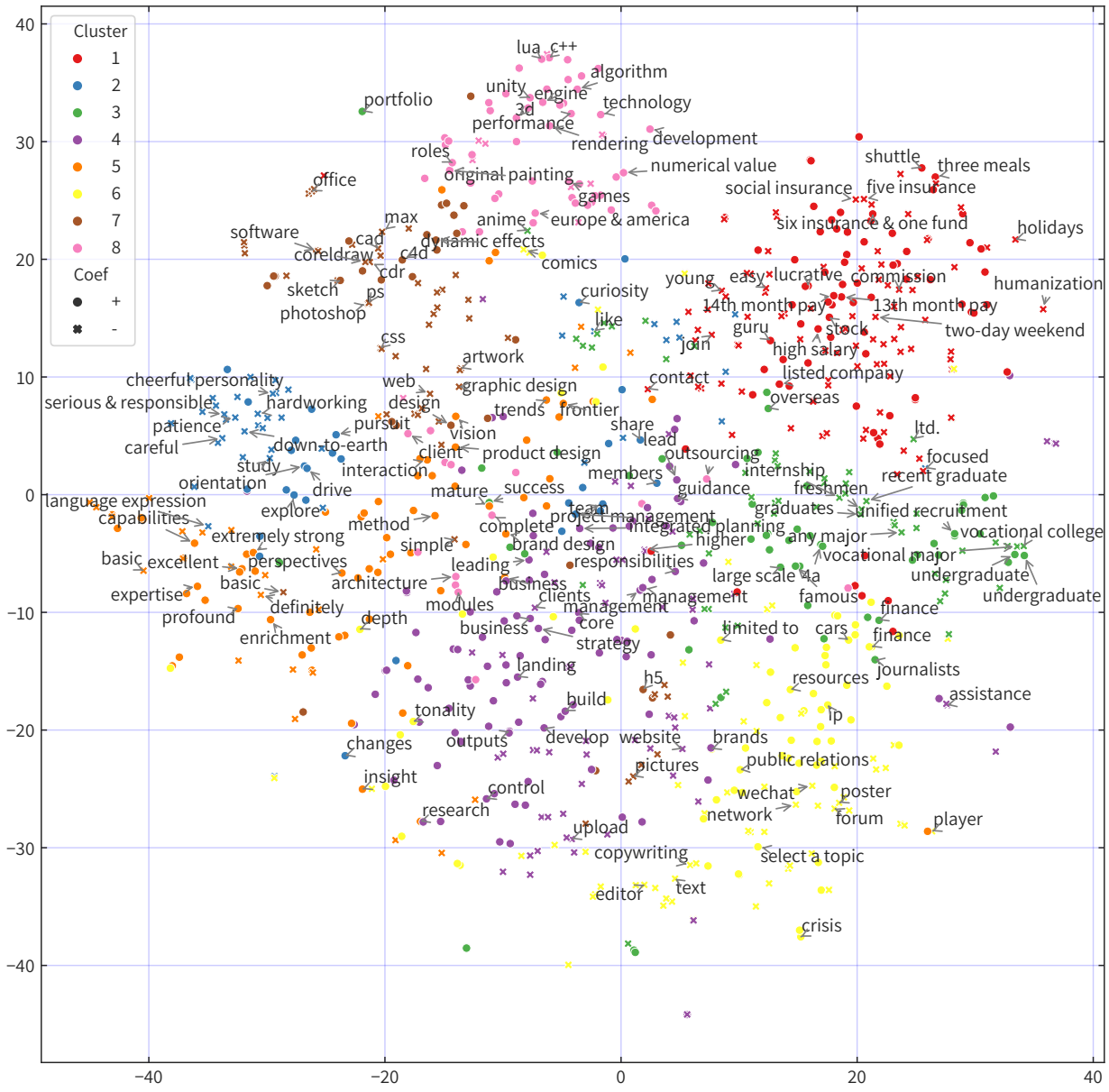


Notes. The figure plots the standard deviation for all nonzero coefficients under subsampling. The corrected standard deviation for each feature is calculated as $\text{sd}(\hat{\zeta})\sqrt{1/10}$ (because ten splits) under the assumption that the estimator's rate of convergence to be $\sqrt{N}$. Changing the number of splits or the rate of convergence does not qualitatively change the results.

