

Uncovering IT Career Path Patterns with Job Embedding-based Sequence Clustering

HAO ZHONG, ESCP Business School, Paris, France

CHUANREN LIU, The University of Tennessee, Knoxville, United States

CHAOJIANG WU, Kent State University, Kent, United States

Extracting typical career paths from large-scale and unstructured talent profiles has recently attracted increasing research attention. However, various challenges arise in effectively analyzing self-reported career records. Inspired by recent advancements in neural networks and embedding models, we develop a novel career path clustering approach and apply it to uncover information technology (IT) career path patterns. Specifically, we construct employment profiles of over 60,000 IT professionals, and form their career path sequences by chaining the job records in each profile. Then we simultaneously learn cluster-wise job embeddings and construct career path clusters. The resultant cluster-wise likelihoods of career paths can quantify their soft bonding with different clusters, and the job embeddings can reveal connections among job titles within each cluster. With both real and simulated data, we conduct extensive experiments with our framework to establish the modeling performance and great improvement over the traditional optimal matching analysis methods. The empirical results from analyzing real data on career paths show that our approach can discover distinct IT career path patterns and reveal valuable insights.

CCS Concepts: • **Information systems** → **Clustering**; • **Computing methodologies** → **Mixture modeling**;

Additional Key Words and Phrases: Career path clustering, sequential job embedding, mixture Markov models, IT workforce

ACM Reference Format:

Hao Zhong, Chuanren Liu, and Chaojiang Wu. 2025. Uncovering IT Career Path Patterns with Job Embedding-based Sequence Clustering. *ACM Trans. Manag. Inform. Syst.* 16, 2, Article 16 (March 2025), 32 pages. <https://doi.org/10.1145/3712705>

1 Introduction

The wide adoption of employment-oriented online services has facilitated novel research investigations on professional career patterns with rich career data. The research on professional careers is particularly important for industries where talents change jobs frequently. For instance, an **information technology (IT)** professional typically transitions between job roles every one or two years, building up a career trajectory comprised of a series of job positions. Personal income

Dr. Hao Zhong was supported by National Natural Science Foundation of China (NSFC) under Grant 72371011.

Authors' Contact Information: Hao Zhong, ESCP Business School, Paris, Île-de-France, France; e-mail: hzhong@escp.eu; Chuanren Liu, The University of Tennessee, Knoxville, Tennessee, United States; e-mail: cliu89@utk.edu; Chaojiang Wu, Kent State University, Kent, Ohio, United States; e-mail: cwu13@kent.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2158-656X/2025/03-ART16

<https://doi.org/10.1145/3712705>

growth is usually linked to transitions in employment throughout one's career development [44]. There is a significant interest in identifying common sequential patterns from a substantial volume of career path records. From a practical perspective, the identification of career patterns has the potential to assist in a range of applications such as job recommendation, career planning, and talent management [27]. For example, knowing the typical career trajectories and transitions can facilitate individuals in making well-informed choices regarding their career advancement. From a societal perspective, identifying career patterns from a large population can assist in tackling issues related to social advancement and workforce diversity. For instance, it can help identify barriers to career advancement and opportunities for social mobility, thus promoting equal access to career opportunities for all individuals of society. Analyzing career patterns can also help uncover disparities in representation among different demographic groups in specific fields or professions, which can further guide efforts to foster diversity and inclusion in the workforce.

Previous research have delved into the problem of career path characterization. Early studies adopted the **optimal matching analysis (OMA)** [43], a sequence analysis method suitable for career path clustering. These methods rely on a two-step process. Firstly, the similarity between each pair of sequences is evaluated using edit distance, which quantifies the cost of editing one sequence into another. In the second stage, a clustering algorithm is utilized to generate sequence clusters by analyzing the pairwise similarity matrix. For instance, Biemann and Wolf [11] employed such methods to identify six career patterns of top management team members, which are determined by their international experience, organizational tenure, and professional experience. Biemann et al. [12] investigated the impact of socio-demographic factors on individuals' adherence to specific career trajectory patterns. A detailed review conducted by Dlouhy and Biemann [17] underscored the importance of longitudinal research that centers on the progression of occupational states in the investigation of career development.

The emergence of online talent networking platforms (e.g., LinkedIn) has sparked a resurgence in exploring extensive and unstructured career path data for research purposes. These platforms are now recognized as "a valuable and reliable source for investigating the human resource-related problems" [54]. However, conventional career path characterization approaches are inadequate for analyzing self-reported data due to several challenges. First, they operate under the assumption that the elements (e.g., job titles) are equally discernible within the sequences. In reality, some job titles are much more similar to each other than other job titles in our data. Conventional methods do not take into account weighted distances when estimating the cost of editing one sequence into another. Second, the weighted distances between sequence elements are unavailable to extend these methods for the task of career path characterization. Analyzing raw career records becomes increasingly challenging as self-reported job titles on online talent networking platforms are much more intricate compared to the job titles in the aforementioned studies. As an example, there are over 48 K unique job titles within our data. Third, as highlighted by Lappas [31], clusters of career trajectories exhibit inherent ambiguity and cannot be easily reduced to standard prototype paths. Therefore, clustering approaches that rely on the aggregated centroids and hard clustering labels may overlook fine-grained information present in the large-scale career trajectory data. We need a quantitative measure to accurately assess the distances between job titles, facilitating clustering analysis of a vast array of job sequences.

To tackle these obstacles, we introduce an approach based on sequential job embedding and mixture Markov models. There are two unique differences in our approach compared with the conventional OMA-based approaches. First, we employ optimization and statistical approaches to learn latent embeddings of job titles. The latent embeddings have the capability to represent the relationship between both the concepts of "functionality" and "responsibility" among raw job titles as demonstrated by Liu et al. [32] in their study. This is achieved by mapping each term within

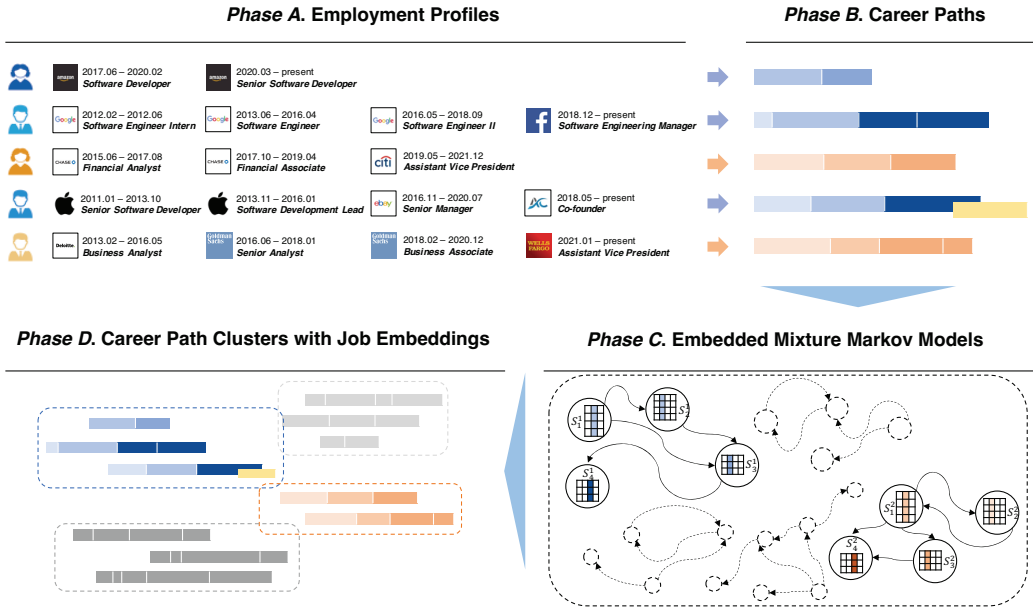


Fig. 1. A conceptual diagram of our career path clustering approach.

job titles to a common Euclidean space, allowing for the modeling of career paths using a Markov assumption. Second, we substitute conventional cluster centroids with a mixture of probabilistic models that are effective at handling the heterogeneous nature of real-world career paths. By incorporating latent embeddings of job titles and a mixture of Markov models, our model effectively tackles the dual challenges presented when analyzing extensive career path data.

We demonstrate the procedure of our method in Figure 1. Essentially, our objective is to identify cohesive groupings of career paths that exhibit the following characteristics: (1) the clusters are distinctive; (2) each cluster is associated with a significant number of individuals and their career trajectories; and (3) the collective clusters can effectively represent the career paths of the majority. Let's start with talents' employment profiles (*Phase A*). An average employment profile typically comprises the individual's historical and current employment information, consisting of job titles and the respective duration of employment for each position held. We arrange the job records chronologically for each profile to form a *career path* (*Phase B*). Note that the duration of each employment record is indicated in our coding scheme as it serves to signify the level of significance of the role within the respective career path. In *Phase C*, we employ the **Embedded Mixture Markov Models (EMMMs)** to simultaneously learn cluster-wise job embeddings and construct career path clusters. The likelihood of each career path's association with a cluster is computed through the proximity measurement of sequentially adjacent job titles in the embedded space. The optimal assignments of career paths are then learned by repeatedly selecting the cluster with the highest estimated likelihood. Meanwhile, job embeddings specific to each career are adjusted to capture unique attributes associated with clusters. In contrast to conventional OMA-based approaches and other categorical sequence clustering methods, our embedding-based model is adept at handling careers that involve sequentially overlapping job positions, as exemplified by the fourth talent's career path in Figure 1. It is a common case for individuals who concurrently hold multiple professional positions throughout their careers. Once the model training process reaches convergence, the cluster-specific likelihoods of career paths explicitly measure their level

of association with the respective clusters (*Phase D*). The resultant job embeddings generated by our model capture underlying relationships among various job titles within individual clusters.

Utilizing our methodology enables us to explore important research questions using extensive career trajectory data. Specifically, we attempt to uncover significant gender differences in the career advancement of the IT workforce. We construct several key career advancement metrics and conduct in-depth analyses to reveal differences in the career scores between men and women. We also show that the career path clusters constructed by our approach could not only greatly differentiate the clusters of career path advancement, but also improve our ability to capture gender career advancement disparity. Overall, our proposed clustering method demonstrates high efficacy in clustering IT workforce career paths and discerning cluster-wise differences in various metrics. Our analysis also demonstrates the flexibility of our clustering approach for the inclusion of individual characteristics such as gender and other information, which is critical to support further research investigations.

2 Literature Review

Our work is mainly associated with the extensive body of literature concerning career path studies. Earlier research in this field employed surveys, interviews, and/or narratives methods to understand career paths. Reich and Kaarst-Brown [41] carried out a case study involving 51 former IT professionals using interviews and surveys to understand their career transitions from IT to non-IT business roles. Later, through analyzing the narratives collected from a cohort of research scientists Cohen [16] identified four typical career discourses of scientists. Idowu and Elbanna [24] employed a qualitative and inductive research methodology, involving multiple data sources, to offer a thorough and insightful exploration of the evolution of the work development trajectory of crowdworkers. However, early methodologies heavily depended on subjective self-expressions of identity and retrospective sense-making and, unfortunately, have not been widely adopted in recent career path studies [50]. Instead, recent researchers have transitioned their focus toward data-driven approaches. For example, Agarwal and Ohyama [3] employed multiple regression analysis techniques to investigate the career paths of scientists in terms of their focus on basic or applied science and their choice between industry and academia. Their objective was to investigate the impact of scientists' varied career trajectories, as well as their individual skills and preferences, on their career decisions and subsequent income growth.

2.1 Career Path Characterization

Among the myriad studies conducted on career trajectories, the research community has shown significant interest in the problem of *career path characterization*. Due to the sequential nature of career paths, the majority of researchers have utilized sequence analysis and clustering methods. Abbott and Hrycak [1] played a significant role as one of the early contributors in this area. Their research was instrumental in introducing the **optimal matching algorithm (OMA)** for determining the prototype careers of German musicians. Later, Blair-Loy [13] used the optimal matching techniques and the complete linkage-based clustering method to classify the careers of female finance executives into four career types. On the other hand, numerous studies have concentrated on the careers of top-level executives. Biemann and Wolf [11] utilized OMA to develop six career patterns for top management (TMT) members from five different nations, considering factors such as international experience, organizational tenure, and professional experience. In a more recent publication, Koch et al. [30] examined the career trajectories of **Chief Executive Officers (CEOs)** in Fortune 100 companies. Similarly, they used OMA in conjunction with hierarchical cluster analysis to identify unique clusters within CEOs' career paths and to gain a better understanding of the factors driving these individuals to attain top management positions. In summary, these

methods tend to categorize career paths in a two-step process. The initial step involves calculating pairwise distances between career sequences, predominantly through the utilization of the OMA technique. Then, various clustering methodologies are utilized on the distance matrix to categorize job sequences into distinct clusters. This two-step method is often successful, particularly in cases where there is low heterogeneity in job titles.

