On Modeling Labor Markets for Fine-Grained Insights

Hendrik Santoso Sugiarto(✉) and Ee-Peng Lim

Singapore Management University, Singapore, Singapore
{hendriks,eplim}@smu.edu.sg

Abstract. The labor market consists of job seekers looking for jobs, and job openings waiting for applications. Classical labor market models assume that salary is the primary factor explaining why job-seekers select certain jobs. In practice, job seeker behavior is much more complex and there are other factors that should be considered. In this paper, we therefore propose the **Probabilistic Labor Model (PLM)** which considers salary satisfaction, topic preference matching, and accessibility as important criteria for job seekers to decide when they apply for jobs. We also determine the user and job latent variables for each criterion and define a graphical model to link the variables to observed applications. The latent variables learned can be subsequently used in downstream applications including job recommendation, labor market analysis, and others. We evaluate the PLM model against other baseline models using two real-world datasets. Our experiments show that PLM outperforms other baseline models in an application prediction task. We also demonstrate how PLM can be effectively used to analyse gender and age differences in major labor market segments.

Keywords: Labor market · Probabilistic labor market modeling · Labor market analysis

1 Introduction

Motivation. Recent technological advances create new jobs while making many existing ones obsolete. This rapid change not only affects job seekers and employers, but also governments which are tasked to address labor shortage or excess problems in the labor market. It is thus ideal to have the labor market analysed quickly to detect trends and events for intervention. Meanwhile, job portals on the Web bring jobs closer to job seekers at the same time collecting a lot of data about the jobs, job seekers and their application behavior. In some cases, the job portals are so large that they could represent sizeable labor markets. The job portal datasets also open up new possibilities for labor market research which are much more efficient than traditional surveys. Labor market surveys are usually conducted sporadically as they incur significant costs and human efforts. They

I. Koprinska et al. (Eds.): ECML PKDD 2020 Workshops, CCIS 1323, pp. 9–25, 2020.
https://doi.org/10.1007/978-3-030-65965-3_1

are not always able to reflect the pace of change in the labor market. Moreover, traditional research methods could only analyse the labor market at the macro-level, limiting its ability to support interventions with focus targets.

Research Objectives. In this paper, we therefore seek to introduce a new labor market model to conduct fine-grained analysis of jobs and job seekers in a labor market. Instead of a salary-only approach, we consider a rich set of variables to model the salary, topic, and accessibility criteria applicants use to decide which jobs to apply. As offer salary information can be found in almost every job, an applicant can easily compare that with his/her own reserved salary before submitting applications. There are also clusters (or topics) of jobs which different groups of applicants show interest in. There are also factors affecting how easy applicants can access the jobs. For each criterion, we consider a set of relevant latent variables (e.g., reserved salary), observed variables (e.g., offer salary), and the inter-variable relationships so as to construct the full labor market model.

The latent variables learned from the new labor model will benefit different market stakeholders. From the labor researcher's standpoint, this solution approach significantly lowers the barrier of analysing labor markets and their behavior. The model can help job seekers to determine their asking salaries for specific type of jobs. Employers can utilize the model to set appropriate salaries to attract talent. Finally, the analysis from this model can be utilized by policy makers in a targeted manner (e.g., immigration policy [9] and education system [7,24] to counter labor shortage/excess.

Overview of Modeling Approach. We first define the observed labor market data as $D = (U, P, A)$. U denotes a set of job seekers, or simply users; P denotes a set of job posts; and $A = \{A_{i,j}\}$ denote job application matrix of dimension $|U| \times |P|$. Every job p_j is assigned an offer salary range $[w_j^{min}, w_j^{max}]$. $A_{i,j} = 1$ when job seeker u_i applies job p_j, and $= 0$ otherwise.

With the observed labor market data, we develop a model called the *Probabilistic Labor Market (PLM) Model*. This model learns several important user variables, namely: (a) user topics, (b) user reserved salary, (c) user effort level, and (d) user optimism, as well as job variables, namely: (a) job topics, and (b) job visibility. The interactions between these latent variables and observed variables lead to multiple criteria behind users applying for jobs. More details about PLM is given in Sect. 3.

By incorporating all the above criteria, we can jointly learn all the PLM latent variables for all users and jobs in the market. This will then enable us to: (a) analyse the values and distributions of latent user and job variables, determine interesting patterns in their values, and correlate them to explain the observed application behavioral data; (b) derive latent labor market segments for dividing the labor market into smaller sectors that facilitate fine-grained analyses; and (c) predict the missing application which could be used for job recommendation.

Contributions. In the paper, we make the following key contributions:

- We develop a novel probabilistic model PLM for modeling labor markets. To the best of our knowledge, this is the first of its kind using observed job and application data to construct a generative labor market model.
- We evaluate PLM against several baseline models in application prediction task and show that PLM yields the best prediction accuracies.
- We apply PLM on real world job and application datasets. The analysis of the learned user and job variables reveals differences between labor markets, differences between labor market segments, and interesting gender/age differences across labor market segments.

Paper Outline. We will first cover some related works in Sect. 2. We present the PLM model in Sect. 3. Section 4 shows the experiment results using real world data respectively. Finally, we apply PLM to conduct labor market analysis in Sect. 5. Section 6 concludes the paper and highlights future works.

