IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 17, 2021, accepted March 8, 2021, date of publication March 17, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066204

On Aggregating Salaries of Occupations From Job Post and Review Data

CHIH-CHIEH HUNG[1] AND EE-PENG LIM[2]
[1]Department of Management Information System, National Chung Hsing University, Taichung 402, Taiwan
[2]Living Analytics Research Centre, Singapore Management University, Singapore 188065

Corresponding author: Ee-Peng Lim (eplim@smu.edu.sg)

This work was supported by the National Research Foundation, Singapore through the International Research Centres in Singapore Funding Initiative.

ABSTRACT The popularity of job websites has significantly changed the way people learn about different occupations. Among the insights offered by these websites are the statistics of occupation salaries which are useful information for job seekers, career coaches, graduating students, and labor related government agencies. Such statistics include the distribution of job salaries of each occupation, such as average or quantiles. However, significant variability in salary (and review salary) can be found among jobs of the same occupation as we gather job post and review data from job websites. Such variability shows the existence of biases, including salary competitiveness in job posts and salary inflation in job reviews. Based on the observation, we aim at developing an approach to derive occupation salary for a job market, named unbiased salary, by aggregating offer salaries from job posts and review salaries from review data and at the same time removing their biases. To achieve this goal, we proposed COC-model to learn unbiased salaries of occupations, competitiveness of companies and inflation of companies efficiently. COC here is an abbreviation of ''Company, Occupation, Company'', which represents two different connections between companies and occupations from job posting site and job review site. COC-model represents the dependency of salary information between companies and occupations in job post data and job review data. It begins with defining three latent variables, say competitiveness, inflation, and unbiased salary, based on their dependencies. Instead of computing these variables iteratively, we formulate the interaction among these three latent variables into a matrix form so that these values could be then efficiently learned in a unified way by a series of matrix operations. Extensive experiments are conducted, including empirical studies about competitiveness and inflation of companies using real dataset and performance testing by synthetic dataset. The experimental results show that COC-model can not only derive unbiased salaries effectively but also help us to understand latent biases in job post and job review data.

INDEX TERMS Salary aggregation, unbiased occupation salary, bias modeling.

I. INTRODUCTION

The popularity of job websites including LinkedIn, Glassdoor, and Indeed has significantly changed the way job seekers find employers, and vice versa. These websites also provide company and user-contributed data from which useful insights about jobs, occupations and companies can be derived. Among these insights are the occupation salaries which could be a key factor job seekers use to decide jobs to apply. The salary information also serves as an important reference to employers who want to attract competent job applicants in order to stay competitive in the job market.

With job and salary related data readily becoming available, salary analysis has been a popular research topic in recent years. One direction of this research is to study important factors that influence the amount of salary. Xu *et al.* proposed a skill popularity topic model which incorporates different job attributes and latent connections between job skills by given job posts [1]. Besides, there are several studies focus on analysing salary discrepancy due to gender. Most of these studies adopt a survey approach while some of them consider a data driven approach. Chamberlain demonstrated that female employees are paid less compared to their

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

male counterpart using the salary data reported by Glassdoor users [2]. Salary prediction is also one of the most popular research topics. Lazar *et al.* used survey data to build a predictive model for annual salary [3]. Jackman *et al.* proposed a framework based on various models, such as regression, artificial neural networks, and random forests, to predict salary by text features extracting from job descriptions [4]. To determine benchmarking salary for each job position, Meng *et al.* modeled the problem as a matrix completion task and developed a matrix factorization model which can predict missing salary information by integrating multiple confounding factors [5]. To provide insights on compensation distribution to job seekers, Chen *et al.* proposed a framework that overcomes issues of data sparseness: the company embeddings are firstly derived, the pairwise similarities of these embeddings are then computed and the these similarities are incorporated as peer company groups into the proposed Bayesian statistical model to predict insights at the company level [6].

Different from the studies in the literature, we differentiate jobs from occupations in this paper as an **occupation** is a collection of job positions that share similar characteristics. For example, *software developer* is an occupation which consists of various job positions including *software analyst*, *software engineer*, and *software developer*, etc. To organize jobs in its job market, each geopolitical region would normally maintain a dictionary of occupations for reference. For example, a comprehensive dictionary of 974 occupations has been defined and maintained by O*Net (www.onetcenter.org/), a program supported by the U.S. Department of Labor/Employment and the Training Administration (USDOL/ETA). We assume that when job seekers search for jobs, they tend to look for jobs associated with certain occupations matching their expertise and preference. As such, understanding the salary distribution at the occupation level in the job market is much more useful than the salary of a specific job position.

It is not a trivial task to obtain occupation level salaries. A conventional approach is to conduct surveys on human resource experts across industries. This approach is usually costly and laborious. Moreover, it may take such an approach too long to detect changes to job market for career coaching and job recommendation applications. Conducting data science analysis on observed job and occupation salary data would therefore be a more promising alternative. Compared with conducting surveys, the data science approach would potentially generate fast turnaround salary results, as long as the appropriate data is available. In this paper, we use both job post data and job review data in our occupation salary analysis. Job post data refers to the salary information of job positions posted by employers, e.g. Indeed (www.indeed.com), whereas job review data refers to those contributed by employees e.g., Glassdoor (www.glassdoor.com). Interestingly, job post data and job review data usually reveal significant variability in occupation level salaries, which shows the existence of biases.

TABLE 1. Example of designers' salaries.

Company Name	Industry	Offer Salary (Job Post Data)	Review Salary (Job Review Data)
Ubisoft	Video Games	7,000	3,513
Koei Temco	Video Games	4,000	3,600
Convertium	Marketing	3,000	3,000
SPH Magazines	Publishing	2,850	3,021

This could be best understood by the example as shown in Table 1. This table shows the average offer and review salaries of Designer in our real dataset. In job post data, Ubisoft offered much more competitive salary for Designer, say 7,000 dollars, than other companies. The level of offer salaries could be affected by the bias caused by company-specific factors, such as industry types, company situation, and so on. On the other hand, there are certain gaps between offer salaries and review salaries for Designers. Especially, Ubisoft has the biggest difference between the offer salary (7,000 dollars) and the review salary (3,513 dollars). Such gaps may be caused by the bias from under-reported salaries in job review data. These biases should be taken into account carefully; otherwise, the standard statistics, such as average and variance, would become less meaningful when we determine occupation-level salaries.

In this paper, a novel approach *COC-model* is proposed, to determine occupation salaries for a job market by aggregating the offer salaries in job post data and review salaries in job review data of an occupation by removing their biases. COC here is an abbreviation of ''Company, Occupation, Company'', which represents the core idea that COC-model could represent and solve the dependency of salary information between companies and occupations in job post data and job review data. In a nutshell, COC-model firstly defines three latent vectors, *competitiveness*, *inflation*, and *unbiased salaries*, and formulates the dependencies between these vectors, so as to define a computational model to simultaneously compute occupation's unbiased salary, company's competitiveness and inflation.

Specifically, we firstly model the biases from the offer salaries from job post data and that of the review salaries in job review data among companies as *competitiveness* and *inflation*, respectively. The competitiveness of a company is defined as the tendency of offering a higher salary for a position in the occupation than other companies.[1] The inflation of a company is defined as the tendency that employees in a company report their salaries higher than that of others corresponding to the same occupation in other companies. The unbiased salary of an occupation, the central tendency of salaries of this occupation, is then defined as the salary after removing the biases of competitiveness and inflation. Conversely, competitiveness and inflation can also be represented in the form of unbiased salary. Through such

[1]The reader should note that the competitiveness of a company here refers to the recruiting competitiveness, observed from offer salaries, which differs from company competition referring to the ability and performance of a firm.

depedencies, these three vectors could be derived iteratively until convergence is achieved; however, the iterative computation would usually suffer from poor efficiency. To overcome this issue, COC-model represents the dependencies into a matrix form. The learning of unbiased salary, competitiveness, and inflation is then formulated as a problem of solving a system of linear equations so that these three vectors could be learnt at the same time.