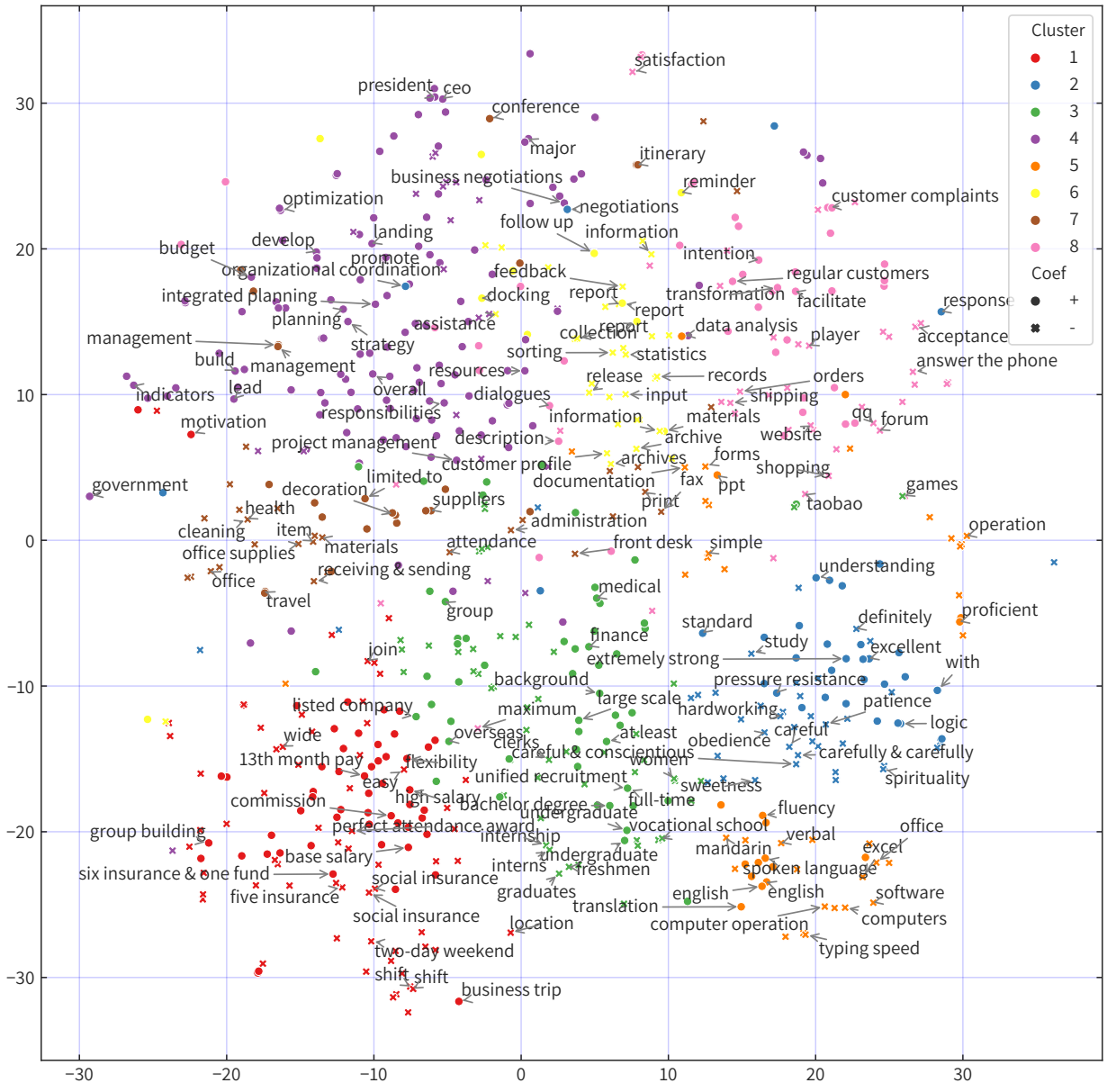
D.3 Feature Clustering

Figure D3: Feature Clustering

(a) Design_Media



(b) Administrative



D.4 Posted Wage Regressions

Figure D4: Lasso Prediction on Posted Wage

