9-2018

Are you on the right track? Learning career tracks for job movement analysis

Meng-Fen CHIANG
Singapore Management University, mfchiang@smu.edu.sg

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

Wang-Chien LEE
Pennsylvania State University

Yuan TIAN
Singapore Management University, ytian@smu.edu.sg

Chih-Chieh HUNG
Tamkang University

Are You on the Right Track? Learning Career Tracks for Job Movement Analysis

Meng-Fen Chiang[1], Ee-Peng Lim[1], Wang-Chien Lee[2], Yuan Tian[1], and Chih-Chieh Hung[3]

[1] Living Analytics Research Centre, Singapore Management University, Singapore
{mfchiang,eplim,ytian}@smu.edu.sg
[2] The Pennsylvania State University, USA
wlee@cse.psu.edu
[3] Tamkang University, Taiwan
oshin@mail.tku.edu.tw

Abstract. Career track represents a vertical career pathway, where one can gradually move up to take up higher job appointments when relevant skills are acquired. Understanding the propensity of career movements in an evolving job market can enable timely career guidance to job seekers and working professionals. To this end, we harvest career trajectories from online professional network (OPN). Our focus lies on obtaining a macro view on career movements at the track granularity. Specifically, we propose a semi-supervised career track labelling framework to automatically assign career tracks for large set of jobs. To contextually label jobs, we collect example jobs with career track labels identified by human resource specialists and domain experts in Singapore. An intuitive idea is to learn the labelling knowledge from the example jobs and then apply to jobs in OPN. Unfortunately, such small amount of labeled jobs presents a great challenge in our attempt to accurately recover career tracks for plentiful unlabelled jobs. We thus address the issue by resorting to semi-supervised learning methods. This research not only reduces the human annotation efforts in maintaining the career track knowledge databases over time across different geographical regions, but also facilitates data science study on career movements. Extensive experiments are conducted to demonstrate the labelling accuracy as well as to gain insights upon obtained career track labels.

Keywords: Label Propagation · Career Movements Analysis · Career Track Labelling.

1 Introduction

Motivation. Job market is constantly evolving, attracting new talents to take up jobs offers. Meanwhile, from career development point of view, when a person enters a job market, there are many career tracks for her to choose. Each career track requires a set of skills which may take a significant period of time and training to acquire. One could gradually take up higher job appointments as

she acquires the relevant skills. Career track thus represents a vertical path of skill specialization. For example, Table 1 depicts an expert crafted framework of career tracks in the infocomm technology sector (ICT)[4]. ICT experts have identified seven career tracks, i.e., infrastructure (IFR), software and applications (SAA), sales and marketing (SAM), professional service (PS), support (SP), security (SC), and data (DT) tracks. As shown, associated with each track is a set of exemplary job titles. A change of job within the same track represents a vertical career move, while that across tracks is known as a lateral career move. A framework covering career tracks and jobs under each track is very useful for career coaching and skill training. Nevertheless, such a manual approach requires experts from various industry domains who are limited by number, and does not scale for massive number of job titles. To address this limitation, we need a solution that can automatically assign career track to jobs.

The ability to automatically assign career tracks to jobs will also enable us to conduct a data science study to characterize career movements of a population of working professionals in a constantly evolving job market. Suppose we have the job history data of a working population, and every job is assigned with a career track label. We now can answer several interesting data science questions, including: *"what is the proportion of people staying in the same career tracks versus switching from one career track to another?", "How is this proportion different for people who work in different industry sectors?", "For people in some career track of some industry sector (e.g., infrastructure track of ICT sector), how easy is it for them to switch to another career track of the same (or different) industry sector?"* To the best of our knowledge, such a data science study of career track changes for a working population has not been conducted previously. On the other hand, the insights of career track changes are of utmost importance to the design of public policies to effectively manage the supply of labor in every industry segments. This further strengthens the need for this research.

Table 1. Career tracks in infocomm technology sector (ICT) include infrastructure (IFR), software and applications (SAA), sales and marketing (SAM), professional service (PS), support (SP), security (SC), and data (DT) tracks.

Track	**Exemplary Job Titles**
IFR	cloud engineer, infrastructure engineer, infrastructure executive, information architect, etc.
SAA	application developer, platform engineer, product manager, system analyst, etc.
SAM	sales executive, channel sales manager, digital marketing executive, etc.
PS	IT consultant, project manager, business analyst, program director, etc.
SP	quality analyst, IT auditor, system administrator, support analyst, etc.
SC	security operations analyst, security executive, security engineer, cyber risk analyst, etc.
DT	data engineer, data scientist, data analyst, business intelligence manager, etc.