With the unbiased occupation salaries, we can conduct data science studies of job market answering meaningful research questions. The first research question is how much a job salary deviates from the corresponding occupation salary. This helps us understand the elasticity of salaries across an occupation. The second research question is about the types of companies with very high and low competitiveness. This helps us discover companies that are more or less aggressive in attracting talents by salary. Finally, we would like to analyse user communities that inflate salaries. We also conduct an experiment using unbiased salary, competitiveness and inflation to show these values are important features for salary prediction. We further conduct extensive experiments using synthetic dataset to evaluate the accuracy of COC model and other baseline models in recovering unbiased salaries of occupations, competitiveness and inflation of companies.

To our best knowledge, this paper is the first work that addresses the existence of the biases in job post data and job review data. By modeling two biases competitiveness and inflation, we can explain the discrepancy between occupational salaries from job post and review data and further define unbiased salary, a central tendency of occupation salaries, by removing these biases. Moreover, an efficient computational approach is proposed to learn unbiased salaries, competitiveness and inflation at the same time. Our model is the first of its kind that adopts a data driven approach which can generate results more quickly than the traditional survey approach. Last but not least, the derived unbiased salaries, competitiveness and inflation could be used to answer meaningful data science questions. We believe these contributions provide sufficient innovations in the field of salary data analysis.

The rest of the paper is organized as follows. Section II covers the past relevant research works. Section III defines the occupation level salary problem, describes the data crawled from two job websites to be used in our analysis, and presents the set of research questions to be studied. We propose our model in Section IV. Our experiments and results are given in Section V. Finally, we conclude in Section VI.

II. RELATED WORKS

A. JOB ATTRIBUTE MODELING

With many job websites offering online recruitment and social network services for professionals, e.g., Indeed and Linkedin, more and more data sources are available for researchers to take the data-driven approach to analyse salaries. Using job posts with job descriptions, job levels, required skills and other attributes, Xu proposed a skill popularity based topic model to generate a job skill network extracted from a corpus of job posts. This topic model incorporates different job attributes and latent connections between skills. This model is then used to rank skills based on their multi-faceted popularity which turns out to be associated with high-pay jobs [1]. Using historical interview data, Lin proposed a latent variable model for assessing job interviews and learning representative perspectives of different job interview processes [7]. Based on Glassdoor data, Luo proposed a framework to analyse textual content in the data to reveal the relation between employee satisfaction and company performance. The result showed that the overall employee satisfaction and corporate performance are positively correlated suggesting that it is important to address employee satisfaction for improving performance [8]. There are several research works on analysing salary discrepancy due to gender. Most of these works adopt a survey approach but some of them consider a data driven approach. Chamberlain demonstrated that female employees lag behind their male counterparts using salaries reported by Glassdoor users [2]. To compute the compatibility between employees and their organizations, authors in [9] proposed a neural network approach for modeling the compatibility for person-organization fit with talent turnover and job performance. Organizational Structure-aware Convolutional Neural Network is firstly proposed to extract organization-aware compatibility features for person-organization fit and recurrent neural network with attention mechanism is used to model temporal information.

B. SALARY PREDICTION

Salary prediction is also a popular research topic. Past research on salary prediction often used survey data from third parties and domain knowledge of experts. Using salary survey data, Lazar developed a prediction model using support vector machine to determine whether a person has an annual salary over $50K [3]. Jackman and Reid proposed to use maximum-likelihood regression, lasso regression, artificial neural networks and random forests to predict job salaries using text-only features in job descriptions [4]. The past salary prediction research focuses on salaries corresponding to job positions instead of occupations. Using an employment website which includes company and salary information, Lin proposed a collaborative topic regression model to predict job salaries by modeling both textual (e.g., reviews) and numerical information (e.g., salaries and ratings) collaboratively [10]. Meng modeled salaries in a job-company matrix with missing salary values. Meng then proposed a matrix-factorization method to predict the missing salaries [5]. Kenthapadi and others proposed a system to analyse salary insights of job titles for different geographic regions, companies of different sizes, and for job seekers of different education background using the knowledge graph internally constructed by LinkedIn as well as the large amount of LinkedIn user data [11]. Authors in [6] proposed a two-step

framework to analyze the company transition data of Linked members to derive company embeddings. The pairwise similarities between companies can be then computed based on these embeddings. Then the proposed Bayesian statistical model is used to predict insights at the company level.

For job search, career planning career counselling applications, it is more useful to group jobs of the same kind as an occupation. Our work therefore focuses on determining occupation salaries. Moreover, occupation level salaries should be derived by some aggregation of job salaries, which has not been studied in the past. Finally, our work addresses biases in salary reports and posted salaries shared by web users and companies.

C. CAREER DEVELOPMENT

The study of career development has become more and more important during a time of rising competition. From human resource perspectives, understanding the important factors of job changes are very helpful for making strategies for manpower management by predicting employees' job change occasion. Taking the experience data of individuals, authors in [12] proposed an approach based on survival analysis to model the talent career paths by taking turnover and career progression into accounts. The proposed approaches formulated the prediction of survival status at a sequence of time intervals as a multi-task learning problem by considering the prediction at each time interval as a task. Considering the career mobility at an individual level, authors in [13] discovered the correlation between one's job change occasions and check-in records. By aggregating one's work experience in Linkedin dataset and check-in records in Foursquare dataset, a prediction model is then proposed to predict whether or not an employee will change her job in the next month. This paper concluded that the job change occasion can be predicted based on the career mobility and daily activity patterns at the individual level. Authors in [14] proposed a hierarchical career-path-aware neural network for learning individual-level job mobility which is helpful to predict who will be the next employer of an employee and how long this employee will work for this employer. This paper also found out an individual's job mobility is related to job duration, company types, and so on. To understand talent flow at an organization level, authors in [15] proposed a predictive model to predict talent flow from job transition data in Linkedin dataset. A time-aware job transition tensor is first constructed and then the dynamic latent factor is designed based Evolving Tensor Factorization model for predicting the future talent flows. Moreover, the influence of previous talent flows and global market are considered for modeling the evolving nature of each company.

D. RATING AGGREGATION

Salary aggregation is closely related to rating score normalization or aggregation. One simple idea is to normalize all the scores given by a reviewer by determining the score range of the reviewer. By normalizing scores from every reviewers, one obtains comparable scores which can be aggregated by the average function [16]. This aggregation approach however assumes that each item must be rated by every reviewer. Liaw, Lim and Wang overcame the above limitation by modeling users of different degrees of leniency which complicate rating outcome. They proposed a novel model to aggregate the ratings to derive fairer ratings of reviewed items [17]. In this work, a bipartite graph of user and item nodes as well as rating edges between users and items is constructed. Our proposed model is inspired by this graph-based approach which describes a pair-wise relationship. Our model tends to model two pair-wise relationships with more factors included.

As salaries reviewed by users and stated in job posts involve different types of biases among users and companies contributing the data, the above aggregation methods cannot be directly applied. New aggregation models should therefore be designed to handle the different biases.

TABLE 2. Symbols.

Symbol	Meaning
$c_i \in C$	company
$o_j \in O$	occupation
$O^p(c_i)$	occupations with jobs posted by company c_i
$O^r(c_i)$	occupations with jobs reviewed by users from c_i
$C^r(o_j)$	companies with users review salaries of o_j
$C^p(o_j)$	companies posting salaries of o_j
$r_{i,j}$	average review salaries of o_j in c_i
$n^r_{i,j}$	number of review salaries of o_j from c_i
$p_{i,j}$	average posted salary of o_j by c_i
$n^p_{i,j}$	number of offer salaries of o_j by c_i
κ_i	competitiveness of c_i
e_i	inflation of c_i
b_j	unbiased salary of o_j

III. PROBLEM FORMULATION

Let the set of companies be $C = \{c_1, c_2, \ldots, c_m\}$ and the set of occupations be $O = \{o_1, o_2, \ldots, o_n\}$. Let $C^p(o_j)$ and $O^p(c_i)$ denote the set of companies posting jobs of occupation o_j, and the set of occupations with jobs posted by company c_i respectively. We also denote the companies with users (usually the employees of these companies) reviewing salaries of jobs of occupation o_j by $C^r(o_j)$, the set of occupations with jobs reviewed by users from company c_i by $O^r(c_i)$.