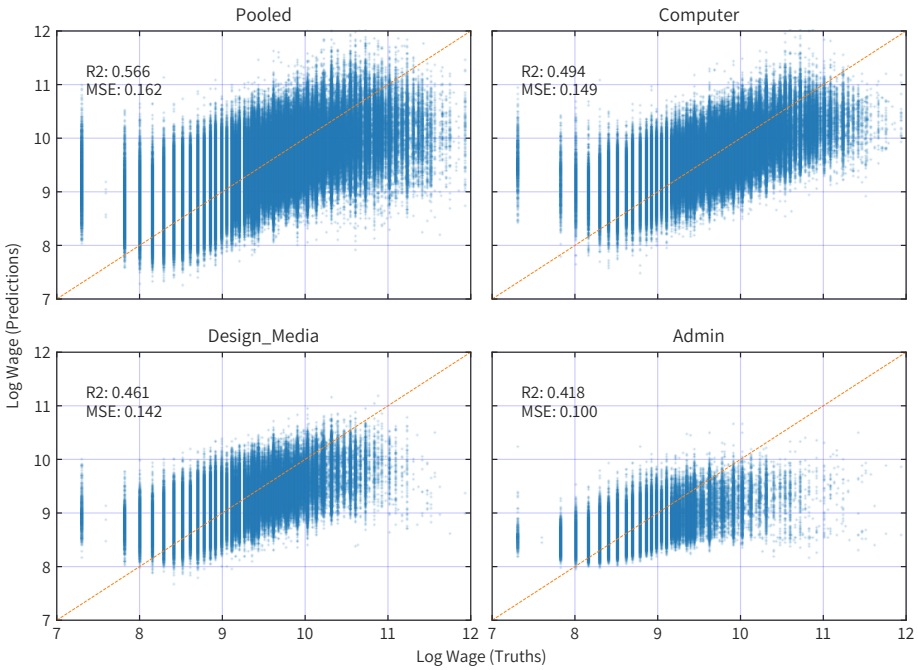


Figure D5: OLS Prediction on Posted Wage Using $\{\Xi_{p=1}^{P}\}$

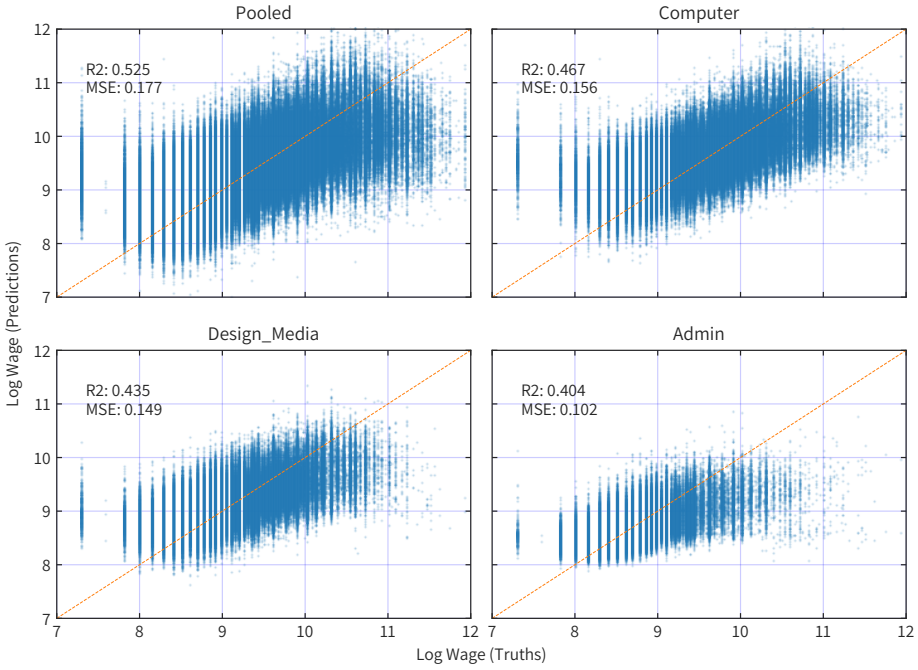
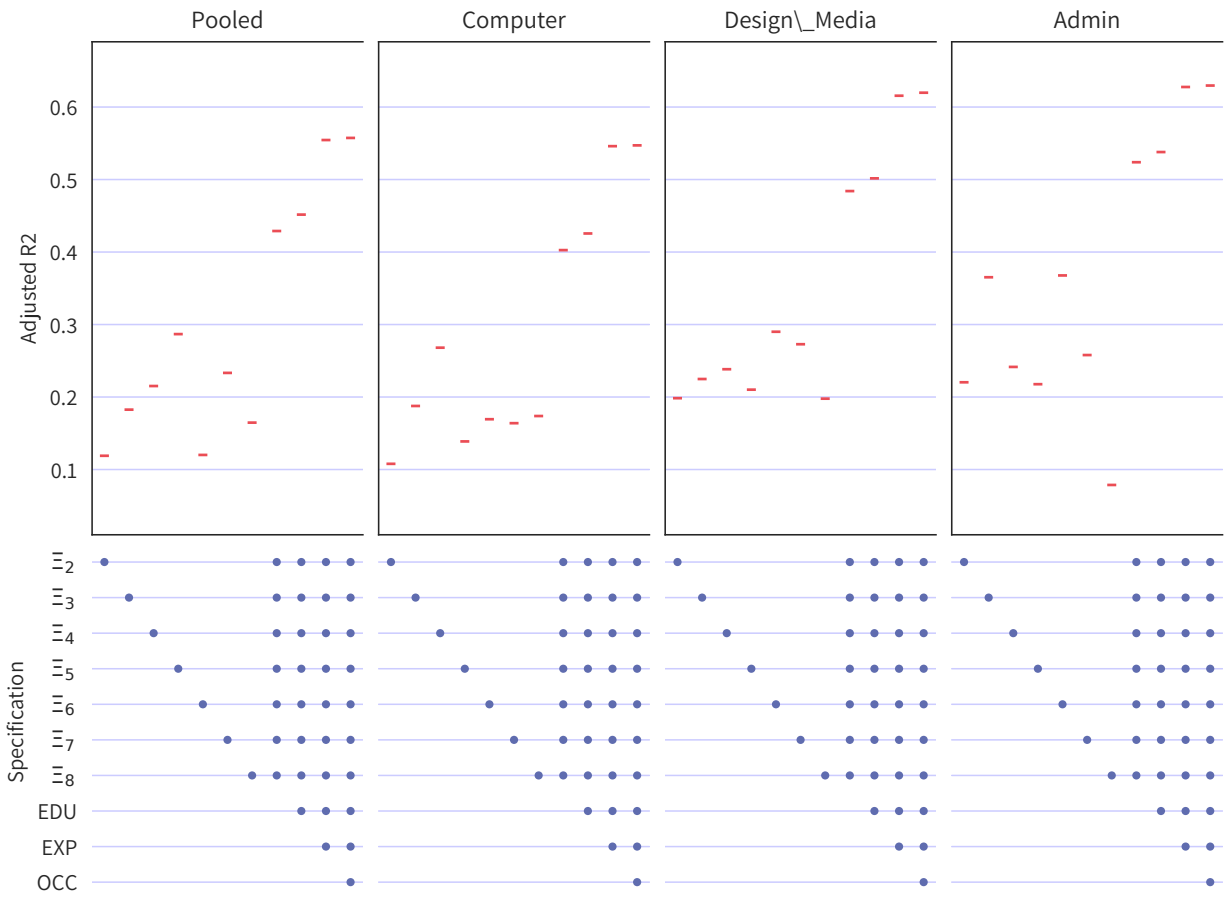