Objectives. We have two key objectives to accomplish in this research. First, we aim to develop methods to automatically assign career track labels to jobs, which will significantly reduce manual efforts to perform the same task for large number

[4] https://www.imda.gov.sg/cwp/assets/imtalent/skills-framework-for-ict/index.html

of jobs. Second, we aim to gain interesting insights about the career movement of a population of working professionals using the automatically assigned career track labels. Both research objectives hinge on solving the following problem:

Problem: (Career Track Labelling) Given a set of jobs consisting of very few of them assigned with career track labels, determine the career tracks for the remaining unlabelled jobs.

Overview of Proposed Research. We address the career track labelling problem by proposed a framework that involves gathering a dataset of jobs among which some are assigned the ground truth career track labels. Given the nature of large majority of unlabelled jobs, our proposed framework adopts semi-supervised learning methods (SSL) to perform this career track labelling [10]. In particular, we aim to create connections among the jobs and to use label propagation algorithms to propagate track labels from the small number of labeled jobs to the remaining jobs. Our framework also includes an evaluation step that compare the semi-supervised label propagation methods with supervised methods. We finally derive insights from the automatically assigned career track labels.

Contributions. The contributions of our work are summarized as follows:

- We formulate the career track labelling problem as a semi-supervised multi-class classification problem and propose a career track labelling framework by exploiting several types of relationships among jobs.
- We demonstrate the effectiveness of the career track labelling framework using a real world dataset in which only very small number of jobs are labeled. We show that label propagation methods significantly outperform supervised learning methods.
- Based on the automatically assigned career track labels, we are able to extract useful insights about the career movement of Singapore job market.

2 Related Work

Job Mobility Framework. Several studies of job mobilities have been reported in the literature on career development [1][2]. Ng et al. defines job mobilities as the career transition patterns [2]. Twelve types of determinants are defined to describe multi-level nature of job mobilities in terms of changes of job status (e.g., upwards, lateral, or downwards), job functions (e.g., same or changed), and employers (e.g., internal or external). Baruch observed a transformation of career system, where job mobility transforms from vertical movement within organization to spiral movement beyond organization boundaries [1]. Another research focus is to model and predict career move [3][4][5][6][7]. Some work takes a survival analysis approache to infer tenure-based decision-making probability [3][7]. Others formulate career choice prediction as a supervised learning problem, assuming that sufficient training data are available [4][5].

Label Propagation. Semi-supervised learning (SSL) is particularly devoted to application domains in which unlabelled data are plentiful [10]. A family of label propagation algorithms have been proposed to tackle SSL problem and are

widely used to predict unobserved node labels on a network in several domains [9][8][12][17][13][14]. Yamaguchi el al. explored the theoretical properties of label propagation algorithms, such as connection to random-walks and convergence [12][13]. Several key questions, such as impact of label distributions and label correlations, were raised to explore the ability and limitations of label propagation [14][9]. Label correlations can be divided into two categories: homophily and heterophily [12]. On homophily networks, nodes with similar labels tend to connect to each other, while the reverse is true on the heterophily networks. LP is a node classification algorithm that exploits homophily effect [8]. On the other hand, OMNI-Prop can leverage on both homophily and heterophily types of label correlations including mixed correlations [12]. Yamaguchi el al. in [13] proposed a novel node classification algorithm, Camlp, not only to exploit label correlations in networks but also to incorporate confidence-awareness in the process of label propagation. Hadiji et al. introduced compressed label propagation (CLP) to efficiently infer missing geo-tags of author-paper-pairs retrieved from online bibliographies [17]. As heterogeneous network datasets become increasing popular and accessible (e.g., researcher networks, Wikipedia entity networks, etc.), label propagation approaches on heterogeneous networks have been developed [15][16]. Deng et al. proposed an efficient K-partite label propagation model that supports heterogeneity with heterophily propagation [16]. Duran and Niepert recently introduced embedding propagation (EP) that learns embeddings of node labels such that the embeddings of all labels of a node are expected to be close to the embeddings of its neighbours [18].

Existing career track knowledge databases. There are a few career track knowledge databases available, which are manually constructed by industry and human resource experts for career analysis and job exploration. O*Net defines a career cluster to be a group of occupations in the same field of work using similar skills[5]. There are altogether 16 career clusters defined in O*Net in which each career cluster covers a set of career pathways. Each career pathway consists of a series of job titles sharing similar skill set and knowledge. Such career pathways are thus very much like the career tracks we are interested in. Skillsfuture is an initiative by Singapore government to identify career tracks in 17 industry sectors relevant to Singapore[6]. Within each industry sector is a set of career tracks. Each career track consists of a number of job titles at different job levels. Skillsfuture also attempts to represent the possible transitions between job titles within and across career tracks. In this paper, we construct a Skillsfuture career track knowledge database consisting Skillsfuture career track labels and job titles assigned with these labels.