We denote the average review salary of an occupation o_j reported by users from c_i as $r_{i,j}$, and the number of reports from c_i reviewing the salaries of o_j as $n^r_{i,j}$. Let the average salary of o_j that c_i offers in job posts for o_j be $p_{i,j}$, and the number of job posts of o_j that c_i offers is denoted as $n^p_{i,j}$. Note that $r_{i,j}$ may not be identical to $p_{i,j}$ as the former is given by users from c_i while the latter is given by c_i. The users and company have different biases. Note that there may be some occupations without any review. In the case of $n^r_{i,j} = 0$, $r_{i,j}$ is undefined. Similarly, $p_{i,j}$ is also undefined when $n^p_{i,j} = 0$.

In this paper, *unbiased salary*, *competitiveness*, *inflation* are three kinds of latent variables. Assume that unbiased salary of each occupation is known. The competitiveness of a company refers to how much the company adjusts its offer

$$M_o = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccccc} o_1 & o_2 & o_3 & o_4 & o_5 \end{array} \\ \begin{pmatrix} 7 & - & - & 5 & 5 \\ 4 & 3 & 3 & - & - \\ 4 & - & 3 & - & - \\ - & 6 & 6 & 5 & 5 \\ - & 6 & 6 & 5 & 5 \end{pmatrix} \end{array} \quad M_r = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccccc} o_1 & o_2 & o_3 & o_4 & o_5 \end{array} \\ \begin{pmatrix} 5 & - & 3 & - & - \\ 5 & 4 & - & - & - \\ 8 & - & - & 7 & 7 \\ - & 5 & 5 & 4 & 4 \\ - & 5 & 5 & 4 & 4 \end{pmatrix} \end{array}$$

FIGURE 1. An illustrative example of offer salaries and review salaries.

salaries above the unbiased salaries in order to attract talents. Similarly, the inflation of a company can be measured by how much the users from this company adjust the review salaries above the unbiased salaries. Our goal is thus to learn these latent variables from the observed offer and review salaries by defining models that capture their interaction leading to the generation of observed data. We now define the problem in this paper as follows:

Unbiased Salary Aggregation Problem: Given $\{p_{i,j}\}$, $\{n^p_{i,j}\}$, $\{r_{i,j}\}$ and $\{n^r_{i,j}\}$ for all $c_i \in C$ and $o_j \in O$ such that $p_{i,j}$'s and $r_{i,j}$'s are known, we want to learn the unbiased salaries b_j's for every occupation $o_j \in O$, the competitiveness κ_i and inflation e_i of every company $c_i \in C$ such that the difference between the predicted and observed salaries, i.e., $|\hat{p}_{i,j} - p_{i,j}|$'s and $|\hat{r}_{i,j} - r_{i,j}|$'s, derived from $\{b_j\}$, $\{\kappa_i\}$ and $\{e_i\}$ is minimized.

In the above problem definition, b_j's are positive values, κ_i is positive when company c_i is competitive, negative when c_i not competitive. Similarly, e_i is positive when c_i has review salary inflated, and negative when c_i has review salary suppressed.

IV. COC: A SALARY AGGREGATION MODEL

A. OVERVIEW

In practice, the offer salaries and review salaries are observable whereas unbiased salary, competitiveness, and inflation are latent. As reflecting how much a company can offer for an occupation, offer salaries could be used to derive competitiveness of companies and unbiased salaries of occupations. Similarly, the review salaries could be used to derive inflation of companies and unbiased salaries of occupations. Obviously, the unbiased salaries are determined by both competitiveness and inflation.

To best understand the interaction among competitiveness, inflation, and unbiased salary, Figure 1 gives an illustrative example of offer salary data M_o and review salary data M_r where c_1, c_2,...,and c_5 are companies and o_1, o_2,... and o_5 are occupations. The (i,j) entry in M_o and M_r refers to the offer and review salary for o_j of c_i, respectively. Consider offer salaries M_o. The offered salary of o_1 suggests that either the salary offered by c_1 (i.e., $M_o(1, 1) = 7$) is too high or that by c_2 and c_3 are too low (i.e., $M_o(2, 1) = 4$ and $M_o(3, 1) = 4$). We can observe that salaries of o_4 and o_5 offered by c_1 tend to agree with that offered by c_4 and c_5. We can say that c_1 is likely to have the same competitiveness as c_4 and c_5 whereas c_2 and c_3 are likely to have less competitive. In this case, the unbiased salary of o_1 is expected to be higher than the average, i.e. $(7 + 4 + 4)/3 = 5$, taking companies' competitiveness into account. On the other hand, consider M_r, the review salaries. The review salaries from c_3 tend to be inflated. Firstly, the review salaries of o_4 from c_3 (i.e. $M_r(3, 4) = 7$) are much higher than the majority of the companies offering the same occupations. Moreover, we can observe that the salary of o_5 reported from employees in c_3 ($M_r(3, 5) = 7$) are much higher than the salaries offered for o_5, i.e., $M_o(1, 5) = 5$, $M_o(4, 5) = 5$, and $M_o(5, 5) = 5$). We can also observe the similar phenomenon from o_1 and o_4. Thus, employees in c_3 are very likely to over-report their salaries, i.e., c_3 has higher inflation. In this case, the unbiased salary of o_1 is expected to be lower than the average, i.e. $(5 + 5 + 8)/3 = 6$, when we take companies' inflation into account.

To derive the values of competitiveness, inflation and unbiased salaries, we firstly model the dependency of these latent variables. Due to the dependency, the value of competitiveness, inflation and unbiased should be learnt in a unified way which could be achieved by a series of matrix operations. At last, the values of these variables could be obtained by solving a linear equation.

B. MODELING RELATIONSHIP AMONG UNBIASED SALARY, COMPETITIVENESS, AND INFLATION

We consider the offer salaries from job post data first. Suppose the competitiveness of all companies (i.e., κ_i's) are known, the unbiased salary of an occupation o_j can be modeled by adjusting offer salary of each occupation by multiplying an adjustment factor and then averaging the adjusted salaries. That is, each offer salary will be adjusted upward or downward to obtain the unbiased occupation salary according to competitiveness of the companies offering o_j. Based on this principle, the **unbiased salary of an occupation** o_j is defined as:

$$\begin{aligned} b_j &= Avg_{c_i \in C^p(o_j)} p_{i,j} \cdot (1 - \alpha\kappa_i) \\ &= \frac{1}{n^p_j} \sum_{c_i \in C^p(o_j)} n^p_{i,j} \cdot p_{i,j} \cdot (1 - \alpha\kappa_i) \end{aligned} \tag{1}$$

where $n^p_j = \sum_{c_i \in C^p(o_j)} n^p_{i,j} = \sum_{c_i \in C} n^p_{i,j}$, and $0 \le \alpha \le 1$.

In Equation 1, the offer salary is adjusted by the adjustment factor $(1 - \alpha\kappa_i)$. A positive (or negative) κ_i sets this factor less than one so that the offer salary will be reduced (or increased)

to obtain the unbiased salary. The parameter α represents the weight given to the competitiveness when the unbiased salary is computed. The larger the α is, the more the company competitiveness will affect the unbiased salary of occupation. For example, consider M_o in Figure 1. Assume that $\{\kappa_1 = 0.5, \kappa_2 = -0.5, \kappa_3 = -0.5\}$ and $\{n^p_{1,1} = 3, n^p_{2,1} = 2, n^p_{3,1} = 1$ are known. Let α be 0.5. The value of b_1 could be obtained by $b_1 = \frac{1}{3+2+1}(3 \times 7 \times (1 - 0.5 \times 0.5) + 2 \times 4 \times (1 - 0.5 \times -0.5) + 1 \times 4 \times (1 - 0.5 \times -0.5)) = 5.125$.