2 Related Works

Much of the past labor market research was derived from the labor economic theory of supply and demand which has been used to determine market equilibrium [3]. Many criticisms have been expressed toward this classical theory because many employers and applicants cannot be matched directly based on this theory and it cannot resolve long-term unemployment [1,10]. Other researchers proposed labor market models to cover wage differentials among similar workers [15]. In recent years, economists have also developed a search theory to study the frictional unemployment [22] and other implications [21].

Nevertheless, classical economics usually only assumes a unified labor market with open competition [6,19]. Alternatively, the theory of labor market segmentation considers partitioning the labor market according to specific criteria such as occupation and location in which participants from one market group cannot easily be included by other market groups [2,8]. In contrast to previous approaches, we propose a model with soft market segmentation based on labor topics. Although our model includes all applicants and jobs in an open competition setting, it distinguishes them by topical groupings and the probability of joining a specific market depends on interest matching between jobs and users.

Furthermore, labor market studies also require extensive experiments or a lot of effort to conduct surveys or census on employers and employees to collect relevant data [4,5,11,12]. In contrast, our proposed probabilistic model utilizes machine learning to learn the labor market situation directly from the interaction between employers and applicants through a job portal. This approach is not only novel but can be built and deployed efficiently. Lately, there are also several studies on labor market from the machine learning perspective. But they are trying to answer different problems. Such as fairness [13], ranking [18], reputation inflation[17], indexing [20], or even data infrastructure [23].

3 Labor Market Modeling

3.1 Probabilistic Labor Market (PLM) Model

In this section, we describe our proposed Probabilistic Labor Market (PLM) Model, the criteria it uses to model the application behavior of users as well as the associated user and job variables. The observed data for learning PLM consists of: (a) a set of users U, (b) a set of jobs P, and (c) a set of applications represented by $A = \{A_{i,j} \in \{0,1\} | u_i \in U, p_j \in P\}$. $A_{i,j} = 1$ when u_i is observed to apply p_j, and $A_{i,j} = 0$ otherwise. In real world settings, we can only observe $A_{i,j} = 1$'s. Each job $p_j \in P$ has an offer salary interval $[w_j^{min}, w_j^{max}]$. While $w_j^{min} < w_j^{max}$ in most cases, it is possible for a job to have $w_j^{min} = w_j^{max}$. As shown in Fig. 1, PLM incorporates salary, topic and accessibility criteria for determining whether a user u_i applies job p_j. The three criteria are represented as the following three probabilities: (a) salary-based probability ($a_{i,j}^s$); topic-based probability ($a_{i,j}^t$); and accessibility-based probability ($a_{i,j}^a$). We then define the probability of u_i applying p_j as $\hat{a}_{i,j} = a_{i,j}^s \cdot a_{i,j}^t \cdot a_{i,j}^a$.

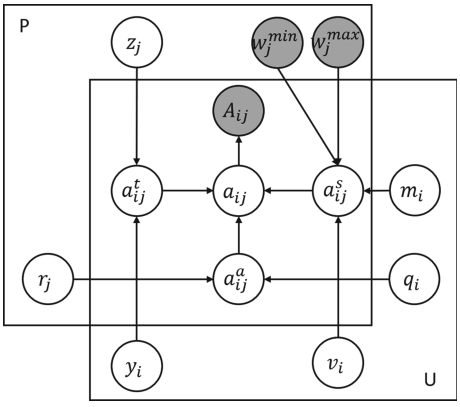

Fig. 1. Probabilistic labor market model

Salary Criteria. The salary criteria is inspired by labor economics. Salary-wise, every user u_i is assumed to have a *reserved salary* v_i. A job p_j is attractive if it offers salary higher than v_i. As each job has an offer salary interval, users may perceive an **effective offer salary** within interval for the purpose of comparison with reserved salary. We thus introduce for each user an optimism variable $m_i \in [0.1]$, to derive effective offer salary $s_{i,j}$ of job p_j with respect to user u_i as follows: $s_{i,j} = m_i \cdot w_j^{max} + (1 - m_i) w_j^{min}$. A user with extreme optimism $m_i = 1$ will use maximum offer salary as effective offer salary, and another user with extreme pessimism $m_i = 0$ will use minimum offer salary instead.

The salary-based probability $a_{i,j}^s$ is then determined by how well the reserved salary v_i is satisfied by the effective offer salary $s_{i,j}$. The more $s_{i,j}$ exceeds v_i,

the more likely u_i is interested in job post p_j, which in turns increases $a^s_{i,j}$. We thus define $a^s_{i,j}$ as: $a^s_{i,j} = \sigma(\frac{s_{i,j}-v_i}{S})$. The sigmoid function $\sigma(x) = 1/(1+e^{-x})$ converts the salary difference into a probability. We apply a simple global scaling S defined by the average difference between maximum and minimum salaries, i.e., $S = \frac{1}{|P|}\sum_{p_j \in P}(w^{max}_j - w^{min}_j)$.

Topic Criteria. Topic-wise, we expect each user to seek jobs matching his or her topical interests. In PLM, we use y_i and z_j to denote topic distributions of user u_i and job p_j respectively. Users should find topic-matching jobs more interesting than non-matching ones. We assume that both users and jobs share the same set of K topics. Probability $a^t_{i,j}$ is then defined by cosine similarity between user and job topic distributions, i.e.: $a^t_{i,j} = cosine(y_i, z_j)$. Cosine similarity is chosen because we want to represent topic matching as a probability (between 0 and 1). Other studies also show that cosine similarity generally performs better than other common measures such as Jansen-Shannon divergence [25].