3 Career Track Labelling Framework

In this section, we first give an overview of the framework proposed for career track labelling. We then introduce datasets collected to identify the career track

[5] O*Net Center. https://www.onetcenter.org/
[6] Skillsfuture. http://www.skillsfuture.sg/

labels for jobs in LinkedIn. Finally, we detail the two major components in our framework.

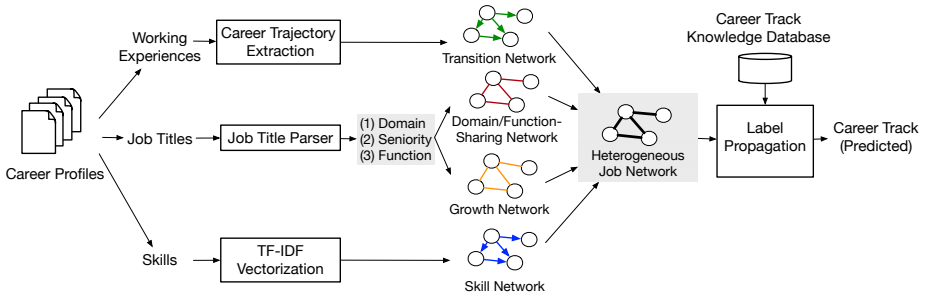
3.1 Framework Overview

Figure 1(a) illustrates the overview of career track labelling framework. The framework consists of two major components: (1) heterogeneous job network construction, and (2) label propagation. In order to exploit the label propagation techniques, the heterogeneous job network construction takes in career profiles in online jobs banks or OPN, i.e., LinkedIn, to derive a heterogeneous network. To fully explore links between jobs, we specifically extract three types of information from career profiles: *career progression trajectories*, *job titles*, and *skills set* of a working professional. Each career trajectory is a sequence of job transitions of a working professional. The entire collection of sequences of job transitions for all working professionals in the dataset is transformed into a *transition network*. For each extracted job title, we develop a job title parser to process each job title into three standardized elements: (1) domain, (2) seniority, and (3) function. We then define the job title *element-sharing link* between jobs to construct *domain/function sharing network* and *growth network*. For a given job title and industry, we extract skills from the job description. By treating each job title as a document and skills as words, we derive a TF-IDF vector. We then compute cosine similarity between any pair of job titles so as to explore skill similarity links and construct a *skill network*. Finally, a heterogeneous job network is constructed by aggregating the above-mentioned link types into one. The label propagation component of the framework propagates career track labels upon various networks attempting to connect unlabelled jobs to the labelled ones.

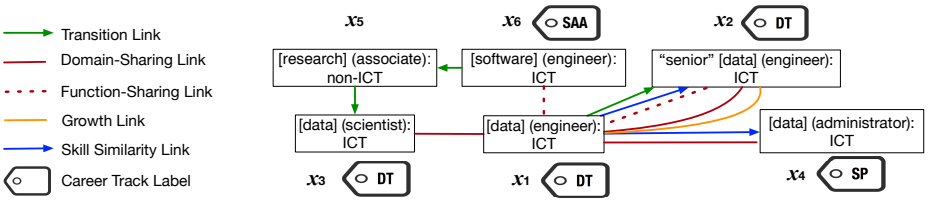
3.2 Data

Skillsfuture Career Track Knowledge Database. In order to facilitate a data science study, and to evaluate career track labelling methods, we construct a knowledge base consisting of ground truth label data extracted from Skillsfuture Singapore's Skills Framework. In this work, we focus on the ground truth labels of jobs in the ICT sector. As mentioned earlier, Skillsfuture Singapore defines seven career tracks in the ICT sector. The statistics of this dataset can be found in Table 2.

Career Trajectory Dataset. To illustrate a data science study of career trajectories using automatically assigned career track labels, we collect all public profiles of LinkedIn users who work in Singapore up to June 2016. To ensure that every job has the required features for track labelling as well as to remove noise, we remove user profiles that contain less than two skills, and job titles that only appear in only one person's career trajectory. Table 2 summarises the key statistics of this dataset.



(a) Framework Overview



(b) An Example of Heterogeneous Job Network

Fig. 1. (a) Framework overview. (b) An example of heterogeneous job network involving six job nodes and five link types. Three title elements are extracted for each job title. For example, senior data engineer consists of [data] as domain, "senior" as seniority, and (engineer) as function. Job titles x_1 and x_2 share five links: (1) job transition from x_1 to x_2, (2) sharing the [data] domain element, (3) sharing the (engineer) function element, (4) sharing [data] domain element and the (engineer) function element, and (5) sharing common skills (assuming both jobs are observed to require Python programming skill).