With the expansion of the IT sector, there has been a notable increase in interests in exploring career paths of IT workforce. Following the same two-step process, Joseph et al. [27] performed an in-depth analysis on characterizing the IT workforce's careers using the OMA approach and a hierarchical agglomerative clustering technique. The study provides an in-depth analysis of the careers and professional experiences of IT professionals, and at the same time, highlighting the research challenge posed by the significant diversity found in the profiles of these individuals. In a similar vein, Lappas [31] also noted that careers of the IT workforce are more diverse than the traditional view of a dual IT career path (technical versus managerial) and are growing in heterogeneity due to a volatile environment. To tackle these challenges, Lappas [31] proposed a novel two-step approach to identify the prototype career paths of the IT workforce. The author initially grouped job titles by utilizing an **Agglomerative Clustering (AC)** algorithm to job description embeddings learned via the **Distributed Memory Model (DMM)**. Then, several career prototypes are discovered via a graph-based clique finder algorithm. Note that our study is connected to Lappas [31]'s work in the following aspects. First, both of our studies share a common focus on career paths clustering of the IT workforce. Second, both employ embedding-based techniques to assess the "similarity" between job titles. Our approach, however, does not necessitate a distinct secondary stage for sequence clustering. Instead, we learn job term/title embeddings and sequence clusters in a concurrent manner.

2.2 Occupational Skills Identification

Although our work is not specifically designed to identify or extract occupation and job skills information, we derive implicit representations (i.e., embeddings) of job titles associated with each career path cluster. We therefore would like to also discuss relevant studies on occupational skills identification. For example, Javed et al. [26] introduced a system named SKILL to address the skills gap in the labor market by accurately detecting and normalizing skills from resumes and job postings. They combined semantic word vectors and a **Markov Chain Monte Carlo (MCMC)** approach for skill detection and sense disambiguation. Bhola et al. [9] addresses the problem of identifying relevant job skills from textual **job descriptions (JDs)**, which often omit crucial skills due to communication gaps between employers and recruiters. They introduced BERT-XMLC, a deep learning framework utilizing pre-trained BERT language models to encode textual JDs and predict relevant skills. Both of these works focus on extracting skills using unsupervised methods, whereas Gujnani and Misra [20] proposed a supervised framework for skill extraction and job recommendation. Specifically, their method combines NLP techniques like NER, PoS tagging, and Word2Vec for skill extraction from text. Implicit skills are inferred using a Doc2Vec model to find similar JDs. Skills are matched to resumes via a bi-directional algorithm that computes an Affinity Score to rank job recommendations. A significant milestone is the comprehensive review by Khaouja et al. [28], which examined 108 studies from 2010 to 2020 on methods and applications for extracting skills from job advertisements. It highlights the evolution from traditional methods like skill counting to advanced AI approaches, such as deep learning and word embeddings, for extracting explicit and implicit skills. Applications include job market analysis, curriculum development, and skill mismatch detection, with emerging areas like gender bias detection and skill demand prediction. The study underscores the need for standardized datasets and advanced modeling techniques to better align labor market needs with workforce training. Another notable study

comes from Hosseinioun et al. [22] which focuses on nested skills. The authors investigated the interdependent structure of skills within labor markets, revealing a hierarchical “skill tree” that connects foundational, general skills to specialized, niche skills. By analyzing extensive U.S. survey data and job transitions, they show how nested skills create pathways for career advancement and wage premiums. It highlights significant disparities across gender and racial/ethnic groups in accessing these nested skill pathways, which are crucial for upward mobility.

Our research aligns with several of the studies mentioned above [e.g., 9, 20] particularly in the application of embedding learnings for occupations or skills. This technique enables more accurate occupation and skill inference, and personalized career guidance by capturing semantic relationships and dependencies between skills and occupations. This improves workforce alignment and adapts dynamically to evolving labor market needs. However, the main distinction of our work lies in its focus on identifying career path patterns and automatically clustering career sequences, which provides valuable career insights. As we will elaborate later, the job title or term embeddings learned during this process also yield interesting findings, albeit as a secondary outcome.

2.3 Categorical Sequence Clustering

In addition to career path studies, our research also contributes to the field of general categorical sequence clustering [4]. In comparison to conventional item clustering, the *categorical sequence clustering* problem is known to present several notable challenges [5]. First, categorical sequences often exhibit high dimensionality, which makes data handling rather complicated. Secondly, and more importantly, calculating the similarity or distance between categorical sequences can be challenging due to potential noise and variations in their lengths. According to Bicego et al. [10], sequential data clustering methods can be classified into three categories. The first is *proximity-based*, which focuses on devising similarity or distance measures between sequences, followed by a standard clustering method based on the sequence distances. They are commonly referred to as raw-data-based methodologies since they typically involve direct processing of raw sequence data. As a result of this characteristic, they encounter various challenges such as sensitivity to noise, amplitude scaling, longitudinal scaling, discontinuities, and temporal drifts [5]. The two-step approach based on OMA as discussed earlier falls into this category. The second category involves a *feature-based* approach, which focuses on extracting temporal features from each sequence. This method then simplifies sequence clustering into a more manageable item or point clustering problem. In contrast to *proximity-based* methods, *feature-based* methods operate with transformed features rather than raw sequences. Therefore, they are more computationally efficient and better for large-scale data sets. The last category encompasses the *model-based* approaches, which assume an analytical model for each cluster and aim to establish the most suitable set of such models that aligns best with the data. Common models in this context include time series models, spectral models, and finite state machines like **hidden Markov models (HMM)**. While model-based approaches are widely favored for their adaptability across various applications, they may suffer from scalability and performance deterioration when faced with intricate data and clusters [5].

Note that model-based approaches have grown in popularity over the years, particularly with HMMs emerging as a focal point in the realm of sequence modeling. In an early study by Smyth [45], the problem of sequence clustering was viewed as a generalization of the standard mixture model approach to clustering in feature space and developed an HMMs-based solution. Likewise, Bicego et al. [10] also employed the HMMs to cluster sequence data. They quantified the similarity between sequences as the proximity of the represented vectors in a new representation space generated using HMMs. Moreover, Xiong et al. [52] proposed a non-HMMs Markov model to approximate the **conditional probability distribution (CPD)** model, which was subsequently employed to develop a novel two-tier Markov model to represent a sequence cluster.

Utilizing this framework, a divisive hierarchical algorithm was implemented to cluster categorical sequences.

Another relevant study was identified in the work by Chen et al. [15] on playlist generation. In contrast to previous studies that concentrated on categorical sequence clustering, the authors opted to address the playlist generation issue by employing a **Latent Markov Embedding (LME)** model. In their approach, sequences (playlists) are treated as Markov chains in a latent space, and each element (song) is represented as a point in the same space. The LME algorithm does not consider sequence elements as isolated entities without metric properties; instead, it provides generalized representations within a Euclidean space. The concept of sequence modeling can be effectively applied to address the problem of sequence clustering. Essentially, our proposed solution in this paper builds upon their idea. In our modeling of sequences, we regard each job term (as opposed to job title) as a data point within the latent space. This approach aims to enhance the representation of semantic relationships among job titles, particularly in cases where they share common or resembling terms.

In summary, our methodology offers distinct advantages compared to prior research when it comes to categorical sequence clustering problems. First, our solution, unlike OMA-based methods, does not require a job title coding scheme and provides greater flexibility through the transformation of job titles into meaningful latent representations. Second, our model integrates the process of sequence clustering with the learning of job term/title embeddings. Our clustering approach is considered “soft” as job sequences are not rigidly assigned to clusters but rather are probabilistically associated with clusters based on the likelihood of the cluster model. Third, our model can effectively manage more complex scenarios of career path clustering, including job sequences with overlapping employment records, through the utilization of the embedding technique. And the job title embeddings derived from this process encapsulate valuable semantic information, proving beneficial for further research endeavors.

In addition, our empirical analysis utilizes the career path clusters to explore gender inequality in the career advancement of IT professionals. Our research thus joins the literature on the discussion of gender differences in IT workforce. Prior literature in IT workforce research has paid some attention to gender differences in IT professionals regarding their commitment [18], turnover [33], roles of workplace culture [51], job embeddedness [42], and retention [48]. Our research contributes to the research of gender differences in IT workforce [e.g., 19, 21] where relatively little attention is paid to career advancement with a handful of exceptions [e.g., 6, 34].

3 Data Description and Summary

Our data are collected from *LinkedIn*, the world’s largest professional networking platform, which enables members to create occupational profiles and network with other professionals. As of February 2024, it had 1 billion members from more than 200 countries and territories worldwide.¹ Our study focuses on IT-related career trajectories through intentionally limiting the *industry keywords* to a predefined list (Figure A.1 in Appendix A). We gather data by randomly collecting LinkedIn user profiles that include any of the industry keywords from this list. Around 80% of users are from within the following four industry sectors: *Information Technology and Services*, *Computer Software*, *Telecommunications*, and *Internet*. This process left us with a dataset containing around 61,000 individuals.

In our study, we explore three specific types of career-related constructs, that is, *job sequence*, *job title*, and *job term*. A *job sequence* pertains to the career trajectory of a professional, delineated as a sequential lineup of job titles. A *job title* is a descriptive phrase that clearly defines the functionality

¹<https://about.linkedin.com/>

Table 1. Top 20 Most Frequent Terms in *RES* and *FUN*

<i>RES</i>	manager, engineer, senior, analyst, consultant, developer, director, lead, specialist, assistant, technician, administrator, president, architect, executive, owner, programmer, principal, associate, founder
<i>FUN</i>	software, technical, support, business, project, systems, development, network, operations, system, service, services, technology, product, engineering, marketing, web, customer, team, program

Table 2. Descriptive Statistics for Our Dataset

Variable	Count	Variable	Mean	Std	Min	Max
# Career paths (talents)	61,505					
# Female talents	16,108					
# Distinct job titles	48,359	# Job titles per career	6.97	3.58	1	28
# Distinct job terms	825	# Job terms per job title	2.55	1.20	1	13
# Distinct <i>FUN</i> terms	572	# Years per career (tenure length)	15.79	6.69	1	29
# Distinct <i>RES</i> terms	253	Employment year			1,990	2,018

and responsibility of a particular role, such as “software engineer” and “project manager”. *Job terms* refer to the specific words that make up a *job title*. Job terms can be classified into two genres [32]: *functionality (FUN)* and *responsibility (RES)*. For instance, the terms “software” and “engineer” contribute to the job title “software engineer”, where “software” is a *FUN* term and “engineer” is a *RES* term. The job titles found in all employment records within our data sample are quite diverse, with a total of 48 K different phrases. These job titles include 825 distinct terms, 572 of which are *FUN* terms and 253 of which are *RES* terms. We present a list of the twenty most commonly found terms in the genres of *RES* and *FUN*, as detailed in Table 1. Examples of *RES* terms include manager, engineer, senior, and analyst, while *FUN* terms include software, technical, systems, and product.

Table 2 presents a comprehensive overview of the descriptive statistics of data sample. On average, a job title consists of 2.55 terms, while a job sequence comprises 6.69 titles. The distribution of job sequence length (number of job titles) (Figure A.2 in Appendix A) reveals that a significant majority of job sequences consist of ten or fewer job titles. And the distribution of career tenure (Figure A.3 in Appendix A) highlights that a predominant portion of careers last between 5 to 15 years. Talents typically have an average career tenure of 15.79 years. The most frequently used job titles are software engineer, software developer, consultant, and manager (as shown by a Word Cloud in Figure A.4 in Appendix A), reflecting our study’s focus on the IT sector.

4 Methodology

4.1 Problem Formulation

We consider a set of career path sequences $\{S_n : n = 1, 2, \dots, N\}$, where $S_n = \langle s_0^n, s_1^n, s_2^n, \dots, s_{|S_n|}^n \rangle$ is the n th sequence, and $s_i^n \in \mathcal{S}$ is the i th job title in the sequence and $\mathcal{S}$ is the set of unique job titles from all career path sequences. Note that the length of each sequence $|S_n|$ varies for different job sequences. Since job titles can be collected on a regular basis (e.g., quarterly or yearly) or just when a job change is detected, the sequence length can reflect the duration of a career or the frequency of job changes, respectively. Given a large number of such career path sequences, we seek to construct M career path clusters $\{C_1, C_2, \dots, C_M\}$, where each cluster can capture similar career path dynamics for a set of talents. The key mathematical notations used in this document are outlined in Table 3.