Accessibility Criteria. Finally, the accessibility-based probability $a^a_{i,j}$ is determined by the effort-level of user u_i in job seeking, denoted by q_i ($q_i \in [0,1]$) and the visibility of the job j, denoted by r_j ($r_j \in [0,1]$). If $q_i = 0$, u_i is known to put in zero effort into job seeking resulting in not applying for any jobs that suit him or her. If $q_i = 1$, u_i will apply for all jobs that suit him or her. Mathematically, we define $a^a_{i,j}$ as: $a^a_{i,j} = q_i \cdot r_j$.

As we want to minimize the difference between model predictions and real applications, we define the objective function of PLM as:

$$F(U,P,A) = \sum_{u_i \in U, p_j \in P} (A_{i,j} - \hat{a}_{i,j})^2$$

To learn PLM well, we sample a subset of negative user-job pairs randomly and denote it by D^-. Specifically, for each positive user-item pair (u_i, p_j) with $A_{i,j} = 1$, we randomly select a set of N_{neg} negative user-item pairs, $(u_i, p_{j'})$'s such that $(u_i, p_{j'}) \notin D^+$ and add to D^-. The *positive-negative ratio* refers to $1/N_{neg}$. In our experiments, we have use 1/5 as the default ratio. Since the user-job matrix is usually sparse, better performance can be achieved by assigning higher ratio. However, higher ratio requires costlier calculation time. This ratio choice allows the model to achieve a reasonable performance within a reasonable time. Combining the negative sampling strategy, the objective function is revised as follows:

$$F(U,P,A) = \sum_{(u_i,p_j) \in D^+} (A_{i,j} - \hat{a}_{i,j})^2 + \sum_{(u_i,p_j) \in D^-} (A_{i,j} - \hat{a}_{i,j})^2$$

3.2 Model Learning

The learning of our model variables $\mathbf{X} = [\mathbf{v}, \mathbf{m}, \mathbf{q}, \mathbf{y}, \mathbf{r}, \mathbf{z}]$ is performed by minimizing the objective function. Specifically, for any model latent variables x, we update it by $x^{next} = x - \gamma \frac{\partial F(x)}{\partial x}$ iteratively. The derivative of F is:

$$\frac{\partial F}{\partial x} = -2 \sum_{(u_i,p_j)\in D^+} (A_{i,j} - \hat{a}_{i,j}) \frac{\partial \hat{a}_{i,j}}{\partial x} - 2 \sum_{(u_i,p_j)\in D^-} (A_{i,j} - \hat{a}_{i,j}) \frac{\partial \hat{a}_{i,j}}{\partial x}$$

By definition, the value of v_i, y_{ik}, z_{jk} should be non-negative. Every time the model updated any of those variables into a negative value, we clip the value back to 0. Similarly, the values of m_i, q_i, and r_j should be between 0 and 1. Therefore we clip the updating of these variables to be between 0 and 1.

The parameter, number of topics (K), has to be empirically determined for every given dataset. In our experiments on real world data, we therefore vary and select an appropriate value for K.

4 Experiments

We obtain two large job application datasets and design experiments to evaluate PLM against other models for the application prediction task.

4.1 Datasets

The main dataset in this paper is taken from the jobs bank of a major Asian city (SJD). This dataset covers job vacancies posted by all registered companies in the city as required by law and applications to these jobs in the year 2015. We acquired this dataset through collaboration with the dataset owner. The dataset consists of three types of data: (a) job posts, (b) applicants, and (c) applications. Every application involves an applicant and the job post he/she applied. The dataset covers jobs from all job sectors and can be accessed by all applicants.

The second dataset is the Wuzzuf Job dataset (WJD) which is available at Kaggle[1]. The jobs and applicants are mainly from Egypt. Similar to SJD, the WJD dataset covers: (a) job posts, (b) applicants, and (c) applications. It is unclear how representative WJD is but its data size is comparable to that of SJD. Most of the jobs in WJD are from the engineering and IT sectors.

Data Pre-processing. We performed the following data pre-processing steps to each dataset. First, we removed job posts and their corresponding applications which involve part-time, internship, and other ad-hoc jobs. Second, we also removed some jobs to ensure the salary information is reliable [14,16]. Those removed involved: (i) empty offer salary information; (ii) $\frac{w_j^{max}}{w_j^{min}} \geq 3$ (unrealistic salary range); (iii) $w_j = \frac{1}{2} \cdot (w_j^{max} + w_j^{min}) < \$500$ (possible hourly/daily/weekly wages); (iv) $|w_j - \mu_w| > 2\sigma_w$ where μ_w and σ_w are mean and standard deviation of w_j's respectively (salary outliers); (v) $w_j < w_o^{Q1} - 1.5 \times IQR_o$ or $w_j > w_o^{Q3} + 1.5 \times IQR_o$ where IQR_o denotes the inter-quartile range of offer salary of jobs sharing the same occupation o as p_j (occupation-specific salary

[1] https://www.kaggle.com/WUZZUF/wuzzuf-job-posts.

outliers). Finally, we filter users and jobs with less than 5 applications and a maximum of 300 applications iteratively until all users and jobs have at least 5 applications and maximum 300 applications to get high quality application data for training. The above filtering removed non-active users and non-popular jobs, as well as users who are spammers or testers the job portal. This filtering also removed possible scam jobs that attracted many users. For WJD, we only consider jobs using Egyptian currency in their offer salaries.