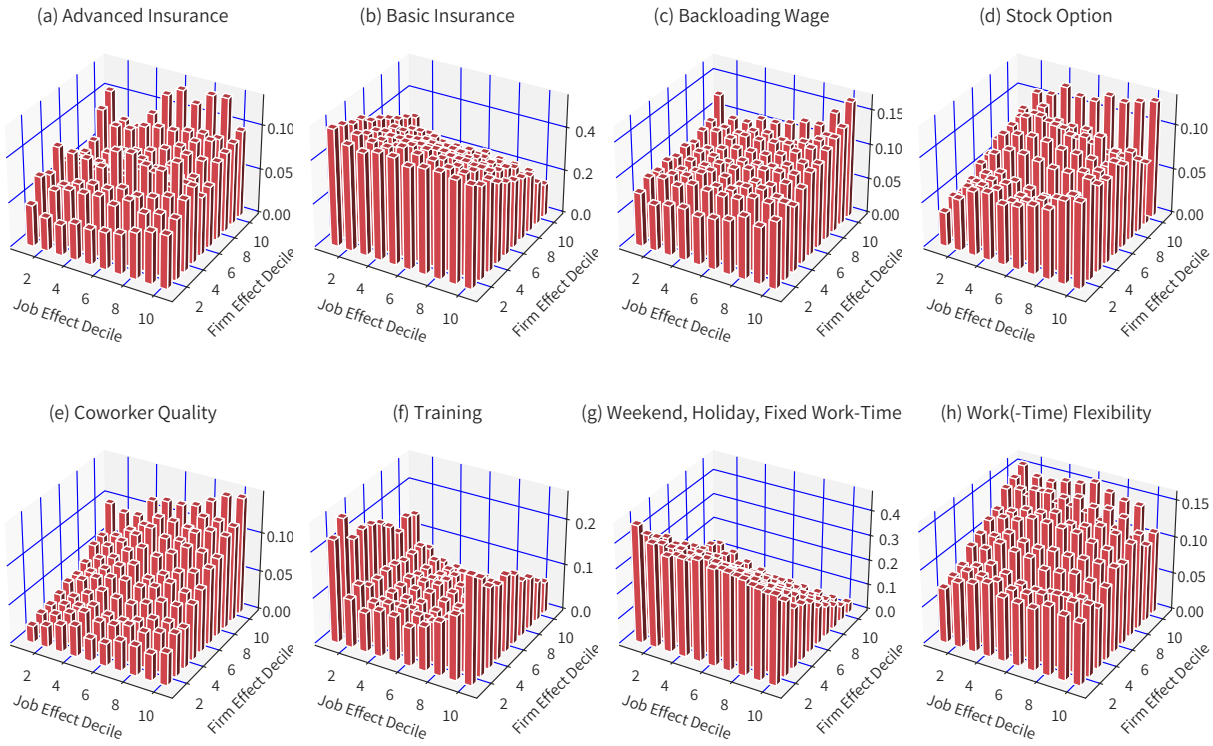
Figure D6: Prediction Power of Skill and Task Clusters in Posted Wage Regression