3.3 Heterogeneous Job Networks Construction

There exist various kinds of links amongst jobs. Each link features some unique semantic, which may be useful for identifying career tracks. We first categorize those links into three types, and discuss their semantics as well as their limitations. We then propose a notation of *heterogenous job network* for the career track labelling problem.

Job Title Element-Sharing Link. Job titles usually contain unstructured text. To extract useful information and remove noises, we develop a job title parser which extracts from each job title three title elements: (1) domain, (2)

Table 2. Data Statistics.

Dataset	All Sectors		ICT Sector
	#Jobs	#Skills	# Jobs
Skillsfuture[7]	unknown	unknown	124
LinkedIn	7,975	2,600	1,252

seniority, and (3) function. For example, "senior data engineer" is parsed into (1) data (2) senior, and (3) engineer elements, correspondingly.

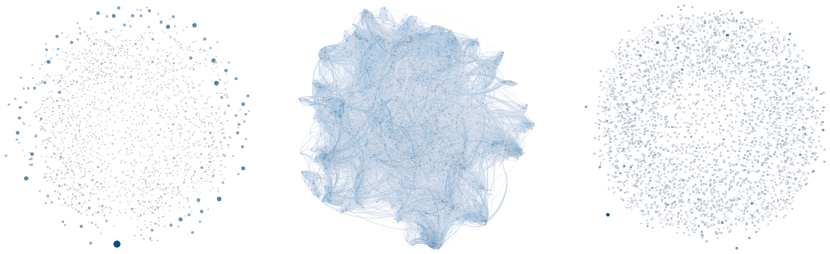
Domain-sharing and *function-sharing* links are examples of job title element-sharing links. The intuition behind domain-sharing (function-sharing) links is that two jobs are likely to belong to the same career track if their job titles share the same domain (function) element. For example, "data engineer" and "data scientist" share the same domain element "data" and are thus likely under the same track (i.e., data track). However, domain-sharing links on its own may not always suggest two jobs belonging to the same career tracks. For example, the two job titles "data engineer" and "data administrator" share the same "data" domain element but they belong to the data track and support track respectively.

Growth link is a link combining both same domain element and same function element. For example, "data engineer" and "senior data engineer" share both "data" domain element and "engineer" function element. The *growth link* is thus introduced between the two job titles.

Transition Link. From the career trajectories of working professionals, we can derive job transition links which may provide additional features for career track label assignment. The intuition is that people are more likely to stay within the same career track instead of to switch between career tracks. For example, suppose many people are observed to move from "platform engineer" to "application developer". The two jobs are likely to belong to the same career track, and knowing the track of one will help to infer that of another. In this example, the two jobs belong to the software and application track. However, some job transitions may involve changes of industries. We may even find a job transition from one industry sector to another industry sector, followed by another job transition back to the first industry sector, as shown in Figure 1(b), i.e., $x_6 \rightarrow x_5 \rightarrow x_3$.

Skill Similarity Link. Job titles can be related by similar skills. For example, both "test specialist" and "quality assurance analyst" are connected by skill similarity links because the skills required in both jobs are similar, e.g., manual testing, and agile project management. A job x_j is linked to another job x_i by skill similarity if x_j is considered as sufficiently similar to x_i by skill. To compute the skill similarity between x_i and x_j, we represent each job as a skill vector, where each vector element represents the TF-IDF value of a skill from the skill dictionary. We then calculate the cosine similarity for each pair of jobs. For each job x_i, we select the k most similar jobs as x_i's out-neighbors. Note that skill similarity link is directed as x_j being one of the k nearest neighbors of x_i does not imply x_i is one of k nearest neighbors of x_j.

Table 3 gives the statistics of the different job networks constructed using various types of links. All these networks share the same set of job nodes (7,975 of them), and same set of job nodes that have been assigned with career track labels (250 of them). Out of the 7,975 job nodes, 1,252 of unlabelled job nodes are from the ICT industry in LinkedIn. Note that only 3.1% of entire set of jobs are labelled, making the career track labelling task extremely challenging.



(a) Domain Network (b) Skill Network (k=10) (c) Growth Network

Fig. 2. Examples of Homogeneous Networks.

Figure 2 depicts the various networks constructed using single link types. The skill similarity link network (k=10) in Figure 2(b) is much denser than networks with job title element-sharing links (domain link in Figure 2(a) and growth link in Figure 2(c).

Table 3. Summary of Networks of Different Link Types.