After pre-processing, we retain for SJD dataset 68,091 jobs (about 26% of all jobs), 33,866 users (about 41% of all users), and 827,380 applications (about 29% of all applications). For WJD, we retain 16,928 jobs (about 80% of all jobs), 66,734 users (about 21% all users), and 1,216,445 applications (about 66% of all applications).

4.2 Application Prediction Task

Task Definition. In this task, we predict the (user,item) pairs that are likely to have applications. This prediction task involves ranking a set of (user,item) pairs (u_i, p_j)'s by application probabilities a_{ij}'s from highest to lowest. The higher the rank, the more likely user u_i applies for job p_j.

Probabilistic Labor Market Prediction Model (PLM): PLM performs application prediction as follows:

$$a_{i,j}^{PLM} = \sigma((s_{i,j}^{PLM} - \hat{v_i}^{PLM})/S) \cdot cosine(\mathbf{y}_i^{PLM}, \mathbf{z}_j^{PLM}) \cdot q_i^{PLM} \cdot r_j^{PLM}$$

Note that the PLM predicts using all topic, salary and accessibility criteria. The variables $s_{i,j}^{PLM}$, ${v_i}^{PLM}$, $\mathbf{y}_i^{PLM}$, $\mathbf{z}_j^{PLM}$, q_i^{PLM}, and r_j^{PLM} are variables under the PLM model defined in Sect. 3.1.

Other PLM Variants: We also introduce several reduced variants of PLM for application prediction. We derive them by dropping one of the salary, topic and accessibility criteria:

- **PLM using Salary and Topic Criteria (PLM(ST)):** This is a PLM variant that assumes that accessibility does not play a part in application decisions. Hence, user efforts and job visibilities are assumed to be identical and set to 1 for all users and jobs respectively.
- **PLM using Salary and Accessibility Criteria (PLM(SA)):** This is a PLM variant that assumes that topical interest does not play a part. Hence, all user and job topic distributions are set to have uniform values $\frac{1}{K}$.
- **PLM using Topic and Accessibility Criteria (PLM(TA)):** This PLM variant assumes that salary is not important and users are always satisfied with any offer salary. Consequently, reserved salaries are set to \$0 and optimisms are set to 1.

Non-PLM Baselines: We also include several other baseline models as follows:

- **Optimism-based (Opt):** This method predicts based on the estimated optimism of user u_i to derive the expected salary for the job p_j:

$$a_{i,j}^{RAvg} = \sigma'(\hat{m}_i^{RAvg} \cdot w_j^{max} + (1 - \hat{m}_i^{RAvg}) w_j^{min})$$

where

$$\hat{m}_i^{RAvg} = 2 \cdot \sigma'(Avg_{A_{i,j}=1} w_j^{max} - w_j^{min}) - 1$$

In the above equations, we use a sigmoid function, σ', which normalizes the input variable by its average over i, i.e., $\sigma'(x_i) = \sigma(\frac{x_i}{(1/|U|)\Sigma_{i'} x_{i'}})$. Note that if the input variable x_i across all baseline methods is always positive, the function $\sigma'(x_i)$ is bounded between 0.5 and 1 (consequently, $0 \leq 2 \cdot \sigma'(x_i) - 1 \leq 1$). On the other hand, if input variable x_i across all baseline methods is not always positive, the function $\sigma'(x_i)$ is bounded between 0 and 1.

- **Salary-based (Sal-A):** This method predicts based on the difference between the average of offer salary upper and lower bounds of job p_j and the reserved salary of u_i derived by averaging the salaries of the applied jobs:

$$a_{i,j}^{Avg} = \sigma'(\frac{1}{2}(w_j^{min} + w_j^{max}) - \hat{v_i}^{Avg})$$

where

$$\hat{v}_i^{Avg} = Avg_{\{A_{i,j}=1\}}(w_j^{min} + w_j^{max})/2$$

- **Salary-based (Sal-M):** This method is similar to Sal-A except a different reserved salary definition.

$$a_{i,j}^{Min} = \sigma'(\frac{1}{2}(w_j^{min} + w_j^{max}) - \hat{v_i}^{Min})$$

where

$$\hat{v}_i^{Min} = Min_{\{A_{i,j}=1\}}(w_j^{min} + w_j^{max})/2$$

- **Topic-based (NMF):** This is a NMF-based model with K latent factors.

$$a_{i,j}^{NMF} = \hat{\mathbf{y}}_\mathbf{i}^{NMF} \cdot \hat{\mathbf{z}}_\mathbf{j}^{NMF}$$

- **Topic-based (LDA):** This is a LDA based model with K topics.

$$a_{ij}^{LDA} = \hat{\mathbf{y}}_\mathbf{i}^{LDA} \cdot \hat{\mathbf{z}}_\mathbf{j}^{LDA}$$

- **User Effort and Job Visibility-based (EV):**

$$a_{ij}^{Pop} = \hat{q_i}^{Pop-q} \cdot \hat{r_j}^{Pop-r}$$

where q_i^{Pop-q} estimates the effort of user u_i by the total number of applications made by u_i, and r_j^{Pop-r} estimates the job visibility of job p_j as the number of applications on p_j. That is:

$$\hat{q}_i^{Pop-q} = 2 \cdot \sigma'(\Sigma_{p_j \in P} A_{ij}) - 1$$

$$\hat{r}_j^{Pop-r} = 2 \cdot \sigma'(\Sigma_{u_i \in U} A_{ij}) - 1$$

Table 1. Application prediction AUCPRC results (real dataset)