4.2 Sequence Clustering with Embedded Markov Models

Our method is designed to construct M clusters of career paths while simultaneously estimates the **Embedded Markov Model (EMM)** parameters for each cluster. To cluster career paths into

Table 3. Mathematical Notations

Variable	Remarks
S	set of career path sequences
N	total number of career path sequences
M	number of clusters
S_n	n th career path sequence
s_i^n	i th job title in the n th sequence
w	job term in job title s_i^n
C_m	m th career path cluster
θ_m	cluster-wise EMM parameters for cluster C_m
D	dimensionality of the embedding space
$v_m(s_i^n)$	embedding of job title s_i^n in cluster m
$v_m(w)$	embedding of job term w in cluster m

M clusters, we randomly assign all paths to M disjoint subsets as initial clusters and then iterate the following two steps until convergence:

Step 1: Estimate the cluster-wise EMM parameters θ_m for all paths in cluster C_m , for $m = 1, 2, \dots, M$. The cluster C_m is defined as:

$$C_m \leftarrow \{S_n : \arg \max_{m'} \Pr(S_n | \theta_{m'}) = m\}. \quad (1)$$

Step 2: Assign each path S_n to the cluster C_m with the highest likelihood, i.e.:

$$m \leftarrow \arg \max_{m'} \Pr(S_n | \theta_{m'}). \quad (2)$$

For better illustration purposes, Figure 2 provides a holistic view of our methodology. We start by discussing the EMM and the probabilistic model $\Pr(S_n | \theta_m)$ ((A) *Transition Probability Computation* in Figure 2). Note that θ_m represents the parameters of cluster-wise EMM, which is estimated to maximize the likelihood:

$$\Pr(\theta_m | C_m) \propto \Pr(C_m | \theta_m) \Pr(\theta_m), \quad (3)$$

where

$$\Pr(C_m | \theta_m) = \prod_{S_n \in C_m} \Pr(S_n | \theta_m). \quad (4)$$

The parameterization with $\{\theta_m : m = 1, 2, \dots, M\}$ and the consequent optimization problem are the primary challenges in our methodology development.

Modeling career paths and job transitions with the Markov assumption is a natural approach, where the probability of the career sequence S_n is decomposed into the product of the transition probabilities $\Pr(s_{i-1}^n \rightarrow s_i^n)$:

$$\Pr(S_n | \theta_m) = \prod_{i=1}^{|S_n|} \Pr(s_{i-1}^n \rightarrow s_i^n | \theta_m). \quad (5)$$

A straightforward next step to estimate transition probabilities is to count transition frequencies between job titles of each cluster's career sequences. However, this naive estimation will deem distinct job titles completely independent. As explained in Section 1, there are important semantic links among complex job titles. To capture these semantic links among job titles, we attempt to

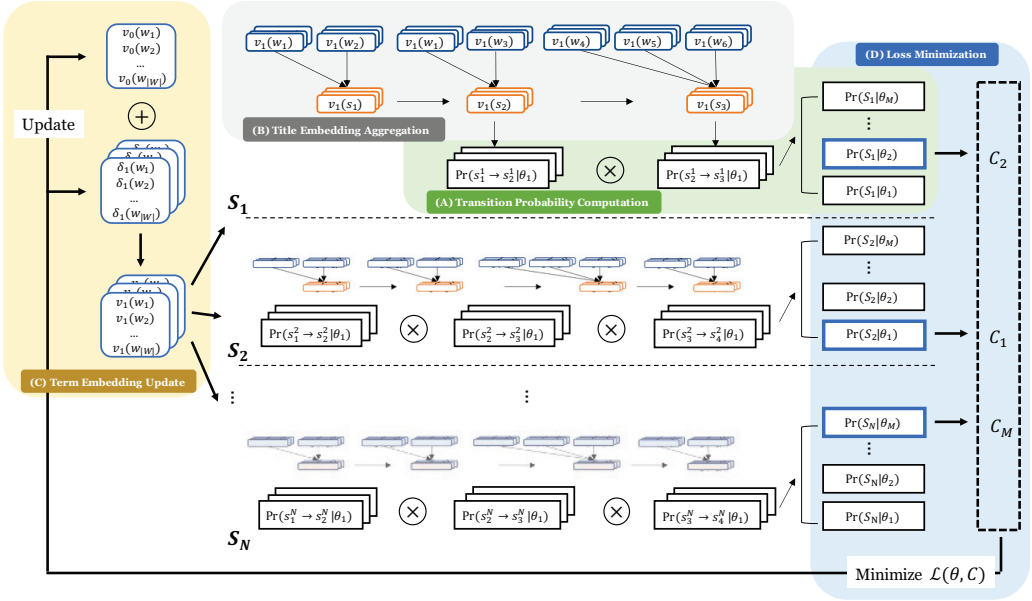


Fig. 2. An illustrative diagram of our EMMs model.

learn their embeddings in an Euclidean metric space. To be specific, we define:

$$\Pr(s_{i-1}^n \rightarrow s_i^n | \theta_m) = \frac{\exp(-\|v_m(s_i^n) - v_m(s_{i-1}^n)\|_2^2)}{Z_m(s_{i-1}^n)}, \quad (6)$$

where

$$Z_m(s_{i-1}^n) = \sum_{s \in \mathcal{S}} \exp(-\|v_m(s) - v_m(s_{i-1}^n)\|_2^2). \quad (7)$$

By this definition, job title embeddings are the optimal locations of job titles in a Euclidean metric space such that the probabilistic likelihood of the sequence cluster C_m in our data is maximized. Finally, using our EMM approach, we have:

$$\Pr(S_n | \theta_m) = \left(\prod_{i=1}^{|S_n|} \frac{\exp(-\|v_m(s_i^n) - v_m(s_{i-1}^n)\|_2^2)}{Z_m(s_{i-1}^n)} \right)^{1/|S_n|}. \quad (8)$$

Note that the probability is normalized by the sequence length $|S_n|$, ensuring that our model will not favor short sequences. The following subsections further elaborate on the details of the job title embedding component to address the unique challenges of career path clustering tasks.

4.3 Embedding of Job Titles via Job Terms

One difficulty in assessing self-reported career profiles is navigating the inherent complexity and ambiguity in raw job titles. Our EMM architecture utilizes a metric embedding approach to address this difficulty. To further reduce the modeling complexity, our method is capable of embedding job titles at the term level. For instance, given job titles “software engineer” and “hardware engineer”, our method learns embeddings of the terms “software”, “hardware”, and “engineer”. The embedding of a job title is then computed by aggregating the embeddings of the job title’s terms, for example, “software engineer” is embedded by aggregating the embeddings of “software” and “engineer”. The key idea behind this type of term-level embedding is that “engineer” can share the

same metric embedding, provided that both job titles indicate semantically similar skills and professional positions. Meanwhile, the embeddings of “software” and “hardware”, respectively, capture the distinctions between the two job positions.

Specifically, we consider each job title s_i^n as a set of terms $\{w \in s_i^n\}$ ((B) *Title Embedding Aggregation* in Figure 2). To compute the embedding $v_m(s_i^n)$ of job title s_i^n in cluster m , we learn and aggregate the term-level embeddings:

$$v_m(s_i^n) = \frac{1}{|s_i^n|} \sum_{w \in s_i^n} v_m(w), \quad (9)$$

where $v_m(w) \in \mathbb{R}^D$ is trainable embedding vector in the EMM learning process and D is the dimensionality of the metric embedding space.

According to Table 2, using term-level embeddings can reduce the number of trainable embeddings from 48,359 (distinct titles) to merely 825 (distinct terms), which is a considerable reduction in modeling complexity. In our empirical studies, we will investigate the performance of the term-level embedding approach from various aspects.

Additionally, our term-level embedding enables us to investigate the responsibility and functionality of professionals, which would be impossible with a title-level embedding approach. As we will discuss in Section 5, *RES* term embeddings computed using our approach can reveal prominent semantic meanings in contrast to other pre-trained language embeddings. In a nutshell, job terms with similar *responsibilities* appear closer in the embedded space.

4.4 Shared Embedding for Multiple EMM Clusters

The term-level embeddings $\theta_m = \{v_m(w) : w \in \mathcal{W}\}$, where $\mathcal{W}$ is the set of unique job terms, are designed to capture the job transition probabilities via a Euclidean metric space for the job sequences in the m th cluster. On the one hand, the cluster-specific embedding space enables the model to better fit the cluster-level job transition dynamics. However, on the other hand, the cluster-independent estimates of the embedding vectors add complexity to the modeling and deteriorate the learning efficiency. To achieve a good balance between the modeling capacity and complexity, we resort to the intuition of Hierarchical Bayes [7]. Specifically, we assume that the cluster-level embeddings $v_m(w)$ follow a Gaussian distribution with the global-level embeddings $v_0(w)$ as the distribution mean. We thus define ((C) *Term Embedding Update* in Figure 2):

$$v_m(w) = v_0(w) + \delta_m(w), \quad (10)$$

where $\delta_m(w)$ is the difference between the term embeddings at the cluster and global levels, respectively. By assuming $\delta_m(w) \sim \mathcal{N}(0, \sigma_m)$, our model penalizes discrepancies in the embeddings of the same term w across different clusters, essentially allowing the clusters to share information for parameter estimation in the clustering phase.

Formally, let $\theta = \{\theta_m : m = 1, 2, \dots, M\}$, our shared embedding approach adopts the prior distribution of the embedding vectors where the probability density is:

$$\Pr(\theta) \propto \prod_{m=1}^M \exp\left(-\frac{\|v_0(w)\|^2}{2\sigma_0^2}\right) \cdot \exp\left(-\frac{\|\delta_m(w)\|^2}{2\sigma_m^2}\right), \quad (11)$$

where the first term assuming $v_0(w) \sim \mathcal{N}(0, \sigma_0)$ is used to control the overall modeling complexity, and the second term assuming $\delta_m(w) \sim \mathcal{N}(0, \sigma_m)$ is used to learn shared embeddings for multiple clusters.

4.5 Overall Optimization and Implementation Details

With shared embeddings among a mixture of EMM clusters, we denote our method as **Embedded Mixture Markov Models (EMMMs)**. To estimate the term-level embeddings $\theta = \{\theta_m = \{v_m(w) : w \in \mathcal{W}\}, m = 1, 2, \dots, M\}$, in the EMMMs, we use the *Negative Log Likelihood* to define the loss function as *(D) Loss Minimization* in Figure 2):

$$\begin{aligned} \mathcal{L}(\theta, C) &= - \sum_{m=1}^M \log(\Pr(C_m | \theta_m)) - \log(\Pr(\theta)) \\ &= - \sum_{m=1}^M \sum_{S_n \in C_m} \log(\Pr(S_n | \theta_m)) - \log(\Pr(\theta)) \\ &= \sum_{m=1}^M \sum_{S_n \in C_m} \frac{1}{|S_n|} \sum_{i=1}^{|S_n|} (\|v_m(s_i^n) - v_m(s_{i-1}^n)\|_2^2 + \log(Z_m(s_{i-1}^n))) - \log(\Pr(\theta)) \end{aligned} \quad (12)$$

where we have $\log(\Pr(\theta)) = -(\frac{M}{2\sigma_0^2} \|v_0(w)\|^2 + \sum_{m=1}^M \frac{1}{2\sigma_m^2} \|\delta_m(w)\|^2)$ according to Equation (11) and after dropping a constant term. The hyper-parameters in the loss function include the number of clusters M , the embedding dimensionality D , and two regularization parameters σ_0 and σ_m .

We provide our algorithm's pseudocode (Algorithm 1 in Appendix B) to optimize the loss function in Equation (12) with stochastic gradient descent procedures implemented using PyTorch [37] and the Adam optimizer [29]. Following the common practice in using these tools, we employ an automatic learning rate scheduler and meanwhile tune the weight decay parameter in Adam optimizer to achieve a similar effect of tuning our hyper-parameters θ_0 and θ_m .

Although the optimization problem in Equation (12) is not convex, it is straightforward to show that Algorithm 1 in Appendix B will converge to a local optimum. At each iteration, we first update $\theta \leftarrow \theta' = \{\theta'_m\}$ with a fixed sequence assignment $C_m = \{S_n : \arg \max_{m'} \Pr(S_n | \theta_{m'}) = m\}$ and current model parameters $\theta = \{\theta_m\}$, where $m = 1, 2, \dots, M$. Note that, given the fixed sequence assignment, our model parameters are learned using the maximum log-likelihood method. Thus, we have $\mathcal{L}(\theta', C) \leq \mathcal{L}(\theta, C)$. In the second step, given the fixed embeddings θ' , each sequence S_n in cluster C_m will be assigned to a new (and possibly different) cluster $C_{m'}$ with the maximal likelihood:

$$\Pr(S_n | \theta'_{m'}) \geq \Pr(S_n | \theta'_m). \quad (13)$$

Thus, we have:

$$\begin{aligned} \mathcal{L}(\theta', C') &= - \sum_{m=1}^M \sum_{S_n \in C_{m'}} \log(\Pr(S_n | \theta'_{m'})) - \log(\Pr(\theta')) \\ &\leq - \sum_{m=1}^M \sum_{S_n \in C_m} \log(\Pr(S_n | \theta'_m)) - \log(\Pr(\theta')) \\ &= \mathcal{L}(\theta', C) \leq \mathcal{L}(\theta, C). \end{aligned} \quad (14)$$

We will provide more empirical convergence analysis in the next section.