Network	transition	function	domain	growth	skill(k=10)	skill(k=30)	skill(k=50)	skill(w)
n	7,975	7,975	7,975	7,975	7,975	7,975	7,975	7,975
l	250	250	250	250	250	250	250	250
u(ICT)	1,252	1,252	1,252	1,252	1,252	1,252	1,252	1,252
u(non-ICT)	6,473	6,473	6,473	6,473	6,473	6,473	6,473	6,473
directed	Yes	No	No	No	Yes	Yes	Yes	Yes
$\|E\|$	26,608	588,733	43,458	14,442	185,305	554,593	929,207	33,530
avg.degree	3.3	73.8	8.0	3.3	23.2	69.5	116.5	4.2
#isoloated nodes	1,238	0	1,805	1,805	0	0	0	3,039

As the above networks share the same job nodes, we can easily union their links so as to construct a *heterogeneous job network* as defined below.

Definition. *A heterogeneous job network, denoted by* $G = (X, \{E^{(r)}\}, \{M^{(r)}\}, Y)$*, consists of* n *nodes in* X*, an adjacency matrix* $M^{(r)}$ *for every link type* $r \in R$*, and* K *distinct track labels in* Y*. We define the adjacency matrix of* r*-type link as an* $n \times n$ *matrix* $M^{(r)}$*, where* $M^{(r)}_{i,j}$*=1 if node* x_i *connects to* x_j *and* $M^{(r)}_{i,j}$*=0 otherwise.* Y *is the* $n \times K$ *partially observed label assignment matrix, where* $y_{i,k}$*=1 if node* x_i *belongs to label* k*, and* $y_{i,k'}$*=0 for* $k' \neq k$*.*

Example. There are two types of nodes: labelled nodes X^L=$\{x_1, \cdots, x_l\}$ and unlabelled nodes X^U=$\{x_{l+1}, \cdots, x_{l+u}\}$, where $l+u$=n. Each labelled node $x_p \in X^L$ is assigned with one track label. In $M^{(r)}$, both $M^{(r)}_{i,j}$ and $M^{(r)}_{j,i}$ equals to 1 when x_i and x_j are connected by an undirected link. For any given job title x_p, we represent its neighbors of r different link types as $N_p = \{N_p^{(1)}, \cdots, N_p^{(|R|)}\}$, where $N_p^{(r)}$ is the set of neighbor (or out-neighbor) nodes in r-th undirected (or directed) link type.

For example, in Figure 1(b), there are five types of links. $N_1^{(1)}=\{x_2\}$ refers to the out-neighbors that x_1 transits to, $N_1^{(2)}=\{x_2, x_3, x_4\}$ refers to neighboring jobs with similar job domain as x_1, $N_1^{(3)}=\{x_2, x_6\}$ refers to neighboring jobs with similar job function as x_1, $N_1^{(4)}=\{x_2\}$ refers to neighboring jobs with same job domain and function as x_1, and $N_1^{(5)}=\{x_2, x_4\}$ refers to neighboring jobs with similar job skills as x_1.

3.4 Career Track Labelling

We formalize the career track labelling problem for a heterogeneous job network as follows:

Problem. (Career Track Labelling Problem) *Given a set of job titles represented as X and a partially observed track label assignments Y, where $X = X^L \cup X^U$, the goal is to rank career tracks for each unlabelled job title $x_p \in X^U$ based on the labelled jobs X^L and the network connectivity.*

Supervised Learning Approach. An intuitive approach to address the track labelling problem is supervised learning. This line of solutions first extracts the features for each job node, and then apply supervised learning algorithms to learn classification functions. For example, we may extract skill features of labelled job titles to train a classifier to predict node labels. Any supervised learning models such as Logistic Regression (LR) can be used. Note that the supervised learning approach has to cope with only 250 labelled job nodes in our dataset.

Label Propagation Approach. The label propagation approach regards career track labelling as a semi-supervised learning problem (SSL). SSL is particularly applicable to network data with very few labelled nodes [10]. A family of label propagation algorithms has been proven effective empirically and theoretically in various problem settings [8][12][13][14].

In this work, we adopt two well-known label propagation algorithms, LP[8] and OMNI-Prop [12]. LP is especially useful when homophily is observed in a given network. OMNI-Prop, on the other hand, is capable of handling networks with both homophily and heterophily effects. The basic scheme of both algorithms is to iteratively perform updates on likelihoods of each career track label assigned to every job node until it converges. The likelihoods of each career track assigned to unlabelled nodes are initialized by arbitrary values. We further elaborate the two algorithms below.

LP. LP essentially looks for a real-value function $f : X \to R$ on G with *harmonic property*, constrained on $f(x_i) = f^L(x_i) \equiv y_i$ for labelled node $x_i \in X^L$ such that Eq. (2) is satisfied.

$$E(f) = \frac{1}{2}\sum_{i,j} M_{i,j}(f(x_i) - f(x_j))^2, s.t. \tag{1}$$

$$f = \underset{f|Y^L=f^L}{\arg\min}\, E(f). \tag{2}$$

where f is a harmonic function. The harmonic property means that the value of f at each unlabelled node $x_i \in X^U$ is the average of f of x_i's neighbors, and it equals to labelled data nodes X^L. Intuitively, the harmonic function wants unlabelled nodes that are nearby in the network to have similar labels.