SJD						WJD				
Without topics						Without topics				
	Opt	Sal-A	Sal-M	EV	PLM(SA)	Opt	Sal-A	Sal-M	EV	PLM(SA)
	0.167	0.151	0.174	0.464	0.482	0.167	0.155	0.176	0.464	0.474
With topics						With topics				
K	NMF	LDA	PLM(ST)	PLM(TA)	PLM	NMF	LDA	PLM(ST)	PLM(TA)	PLM
3	0.425	0.474	0.486	0.595	**0.623**	0.580	0.519	0.465	0.624	**0.640**
5	0.560	0.494	0.571	0.664	**0.686**	0.671	0.545	0.568	0.690	**0.702**
10	0.673	0.495	0.705	0.757	**0.771**	0.752	0.629	0.712	0.774	**0.779**
15	0.720	0.481	0.760	0.794	**0.806**	0.787	0.664	0.770	0.809	**0.813**
20	0.758	0.466	0.796	0.817	**0.829**	0.808	0.700	0.799	0.827	**0.831**
25	0.775	0.459	0.820	0.835	**0.845**	0.825	0.709	0.822	0.841	**0.845**
30	0.788	0.438	0.836	0.846	**0.855**	0.838	0.726	0.836	0.851	**0.855**

4.3 Application Prediction Results

We conduct 5-fold cross validation in which 20% of positive and negative samples are withheld for testing, and the remaining 80% are used for model training. We measure the prediction results by Precision@N and Recall@N at different N so as to report the Area Under the Precision-Recall curve (AUCPRC).

Results. The average AUCPRC results over the 5-fold experiments are shown in Table 1. For the SJD dataset, PLM outperforms all other models across different number of topics, and PLM (TA) yields the second best results. NMF yields the best result among the non-PLM models. LDA performance does not increase anymore beyond $K = 10$. In general, topic-aware models outperform all non topic-aware ones, including PLM(SA) (the best non topic-aware model). This suggests that application prediction is less accurate without knowing the user's and job's topic. PLM, PLM(ST), PLM(TA), and NMF improves their AUCPRC as K increases. We however witness a diminishing improvement as K increases. For example, PLM improves by 0.077 from $K = 5$ to $K = 10$, but only 0.01 from $K = 25$ to $K = 30$. $K = 25$ is then used in subsequent analysis.

Similarly, for the WJD dataset, PLM outperforms all other models across different numbers of topics. Again, NMF yields the second best results. All topic-aware models beat all non topic-aware models and the performance results of all topic-aware models improve as K increases. We also observe the improvement diminishing as K increases.

5 Labor Market Analysis Using PLM

In this section we demonstrate how PLM model is used to compare the SJD and WJD labor markets by the learned latent variables, and to analyse job seekers of different gender and age groups across different market segments. For the job seeker analysis, only SJD dataset is used as it covers jobs across wider sectors than the WJD dataset. Furthermore, based on the results of latent variable

recovery experiment by using synthetic data (not shown here because of page limitation), PLM also performs significantly better than any other alternative baselines in recovering the latent variables. Therefore, we can confidently utilize the learned latent variables from PLM to analyze the labor market. In the following, we use PLM with 25 topics (i.e., $K = 25$) which yields fairly accurate application prediction results in Sect. 4.

5.1 Market Analysis and Comparison

One of the key objectives of PLM is to learn the latent variables of users and jobs. These include the reserved salary (v_i), optimism (m_i) and effort-level (q_i) of each user u_i, and the visibility (r_j) of job p_j. We now compare these variables between WJD and SJD markets. Note that topics are not included in this comparison because they are separately learned for the two datasets. As the two markets adopt different currencies and the reserved salaries of WJD are generally much lower than that of SJD, we focus on comparing the reserved salary distributions of the two markets relative to their average market offer salary. Therefore, we first scale the reserved salaries by the mean of the maximum offer salary of the market ($\frac{v_i}{Avg_{p_j \in P} w_j^{max}}$). Maximum offer salary is used here instead of mid offer salary since both markets have more applicants with high optimism.

Figure 2 shows the boxplots of these variables. The triangle symbol (▲) indicates the average value. The figure shows that the SJD labor market observes higher normalized reserved salary values than the WJD labor market. SJD also has a more balanced distribution than WJD which has a high concentration of users with low reserved salaries.

For optimism, SJD observes a slightly higher average optimism among its users than WJD. On the other hand, users from WJD put up higher effort level than users from SJD. Above observations together reveal that WJD is a tougher labor market than SJD. Finally, we could not find any obvious differences in job visibility distribution between the two markets.

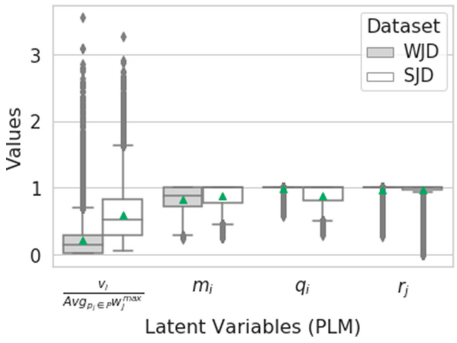

Fig. 2. Distribution of latent variables

5.2 Topic-Specific Labor Segments

We now analyse the topics of SJD dataset being learned by PLM to determine its major labor segments. Each market segment consists of a group of users interested in a cluster of jobs sharing the same topic. These topic-specific labor segments are "soft" as they are not defined by any observable market variable.