5 Career Path Clustering in the IT Sector

In our experimental analysis, we first conduct experiments using synthetic datasets. This aims to provide a thorough examination of different aspects of our model and validate the efficiency of our method. The results of the experiments on synthetic datasets show that our model is able

to achieve a high clustering performance, compared with other conventional sequence clustering methods (e.g., OMA). More details are available in Appendix D.

In the following sections, we apply our model on a representative sample of real-world IT professionals and present several interesting findings for discussion. Similar to other conventional clustering models (e.g., K-means), one critical hyperparameter in our model is the number of clusters. We adhere to a thorough process delineated in Appendix C.1 to determine the most suitable number of clusters for our dataset. The resulting number of clusters is $K = 8$. We now proceed to a series of analyses to understand the characteristics of these clusters of career paths.

5.1 Divergence of Clusters

A key benefit of our method is the flexibility of job title representations achieved through the utilization of embedding techniques. Based upon the right assignment of career paths to clusters, our model tends to bring closer the job titles from the same sequence in the high-dimensional embedded space. Simply put, if our clustering method is successful, the distance covered by a career path within its designated cluster should be shorter than the distance covered if it were assigned to a different cluster. In this context, we formulate the following model evaluation metric. Based on the clustering results, we compute job title embeddings in each cluster using Equation (9). Then, we calculate the “length” of a job sequence in the embedded space as the sum of *euclidean distances* between successive job titles. The “average length” of job sequences is thus defined as L_m in cluster C_m . On the other hand, for each cluster C_m , we estimate the “dispersion” of job titles as the average distance of each point (representing a job term) to the cloud center, denoted as D_m . Finally, we define the cluster’s *career path divergence score* (CPD_m) as the average sequence length normalized by the job title dispersion, that is:

$$CPD_m = \frac{L_m}{D_m}. \quad (15)$$

It is noteworthy that a lower divergence score suggests more effective clustering process, as the career paths assigned to each cluster are *shorter* in the embedded space.

As a benchmark, we construct another model that groups job sequences into a *single* cluster C^s and generates new embeddings for the same job titles. Our goal is to demonstrate the effectiveness of our clustering approach in accurately categorizing job sequences into appropriate clusters, as opposed to the scenario where no clustering is applied (referred to as the single-cluster case). This assertion is affirmed if we observe a notable reduction of the job sequences length within the embedded spaces. We calculate the *career path divergence score* of the same sequence assignments using the eight-cluster embeddings (denoted as CPD_m) and the single-cluster embedding (denoted as CPD_m^s), respectively. In Figure 3, we compare CPD_m and CPD_m^s for each cluster as well as their respective average CPD , represented by the horizontal lines. The figure reveals that grouping job sequences into appropriate clusters significantly reduces divergence scores, demonstrating the efficacy of our model.

Next, we utilize PCA [2] as a dimensionality reduction technique to transform high-dimensional embeddings into two-dimensional vectors. This process enables us to visualize career paths as trajectories within the embedding space. As a case study, we select one illustrative career path (identified with high log-likelihood) from each cluster and plot their projections on the two-dimensional plane in Figure 4, where each point represents a projected job title. These career paths are depicted on two separate sets of two-dimensional planes, corresponding to the results derived from (i) the single-cluster model and (ii) the eight-cluster model. The figure in the top-left corner depicts the outcomes of the single-cluster model for the eight exemplary career paths, while the subsequent eight figures display the separate clusters identified through the eight-cluster model. It is evident

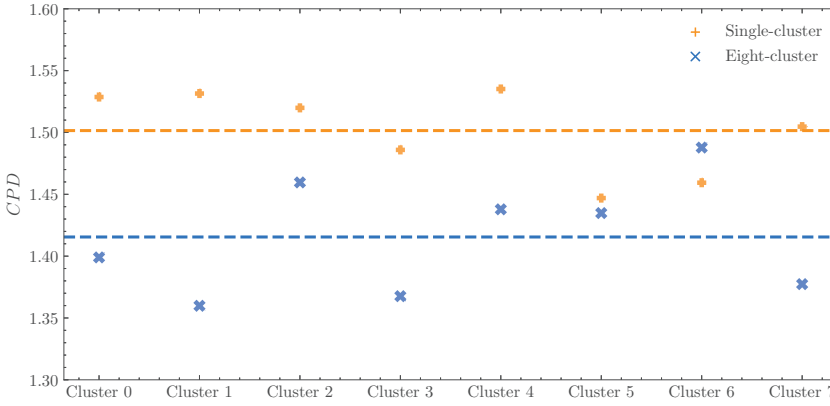


Fig. 3. Divergence scores from the eight-cluster and single-cluster models.

that the connected dots in the first figure are considerably more scattered compared the dots in the other eight figures. This example once more showcases the effectiveness of our model in learning cluster-specific job title representations and producing proper job sequence assignments.

5.2 Cluster-wise Job Title Distribution

To better understand the unique characteristics of career paths inside each cluster, we visualize the majority of job titles for each cluster using *Word Cloud* plots in Figure 5 (in contrast to the holistic view in Figure A.4 in Appendix A). Upon initial observation, we have identified commonalities among Cluster 0, 3, and 6, as they all have a substantial amount of job titles associated with the role of a *manager*. With further analysis, we observe that Cluster 0 comprises individuals who possess extensive experience in software development, whereas Cluster 3 primarily comprises individuals with a background in network engineering. In Cluster 6, there is a notable trend of career progression from the role of a *software engineer* to that of a *project manager*. On the other hand, the job titles in Clusters 2 and 4 are more distinctive compared to others. Cluster 2 offers various career paths that lead to upper management positions such as *president* and *executive*. Cluster 4 is distinctive as it includes a higher concentration of job titles associated with the roles of a *business analyst* and *consultant*.

In addition, Clusters 1, 5, and 7 appear to be similar as they contain more job titles related to *software engineer* and *senior software engineer*. To further examine the differences among the three clusters, we have utilized a heatmap to visualize the distributions of *starting* job titles, as illustrated in Figure 6. An interesting observation is that, within Cluster 1, individuals typically start their professional career in the capacity of an *intern*, subsequently advancing to positions of *software engineer* and *senior software engineer*. Clusters 5 and 7 are less distinguishable, showing a higher similarity in career trajectories that originate from a background as a *software engineer*. In this regard, we have created a Sankey diagram (Figure C.3 in Appendix C.2), which illustrates all career paths stemming from the initial role of a *software engineer*. It is observed that most professionals starting with the job title of a *software engineer* are categorized within Clusters 5 and 7, consistent with the pattern illustrated in Figure 6. Additional scrutiny reveals that Cluster 7 pertains to individuals who persist in a regular *software engineer* position for an extended duration. In contrast, Cluster 5 captures those who attain promotions to *senior software engineer* or *lead software engineer* within a shorter timeframe. Cluster 6 also includes a significant number of *software engineer* titles, with employees who have been promoted to *senior software engineer* after a relatively longer

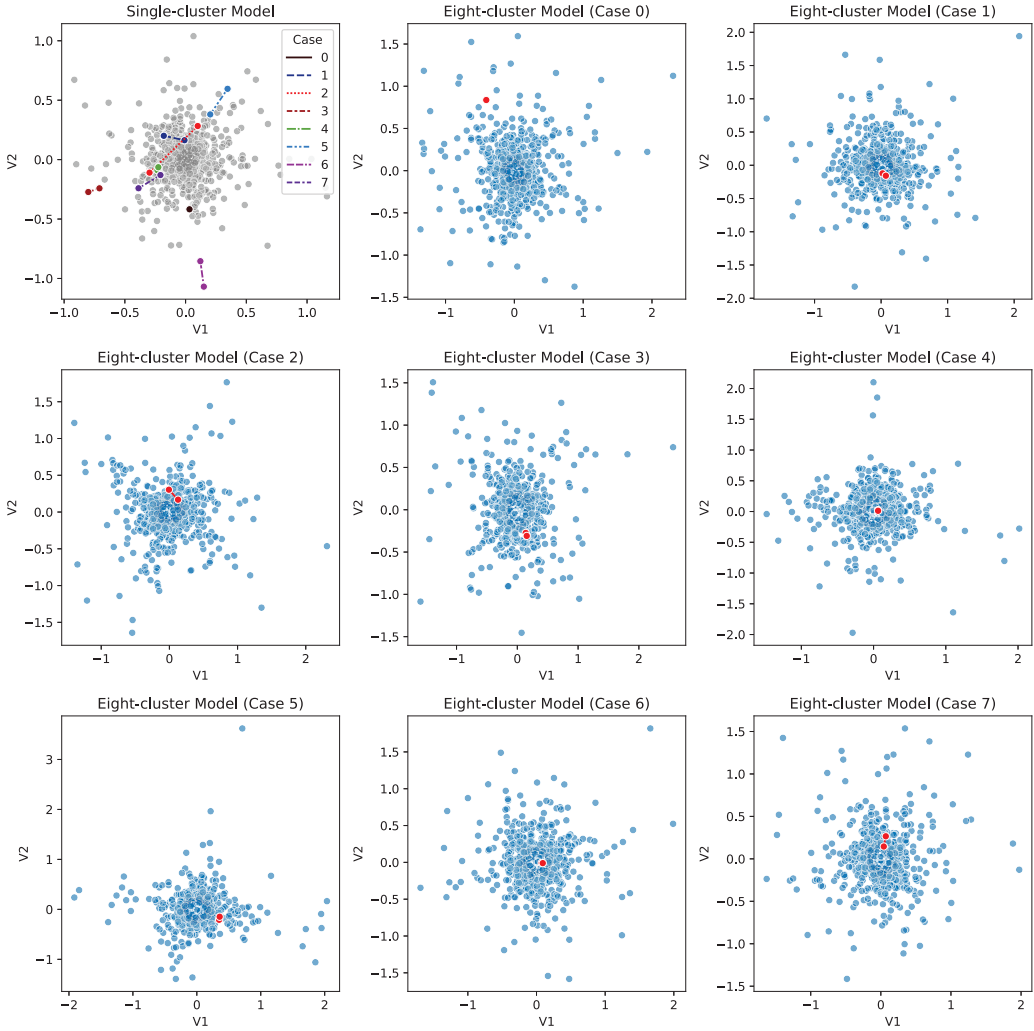


Fig. 4. Visualization of career paths on two-dimensional planes.

period. Our findings highlight the effectiveness of our approach in delineating distinct career paths, even in cases where employees share similar job titles.

5.3 Job Term Embeddings

In addition to job sequence clustering, our model is designed to generate global job term embeddings $v_0(w)$ that contain shared knowledge of job titles among clusters. In this section, we will delve deeper into the job of *RES*. In general, job titles like *intern* and *assistant* are considered *junior-level* positions; *associate* and *manager* are *senior-level* ones; and *CEO*, *CFO*, *CTO*, and *partner* associate with *top(executive)-level* positions. The purpose of this investigation is to examine if the job embeddings derived from our model can reveal implicit insights into job hierarchy. We begin by scattering and labeling several representative job terms based on their two-dimensional projections derived from PCA in $v_0(w)$ as illustrated in Figure 7(a). The visualization indicates that the *top-level* terms are centralized in the plot, surrounded by the *junior-level* terms. The distribution of

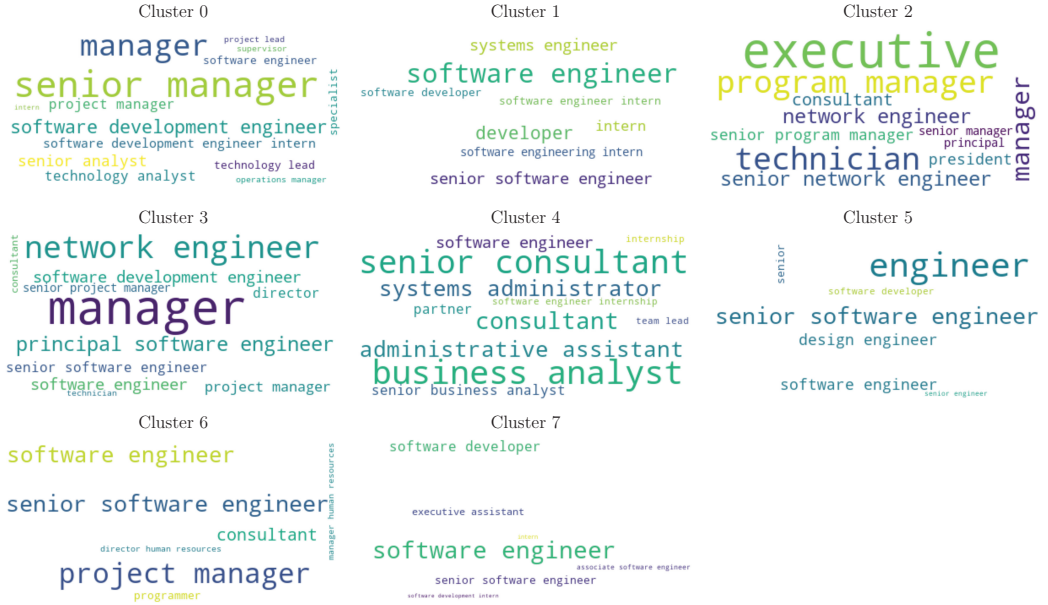


Fig. 5. Word cloud of job titles by cluster.