The procedure of computing f is known as *harmonic energy minimization*. Let the adjacent matrix M be a composition of four blocks as follows:

$$M = \begin{bmatrix} M^{LL} & M^{LU} \\ M^{UL} & M^{UU} \end{bmatrix} \tag{3}$$

where M^{UL} represents the adjacency matrix from unlabelled node X^U to labelled nodes X^L. Let $f = \begin{bmatrix} f^L \\ f^U \end{bmatrix}$, where f^U denotes the values on the unlabelled nodes. The harmonic solution $\Delta f = 0$ subject to $Y^L = f^L$ is given as follows:

$$f^U = (D^{UU} - M^{UU})^{-1} M^{UL} f^L \tag{4}$$

where $D = \text{diag}(d_i)$ is the diagonal matrix with entires $d_i = \sum_j M_{i,j}$. Finally, $y_{i,k}$ is obtained for each unlabelled node x_i by assigning f^U.

OMNI-Prop. In OMNI-Prop, self-score and follower-score are defined for a node and label pair. Self-score $y_{i,k}$ refers to how likely node x_i has label k, while follower-score $z_{j,k}$ refers to how likely the in-neighbors of node x_j have label k. In OMNI-Prop, self-scores (and follower-scores) are iteratively updated using the evidences from out-neighbors (in-neighbors) until Y (Z) converges. Every $y_{i,k}$ is updated by aggregating the follower-scores from its out-neighbors as follows:

$$y_{i,k} = \frac{\sum_{j=1}^{n} M_{i,j} z_{j,k} + \lambda b_k}{\sum_{j=1}^{n} M_{i,j} + \lambda} \tag{5}$$

where b_k is the prior belief about its label and $\lambda \geq 0$ is the strength parameter. Similarly, $z_{j,k}$ is also updated by aggregating self-scores from its in-neighbors as follows:

$$z_{i,k} = \frac{\sum_{i=1}^{n} M_{i,j} y_{i,k} + \lambda b_k}{\sum_{i=1}^{n} M_{i,j} + \lambda}. \tag{6}$$

If node x_i has lots of in-neighbors with label k, follower-score $z_{i,k}$ becomes larger, which in turn increases the self-score of every in-neighbor of x_i.

The determination of label in LP and OMNI-Prop is the same. Once $y_{i,k}$ is obtained for each unlabelled node x_i, the label for node x_i is determined by $\hat{k} = \max_k y_{i,k}$. For evaluation purpose, we can rank unlabelled nodes by their $y_{i,\hat{k}}$ with respect to $\hat{k}$ and return the top-k job titles in X^U for quality analysis and further studies.

4 Experiment

In this section, we perform experiments to answer the following questions:

- Q1: How well does label propagation perform compared to supervised learning baselines?
- Q2: How can we use the automatically predicted career track labels to analyze the career movement of a population of working professionals?

4.1 Experiment Setup

Other than using LP and OMNI-Prop label propagation methods, we also include several baseline methods for comparison in our experiments.

Baselines.

1. Majority: The Majority method always returns the most frequent class labels in the training set.
2. K-Nearest Neighbors: kNN assigns the class label that appears most frequently in x_p's neighborhood to each node $x_p \in X^U$.
3. Supervised Learning Methods: These methods represent each job title as a skill vector with TF-IDF elements. We train classifiers using Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF).
4. Semi-Supervised Learning Methods: We adopt LP and OMNI-Prop (OMNI) to perform labels propagation in single link networks. We use the default parameter settings of OMNI (λ=1, and uniform label distribution for b_k) for all experiments.

Evaluation. We evaluate the performance of career track labelling based on limited ground-truth available. Note that the labelled jobs in data, security, and support tracks are very few. We thus merge them into the "Others" track. Our goal is to determine the label of each job in corpus (7,975 jobs) into one of the 5 career tracks, including (1) infrastructure (IFR), (2) software and applications (SAA), (3) sales and marketing (SAM), (4) professional service (PS), and (5) others (O). To achieve this, we collect annotations of job titles in "ICT" industry by three domain experts. Among them, only 250 job titles in "ICT" industry are annotated with full consensus among the experts. The label distribution of the 250 job titles is summarized in Table 4.

Experiment 1. In the first experiment, we conduct a leave-one-out prediction task with 249 labeled job titles used as the seed set and the remaining one labeled job title for prediction. We record the precision@1 for each job title as target and take the average precision@1 from the 250 test instances.