For each topic l, we include a user u_i under the topic l labor segment if $cosine(\mathbf{y}_i, \mathbf{t}_l) \geq 0.5$. Similarly, we include a job p_j under the labor segment if $cosine(\mathbf{z}_j, \mathbf{t}_l) \geq 0.5$. Here we use the original definition of PLM, where $cosine(\mathbf{z}_j, \mathbf{t}_l)$ is the degree of matching between job p_j and the topic l ($\mathbf{t}_l$ is a one-hot K dimensional vector for topic l). With this rule, each user or job can also belong to exactly one topic-specific labor segment. We use U_l and P_l to denote the users and jobs in this topic-l labor segment respectively.

While we have $K = 25$ topics, we focus on a few more popular topic-specific labor segments with number of users and jobs $|U_l|+|P_l| > 2000$. Table 2 show the top 13 topic-specific labor segments and their representative jobs. We manually assign for each topic a label to summarize jobs in that segment. Table 2 shows that the major topic-specific labor segments have clear topics. Across these 13 major labor segments, Trading & Investment is the only segment having more users than jobs, i.e., $|U_l| > |P_l|$. The other market segments have a distinctive shortage of supply of manpower as there are more available jobs than suitable applicants who can fill them.

5.3 Labor Segment Level User Analysis

In this section, we analyse reserved salary, optimism and effort-level of users in each of the major topic-specific labor segments of SJD labor market. Figure 3 shows distributions of these variables. The median and average values of each distribution are indicated by line (−) and triangle (▲) symbols respectively.

The distributions of optimism and effort-level for all these labor segments are skewed towards high values. This suggests that users have high optimism and high effort. The Financial Management and PM+Design & Architecture labor segments have the most optimistic users, and the Clerical labor segment has the least optimistic users. Effort-level wise, users from the Finance management, Accounting and PM+Design & Architecture segments seem to put in highest efforts in job seeking. On the other hand, users from the Education+Programming segment seems to put in less effort.

The distribution of reserved salary for all these labor segments are skewed towards lower values. It means the majority of people expect lower reserved salaries. Only few people expect very high reserved salaries across different labor segments. The clerical segment has the lowest median and mean reserved salary, while Education + Programming, Information Technology, Project Management + Design & Architecture segments have higher median and mean reserved salaries.

Table 2. Major topic-specific labor segments

Topics (l)	Top dominant jobs	$\|U_l\|$	$\|P_l\|$
Clerical	Admin Assistant, Admin Clerk, Receptionist (General), Admin Executive, Administrator, Customer Service Officer, Call Centre Agent, Sales Coordinator	2634	4554
Secretarial & Personal Assistant (PA)	Admin Assistant, Human Resource Executive, Secretary, Human Resource & Admin Officer, Assistant, Personal, Human Resource Asst, Receptionist (General), Admin Exec'	2234	4103
Financial Management	Accountant, Finance Manager, Assistant Finance Manager, Accounts Executive, Analyst, Financial, Controller, Financial, Senior Accountant (General), Accounting Manager	1717	3507
Marketing & Public Relation (PR)	Manager, Marketing, Marketing Executive, Brand Manager, Assistant Marketing Manager, Regional Marketing Manager, Marketing Communications Manager, Marketing Communications Exec, Senior Marketing Exec	1631	2563
Accounting	Accounts Executive, Accounts Assistant, Accountant, Account Executive, Finance Executive, Account Assistant, Accounts Officer, Accountant, Assistant	1152	2939
Human Resource (HR)	HR Executive, HR Manager, HR Business Partner, HR & Admin Officer, Senior HR Executive, HR Assistant, HR & Admin Manager, HR Assistant Manager	1268	1826
Research & Lab	Research Assistant, Research Officer, Clinical research coord, Laboratory Technician, Medical Technologist, Researcher, Chemist, Laboratory Assistant	1216	1645
Project Management + Design & Architecture	IT Project Manager, IT Manager, Designer, Graphic, Project Manager, Svc Delivery Manager, Architectural Designer, Designer, Interior, Architectural Asst	1100	1597
Trading & Investment	Analyst, Associate, Trader, Mgmt Trainee, Invt Analyst, Risk Analyst, Commodities Trader, Business Analyst	1629	975
Supply Chain	Resident Engineer, Purchasing Executive, Purchaser, Buyer, Marine Superintendent, Logistics Executive, Technical Superintendent, Procurement Executive	1001	1572
Business Software	Business Analyst, Application Support Analyst, Information Technology Business Analyst, Associate, Senior Business Analyst, Analyst, System Analyst, Engineer, Software	754	1720
Information Technology	System Administrator, Art Director, IS Engineer, IT Project Manager, IT Manager, Desktop Support Engineer, Compliance Officer, Analyst	844	1512
Education + Programming	Teacher (Int School), Java Dev, Sr Engineer, Software, Sr Java Developer, Project Manager, Engineer, Software, Application Developer, Commercial School Teacher	839	1200