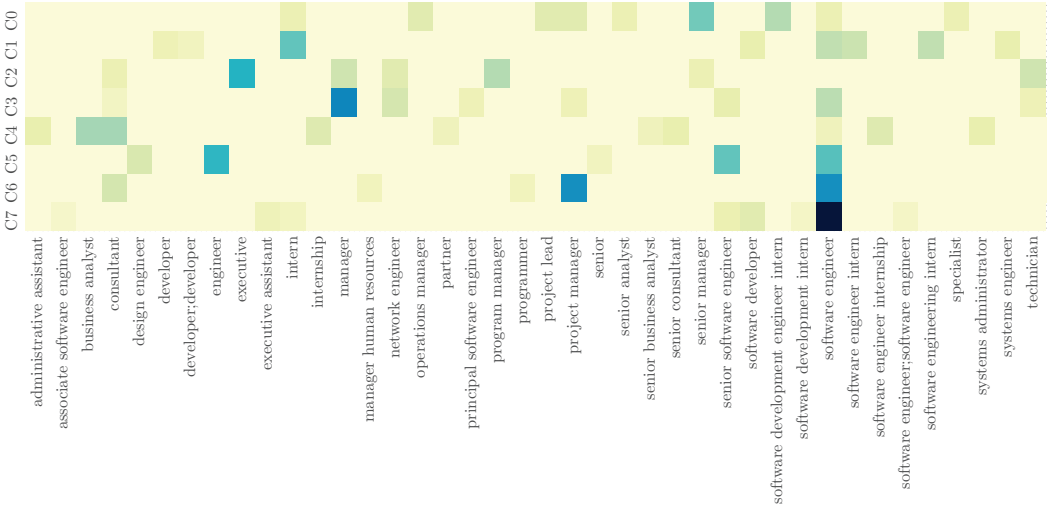


Fig. 6. Distribution of starting job titles in different clusters.

these job terms reveals their respective hierarchical levels. For comparison, we use a pre-trained word embedding model *glove-wiki-gigaword-50* to produce new embeddings of these same job terms. This model is based on GloVe [38] and was trained using the Wikipedia 2014 + Gigaword 5 dataset with an embedding dimensionality of 50.² In a similar manner, we distribute these job terms on another two-dimensional plane based on their PCA transformations in Figure 7(b). In this plot, there is no clear hierarchical pattern observed in how these job titles are distributed.

²<https://nlp.stanford.edu/projects/glove/>

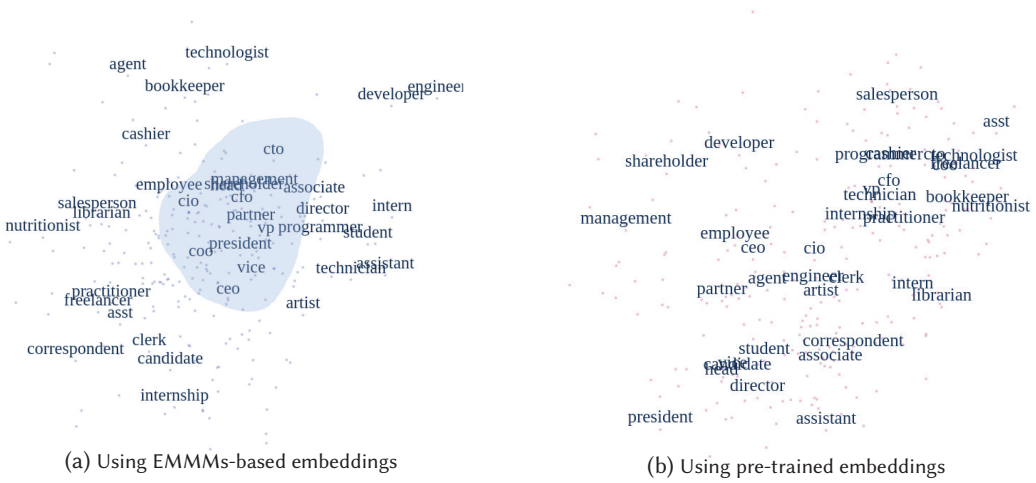


Fig. 7. Scatter plots of the *RES* terms on two-dimensional planes.

6 Cluster and Gender Differences in Career Advancement

In this section, we aim to further study the efficacy of our clustering method with an emphasis on relating to gender differences in career advancement of the IT workforce. We first implement a semi-supervised learning method to find the scores of several important aspects of the job titles, including title change frequency, technicality, hierarchy, independence, and entrepreneurship. The first score is the average year per title change, which measures the average duration of a title before a change to the next. This score is the reciprocal of the number of title changes per year. A longer average duration between consecutive titles indicates a less frequent change of titles on a yearly basis. The second score is on the technicality of the job title, which ranges from 1 (a pure technical title) to 0 (purely administrative or management job title). The third score is on the hierarchy of a job title, which ranges from 5 (executive) to 1 (entry-level), with 4 (director), 3 (manager), and 2 (senior) in between. The hierarchy score reflects the level of the position of an individual in a company. The fourth score is on measuring the independence of a job title, which ranges from 1 (a contractor or consultant) to 0 (full-time employee). The independence score reflects the extent of IT worker operates in a self-employed fashion. The fifth and last score is on measuring entrepreneurship, which ranges from 1 (founder or co-founder of a company) to 0 (employee). The entrepreneurship score reflects not only workers' experience in founding new businesses but also a distinct career choice.

The semi-supervised score construction, although not perfect, can support our investigation by providing an expedient and efficient way of measuring IT career advancement. The semi-supervised learning process works as follows. The first step involves identifying the top 300 most frequent or prominent job titles, denoted as S_F . Then, all authors of this paper assign *ratings* to each aspect of the job titles. And the *rating* with the highest number of votes for each aspect of the job title is selected as its *expert score*. To obtain scores for job titles in S_{-F} (not in S_F), we compute their scores for each aspect by taking the weighted average of the *expert scores* of the most similar job titles in S_F (based on cosine similarity). The process can be broken down into the following steps: (1) we retrieve the embeddings for all job titles that were learned from the EMMs model; (2) we calculate the cosine similarity between the embedding of each job title within S_{-F} and all of the job title embeddings found in S_F ; (3) we use the cosine similarity measures as weights and proceed to compute the weighted average of the corresponding expert scores, which then acts as

the estimated score for each job title within S_F . This semi-supervised learning process enables us to generate five aspect scores for each job title in the dataset.

To explore the gender and cluster differences of the five metrics, we then use statistical procedures to compare the different career advancement metrics across either genders (see Figure 8), clusters (see Figure 8), or both (see Figure 9). We construct 95% confidence intervals for each subgroup of individuals, and all confidence intervals are adjusted to have a family-wise confidence interval of 95%, which means there is a family-wise confidence level of 95% for any number of confidence intervals to be valid. To be more specific, the two confidence intervals of female and male groups can simultaneously be correct with 95% confidence; the 8 confidence intervals of different clusters are guaranteed to be simultaneously correct with 95% confidence; and the 16 confidence intervals of the exhaustive combination of gender and cluster levels are controlled to be simultaneously valid with 95% confidence. The sole difference between the three is lying on the number of simultaneous confidence intervals. We use the conservative Bonferroni procedure so that the family-wise chances of getting false confidence intervals are controlled to be within 5% [35]. We will now examine in the following the results for each metric. It is important to note that all results and interpretations are based on our curated sample. Therefore, any effort to apply these insights beyond this sample should be approached with careful consideration of their validity and applicability.

6.1 Number of Years Per Title Change

First, we observe that male employees tend to have a shorter duration (0.09 years or 1 month) between changes in job titles, as indicated in Figure 8. In other words, on average, men change job titles slightly more frequently than women. This metric is also shown to be quite different for the 8 clusters. For example, the longest and shortest duration between job titles is approximately 2.4 and 2.1 years, about 3.5 months apart.

Second, by using multiple comparison procedures, we are able to group clusters based on pairwise comparisons into three groups: cluster 0 in Group I, cluster 6 in Group II, and everything else in Group III. In this example, Group I has the longest duration, Group II has the shortest duration, and all the remaining clusters have about the same mid-range duration. When gender information is also used in the clustering process, there seems to be further separation in this metric. The longest and shortest duration between job titles is approximately 2.5 and 1.9 years, 0.6 years, or 7 months apart. Multiple comparison procedures give 4 groupings. From longest to shortest, they are cluster 4 in Group I, clusters 3, 5, 6, and 7 in Group II, clusters 1 and 2 in Group III, and cluster 0 in Group IV. Note that the clustering numbers with or without gender information are independent of one another. In fact, a two-way tabulation of the frequencies by the two clustering shows quite evenly dispersed distribution, indicating the two clustering results are drastically different for individuals.

Lastly, we show the merit of the clustering approach by trying to discern the differences in an average year-per-title change with both gender and clusters considered. For more apparent comparisons, we generate a random clustering by randomly assigning an individual to a cluster. We should expect no statistical differences across random clusters, which is indeed observed in the first column of Figure 9. Our clustering approach, in contrast, effectively discerns the significant difference among the clusters (as seen in Figure 8), and further differences among the mixture of gender and clusters sub-groups (as seen in the second and third column of Figure 9). When gender information is not used in clustering (second column), the average year per title change metric, while more saliently different than in the random clustering, largely overlaps with each other from one group to another gender-cluster sub-group. Interestingly, when gender information is considered in the clustering process, the gender-inclusive clustering results (third column) reveal a

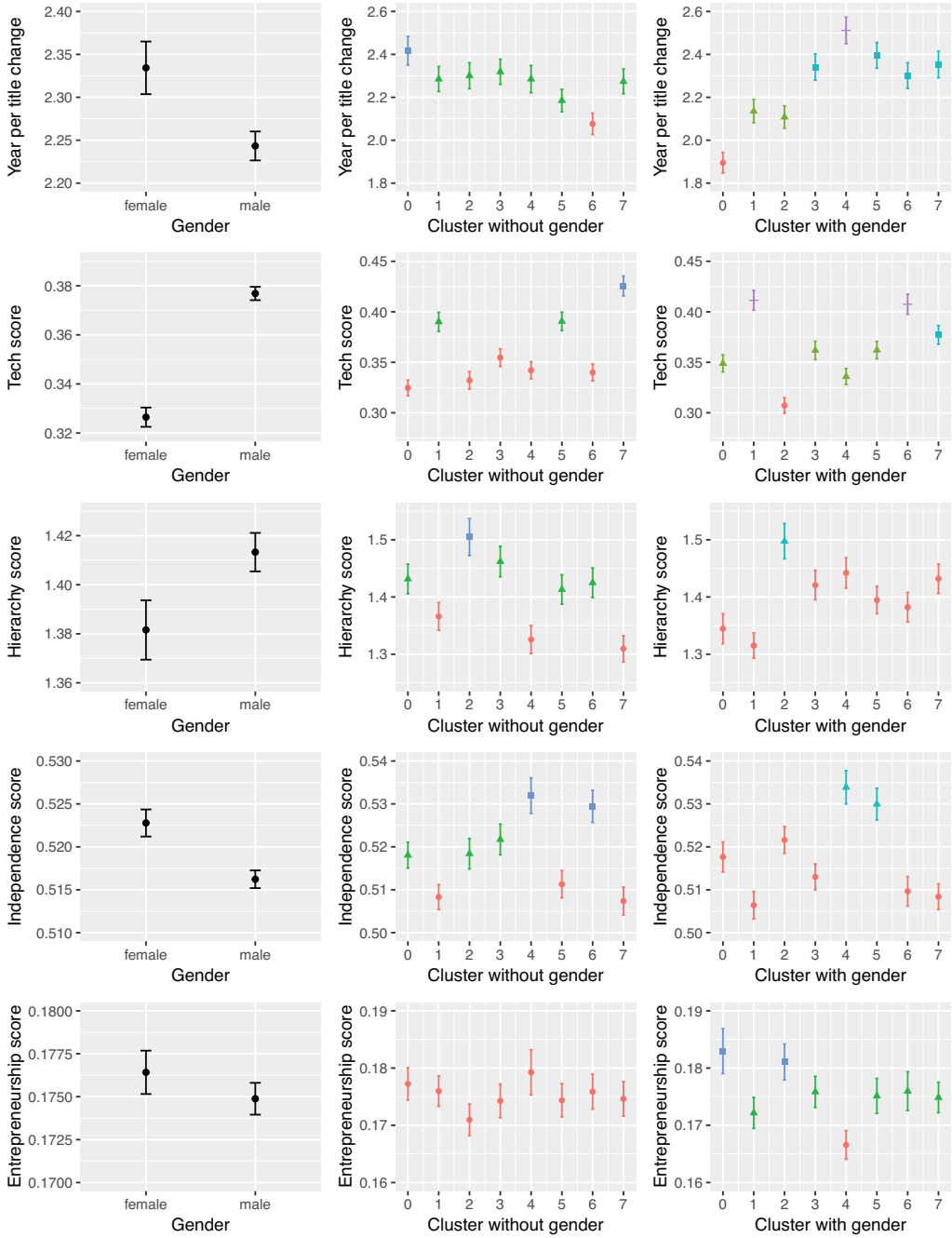


Fig. 8. Differences of all metrics across either gender or cluster.

distinct gender-specific difference within each cluster. Male employees in Cluster 0 have the shortest duration between job title changes, while the female employees in Cluster 3 and 4 tend to have the longest duration. This finding provides a more nuanced understanding of the difference in job duration among employees than simply looking at averages across different genders (see Figure 8).