Experiment 2. The second experiment is to use all 250 labeled job titles as training instances (for the supervised learning approach) or seeds (for the label propagation approach) to predict the career track labels of 1,252 remaining unlabelled ICT jobs. To measure the prediction accuracies, we recruit annotators to judge the top-k job titles of each career track returned by each method. We then compute precision@k for each track with k ranging from 10 to 100.

4.2 Result Analysis

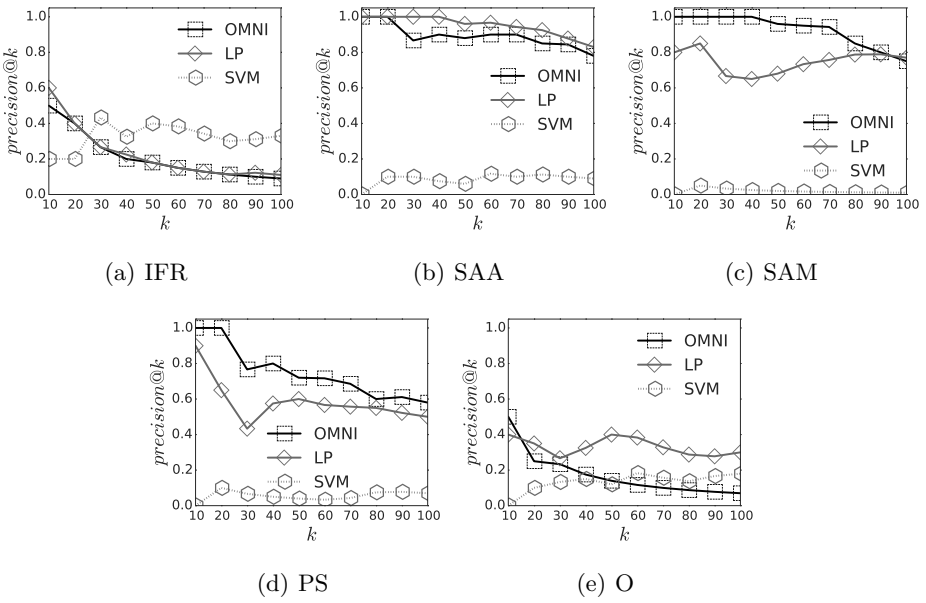
Results by Link Type. Table 4 shows the results of the the first experiment. We see that domain-sharing and growth links are effective in career track

Table 4. Career track labelling accuracy: avg. prec@1 $\geq$ 90% are highlighted.

Label Distribution			Num. Labelled Job Titles in Seed Set					
			Total	"IFR"	"SAA"	"SAM"	"PS"	"O"
			250	59	56	49	47	39
Method		**Network**	**avg. prec@1**	**prec@1**				
Unsupervised	Majority	NA	0.048	0.24	0.00	0.00	0.00	0.00
	kNN	transition	0.57	0.30	0.78	0.65	0.31	0.80
		function	0.63	0.58	0.79	0.53	0.59	0.64
		domain	**0.92**	0.68	1.00	1.00	0.93	1.00
		growth	**0.92**	0.61	1.00	1.00	1.00	1.00
		skill (k10)	0.72	0.77	0.83	0.59	0.39	1.00
		skill (k30)	0.34	0.81	0.00	0.60	0.27	0.00
		skill (k50)	0.29	0.76	0.00	0.50	0.21	0.00
		skill (w)	0.86	0.33	1.00	0.96	1.00	1.00
Supervised	LR	NA	**0.92**	0.9	0.96	1.0	0.89	0.85
	SVM	NA	**0.96**	0.98	1.0	1.0	0.89	0.92
	RF	NA	**0.93**	0.97	0.98	0.96	0.87	0.87
SSL	LP	transition	0.55	0.39	0.71	0.48	0.31	0.88
		function	0.63	0.58	0.79	0.53	0.59	0.64
		domain	**0.92**	0.68	1.00	1.00	0.93	1.00
		growth	**0.92**	0.61	1.00	1.00	1.00	1.00
		skill (k10)	0.80	0.86	0.85	0.74	0.74	0.79
		skill (k30)	0.66	0.82	0.70	0.64	0.57	0.55
		skill (k50)	0.75	0.87	0.80	0.63	0.47	1.00
		skill (w)	0.86	0.33	1.00	0.96	1.00	1.00
	OMNI	transition	0.45	0.30	0.86	0.70	0.41	0.00
		function	0.63	0.58	0.79	0.53	0.59	0.64
		domain	**0.92**	0.68	1.00	1.00	0.93	1.00
		growth	**0.92**	0.61	1.00	1.00	1.00	1.00
		skill (k10)	0.87	0.75	0.84	0.85	0.90	1.00
		skill (k30)	0.60	0.72	0.63	0.75	0.92	0.00
		skill (k50)	0.58	0.76	0.57	0.70	0.87	0.00
		skill (w)	0.86	0.32	1.00	1.00	1.00	1.00