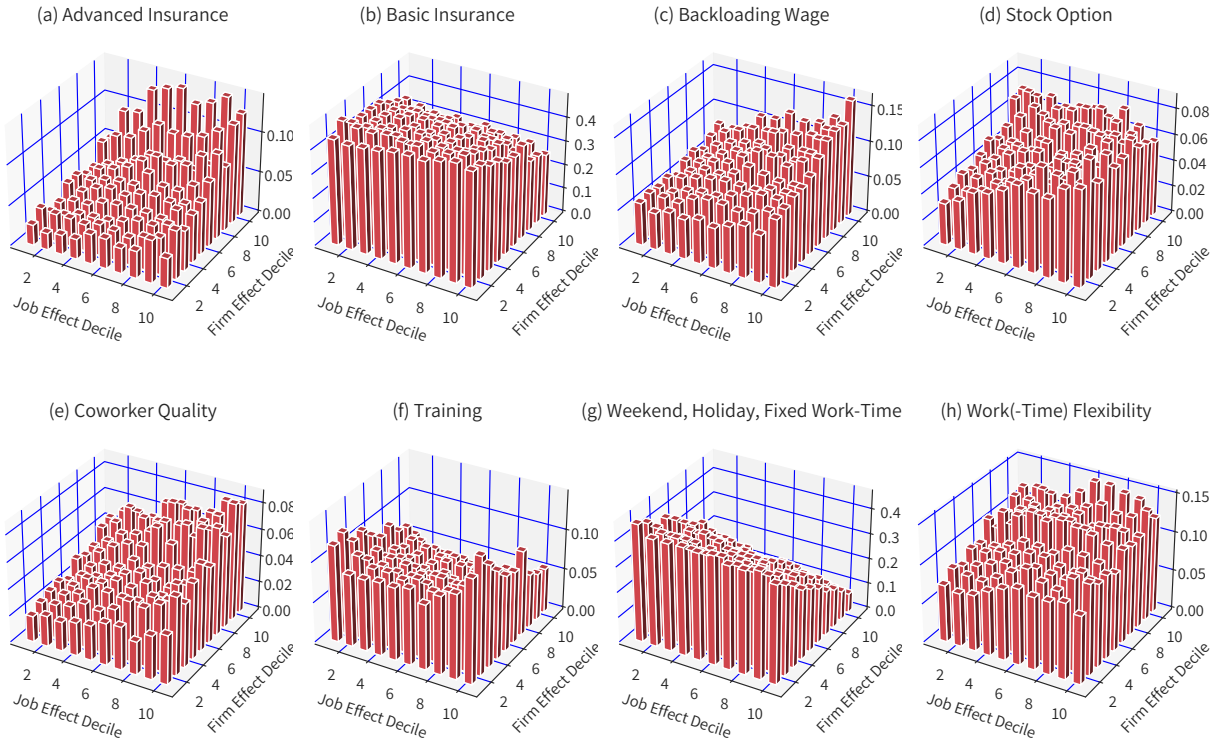
D.5 Compensation Occurrence

Figure D7: Compensation Occurrence

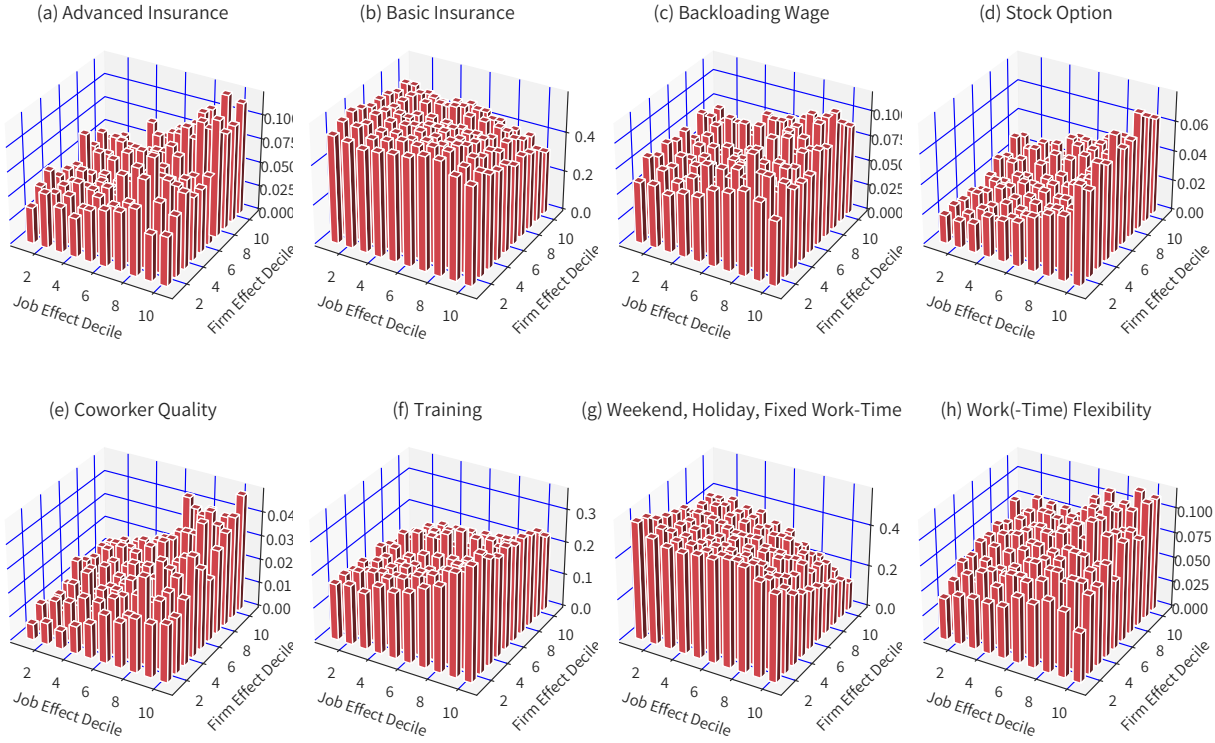
(a) Computer



(b) Design_Media



(c) Administrative



Appendix E. A Simple Model Of Sorting And Efficiency Compensation

We now build a simple model with firm-worker sorting and efficiency compensation and show that the results derived can be consistent with the results we find in our data and other similar results found in the recent literature. In comparison to the canonical compensation differential model, in our model workers and firms are not heterogeneous in their tastes of and cost functions of various peculiarity and nonpecuniary compensations, rather they are heterogeneous in their productivity like the typical assignment model.[84] The variation of the firms' provision of a certain compensation comes from different efficiency (or inefficiency) levels of the compensation in different firm-worker pairs. In fact, the traditional mechanism of compensation differential still exists but is now offset by a new efficiency mechanism when the compensation is efficient or magnified when the compensation is inefficient. The level of this new efficiency channel depends on the level of firm-worker sorting. We show that this model can thus generate flexible results on the wage impact of compensation provision that could be both consistent and inconsistent with the predictions of the traditional compensating differential model. We also show that this new theory can generate important implications for understanding labor market inequalities.