Assuming that the unbiased salaries of occupations o_j's are known and denoted by b_j's, the **competitiveness of a company** c_i can be derived by computing the average relative difference between the offer salaries of c_i for occupation o_j ($p_{i,j}$) and the unbiased salaries b_j, for all occupations o_j's with $n^p_{i,j} > 0$. The competitiveness of a company is then defined as:

$$\begin{aligned}\kappa_i &= Avg_{o_j \in O^p(c_i)} \frac{p_{i,j} - b_j}{p_{i,j}} \\ &= 1 - \frac{1}{n^p_i} \sum_{o_j \in O^p(c_i)} \frac{n^p_{i,j} \cdot b_j}{p_{i,j}} \end{aligned} \tag{2}$$

where $n^p_i = \sum_{o_j \in O^p(c_i)} n^p_{i,j} = \sum_{o_j \in O} n^p_{i,j}$.

The competitiveness index κ_i is zero if the unbiased salary of occupation o_j is identical to the offer salaries of all jobs of o_j offered by c_i. Otherwise, κ_i is positive (or negative) if the unbiased salary is lower (or higher) than the average of offer salaries of all jobs offered by c_i. For example, consider c_3 in Figure 1. Given $\{b_1 = 6, b_3 = 5\}$ and $\{n^p_{3,1} = 3, n^p_{3,3} = 2\}$, the competitiveness of c_3 could be derived by $\kappa_3 = 1 - \frac{1}{3+2} \times (\frac{3 \cdot 6}{4} + \frac{2 \cdot 5}{3}) = -0.57$. The value of competitiveness could reflect that c_3 is relatively less competitive.

On the other hand, we consider the review salaries from review data. From review salaries, we could also define unbiased salary in terms of inflation. Assume that the inflation indices of companies e_i's are known. The **unbiased salary of an occupation** o_j is then defined as:

$$\begin{aligned} b_j &= Avg_{c_i \in C^r_{i,j}} r_{i,j} \cdot (1 - \beta e_i) \\ &= \frac{1}{n^r_j} \sum_{c_i \in C^r_{i,j}} n^r_{i,j} \cdot r_{i,j} \cdot (1 - \beta e_i) \end{aligned} \tag{3}$$

where $n^r_j = \sum_{c_i \in C^r(o_j)} n^r_{i,j} = \sum_{c_i \in C} n^r_{i,j}$, and $0 \leq \beta \leq 1$ is a user-defined parameter.

For example, consider M_r in Figure 1. Given $\{e_1 = -0.2, e_2 = -0.1, e_3 = 0.5\}$, $\{n^r_{1,1} = 3, n^r_{2,1} = 2, n^r_{3,1} = 1\}$ and $\beta = 0.5$, we can derive that $b_1 = \frac{1}{3+2+1}(3 \times 5 \times (1 - 0.5 \times (-0.2)) + 2 \times 5 \times (1 - 0.5 \times (-0.1)) + 1 \times 8 \times (1 - 0.5 \times 0.5) = 5.5$.

We can see that the unbiased salary of o_1 here is 5.5 which is smaller than the average of review salary of o_1 (i.e., $\frac{5+5+8}{3} = 6$)when we remove the bias of inflation of companies. We can also find that the unbiased salary of o_1 obtained from review salary here (5.5) is different than that from offer salary (5.125). This shows the unbiased salary of o_1 should be obtained by consolidating in a sophisticated way.

Similarly, we can define the **inflation index of a company** c_i as

$$\begin{aligned} e_i &= Avg_{o_j \in O^r(c_i)} \frac{r_{i,j} - b_j}{r_{i,j}} \\ &= 1 - \frac{1}{n^r_i} \sum_{o_j \in O^r(c_i)} \frac{n^r_{i,j} \cdot b_j}{r_{i,j}} \end{aligned} \tag{4}$$

where $n^r_i = \sum_{o_j \in O^r(c_i)} n^r_{i,j} = \sum_{o_j \in O} n^r_{i,j}$.

For example, suppose $b_1 = 6, b_3 = 5$ and $n^r_{1,1} = 3$, $n^r_{1,3} = 4$, we then derive $e_1 = 1 - \frac{1}{3+4}(\frac{3 \times 6}{5} + \frac{4 \times 5}{3}) = -0.46$. Consistent to our observation, a negative inflation of c_1 show that employees in c_1 tend to under-report their salaries.

These equations above model the dependency between unbiased salary and competitiveness, and that between unbiased salary and inflation. We can observe that competitiveness and inflation are also related to each other since unbiased salary is an intermediate latent variable correlated to both of them. Formally, the unbiased salaries should satisfy the following two equation systems: (i) Equations 2 and 1, and (ii) Equations 4 and 3. It is necessary to develop an efficient algorithm to derive these three latent variables.

C. SOLVING UNBIASED SALARY, COMPETITIVENESS, AND INFLATION

The goal of this paper is to derive unbiased salary, competitiveness, and inflation to best fit the offer salaries and review salaries. A naive way could be to randomly initialize the unbiased salaries at first, and then iteratively adjust the unbiased salaries, competitiveness and inflation so that the derived offer and review salaries can be as close as the observable offer and review salaries respectively by the above equations. However, such an iterative approach is inefficient due to the complicated dependency.

Here, we propose COC-model which aims at learning these three latent variables together. By representing the above equations in matrices, the dependencies among the three latent variables are simplified into a matrix equation that can be solved easily. Therefore, unbiased salaries, competitiveness, and inflation could be learnt analytically.

Given the set of companies $C = \{C_1, C_2, \ldots, C_m\}$ and occupations $O = \{o_1, o_2, \ldots, o_n\}$, the unbiased salaries of occupations can be represented as $\vec{b} = [b_1, b_2, \ldots, b_n]^T$, the competitiveness indices can be represented as $\vec{\kappa} = [\kappa_1, \kappa_2, \ldots, \kappa_m]^T$, and the inflation indices can be represented as $\vec{e} = [e_1, e_2, \ldots, e_m]^T$. First, the second term of Equation 2 is a linear combination of $\vec{b}$ where the coefficient of each b_j is $\frac{n^p_{i,j}}{n^p_i \cdot p_{i,j}}$. Let $\mathbb{1}$ be an $m \times 1$ matrix with all entries 1. By defining an $m \times n$ matrix $W = [W_{i,j}] = [\frac{n^p_{i,j}}{n^p_i \cdot p_{i,j}}]$, the vector $\vec{\kappa}$ can be rewritten into:

$$\vec{\kappa} = \mathbb{1} - W\vec{b} \tag{5}$$

Equation 1, following the same logic above, can be represented into a linear combination of κ_i as

$$b_j = \frac{1}{n_j^p}\sum_i n_{i,j}^p p_{i,j}(1-\alpha\kappa_i) = \frac{1}{n_j^p}\sum_i n_{i,j}^p p_{i,j} - \frac{1}{n_j^p}\sum_i n_{i,j}^p p_{i,j}(\alpha\kappa_i) \quad (6)$$

Therefore, by defining an $m \times n$ matrix $U = [U_{i,j}] = [\frac{n_{i,j}^p p_{i,j}}{n_j^p}]$, the vector $\vec{b}$ can be rewritten into:

$$\vec{b} = U^T\mathbb{1} - \alpha U^T\vec{\kappa} \quad (7)$$

Similarly, the vector $\vec{e}$ can be rewritten into:

$$\vec{e} = \mathbb{1} - R\vec{b} \quad (8)$$

where $R^{m\times n} = [R_{i,j}] = \frac{n_{i,j}^r}{r_{i,j}n_i^r}$. In terms of review salaries, the vector $\vec{b}$ can be represented into the other equation:

$$\vec{b} = Q^T\mathbb{1} - \beta Q^T\vec{e} \quad (9)$$

where $Q^{m\times n} = [Q_{i,j}] = \frac{n_{i,j}^r r_{i,j}}{n_j^r}$.