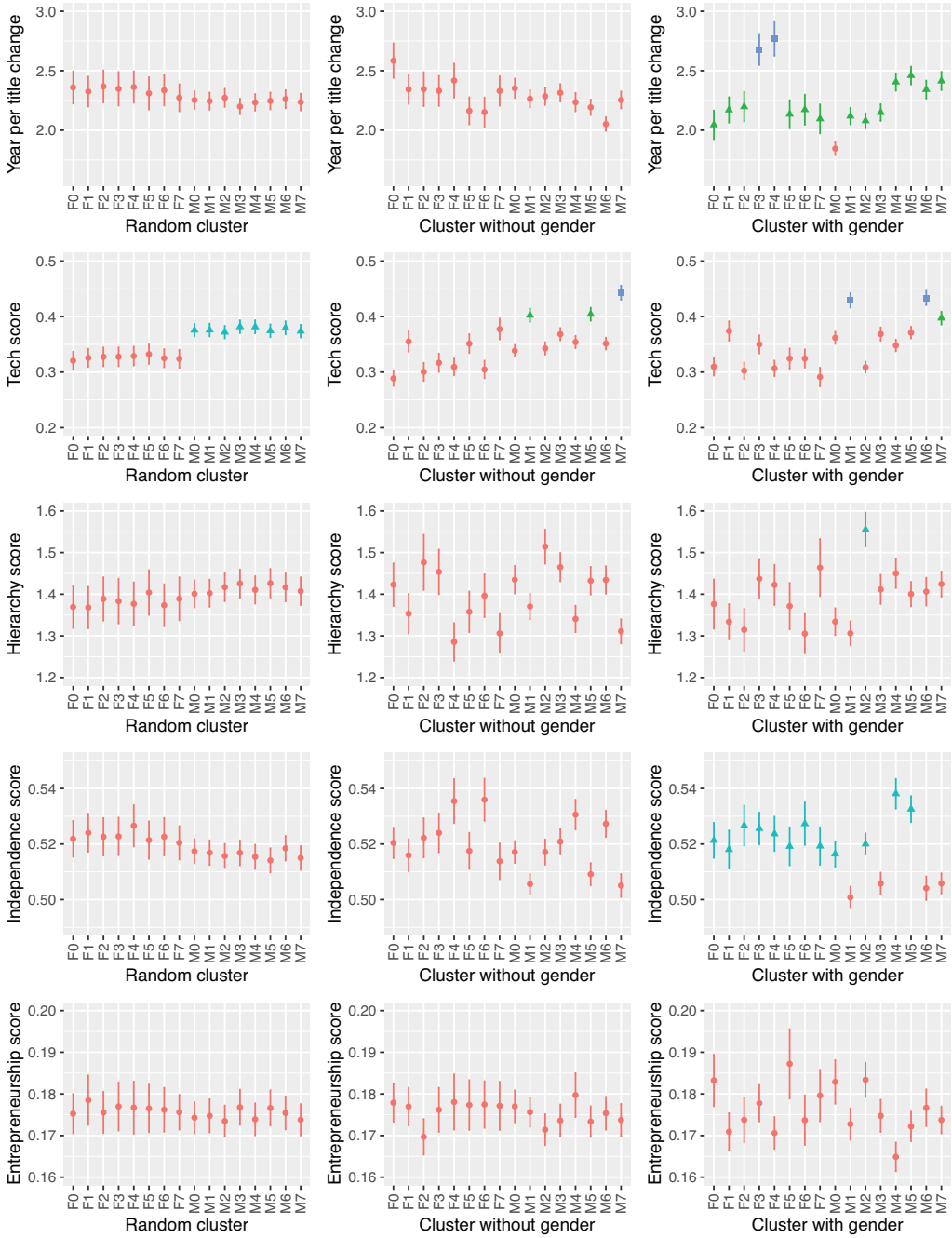


Fig. 9. Differences of all metrics across both gender and cluster.

6.2 Other Differences in Career Advancement

Current tech score. As seen in Figure 8, male employees tend more likely to work in a technical job post than female employees, the fact of which is well-known in the IT industry. In Figure 8, there is also a clear separation on the tech score across the 8 clusters, both with or without gender

in the clustering process. The gender difference is so salient in tech score that even in the random clustering results, there is a clear separation of male and female groups (see Figure 9). Additional groups are identified when both gender and cluster are considered in the two-way multiple comparisons. However, there seems to be little meaningful difference in the two clustering (with or without gender in the clustering process) results in terms of grouping and separation.

Current hierarchy score. The hierarchy score exhibits similar patterns as the tech score. In general, male IT workers tend to be in a higher hierarchy than female workers. In Figure 8, clustering processes also clearly separate hierarchy scores into groups. While the grouping numbers are fewer when gender is considered in the clustering process, the separation of groups is mainly driven by the heavy overlapping intervals of clusters 5 and 6, but not by the range of differences in the clusters. When both gender and clusters are considered (see Figure 9), the hierarchy score shows no clear separation because of the large overlapping intervals across different gender-cluster groupings. The only exception is the males in cluster 2 (clustering with gender information), who exhibit a much higher hierarchy score. Further investigation finds that this group is distinct from other groups because it contains a higher number of individuals who currently sit in the highest-ranking positions, such as executive roles.

Current independence score. The independence score seems to tell a different story for female IT workers. On average, female IT workers seem to be more likely in a consulting or contractor role than their male counterparts. This may be due to the fact that female workers unfairly bear more family obligations and face extra challenges to establish a long-term career instead of working independently. Similarly, in Figure 8, our clustering processes also separate independence scores into clear groups. While the grouping numbers are fewer when gender is considered in the clustering process, the separation of groups is mainly driven by the heavy overlapping intervals of cluster 3, but not the range of differences in the clusters. When both gender and clusters are considered (see Figure 9), the independence score shows no clear separation in clustering results without the inclusion of gender because of the large overlapping intervals across different gender-cluster groupings. When gender is considered in clustering, the clusters are distinctly put into two groups with clear separations, with the help of slightly larger range separation and less overlapping.

Current entrepreneurship score. The entrepreneurship score seems to be similar to the independence score for female IT workers. As seen in Figure 8, female IT workers seem more likely to find a company than their male counterparts. Without considering gender in clustering, the results are not discernibly different from one cluster to another. However, when gender is considered in clustering, more separation is clear—both larger range separation and less overlapping. When both gender and clusters are considered (see Figure 9), the entrepreneurship score shows no clear separation in clustering results regardless of the inclusion of gender in clustering. This is due to the large overlapping intervals across different gender-cluster groupings. When gender is considered in clustering, while the range separation helps a little, the overlapping across groups makes the separation impossible.

7 Discussions

7.1 Contributions to IS Literature

Our study contributes to the IS literature of IT-related career studies [25, 27, 31, 41]. First, we advance the study of IT-related careers by developing a novel career path clustering model, which is a non-trivial improvement over the case-based studies [41] and conventional sequence similarity-based clustering methods [27, 31]. Compared with other methods, our model can better address the high-heterogeneity issue of self-reported career records and perform career path clustering in a *fuzzy* manner. Furthermore, our method is applicable to other important sequence-related IS research, such as digital trace analysis [39] and business process mining [14]. In addition, our

research also contributes to sparse research of career advancement in IT workforce [6, 34] by uncovering some interesting gender differences. Our work also fits into the computational design science research [36, 40]. The computational design science research is concerned with “solving business and societal problems by developing computational models and algorithms” [40]. Our study develops a novel sequence clustering algorithm using sequential embedding and mixture Markov models to address the career path clustering problem. Our contributions are rooted in algorithm-centric advancement and application-oriented solutions.

7.2 Methodological Implications

Our study reveals two primary methodological implications. The first is the embedding-based design in modeling career paths. This technique stems from the recent advancement of representation learning methods, which are the core of deep learning research [8]. The embedding-based design can distill nuanced subtleties of job titles and further serve other research purposes. Our work is another innovative example of this useful technique applied to IS research problems. Another important implication of our methodological contribution relates to the use of mixture models for sequence clustering. This method is advantageous by generating fuzzy clusters and capturing more fine-grained information among different clusters. Our study demonstrates the superiority of such models and thus opens more doors to the application of mixture models for unsupervised learning problems in IS research. Given these implications, our method can be generally applied to various other sequence clustering problems. It is particularly beneficial if the sequence includes a diverse set of tokens and lacks uniformity. These scenarios often arise when analyzing sequences generated by human behavior. Examples of applications may include clustering customer purchase behavior and analyzing student learning patterns.

7.3 Clarification on Ethical Concerns

It is important to clarify that the methodology employed in our study focuses on sequence clustering analysis, which is primarily exploratory in nature. Our aim is to glean insights and patterns from the data rather than making individualized decisions, which is fundamentally distinct from predictive or prescriptive modeling. Ethical issues like biases and privacy concerns have been significantly alleviated. For example, given its design, our model does not inherently favor a specific career trajectory or pattern. The distribution of the clusters is primarily determined by the data they are derived from with the help of our model. In addition, our method would not raise additional privacy concerns. The result of our analysis is the identification of “clusters” (career path patterns) where no personal details or individual identities will be disclosed.

7.4 Limitations and Future Research

There are several limitations in our study, which can serve as a springboard for future research. First, our clustering model, like others, requires several user-specified inputs. While users are able to determine the number of clusters using the Elbow method, future research could investigate more automatic and principled choices. Meanwhile, the final clustering output is dependent on initial clusters, which are best generated using pilot models (e.g., OMA). Second, our model does not directly incorporate or differentiate between job term genres, that is, *functionality* and *responsibility*. The development of algorithms for genre-aware career path clustering is an intriguing and promising area of research. Third, our current method overlooks other job variables when learning career patterns, such as employers, industry sectors, and talents’ demographics. A more in-depth study and understanding of career patterns through the use of heterogeneous information has notable benefits for both academics and practitioners. Fourth, in this study, we have conducted a comprehensive analysis of the differences in career advancement between genders. Unfortunately,

we were unable to perform similar analysis on other factors such as education and social connections due to limited data availability. A potential future direction involves applying our method to identify **diversity, equity, and inclusion (DEI)** biases among IT professionals, once the necessary dataset becomes available. Lastly, we note that any implications derived from the dataset come from a curated dataset (i.e., a large sample of IT professionals) that is methodologically fit for our proposed method. One may exercise caution when trying to interpret any implications in context outside of our samples. For example, an industry outside of IT may exhibit different gender differences in terms of career path progression. Exploring these areas in future research could provide valuable insights into variations of individual's career advancement. Bioinformatics research also deals with sequence clustering which might be relevant to our research. Future research may distinguish or apply our approach to other fields such as gene sequence clustering.³ Interested readers may refer to review articles such as [53, 55].

8 Conclusion

In this study, we developed a novel sequence clustering approach and applied it to uncover patterns in IT career paths. Specifically, we construct an embedded Markov model for each career path cluster. The optimal assignment of career paths to the clusters is obtained by simultaneously learning cluster-wise job embeddings and optimizing the mixture Markov models. Experimental results demonstrated our model's efficacy and revealed interesting career patterns in the IT workforce. In addition, we showed that our approach can enable novel research investigation on various career advancement metrics across genders and career path clusters. The improved career path clusterings are valuable for tasks like talent recruiting and career planning. The acquired insights are also useful for educational institutes to design appropriate training programs and provide people with the skills they require to successfully navigate through long and complex career paths.