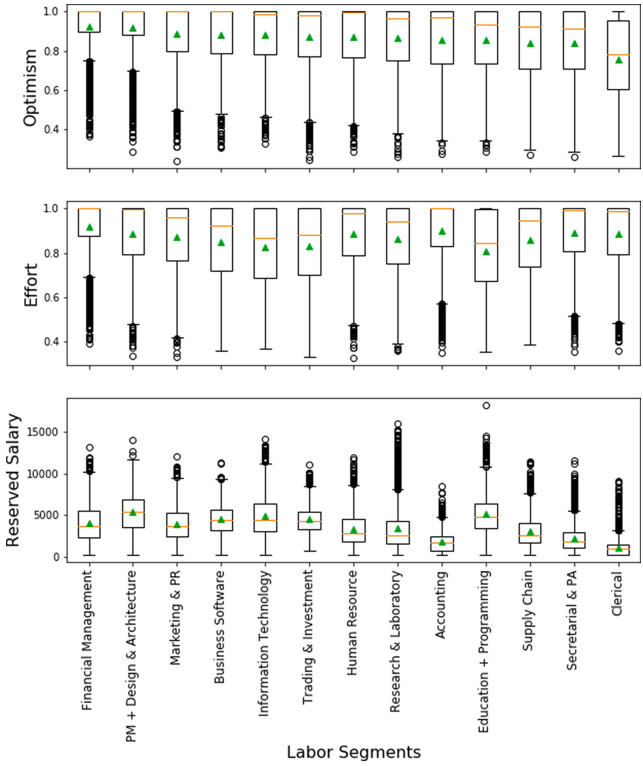

Fig. 3. User variables of labor segments

5.4 User Analysis by Gender and Age

User Analysis by Gender. Next, we study gender differences in the major labor segments as shown in Fig. 4. Female-male applicant proportions across all labor segments is almost equal (49 : 51). This proportion is represented by dotted black line. The bar chart indicates the percentage of female applicants in each labor segment (the rest is filled by male applicants). The labor segments are sorted by increasing female dominance. Labor segments such as Clerical, Secretarial & PA, Accounting and Human Resource are more preferred by female applicants. In contrast, PM + Design & Architecture, Information Technology, and Trading & Investment are dominated by male applicants.

According to the male-female median reserved salary ratios $\frac{v^{male}}{v^{female}}$ indicated by the blue squares, male users enjoy higher median reserved salary than females across all the major labor segments (except for Accounting segment). In particular, for the clerical labor segment which females dominate, male users have overall reserved salary more than 50% higher than that of female users. Female applicants appear to expect less reserved salary than male applicants.

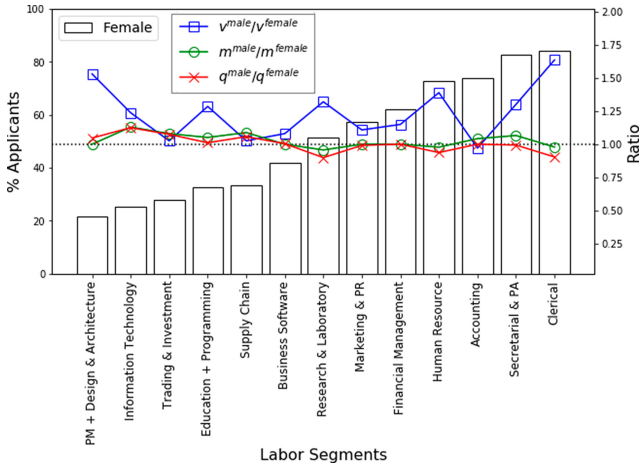

Fig. 4. Gender specific latent variables in major labor segments

Moreover, we also observe that male users have higher optimism (indicated by green circles) and effort (indicated by red crosses) in labor segments such as IT and Supply Chain. On the other hand, female users have higher optimism and effort in other labor segments such as Research & Laboratory and HR. However, the gaps in terms of optimism and effort values between female and male users are not as big as reserved salary.

User Analysis by Age. We now examine the age differences in the major labor segments. Specifically we only focus on the profile differences between users below 30 and above 30 in Fig. 5. The below-30 group accounts for 38% of all applicants as indicated by the black dotted line. Trading & investment, research & laboratory, marketing & PR, clerical, and secretarial & PA labor segments are preferred by younger applicants, or they may be more suited for younger applicants. PM + Design & Architecture, Education + Programming, and several others segments are preferred by older applicants.

We observe that median reserved salary for older users (indicated by blue squares) is generally higher than that of younger applicants across all the major labor segments (except in Accounting and Clerical segments where median reserved salaries are approximately equal). The above observations are reasonable as older applicants usually expect higher salaries. Accounting and clerical segments are likely to be age-neutral.

We also observe that older applicants have higher optimism (shown as green circles) and effort (shown as red crosses) in IT and Supply Chain. On the other hand, younger applicants have higher optimism and effort in other labor segments such as Research & Laboratory.

While the above analysis only involves gender and age, similar analysis can be performed for user groups defined based on other attributes such as race, and education. This allows us to understand differences between other user groups

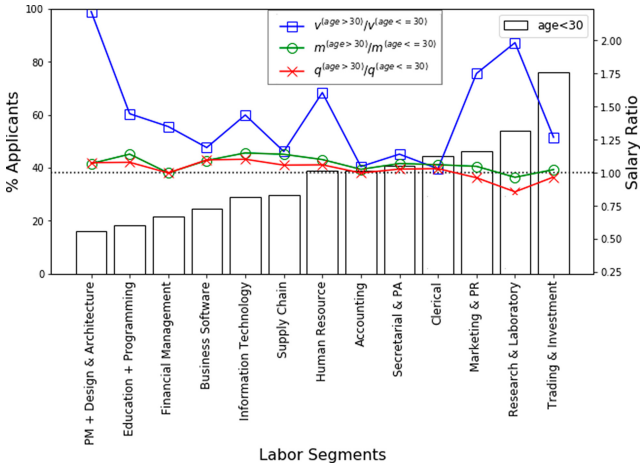

Fig. 5. Age specific latent variables in major labor segments

in the labor market or labor market segments. We shall leave these studies to future work.