In the above equations, there are altogether three unknowns $\vec{\kappa}$, $\vec{e}$ and $\vec{b}$. Each entry in $\vec{b}$ should be positive since the unbiased salary of each occupation should be positive whereas no range limit is imposed on $\vec{\kappa}$ and $\vec{e}$. Therefore, the next step is to infer $\vec{b}$ by replacing $\vec{\kappa}$ and $\vec{e}$ in terms of $\vec{b}$. Once $\vec{b}$ is derived, $\vec{\kappa}$ and $\vec{e}$ can be also computed.

By replacing $\vec{\kappa}$ in Equation 5 by Equation 9, and $\vec{e}$ in Equation 9 by Equation 8, we obtain:

$$\begin{cases}\vec{b} = U^T\mathbb{1} - \alpha U^T(\mathbb{1} - W\vec{b}),\\ \vec{b} = Q^T\mathbb{1} - \beta Q^T(\mathbb{1} - R\vec{b})\end{cases} \quad (10)$$

By subtracting the second equation by the first equation in Equation 10, we can obtain that:

$$(U^T - Q^T - \alpha U^T + \beta Q^T)\mathbb{1} + (\alpha U^T W - \beta Q^T R)\vec{b} = 0 \quad (11)$$

Therefore, the vector $\vec{b}$ can be expressed as:

$$A\vec{b} = C \quad (12)$$

where $A = \alpha U^T W - \beta Q^T R$ and $B = (1-\beta)Q^T\mathbb{1} - (1-\alpha)U^T\mathbb{1}$

Here, we have simplified the dependency of three latent variables into one single linear equation system. As we weave correlation between unbiased salary and competitiveness/inflation into the matrices A and B, the meaning of finding a $\vec{b}$ is to find the unbiased salary which can satisfy the constraints with respect to $\vec{\kappa}$ and $\vec{e}$. However, it is not always possible to find a solution $\vec{b}$ to satisfy $A\vec{b} = B$. In this case, we would like to find a vector $\vec{b^*}$ so that $A\vec{b^*} = B$ and the error $L(\vec{b}, \vec{b^*}) = |\vec{b} - \vec{b^*}|$ is minimized. If there exists $\vec{b}$ such that $A\vec{b} = B$, then $\vec{b^*} = \vec{b}$. From the linear algebra view, $A\vec{b^*}$ should be the projection of B to column space of A. By this concept, we can use a pseudo-inverse matrix of A to find $\vec{b^*}$ in a close form: $\vec{b*} = (A^TA)^{-1}A^TB$. Since unbiased salary should be positive, we can further use non-negative least squares approaches, such as Lawson—Hanson algorithm and Fast NNLS, to solve $\vec{b^*}$ to ensure its non-negativity [18]–[20]. Once $\vec{b}$ is obtained, $\vec{\kappa}$ and $\vec{e}$ could be derived by replacing $\vec{b}$ in Equations 5 and 8 respectively.

TABLE 3. Statistics of real datasets.

	JobsBank: $x = p$	GlassDoor: $x = r$
Company Count	398	398
Job Post Count	2919	-
Job Review Count	-	10867
Company-Review-Occupation Count	-	2914
Company-Post-Occupation Count	2919	-
Mean Occupation Count per company	7.8	27.3
SD Occupation Count per company	11.44	99.88
Mean Company Count per occupation	29.79	105.75
SD Company Count per occupation	99.54	318.89

V. EXPERIMENTAL RESULTS

A. EXPERIMENT SETUP

There are two main goals to achieve in our experiments. The first goal is to apply the proposed model on the real world data and to determine if the results is consistent with the common knowledge about a job market. The second is to demonstrate how well our proposed model can predict the ground truth occupations' unbiased salaries, companies' competitiveness and inflation using a synthetic dataset, and how it behaves under different parameter settings.

Baselines: Since our model is the first work to derive occupation salaries by aggregating job post data and job review data, there is no existing work to be compared. Therefore, we compare our proposed method, denoted as COC against two baselines, say AVG and LQ, which are defined as follows:

AVG simply derives the unbiased salary ($\hat{b}^{Avg}$) of an occupation by averaging all the offer salaries and review salaries for this occupation. That is,

$$\hat{b}_j^{Avg} = \frac{\sum_{c_i\in C^r(o_j)} p_{i,j} + \sum_{c_i\in C^p(o_j)} r_{i,j}}{n_j^r + n_j^p} \quad (13)$$

where n_j^r and n_j^p are the number of review and offer salaries of occupation o_j, respectively. The competitiveness $\hat{\kappa}_i^{Avg}$ and inflation $\hat{e}_i^{Avg}$ of a company are derived by Equations 2 and 4, respectively.

LQ is a score aggregation approach proposed in [17] to compute the unbiased quality scores of items and leniency scores of reviewers from the observed ratings from reviewers to items. We adapt LQ by treating occupations as items and companies as reviewers. LQ considers a bipartite graph with companies and occupations as two classes of vertices. Unlike Equation 5 and 7, LQ defines the competitiveness vector by the equation $A\vec{\kappa^{Lq}} = B$ where $A = I - \alpha WU^T$ and $B = \mathbb{1} - WU^T\mathbb{1}$ and the unbiased salaries can be then derived by using this equation $\vec{\kappa^{Lq}} = \mathbb{1} - W\vec{b}$. The inflation e_i^{Lq}'s can be derived in the similar way.

TABLE 4. Competitiveness and inflation of *developer* with average offer and review salaries.

Company	Industry	κ_i^{coc}	e_i^{coc}	κ_i^{avg}	e_i^{avg}	$\mu_{p,cp}$	$\mu_{r,cp}$	$\mu_{p,id}$	$\mu_{r,id}$
Optimum Solutions	IT Services	0.23	-0.16	-0.06	0.14	6050	4918	5589	4850
Credit Suisse	Investment Banking	0.19	-0.05	0.05	0.16	12185	7608	9769	8143
Isentia Brandtology	Advertising	-0.45	-1.17	-0.90	-0.50	4550	2830	3800	3332
SAP Asia	Computer HW/SW	0.28	0.02	0.06	0.17	7083	6401	6250	5688
Barclays Bank	Banks	0.33	-0.62	0.05	0.35	8729	7520	8652	7080

Metrics: We use mean absolute error and Kendall coefficient to evaluate model effectiveness in recovering the ground truth on synthetic dataset only.

Mean Absolute Error (MAE): Given two vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$, the mean absolute error between $\vec{x}$ and $\vec{y}$ is defined as $MAE(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$. The smaller the MAE value is, the more similar the given two vectors are. In our experiments, MAE is used to determine the accuracy of recovering the unbiased occupation salaries b_j's.

Kendall Coefficient: Given two vectors $\vec{x}$ and $\vec{y}$, Kendall coefficient measures the similarity of the ordering of elements from two vectors, which counts the number of pairs for which $\vec{x}$ and $\vec{y}$ agree on their relative ranks. Given two vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$, any pair (i, j) is concordant if either $x_i > x_j$ and $y_i > y_j$ hold or $x_i < x_j$ and $y_i < y_j$ hold. Otherwise, it is dis-concordant. Kendall coefficient τ is to compute the average of the difference between the number of concordant pairs and that of dis-concordant ones, which can be formulated as:

$$\tau(\vec{x}, \vec{y}) = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) \cdot sgn(y_i - y_j) \quad (14)$$

where $sgn(\cdot)$ is the sign function. The larger this coefficient is, the more similar the relative rank between two given vectors is.

In our case, two vectors $\vec{x}$ and $\vec{y}$ would be the pair of the competitiveness (or inflation) vector generated by COC and the ground-truth competitiveness (or inflation) vector. We use Kendall coefficient to determine the accuracy of recovering the ordering of competitiveness and inflation as the actual values are less important for the two bias variables.