References

- [1] Andrew Abbott and Alexandra Hrycak. 1990. Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology* 96, 1 (1990), 144–185.
- [2] Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 4 (2010), 433–459.
- [3] Rajshree Agarwal and Atsushi Ohyama. 2013. Industry or academia, basic or applied? Career choices and earnings trajectories of scientists. *Management Science* 59, 4 (2013), 950–970.
- [4] C. C. Aggarwal and C. K. Reddy. 2013. *Data Clustering: Algorithms and Applications*. Taylor & Francis.
- [5] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering—a decade review. *Information Systems* 53, October–November (2015), 16–38.
- [6] Deborah J. Armstrong, Cynthia K. Riemenschneider, and Laurie G. Giddens. 2018. The advancement and persistence of women in the information technology profession: An extension of Ahuja's gendered theory of IT career stages. *Information Systems Journal* 28, 6 (2018), 1082–1124.
- [7] Neeraj Arora, Greg M. Allenby, and James L. Ginter. 1998. A hierarchical Bayes model of primary and secondary demand. *Marketing Science* 17, 1 (1998), 29–44.
- [8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.
- [9] Akshay Bhola, Kishalay Halder, Animesh Prasad, and Min-Yen Kan. 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5832–5842.
- [10] Manuele Bicego, Vittorio Murino, and Mário A. T. Figueiredo. 2003. Similarity-based clustering of sequences using hidden Markov models. In *Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 86–95.
- [11] Torsten Biemann and Joachim Wolf. 2009. Career patterns of top management team members in five countries: An optimal matching analysis. *The International Journal of Human Resource Management* 20, 5 (2009), 975–991.

³We thank an anonymous reviewer for this intriguing suggestion.

- [12] Torsten Biemann, Hannes Zacher, and Daniel C. Feldman. 2012. Career patterns: A twenty-year panel study. *Journal of Vocational Behavior* 81, 2 (2012), 159–170.
- [13] Mary Blair-Loy. 1999. Career patterns of executive women in finance: An optimal matching analysis. *American Journal of Sociology* 104, 5 (1999), 1346–1397.
- [14] Dominic Breuker, Martin Matzner, Patrick Delfmann, and Jörg Becker. 2016. Comprehensive predictive models for business processes. *MIS Quarterly* 40, 4 (2016), 1009–1034.
- [15] Shuo Chen, Josh L. Moore, Douglas Turnbull, and Thorsten Joachims. 2012. Playlist prediction via metric embedding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 714–722.
- [16] Laurie Cohen. 2006. Remembrance of things past: Cultural process and practice in the analysis of career stories. *Journal of Vocational Behavior* 69, 2 (2006), 189–201.
- [17] Katja Dlouhy and Torsten Biemann. 2015. Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior* 90, October (2015), 163–173.
- [18] Jen-Ruei Fu and Jessica H. F. Chen. 2015. Career commitment of information technology professionals: The investment model perspective. *Information & Management* 52, 5 (2015), 537–549.
- [19] Elena Gorbacheva, Jenine Beekhuyzen, Jan vom Brocke, and Jörg Becker. 2019. Directions for research on gender imbalance in the IT profession. *European Journal of Information Systems* 28, 1 (2019), 43–67.
- [20] Akshay Gugnani and Hemant Misra. 2020. Implicit skills extraction using document embedding and its use in job recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13286–13293.
- [21] Kevin Harmon and Eric Walden. 2021. Comparing three theories of the gender gap in information technology careers: The role of salience differences. *Journal of the Association for Information Systems* 22, 4 (2021), 1099–1145.
- [22] Moh Hosseinioun, Frank Neffke, Letian, Zhang, and Hyejin Youn. 2024. Nested Skills in Labor Ecosystems: A Hidden Dimension of Human Capital. arXiv:2303.15629. Retrieved from <https://arxiv.org/abs/2303.15629>
- [23] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (1985), 193–218.
- [24] Ayomikun Idowu and Amany Elbanna. 2022. Digital platforms of work and the crafting of career path: The crowd-workers’ perspective. *Information Systems Frontiers* 24, 2 (2022), 441–457.
- [25] Magid Igbaria and Sidney R. Siegel. 1993. The career decision of information systems people. *Information & Management* 24, 1 (1993), 23–32.
- [26] Faizan Javed, Phuong Hoang, Thomas Mahoney, and Matt McNair. 2017. Large-scale occupational skills normalization for online recruitment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. 4627–4634.
- [27] Joseph, Boh, Ang, and Slaughter. 2012. The career paths less (or more) traveled: A sequence analysis of it career histories, mobility patterns, and career success. *MIS Quarterly* 36, 2 (2012), 427.
- [28] Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A survey on skill identification from online job ads. *IEEE Access* 9 (2021), 118134–118153.
- [29] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- [30] Michael Koch, Bernard Forgues, and Vanessa Monties. 2017. The way to the top: Career patterns of Fortune 100 CEOs. *Human Resource Management* 56, 2 (2017), 267–285.
- [31] Theodoros Lappas. 2020. Mining career paths from large resume databases: Evidence from IT professionals. *ACM Transactions on Knowledge Discovery from Data* 14, 3 (2020), 1–38.
- [32] Junhua Liu, Chu Guo, Yung Chuen Ng, Kristin L. Wood, and Kwan Hui Lim. 2019. IPOD: Corpus of 190, 000 industrial occupations. arXiv:1910.10495. Retrieved from <https://arxiv.org/abs/1910.10495>
- [33] Janice Lo. 2015. The information technology workforce: A review and assessment of voluntary turnover research. *Information Systems Frontiers* 17, 2 (2015), 387–411.
- [34] Kimberly McGee. 2018. The influence of gender, and race/ethnicity on advancement in information technology (IT). *Information and Organization* 28, 1 (2018), 1–36.
- [35] John Neter, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. 1996. Applied linear statistical models. (1996).
- [36] Balaji Padmanabhan, Xiao Fang, Nachiketa Sahoo, and Andrew Burton-Jones. 2022. Machine learning in information systems research. *MIS Quarterly* 46, 1 (2022), iii–xix.
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035.
- [38] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

- [39] Brian T. Pentland, Emmanuelle Vaast, and Julie Ryan Wolf. 2021. Theorizing process dynamics with directed graphs: A diachronic analysis of digital trace data. *MIS Quarterly* 45, 2 (2021).
- [40] Arun Rai. 2017. Editor's comments: Diversity of design science research. *MIS Quarterly* 41, 1 (2017), iii–xviii.
- [41] Blaize Horner Reich and Michelle L. Kaarst-Brown. 1999. Seeding the line: Understanding the transition from IT to Non-IT careers. *MIS Quarterly* 23, 3 (1999), 337–364.
- [42] Sherry Ryan and Gina Harden. 2014. Job embeddedness of information technology professionals: The effects of gender. *Journal of Computer Information Systems* 54, 4 (2014), 52–59.
- [43] D. Sankoff, J. Kruskal, and J. Nerbonne. 2000. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Cambridge University Press.
- [44] Kathryn L. Shaw. 1987. Occupational change, employer change, and the transferability of skills. *Southern Economic Journal* 53, 3 (1987), 702–719.
- [45] Padhraic Smyth. 1996. Clustering sequences with hidden Markov models. *Advances in Neural Information Processing Systems* 9 (1996), 648–654.
- [46] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2016. *Introduction to Data Mining*. Pearson Education India.
- [47] Robert L. Thorndike. 1953. Who belongs in the family. In *Psychometrika*. Citeseer.
- [48] Eileen M. Trauth, Jeria L. Quesenberry, and Haiyan Huang. 2009. Retaining women in the US IT workforce: Theorizing the influence of organizational factors. *European Journal of Information Systems* 18, 5 (2009), 476–497.
- [49] Jarkko Venna and Samuel Kaski. 2001. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Georg Dorffner, Horst Bischof, and Kurt Hornik (Eds.). Vol. 2130. Springer Berlin Heidelberg, 485–491.
- [50] Claartje J. Vinkenburg and Torsten Weber. 2012. Managerial career patterns: A review of the empirical evidence. *Journal of Vocational Behavior* 80, 3 (2012), 592–607.
- [51] Rose Wengling and Steven Thomas. 2009. Workplace culture that hinders and assists the career development of women in information technology. *Information Technology, Learning & Performance Journal* 25, 1 (2009).
- [52] Tengke Xiong, Shengrui Wang, Qingshan Jiang, and Joshua Zhexue Huang. 2011. A new markov model for clustering categorical sequences. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 854–863.
- [53] Rui Xu and Donald C. Wunsch. 2010. Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering* 3, 1 (2010), 120–154.
- [54] Xiaoying Xu, Hanlin Qian, Chunmian Ge, and Zhijie Lin. 2020. Industry classification with online resume big data: A design science approach. *Information & Management* 57, 5 (2020), 103182.
- [55] Quan Zou, Gang Lin, Xingpeng Jiang, Xiangrong Liu, and Xiangxiang Zeng. 2020. Sequence clustering in bioinformatics: An empirical study. *Briefings in Bioinformatics* 21, 1 (2020), 1–10.

Appendices

A Data Summary

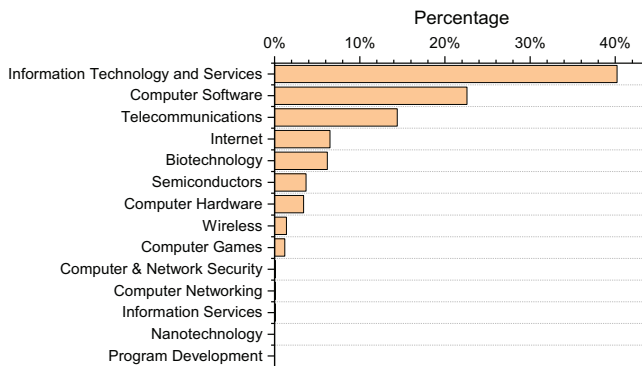


Fig. A.1. Percentage distribution of talents by industry.

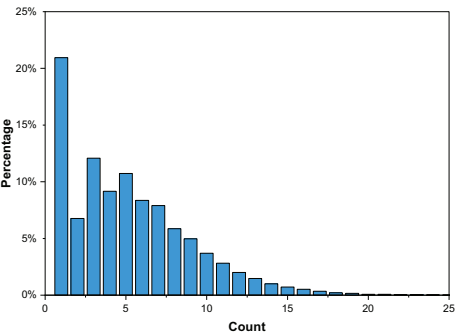


Fig. A.2. Job sequence length.

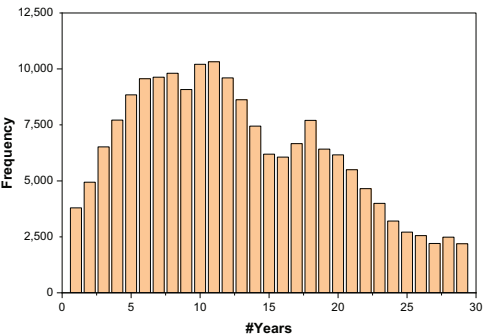


Fig. A.3. Career tenure length.



Fig. A.4. Word cloud of all job titles in our data sample.

B Algorithm's Pseudocode

ALGORITHM 1: Embedded Mixture Markov Models (EMMMs)

Input: The set of sequences $\{S_n : n = 1, 2, \dots, N\}$; the number of clusters M ; the embedding dimensionality D ; the number of iterations T ; the number of epochs E .

Output: Cluster-wise term embeddings θ_m and probability log-likelihoods $\Pr(S_n|\theta_m)$.

- 1 Determine the initial clusters of all career sequences and randomly initialize $v_0(w)$ and $\{\delta_m(w) : m = 1, 2, \dots, M\}$, where $w \in \mathcal{W}$;
- 2 **for** $i=1$ **to** T **do** // T is the number of iterations
 - 3 **for** $m=1$ **to** M **do** // M is the number of clusters
 - 4 **if** $i=1$ **then**
 - 5 Initialize an EMM model with $v_0(w)$ and $v_m(w)$;
 - 6 **else**
 - 7 Update $v'_0(w) \leftarrow v_0(w) + \frac{1}{M} \sum_m \delta_m(w)$;
 - 8 Update $\delta'_m(w) \leftarrow \delta_m(w) - \frac{1}{M} \sum_m \delta_m(w)$;
 - 9 **end**
 - 10 **for** $e=1$ **to** E **do** // E is the number of epochs
 - 11 Randomly generate a batch;
 - 12 Compute the embedding of sequence S_n by aggregating $v_m(w) = v_0(w) + \delta_m(w)$ for any job term w in the sequence;
 - 13 Compute the overall loss $\mathcal{L}(\theta, C)$ according to Equation (12);
 - 14 Optimize $v_0(w)$ and $\delta_m(w)$ using backpropagation;
 - 15 **end**
 - 16 **end**
 - 17 **end**
 - 18 Compute $\Pr(S_n|\theta_m)$ for each sequence S_n in any cluster C_m using Equation (8).