In the economy there is a continuum of workers with same utility function $U(C, a, h)$ but heterogeneous productivity $q \in [0, 1]$, where C is the monetary consumption, $a \in \{0, 1\}$ is the indicator of if the worker receiving a nonpecuniary amenity, e.g. the firm's provision of a certain insurance, and h is the level of a nonpecuniary disamenity, e.g. the requirement on additional working hour. To simplify the analysis, we assume the utility function takes an additively separable form,

$$U(C, a, h) = C + \phi_a a - \frac{h^{1+\phi_h}}{1+\phi_h} \tag{5}$$

, with $\phi_a, \phi_h > 0$. We assume that there is no income other than wage, thus $C = w(q)$, and the level of amenity a and disamenity h are decided by the firm that the worker matched in the equilibrium.

The economy also has a large continuum of potential firms that are ex-ante homogenous and want to hire workers. To facilitate sorting or separation, we assume that these firms face the same O-Ring type production technology, i.e. a production function with complementarity across all labor positions.[85] In addition, we extend the original O-Ring production function with efficiency compensation so that the labor productivity and the output Y_j of a firm j will also depend on the firm's choices on its supply of the compensations. In particular, the production

[84] A more general analysis with heterogeneous worker preferences (and heterogeneous firm cost functions of compensation provision) requires to extend the problem to multidimensional matching. However, this will significantly increase the analysis tractability, and thus we leave it for future research.

[85] An alternative way to establish sorting is to generate two-dimensional matching by assuming an exogenous distribution of firms with heterogeneous productivity and assuming a pair-wise production function with complementarity. One advantage of our one-dimensional matching setting is that we can easily generate endogenous firm size and have the firm size correlated with firm level productivity, as what we find the data and in the literature. The choice of this alternative setting and our setting does not affect our main results on compensation provision. More generally one can have both two-side heterogeneity and endogenous firm size, see Eeckhout and Kircher (2018).

function takes the form:

$$Y_j = AN_j^{1+\alpha} \prod_{i=1}^{N_j} q_i e(a,h) \tag{6}$$

, where A is a common productivity, N_j is the total number of worker the firm hires, and $e(\cdot)$ is an efficiency function which is increasing and strictly concave in both arguments. To simplify the exposition, we also assume the function $e(\cdot)$ takes an additively separable form:

$$e(a,h) = 1 + \gamma_a a + \frac{h^{\gamma_h}}{\gamma_h}$$

, where $\gamma_a, \gamma_h \in [0,1)$ control the decreasing return for these two efficiency compensation. This also helps us to illustrate that even when the decisions on the level of different amenities and disamenities are irrelevant by themselves, their occurrence could be still correlated through their relationships with the firm productivity. We further assume that for the amenity a, firm will pay the cost, which is a per-worker cost κ multiplied by the total number of worker N if $a = 1$. In comparison, for the disamenity h, a firm does not pay any direct cost but need to pay a higher wage w to compensate the loss in worker's utility. For analytical tractability, we also assume a fixed N for all firms but will show later that relaxing N to be another endogenous firm choice does not change our results.

The competitive equilibrium in this economy is defined as an assignment of worker types to firms and a utility schedule, $u(q)$, such that (i) given the utility schedule, all active firms maximize their profits by employing their workers in a way consistent with the assignment and by choosing the wages for their workers, $w(q)$, and the levels of two (dis)amenities, a and h, and that (ii) the labor market clears for workers of all productivity levels. The competitive equilibrium here coincides with the stable matching of the assignment problem: the workers' utility schedule and the firms' profit schedule is on the possibility frontiers and there does not exist other assignments that can generate larger payoffs.

We then characterize the competitive equilibrium by analyzing the firm's profit maximization problem:

$$\begin{aligned} \max_{\{q_i\}_{i=1}^N, a, h, w(q)} \; & AN^{1+\alpha} \prod_{i=1}^{N} q_i e(a,h) - \sum_{i=1}^{N} w(q_i) - a\kappa N \\ \text{s.t.} \quad & w(q) + \phi_a a - \frac{h^{1+\phi_h}}{1+\phi_h} \geq u(q) \quad \forall q \in \{q_i\}_{i=1}^N \end{aligned} \tag{7}$$