B. EMPIRICAL DATA SCIENCE STUDY ON REAL DATASET

Data Description: Our data collection begins with gathering relevant job salary data from JobDB, a job post website in Singapore, and Glassdoor (glassdoor.com), a job review website. In job post data, companies post their job openings. Each job post includes a company name, a job title, the offer salary range (monthly), and others. In job review data, each review includes a company name, a job title, a range of review salaries (in terms of minimum, average, and maximum), and the number of users providing their reviews. For simplicity, we derive the offer salary of a job as the average between the minimum and maximum salaries, and the offer salary of an occupation to be the average of all offer salaries of jobs associated with the occupation.

To address the heterogeneity between these two data sources, we associate all job positions/titles with the corresponding occupations and resolve the different company names. We standardize all salaries in our experiments to be the average monthly salaries. We also match the companies between the two websites using an accurate company name matching algorithm.[2] We only analyse companies that post jobs and receive reviews. Note that every job title is decomposed into domain, position and function. For example, the job title ''senior financial manager'' has ''senior'' position word, ''financial'' domain word and ''manager'' function word covering the seniority, domain and job function respectively. Due to large variability in job titles, we use only the function word of job title as occupation. The statistics of this combined real dataset is shown in Table 3.

It is worth mentioning that the data size would be decided by several factors. Data matching is one of the most determinant factor. since most job review data are user-generated data, a unique job title or company name may have many alias names. After matching job post data and job review data with the same job title and company name, the size of matched data would be much smaller than the raw data. Identifying unique company names from many aliases would be also an interesting research topic; however, this is beyond the scope of this paper. On the other hand, the availability of job post data is also an dominant factor. Job post data cannot be always owned by the government website as our data source did. In practice, job post data are owned by job search films. These films usually set barrier for data crawling to protect their contents Thus, a large amount of job post data is not usually available.

Case Example Analysis: We now select a target occupation that can be found in different industries and compare its unbiased salary with observed post and review salaries. The occupation targeted in this study is *developer*. Different companies would hire developers for different functions such as web development, mobile development, and so on. Therefore, developer should be a good target to study the competitiveness and inflation of different companies. Based on our COC-model, the unbiased salary of developer is \$6698.

As shown in Table 4, the unbiased salary of developer is different from average post and review salaries in five companies each from a different industry. $\mu_{x,y}$ refers to the average of company-specific ($y = cp$) or industry-specific ($y = id$) review ($x = r$) or offer salaries ($x = p$). The Table shows that

[2]For brevity, we leave out the details of this matching algorithm.

developer occupation has different salary standards ($\mu_{p,id}$ and $\mu_{r,id}$) in different industries. For example, developer from Advertising has the lowest average salaries $\mu_{p,id}$ ($3800) and $\mu_{r,id}$ ($3332) compared with the unbiased salary of $6698. In contrast, Investment Banking (IB) gives developer highest average offer salary $\mu_{p,id}$ ($9769) and review salary $\mu_{r,id}$ ($8143).

A higher competitiveness reflects that this company usually posts opportunities of developer with higher salaries than others. For example, SAP Asia has a positive $\kappa_i^{coc} = 0.28$ and is more competitive than other companies (except Barclay). This is consistent with Barclay's average offer salaries $\mu_{p,cp} = \$7083$ higher than the industry average $\mu_{p,id} = \$6250$. On the other hand, a lower inflation usually reflects that the company getting reviews that under-report their salaries. For example, *Isentia Bradtology* has a negative $e_i^{coc} = -1.17$ suggested by its $\mu_{r,cp} = \$2830$ lower than $\mu_{r,id} = \$3332$.

Consider Barclays with competitiveness and inflation different between COC and AVG models. As shown in Table 5, Barclays offered most jobs with higher offer salaries $\mu_{p,cp}$ than both unbiased salaries b_j and average offer salaries $\mu_{p,id}$, especially for manager. Hence, Barclay deserves high competitiveness κ_i^{coc} which is more reasonable than the small $\kappa_i^{avg} = 0.05$ returned by the AVG model. On the other hand, the analyst in this bank sees lower review salaries (around $5000) compared with the unbiased salary and offer salaries (around $9000). Since the review salaries of other occupations are slightly higher than the average salaries in the industry or the unbiased salary, this leads to the negative inflation (-0.62) for Barclays.

TABLE 5. Occupations and Salaries (in $) in Barclays Bank.

Occupation	$\mu_{p,cp}$	$\mu_{r,cp}$	$\mu_{p,id}$	$\mu_{r,cp}$	b_j
Analyst	8915×10	5054×252	7363	5203	5638
Auditor	9550×3	11333×20	8787	11000	8901
Developer	8729×2	7520×5	8652	7080	6698
Manager	11041×3	8305×3	7002	5803	7870

Data Science Research Questions: Finally, we can also answer two research questions about 1) the types of companies that demonstrate high and low competitiveness and 2) user communities that are with high and low inflation. We focus on the top-five industries in Singapore based on the number of job posts in the industry. Based on κ^{coc}, the industries are ordered as: Investment Banking > Consulting > Banks > Computer HW/SW > IT services. The non-technology industries are more competitive than technology industries. This does not come as a surprise as Singapore is a major financial hub in Asia. Regarding inflation, the industries are ordered as: Consulting > Investment Banking > Computer HW/SW > IT services > Banks. Most companies have under-report review salaries resulting in mostly negative inflation. Interestingly, Consulting industry is the exception as their review salaries appear to be higher. This may reflect the more aggressive review salary sharing among users for this industry.

Offer Salary Prediction: Besides the case studies, another means of evaluation is to see if the features generated by COC-model, i.e., unbiased salaries, competitiveness and inflation, can improve some downstream application tasks. The objective of this experiment is to evaluate how COC-model outputs can improve the task of offer salary prediction. We define offer salary to be at the occupation level, that is, salary of a specific occupation based on job posts by a specific company. In this task, we define such a salary by averaging the salaries of job posts of the same occupation from the company.

Offer salary prediction can be formulated as a matrix-completion problem, and there are two well known approaches: Non-negative Matrix Factorization (NMF) [21] and Factorization Machine (FastFM) [22]. NMF takes a matrix with each entry representing the offer salary of an occupation from a company. FastFM takes company and occupation features as its input and the offer salary as its prediction target. FastFM uses one-hot-encoded company and occupation features. Formally, suppose that company c_i offers occupation o_j by salary $p_{i,j}$, the feature vector would be $\vec{x} =< x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+n} >$ where $x_k = 1$ if $k = i$ and $k = i + j$ and the others are 0, and the target would be the offer salary $p_{i,j}$. FastFM(AVG) takes not only one-hot-encoded company and occupation features but also the average salary of each occupation and the average salary of all occupations a company offered. Formally, the feature vector would be $\vec{x} =< x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}, x_1^c, \ldots, x_m^c, x_1^o, \ldots, x_n^o >$ where $< x_1, \ldots, x_{m+n} >$ is one-hot-encoded company and occupation features, $x_i^c = Avg_{o_j \in O_{c_i}^p}(p_{i,j} + r_{i,j})$, and $x_j^o = Avg_{c_i \in O_{c_j}^p}(p_{i,j} + r_{i,j})$. FastFM(COC) takes one-hot-encoded company and occupation, the unbiased salaries, competitiveness, and inflation generated from COC-model. Formally, the feature vector would be $\vec{x} =< x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}, \kappa_1, \ldots, \kappa_m, e_1, \ldots, e_m, b_1, \ldots, b_n >$ where $< x_1, \ldots, x_{m+n} >$ is one-hot-encoded company and occupation features, κ_i and e_i are the competitiveness and the inflation of the i-th company and b_i is the unbiased salary of the i-th occupation.