6 Conclusion

We have developed a probabilistic model called PLM to study labor market directly using observed data. This model combines salary requirement, topic matching, and job accessibility are the three main criteria for users to select jobs to apply for. PLM also learns user and job factors useful for data science analysis. Our experiments show that PLM outperforms other baseline models in prediction tasks. Moreover, we also demonstrate the strength of the model in analyzing various aspects of the labor market.

The immediate applicability for the social good lies in the learned latent variables. These information can be utilized by a job seeker to compare his/her personal latent variables (e.g. reserved salary, effort, optimism) with his/her competitors'. The employers can also compare their salary competitiveness with their potential applicants' reserved salary. Furthermore, the policy maker can also utilize labor topics analysis to tackle labor shortage or even gender gap in a targeted manner (i.e. specific labor segments).

There are some limitations of this study that can be improved in future work. More advanced versions of the PLM will be developed to cope with the long tailed data distribution. The performance of the model can also be improved by considering different negative sampling strategies. We also plan to extend the model to conduct analysis at the user or job level to generate even more fine-grained insights. Moreover, the learned latent variables from PLM can be utilized and aligned into labor economics problems such as labor supply, demand, elasticity,

and the equilibrium state of each market. PLM can also be extended to model the labor segments more accurately using textual features of job descriptions.

Acknowledgment. This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Arulampalam, W., Booth, A.L., Taylor, M.P.: Unemployment persistence. Oxford Econ. Pap. **52**(1), 24–50 (2000)
2. Bauder, H.: Labor Movement: How Migration Regulates Labor Markets. Oxford University Press, Oxford (2006)
3. Becker, G.S.: Economic Theory. Routledge, Abingdon (2017)
4. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. Polit. Anal. **20**(3), 351–368 (2012)
5. Borjas, G.J.: The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. Q. J. Econ. **118**(4), 1335–1374 (2003)
6. Cain, G.G.: The challenge of segmented labor market theories to orthodox theory: a survey. J. Econ. Lit. **14**(4), 1215–1257 (1976)
7. Cappelli, P.H., Gaps, S., Shortages, S., Mismatches, S.: Evidence and arguments for the United States. ILR Rev. **68**(2), 251–290 (2015)
8. Dickens, W.T., Lang, K.: The reemergence of segmented labor market theory. Am. Econ. Rev. **78**(2), 129–134 (1988)
9. Guzi, M., Kahanec, M., Kurekova, L.M.: How immigration grease is affected by economic, institutional, and policy contexts: evidence from EU labor markets. Kyklos **71**(2), 213–243 (2018)
10. Hall, R.E.: Employment efficiency and sticky wages: evidence from flows in the labor market. Rev. Econ. Stat. **87**(3), 397–407 (2005)
11. Heim, B.T.: The incredible shrinking elasticities: married female labor supply, 1978–2002. J. Hum. Resour. **42**(4), 881–918 (2007)
12. Horton, J.J., Chilton, L.B.: The labor economics of paid crowdsourcing. In: The 11th ACM Conference on Electronic Commerce, pp. 209–218 (2010)
13. Hu, L., Chen, Y.: A short-term intervention for long-term fairness in the labor market. In: The 2018 World Wide Web Conference (2018)
14. Joinson, A.N., Woodley, A., Reips, U.D.: Personalization, authentication and self-disclosure in self-administered internet surveys. Comput. Hum. Behav. **23**(1), 275–285 (2007)
15. Kaufman, B., Hotchkiss, J.: The Economics of Labor Markets. Harcourt College Publishers (1705)
16. Kenthapadi, K., Ambler, S., Zhang, L., Agarwal, D.: Bringing salary transparency to the world: computing robust compensation insights via linkedin salary. In: ACM Conference on Information and Knowledge Management (CIKM) (2017)
17. Kokkodis, M.: Reputation deflation through dynamic expertise assessment in online labor markets. In: The 2019 World Wide Web Conference (2019)

18. Kokkodis, M., Papadimitriou, P., Ipeirotis, P.G.: Hiring behavior models for online labor markets. In: ACM International Conference on Web Search and Data Mining (2015)
19. Machin, S., Manning, A.: A test of competitive labor market theory: the wage structure among care assistants in the south of England. ILR Rev. **57**(3), 371–385 (2004)
20. Maltseva, A.V., Makhnytkina, O.V., Shilkina, N.E., Soshnev, A.N., Evseev, E.A.: A multilevel index model of labor market dysfunction. In: International Conference on Engineering and MIS (2019)
21. Mortensen, D.T., Pissarides, C.A.: Job creation and job destruction in the theory of unemployment. Rev. Econ. Stud. **61**(3), 397–415 (1994)
22. Pissarides, C.A.: Equilibrium in the labor market with search frictions. Am. Econ. Rev. **101**(4), 1092–1105 (2011)
23. Pitts, R.K.: Spatio-temporal labor market analytics: Building a national web-based system. In: 1st International Conference and Exhibition on Computing for Geospatial Research & Application (2010)
24. Waring, P., Vas, C., Bali, A.S.: The challenges of state intervention in Singapore's youth labour market. Equality Diversity Inclusion: Int. J. **37**, 138–150 (2018)
25. Wartena, C.: Distributional similarity of words with different frequencies. In: Dutch-Belgian Information Retrieval Workshop (2013)