. The profit possibility frontier for the firm is $v(q_1, \ldots, q_N, u)$ with the utility compatibility constraint holding in equality and $a, h, w(q)$ chosen optimally. The derivative of v with respective to each single q_i is $\frac{dv}{dq_i} = AN^{1+\alpha} e(a,h) \prod_{i' \neq i}^{N} q_{i'} e(a,h) > 0$, and thus v is type increasing. It is then easy to see that the cross partial derivatives $\frac{d^2 v}{dq_i dq_{i'}} > 0$ and $\frac{d^2 v}{dq_i du} = 0$, and thus the equilibrium allocation in the economy satisfies positive assortative matching (PAM) and in our case this means all workers employed by any single firm will have the same type, q.[86] Under perfect

[86] In the original O-Ring mode there is perfect transferable utility, and thus the complementarity across different labor inputs in the O-Ring production function, i.e. $\frac{d^2 v}{dq_i dq_{i'}} > 0$, will be enough ensure that in the equilibrium assignment is PAM. However, the amenity and disamenity terms in the worker's utility function generate imper-

segregation, the firm's problem in Equation (7) now can be written as

$$\max_{q,a,h} AN^{1+\alpha}q^N(1+\gamma_a a+\frac{h^{\gamma_h}}{\gamma_h})-N\left(u(q)-\phi_a a+\frac{h^{1+\phi_h}}{1+\phi_h}\right)-a\kappa N \tag{8}$$

. Because the amenity a is a discrete choice, there is a productivity threshold q_a such that the firm that hires workers with q_a will be indifferent between providing or not providing amenity a. In particular the optimal provision strategy will be

$$a=\begin{cases}1, & \text{if } q\geq q_a\\ 0, & \text{if } q<q_a\end{cases}, \text{ and } AN^{\alpha}q_a^N\gamma_a+\phi_a=\kappa \tag{9}$$

This threshold is decided from the equation $AN^{\alpha}q_a^N\gamma_a+\phi_a=\kappa$, i.e. the marginal benefit of providing a equals the marginal cost of providing a. Note that if the amenity a is not efficient at all, i.e. $\gamma_a=0$, then it can return back to the canonical compensation differential where the dispersion of preference ϕ_a and of the cost κ generate sorting between workers and firms. As a result, more productive firm with more productive workers are more likely to offer these efficient compensations. The importance of this channel can be seen more clearly when the cost of the amenity is increasing in the level of the worker, as the case of many insurance and fund. If we assume the per-worker cost of a is actually $p\kappa$, then the cost of providing such amenity increases in firms which employ high q workers and pay high wage w. As a result with the traditional compensating differential mechanism alone, high rank firms are less likely to pay for such amenity, which is inconsistent with the empirical facts found in the literature. However, when a non-wage compensation is efficient and there is enough productivity dispersion across firms and workers, the differences in the efficiency effect could dominate and derive the discrepancy in compensation provision. However, with the efficiency channel, the increase in q has increased marginal benefit from efficiency effect more than covering the increased provision cost, generating positive relationship between firm productivity and compensation provision.

The first order conditions for the rest of two maximization choices q and h are

$$AN^{1+\alpha}q^{N-1}e(a,h)=u'(q), \tag{10}$$

$$AN^{\alpha}q^N h^{\gamma_h-1}=h^{\phi_h} \tag{11}$$

respectively. It's clear from Equation (11) that the optimal level of disamenity $h=\left(AN^{\alpha}q^N\right)^{\frac{1}{1+\phi_h-\gamma_h}}$ is also increasing in productivity q due to the same efficiency reason as amenity a. Therefore, more productive firm will also require high level of disamenity h. However, in this case, this disamenity will be fully compensated by the increase in the wage. To obtain the market wage, we first derive the market utility profile by replacing the optimal efficiency level $e(a,h)$ and

fectly transferable utility and as a result an additional condition $\frac{d^2v}{dq_i du}\geq 0$ is required for segregation assignment to be achieved in the equilibrium. For more details about the sufficient conditions for monotone matching in an economy with assignment problem see Legros and Newman (2007)

then integrate Equation (10) over the entire distribution of worker productivity:

$$u(q) = \begin{cases} \frac{\left(\bar{A}q^N\right)^{1+\omega}}{(1+\omega)(1+\gamma_h)} + (1+\gamma_a)\bar{A}q^N + u_a, & \text{if } q \geq q_a \\ \frac{\left(\bar{A}q^N\right)^{1+\omega}}{(1+\omega)(1+\gamma_h)} + \bar{A}q^N + u_0, & \text{if } q < q_a \end{cases} \tag{12}$$

, where $\bar{A} \equiv AN^{\alpha}$, $\omega = \frac{1+\gamma_h}{1+\phi_h-\gamma_h}$, and u_0 and u_a is the constant of integration. In fact u_0 will be the utility or wage that workers of $p = 0$ obtain, and thus is pinned down by free-entry condition such that $u_0 = 0$. Similarly, u_a is pinned down by the firm indifference at p_a such that $u_a = \phi_a - \kappa$. Finally, the market wage profile can thus be derived from Equation (12) as

$$w(q) = \begin{cases} \bar{A}q^N + \gamma_a\bar{A}q^N - \kappa + \frac{\left(\bar{A}q^N\right)^{1+\omega}}{(1+\omega)(1+\gamma_h)} + \frac{\left(\bar{A}q^N\right)^{\omega}}{1+\gamma_h}, & \text{if } q \geq q_a \\ \bar{A}q^N + \frac{\left(\bar{A}q^N\right)^{1+\omega}}{(1+\omega)(1+\gamma_h)} + \frac{\left(\bar{A}q^N\right)^{\omega}}{1+\gamma_h}, & \text{if } q < q_a \end{cases} \tag{13}$$