We divide the offer salary data into training and test sets, with an 80-20 split. Figure 2 shows the MAE with feature combinations, C, I, S and CIS denoting competitiveness only, inflation only, unbiased salary only, and all-features derived using AVG and COC models. The plain NMF performs the worst among different approaches and FastFM performs better than FastFM(AVG) in most cases. Since average model may not generate representative latent biases and unbiased salaries, FastFM(AVG) could not predict offer salaries well. On the other hand, FastFM(COC) can achieve the lowest MAE in most cases. The reason may be that COC-model can more accurately capture latent biases and unbiased salaries which are proved to be useful in this task. It can be seen that Fast(COC) performs worst than FastFM when the inflation is

FIGURE 2. MAE with different prediction approaches.

added as an extra feature. It shows that inflation alone may not be helpful for this prediction task. This may be due to people likely to under-report their salaries, as shown in our data science study.

C. EXPERIMENTS ON SYNTHESIS DATASETS

Most real-life datasets do not provide ground truth of the competitiveness and inflation of companies and unbiased salaries of occupations. Therefore, the synthetic dataset is generated to allow us to verify effectiveness and study effects of parameter setting on our proposed COC model.

Data Generation: The synthetic datasets are generated based on two distributions: one controls the number of post and review salaries, which simulates the scenario that different companies of different sizes; and another controls the assignment of competitiveness and inflation, which are used to generate the post and review salaries from unbiased salaries.

The generation of each synthetic dataset requires a few input parameters, namely: (a) number of companies (default=1000), (b) number of occupations (default=50), (c) value ranges of competitiveness and inflation (default = [−0.5,0.5]), (d) value range of offer salary and review salary (default = [0.2,1]), and (e) company-occupation edge removal strategy and its associated parameters. With these input parameters, the data generation consists of three steps:

Step 1: We construct a complete tripartite graph COC-graph to represent the relationship between companies and occupations, where three disjoint sets of vertices represent companies, occupations, and companies, and the edges between first (second) companies and occupations represent offer (review) salary. The number of contributors in each edge of this graph is assigned by a random value in range $[1, 20]$. Since not all companies have the same different job openings, we remove some edges from this graph to simulate a company does not have a specific job opening. The way for edge removal is to remove edges for offer and review salaries based on two uniform random numbers *drop_rate_p* and *drop_rate_r*, respectively. In our experiments, both *drop_rate_p* and *drop_rate_r* are assigned the same default value 0.8.

Step 2: Competitiveness κ_i and inflation e_i are assigned to company vertices in the COC-graph uniformly in the range of $[-0.5, 0.5]$. The occupations o_j's are also assigned unbiased salaries b_j's uniformly sampled from the range of $[0.2, 1]$.

Step 3: The review salary of an occupation o_j of c_i is assigned by $r_{i,j} = (1 + \kappa_i) \times b_j$, and the offer salary of an occupation $o_{j'}$ of c_i is assigned by $p_{i,j'} = (1 + e_i) \times b_{j'}$.

To guarantee the reliability of experimental results, we conduct experiments based on the principles. First of all, the values of competitiveness and inflation are generated independently. Thus, we can simulate the scenario that a company may have different offer salary and review salary for the same occupation as we observed in the real dataset. The unbiased salary of each occupation is determined by the normalized range of all the salary information in our real dataset. On the other hand, since the synthetic data are generated with many random variables, the experimental results vary even in the same setting. To reduce the impact of the biases generated from a few extreme cases, we use each setting of random parameters to generate 50 datasets and the metric values reported in this paper is obtained by averaging results from datasets.

Recovery of Unbiased Salaries: We first examine the accuracy of recovering unbiased salaries of occupations for datasets generated by *uniform edge-removal strategy* by varying the parameter *drop_rate_r* from 0.1 to 0.9 while keeping the remaining parameters unchanged with default setting. Since we care about both the actual value and relative rank of recovered unbiased salaries against those of ground truth, we show the accuracy using Kendall coefficient and MAE.

Figure 3 shows Kendall coefficients and MAE of unbiased salaries with *drop_rate_r* varied and *drop_rate_p* set to be 0.8. Figure 3(a) shows that AVG can achieve higher Kendall coefficient when *drop_rate_r* < 0.4. However, both LQ and COC outperform AVG when *drop_rate_r* > 0.6. Moreover, Kendall coefficient of COC only drops slightly in both cases when the value *drop_rate_r* increases whereas that of LQ drops significantly. This result show that COC can derive the unbiased salaries with more correct ranks than AVG and LQ.

Figure 3(b) shows MAE under different *drop_rate_r* settings. The figure shows that COC outperforms both AVG and LQ, and LQ is slightly worse than COC. As *drop_rate_r* increases, the MAE of AVG becomes much worse than those of LQ and COC. The results above can conclude that unbiased salary recovery evaluation, LQ and COC outperform AVG in most cases, and COC outperforms LQ.

In summary, COC generally outperforms LQ by MAE and enjoys similar Kendall Coefficient with LQ. COC is also more consistent across different drop rates. Interestingly, AVG could perform well with more post and review salaries, which is expected based on the law of large number. Given that the real data is expected to have higher drop rate, we expect COC to be a superior model compared with LQ and AVG.

Recovery of Competitiveness and Inflation: Figure 4(a) shows the Kendall coefficients of competitiveness derived by AVG, LQ, and COC. The Kendall coefficient of AVG is much lower than that of COC and LQ. The values of

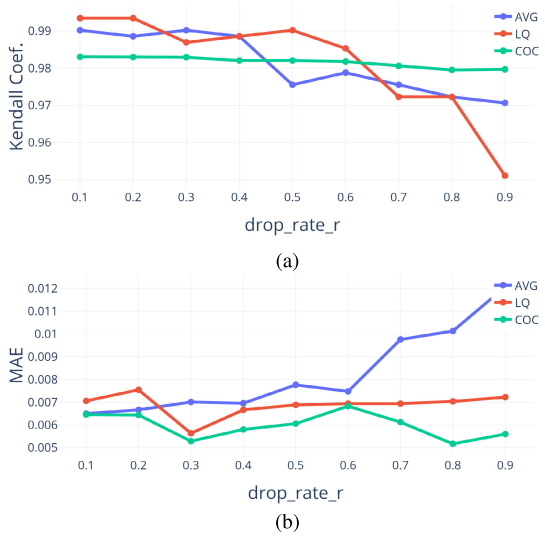


FIGURE 3. Kendall coefficient and MAE of unbiased salaries with edge removal.

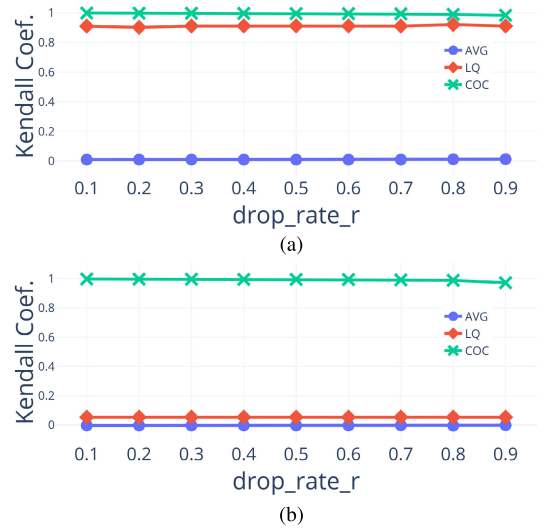


FIGURE 4. Kendall coefficient and distribution error with *drop_rate_r* varied.

Kendall coefficient for LQ and COC are quite close where their difference are within 5%. In Figure 4(b) shows Kendall coefficients of inflation derived by AVG, LQ, and COC. COC still yields the highest Kendall coefficient. In this case, the Kendall coefficient of AVG and LQ are much lower than that of COC. The reason is that LQ uses the way similar to COC to derive competitiveness and then uses average to derive inflation. Putting the results above together, COC again is the best model.