C Career Path Clustering

C.1 Determining the Number of Clusters

Similar to many other conventional clustering models (e.g., K-means), one critical hyperparameter in our model is the number of clusters. Typically, people use the **Silhouette Coefficient (SC)** or **Akaike's Information Criterion (AIC)** to determine the appropriate number of clusters [46]. The calculation of the SC relies on a predetermined distance metric such as euclidean distance. However, the "distances" between clusters produced by our model are undefined, thus disqualifying SC as a suitable option. On the other hand, AIC is also unfavorable in our case as the magnitudes of the resulting log-likelihood are in significantly different scales of M . Instead, we suggest using a criterion similar to AIC, which involves calculating the negative log-likelihood (NLL) averaged by M , that is, *Averaged NLL* = NLL/M . Like AIC, this criterion aims to strike a balance between the model's goodness of fit and its simplicity. The clustering model with the lowest *Averaged NLL* is the preferred choice. Figure C.1 showcases the curve of *Averaged NLL* resulting from our model training process. We find that *Averaged NLL* of the model significantly decreases as the number of clusters approaches $M = 8$ and then gradually converges towards zero. Following the Elbow method [47], we determine $M = 8$ as the proper number of clusters. Then, we proceed with training our model to categorize career paths into eight clusters, each associated with an individual EMM model. The embedding dimensionality is set $D = 32$. We show the learning curves of these models in Figure C.2. As can be seen, NLL s of our EMMM model decline rapidly with an increase in iteration steps.

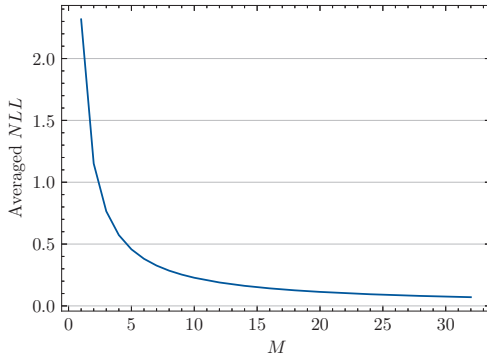
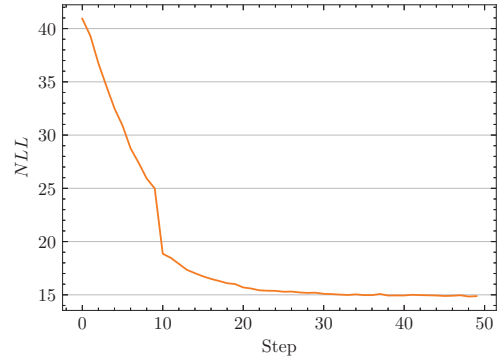
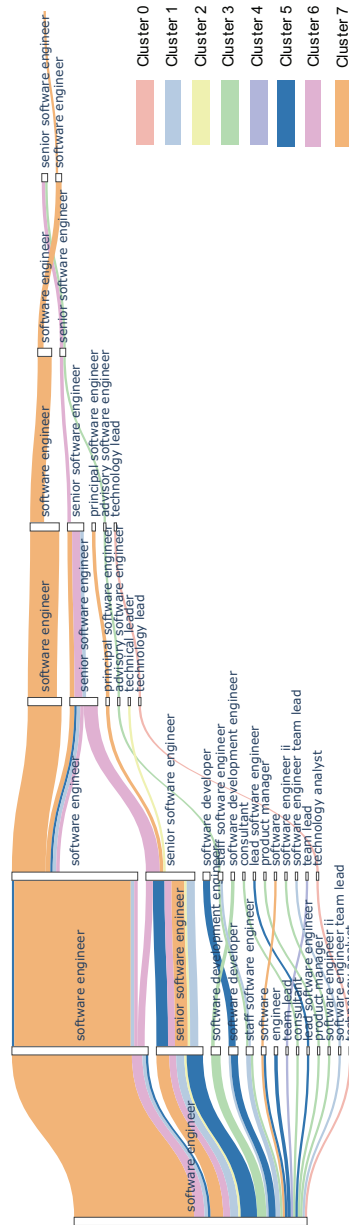
Fig. C.1. Averaged NLL with varying M .

Fig. C.2. Learning curve.



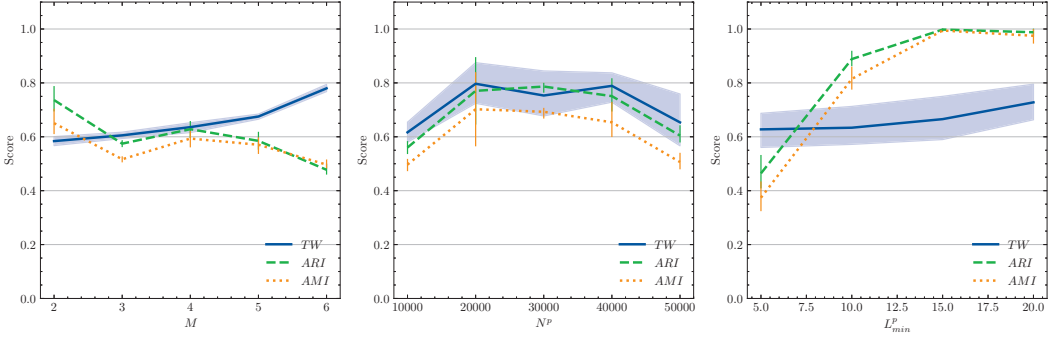


Fig. D.1. Model performance by varying M , N^P , and L_{min}^P on the synthetic data.

is surrounded by N^m sampled points (job terms), which are subsequently utilized to create job titles for each cluster. Job titles are generated within a specified length range (L_{min}^t to L_{max}^t) through the following process. First, an initial point (term) is randomly selected within the cluster. Second, subsequent points (terms) are chosen by prioritizing nearby neighbors in the traversal process. The favoring of geographical proximity corresponds to the sequential adjacency of job titles along career paths. The quantity of job titles within each cluster is determined by a predefined hyperparameter N^t . In a similar vein, we create career trajectories by linking together randomly-chosen job titles, with the specified range of job sequence length (L_{min}^p to L_{max}^p) and the number of sequences (N^p). Note that each cluster consists of an equal quantity (N^p) of career paths in our current configuration. Eventually, through meticulous parameter adjustments, we can generate synthetic datasets tailored to meet specific research for various experiments.

D.2 Performance Metrics

We first assess the model's ability to accurately reconstruct the true clusters, for which two metrics are used: **adjusted Rand index (ARI)** and **adjusted mutual information (AMI)**. ARI is a corrected-for-chance version of the *Rand index* [23], which computes a similarity measure between two clusterings by considering all sample pairs and counting pairs assigned to the same or different clusters in the predicted and true clusterings. AMI is another metric for comparing two labels of the same dataset that is independent of the labels' absolute values. Furthermore, we can assess the model's ability to preserve the local structure of the actual embeddings by employing the **Trustworthiness (TW)** metric [49].

D.3 Sensitivity Analysis

Our model depends on the *initial* groupings of job sequences. We begin our experimental analysis by using the original groupings as the starting clusters. We now analyze how the model's performance changes with varying values of M . By changing M from 2 to 6, we generate several synthetic datasets. Our model needs to reconstruct the original M clusters and the local structure of term embeddings for each dataset. The first chart in Figure D.1 demonstrates that with an increase in M , there is a slight decline in the model's ability to restore the original clusters. This finding aligns with our initial impression that the task becomes increasingly difficult as the required M increases. Meanwhile, as M increases, there is an upward trend in TW , indicating that our model can effectively recover the local structure of the true term embeddings.

Now, let us examine N^p , which represents the total number of possible job sequences. We aim to evaluate the resilience of the model when exposed to different amounts of training data. Let us

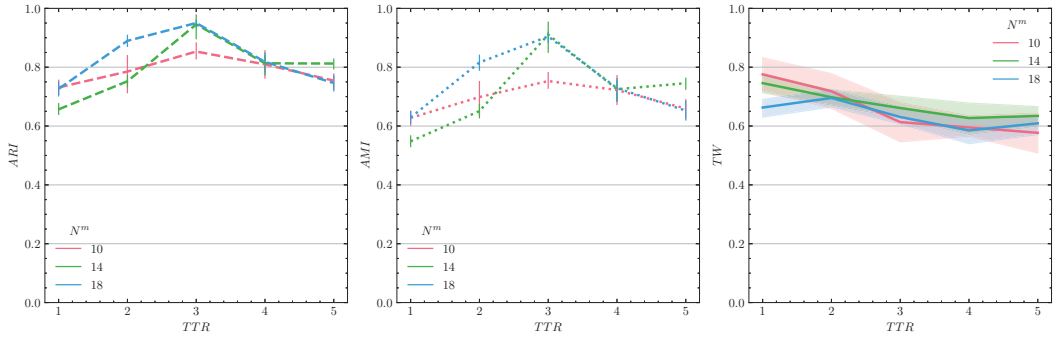


Fig. D.2. Model performance by varying TTR on the synthetic data.

begin by acknowledging the challenges that can ensue from an inadequate or excessive number of job sequences. When we have a limited number of job sequences, we suffer from the issue of data insufficiency. This hinders the model from capturing detailed nuances of job titles and terms, thereby increasing the level of difficulty to restore the original clusters. On the other hand, an excessively large sample size can present challenges for the model as a result of the substantial heterogeneity within the data. Interestingly, the experimental results illustrated in the second chart of Figure D.1 corroborate our initial speculation. It shows that when the value of N^p is either too low or too high, there is a slight decline observed in the three metrics (ARI, AMI, and TW).

As our model aims to group job sequences, it is also important to analyze the model's sensitivity to variations in job sequence length. Technically, our model should favor longer sequences as they capture a greater degree of sequential dynamics among job titles. The final chart in Figure D.1 illustrates a significant improvement our model's ability to accurately reconstruct the original clusters as the minimum sequence length L_{min}^p increases.

Remember that our job sequences are formed by combining multiple job titles, each of which consists of job terms. With a specific sequence length, we would also like to assess the model's robustness concerning the size of the job title set (N^t) and the size of the job term set (N^m). Since job title embeddings are aggregated by averaging job term embeddings, we propose to focus on examining the *proportional* magnitude of the two sets, which we defined as $TTR = N^t / N^m$. Our data simulation process starts by constructing the job term set with sizes of 10, 14, and 18, followed by generating multiple sets of job titles by adjusting TTR within the range of 1 to 5. We perform experiments on these simulated datasets and present our evaluation findings in Figure D.2. Interestingly, when TTR is too low or too high, the model's capacity to restore original clusters (as measured by ARI and AMI) declines. We argue that, when $TTR = 1$ (i.e., the number of titles equals the number of terms), the model training suffers from inadequate job titles to capture all the subtleties associated with the job terms. When TTR is high (i.e., when many more job titles are generated from a fixed set of job terms), the resulting job titles are likely to be similar in nature. This can potentially hinder the ability of our model in learning job term embeddings.

D.4 Model Comparison with OMA

We have completed a sensitivity analysis on different elements of the model. Nevertheless, we are still uncertain about the performance of our model when compared to other existing models, such as OMA. Therefore, we conduct an additional experimental analysis using synthetic datasets to evaluate and compare our model with the OMA model.

In our study, the **OMA** model is implemented using *TraMineR* package of the R programming language. For the implementation of our model, we conducted our experiments with two

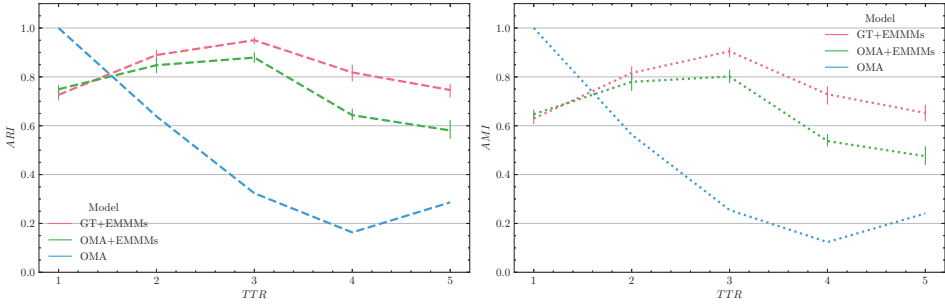


Fig. D.3. Three models' performance by varying TTR on the synthetic data.⁴

initializations of clusters. The first is to use **ground-truth (GT)** clusters, which is expected to deliver the highest achievable level of performance. This model is referred to as **GT+EMMMs**. The second version utilizes the clusters derived from OMA as initial groupings for the sequences. The objective is to assess the capability of our model in capturing nuances overlooked by the OMA model. We denote this model as **OMA+EMMMs**. Figure D.3 presents the ARI and AMI of these models across different TTR s. Since OMA does not produce embeddings, it is not feasible to calculate TW , which therefore is excluded from the evaluation. According to the results, when $TTR = 1$, the OMA can simply achieve the best performance. However, as TTR increases (more in line with the real-world conditions), the OMA's performance deteriorates significantly, whereas the two EMMMs-based models maintain their superiority. In summary, the results demonstrate that OMA+EMMMs is comparable to GT+EMMMs and consistently surpasses the OMA. These findings reinforce our claim concerning the efficacy of EMMMs-based models.

Received 6 March 2024; revised 26 December 2024; accepted 1 January 2025

⁴Since OMA does not produce embeddings, TW is not included as it is not feasible for the calculation.