. The first term $\bar{A}q^N$ is the wage function when there is no non-wage compensation, which, perhaps not surprisingly, increases in productivity p, indicating that firms with high q workers also provide higher wage. The second part $\gamma_a\bar{A}q^N - \kappa$ in the case $q \geq q_a$ is a combination of the increase in wage due to efficiency effect and the compensation differential on a. Note when $q = q_a$ this term is $-\phi_a$, i.e. the workers' utility benefits from firms' provision of a is fully compensated from the reduction in wage, and thus for some range of the productivity $p > p_a$, it will generate a wage plunge comparing to firms that have a close productivity of workers but do no provide amenity a. However, going up the firm rank, the efficiency effect will increase in productivity q, and it's possible that the provision of the amenity a are not compensated from wage reduction at all but actually generate wage gain for the workers. The last two terms, $\frac{\left(\bar{A}q^N\right)^{1+\omega}}{(1+\omega)(1+\phi_h)}$ and $\frac{\left(\bar{A}q^N\right)^{\omega}}{1+\phi_h}$, are the increase in wage due to efficiency effect and compensation differential on h. In this case because h is an amenity and high rank firms demand more h, this generates a positive compensating differential on the wage. Therefore, our model shows that in a standard setting of firm-worker sorting or segregation, an efficient non-wage compensation, whether amenity or disamenity, could further enlarge market wage inequality, especially when the dispersion of productivity is large across firms. In other words, non-wage compensation can be not only an unobserved and overlooked labor market inequality, but also potential drivers behind the increasing in observed wage or earning inequality.

Next we discuss two more implications of our model on the empirical estimation of compensating differential. First, note for q close to q_a, one can find clear evidence of compensation differential by examining wage difference and controlling for worker characteristics. However, if a is a multiple discrete choice or close to an continuous choice, higher level of a will be correlated with higher wage, counteracting to the force of equalizing differential and thus confounding the estimation results. Also, if we assume that at some point in time the government mandate the firms' provision of amenity a (again for example the enforcement of a certain insurance) but the enforcement is not perfect. This will not change anything for all firms with worker productivity $q \geq q_a$, but it now requires all firms with $q < q_a$ also provide a. As a result, the compensation in wage reduction for these firms will be $\gamma_a\bar{A}q^N - \kappa$ and this reduction is larger for firms with lower q, providing larger incentive for these firms to circumvent the

provision of a. The empirical estimation will thus find strong evidence of compensation differential for this firms, but again the higher the labor market with firms and workers with high productivity, the smaller will be the equalizing effect and in some cases be even negative. Second, given that firms gathered with high productivity workers are likely to both provide high levels of efficiency compensations and require high levels of efficiency disamenities, a worker deviated from in the common utility structure for certain amenities will have to also deviate from the current optimal matching and be subject to changes in receiving other compensations. For example, if a middle-age female worker with a high productivity q gives a birth and thus has a large increase in disamenity in working hour (i.e. an increase ϕ_h), she has to go down the job ladder and match to a firm in which workers have productivity lower than q. Moreover, the level of other non-wage compensations might be also downgraded if these compensations are efficient. Therefore, in additional to the part of usual equalizing differential, the cost of such compensation differential will also incorporate the wage decline due to a worse matching and the utility decline due to a less generous package of other compensations.

Finally, we consider the case when N is also a choice of the firm. The additional first order condition with respect to N in this case is

$$AN^{\alpha}q^{N}e(a,h)(1+\alpha+N\ln(q))=w+ac \tag{14}$$

. Further differentiating Equation (14) with respect to q and evaluated at the optimal level, we obtain the optimal choice on firm size:

$$N(p)=\frac{1+\alpha}{-\ln(p)} \tag{15}$$

. This result shows that the firm size increases in p and is irrelevant to the choices of amenities. Therefore, all the relationships we have found between productivity and amenity provision can be now directly translate to the firm size.

Appendix Reference

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2020). The evolution of work in the united states. *American Economic Journal: Applied Economics 12*(2), 1–34.

Eeckhout, J. and P. Kircher (2018). Assortative matching with large firms. *Econometrica 86*(1), 85–132.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

He, C., K. Mau, and M. Xu (2021). Trade shocks and firms hiring decisions: Evidence from vacancy postings of chinese firms in the trade war. *Labour Economics*, 102021.

Legros, P. and A. F. Newman (2007). Beauty is a beast, frog is a prince: Assortative matching with nontransferabilities. *Econometrica 75*(4), 1073–1102.

Turrell, A., B. J. Speigner, J. Djumalieva, D. Copple, and J. Thurgood (2019). Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings. Technical report, National Bureau of Economic Research.