VI. CONCLUSION

In this paper, we proposed the COC-model to derive occupation-level unbiased salaries for a job market by aggregating job post and job review data. To distill unbiased salaries, the biases of offer salaries in job posts and that of review salaries in job reviews are modeled as company competitiveness and inflation. We start from defining unbiased salary, competitiveness, and inflation based on their inter-dependency. These dependencies are then represented as a system of equations involving matrices and the equations can be solved by linear algebra techniques easily. Extensive empirical studies of job salaries and companies are conducted by both the real dataset and the synthetic dataset. On the real world data, the proposed model can derive the results which are consistent with the common knowledge about a job market. On the synthetic dataset, we also demonstrate the proposed model can predict the ground truth occupations' unbiased salaries, companies' competitiveness and inflation effectively.

ACKNOWLEDGMENT

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. The authors would like to express their gratitude to Dr. Ying-Ju Chen, Assistant Professor with the University of Dayton, USA, for giving us constructive comments on revising this article.

REFERENCES

[1] T. Xu, H. Zhu, C. Zhu, P. Li, and H. Xiong, "Measuring the popularity of job skills in recruitment market: A multi-criteria approach," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 2572–2579.
[2] M. Chamberlain, "Demystifying the gender pay gap: Evidence from glassdoor salary data," Glassdoor, Mill Valley, CA, USA, Tech. Rep., 2016. [Online]. Available: https://www.classlawgroup.com/wp-content/uploads/2016/11/glassdoor-gender-pay-gap-study.pdf
[3] A. Lazar, "Income prediction via support vector machine," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2004, pp. 1–4.
[4] S. Jackman and G. Reid, "Predicting job salaries from text descriptions," Ph.D. dissertation, Dept. Statist., Univ. Brit. Columbia, Vancouver, BC, Canada, 2013.
[5] Q. Meng, H. Zhu, K. Xiao, and H. Xiong, "Intelligent salary benchmarking for talent recruitment: A holistic matrix factorization approach," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 337–346.
[6] X. Chen, Y. Liu, L. Zhang, and K. Kenthapadi, "How Linkedin economic graph bonds information and product: Applications in Linkedin salary," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD* 2018, Y. Guo and F. Farooq, Eds. London, U.K.: ACM Press, Aug. 2018, pp. 120–129. [Online]. Available: https://doi.org/10.1145/3219819.3219921
[7] D. Shen, H. Zhu, C. Zhu, T. Xu, C. Ma, and H. Xiong, "A joint learning approach to intelligent job interview assessment," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.
[8] N. Luo, Y. Zhou, and J. J. Shon, "Employee satisfaction and corporate performance: Mining employee reviews on glassdoor. com," in *Proc. Int. Conf. Inf. Syst. (ICIS)*, 2016, pp. 1–16.
[9] Y. Sun, F. Zhuang, H. Zhu, X. Song, Q. He, and H. Xiong, "The impact of person-organization fit on talent management: A structure-aware convolutional neural network approach," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. Anchorage, AK, USA: ACM, Aug. 2019, pp. 1625–1633. [Online]. Available: https://doi.org/10.1145/3292500.3330849
[10] H. Lin, H. Zhu, Y. Zuo, C. Zhu, J. Wu, and H. Xiong, "Collaborative company profiling: Insights from an employee's perspective," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 1417–1423.
[11] K. Kenthapadi, S. Ambler, L. Zhang, and D. Agarwal, "Bringing salary transparency to the world: Computing robust compensation insights via LinkedIn salary," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2017, pp. 447–455.
[12] H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao, "Prospecting the career development of talents: A survival analysis perspective," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, Aug. 2017, pp. 917–925, doi: 10.1145/3097983.3098107.

[13] H. Xu, Z. Yu, H. Xiong, B. Guo, and H. Zhu, "Learning career mobility and human activity patterns for job change analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, C. C. Aggarwal, Z. Zhou, A. Tuzhilin, H. Xiong, and X. Wu, Eds. Atlantic City, NJ, USA: IEEE Computer Society, Nov. 2015, pp. 1057–1062, doi: 10.1109/ICDM.2015.122.

[14] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong, "A hierarchical career-path-aware neural network for job mobility prediction," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. Anchorage, AK, USA: ACM Press, Aug. 2019, pp. 14–24, doi: 10.1145/3292500.3330969.

[15] L. Zhang, H. Zhu, T. Xu, C. Zhu, C. Qin, H. Xiong, and E. Chen, "Large-scale talent flow forecast with dynamic latent factor model?" in *Proc. World Wide Web Conf. (WWW)*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. San Francisco, CA, USA: ACM Press, May 2019, pp. 2312–2322. [Online]. Available: https://doi.org/10.1145/3308558.3313525

[16] H. R. Arkes, "The nonuse of psychological research at two federal agencies," *Psychol. Sci.*, vol. 14, no. 1, pp. 1–6, Jan. 2003.

[17] H. W. Lauw, E.-P. Lim, and K. Wang, "Quality and leniency in online collaborative rating systems," *ACM Trans. Web*, vol. 6, no. 1, pp. 1–27, Mar. 2012.

[18] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems* (Classics in Applied Mathematics), vol. 15. Philadelphia, PA, USA: SIAM, 1995,

[19] J. Chen, C. Richard, J. C. M. Bermudez, and P. Honeine, "Nonnegative Least-Mean-Square algorithm," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5225–5235, Nov. 2011.

[20] V. Franc, V. Hlaváč, and M. Navara, "Sequential coordinate-wise algorithm for the non-negative least squares problem," in *Computer Analysis of Images and Patterns*, A. Gagalowicz and W. Philips, Eds. Berlin, Germany: Springer, 2005, pp. 407–414.

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 556–562.

[22] I. Bayer, "fastfm: A library for factorization machines," *J. Mach. Learn. Res.*, vol. 17, no. 184, pp. 1–5, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-355.html

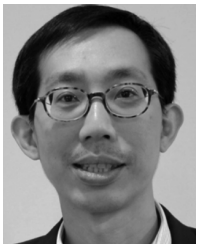
[23] X. J. S. Ashok, E.-P. Lim, and P. K. Prasetyo, "JobSense: A data-driven career knowledge exploration framework and system," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1411–1416.

[24] R. J. Oentaryo, E.-P. Lim, X. J. S. Ashok, P. K. Prasetyo, K. H. Ong, and Z. Q. Lau, "Talent flow analytics in online professional network," *Data Sci. Eng.*, vol. 3, no. 3, pp. 199–220, Sep. 2018.

[25] C. Qin, H. Zhu, T. Xu, C. Zhu, L. Jiang, E. Chen, and H. Xiong, "Enhancing person-job fit for talent recruitment: An ability-aware neural network approach," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2018, pp. 25–34.

[26] H. Xu, Z. Yu, B. Guo, M. Teng, and H. Xiong, "Extracting job title hierarchy from career trajectories: A Bayesian perspective," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3599–3605.

CHIH-CHIEH HUNG received the M.S. and Ph.D. degrees from National Chiao Tung University, Taiwan, in 2005 and 2011, respectively. He is currently an Assistant Professor with the Department of Management Information System, National Chung Hsing University, Taiwan. He has published papers in several prestigious conferences, such as IEEE International Conference on Data Engineering (ICDE), IEEE International Conference on Data Mining (ICDM), ACM Conference on Information and Knowledge Management (ACM CIKM), and prestigious journals, such as IEEE TRANSACTIONS ON KKNOWLEDGE AND DATA ENGINEERING, IEEE SMC, and *VLDB Journal*. He was a recipient of the Best Paper Award from the ACM Workshop on Location-Based Social Network 2009. His research interests include data mining, mobile and pervasive computing, big data analytics, and artificial intelligence.



EE-PENG LIM received the B.Sc. degree in computer science from the National University of Singapore, and the Ph.D. degree from the University of Minnesota, Minneapolis, in 1994. He is currently the Lee Kong Chian Professor with the School of Information Systems, Singapore Management University (SMU). He is also the Co-Director of the Living Analytics Research Center, jointly established by SMU and the Carnegie Mellon University. He serves on the Steering Committee of the International Conference on Asian Digital Libraries (ICADL), Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and International Conference on Social Informatics (Socinfo). His research interests include social network and web mining, information integration, and digital libraries.