labelling. Skill similarity networks on the other hand suffer from more noises when k becomes larger, because more low-similarity jobs are forced to be connected. Skill similarity networks with threshold on similarity weight (skill (w)) empirically perform better than other skill similarity networks except for the infrastructure track. Transition networks consistently perform poorly across different methods. This is reasonable because job transitions may involve changes of career track or industry, which downplays homophily effect in transition networks. Majority essentially assigns the largest career track to every test instance, resulting in the worst avg.prec@1 result of 0.048. The supervised methods in the leave-one-out Experiment 1 perform extremely well against label propagation algorithms. In particular, SVM achieves the best avg.prec@1 results, i.e., 0.96.

Results by Track. In Experiment 2, we use annotations to determine the correct labels of the top ranked track labels returned by each method so as to derive its precision@k. Figure 3 shows the precision@k using two types of links: skill (k10) and domain links. This offers three key insights compared to Table 4. First, we notice that "IFR" and "O" tracks are relatively difficult to predict. On the other hand, perfect labelling is achieved in "SAA" and "SAM" tracks at k=10 by label propagation algorithms, LP and OMNI, respectively. Second, although the supervised learning approach, especially SVM, performs extremely well in the leave-one-out Experiment 1 in Table 4; it, however, fails to generalize



(a) IFR (b) SAA (c) SAM

(d) PS (e) O

Fig. 3. Comparison of Career Track Labelling Accuracy for Top-100 ICT Jobs.

to a larger set of unlabelled jobs in almost every track except "IFR" as shown in Figure 3. This is because supervised learning methods are prone to overfitting for extremely small set of training data. Third, OMNI is more robust compared to LP across different career tracks. Given the nature of weak supervision in our career track labelling problem, semi-supervised learning approach (LP and OMNI) is thus more robust than the best supervised learning method, i.e., SVM.

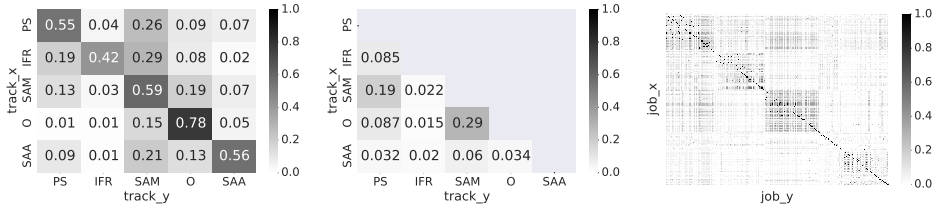
Error Analysis. To showcase the prediction quality, we chose one of the robust label propagation methods, e.g., OMNI, to perform track labelling task on networks with skill ($k10$) and domain links. Table 5 reports the top-5 job titles with automatically assigned career track labels in three major tracks with human judgements, where Hit=1 stands for correct prediction and Hit=0 otherwise. The top-5 predicted jobs for "SAA" and "SAM" are completely correct, while only 40% hit rate is achieved in "IFR" track. "system consultant" is annotated as PS track instead, while in fact it shares the common skill, "servers", in IFR track. Similarly, "system administrator" is annotated as support track (O), while it also shares the common skill, "servers", in IFR track.

4.3 Career Movements Characterization in Singapore

We adopt the same settings in Figure 3 and conduct analytics studies on a sample set of jobs observed among a population of working professional. Our goal is to characterize career movements using the predicted career track labels particularly in three aspects: (1) *movement preferences between career tracks*, (2)

Table 5. Top-5 job titles in predicted career track. (OMNI, skill (k10), domain links).

	"IFR"		"SAA"		"SAM"	
k	Job Title	Hit	Job Title	Hit	Job Title	Hit
1	senior network administrator	1	senior product developer	1	sales administrator	1
2	network administrator	1	staff product developer	1	senior sales specialist	1
3	senior system consultant	0	product developer	1	sales specialist	1
4	system consultant	0	senior product designer	1	principal sales engineer	1
5	system administrator	0	product designer	1	sales engineer	1



(a) Movement Preferences (b) Track Skill Similarity (c) Job Skill Similarity

Fig. 4. Career Movements Analysis.

track-level movement challenges, and (3) *job-level movement challenges* due to skill gap. To control the quality of derived career track labels for unlabelled jobs, we rank all jobs by label likelihoods and select the top 30% job titles in "ICT" industry, resulting in 375 job titles with predicted track labels. Altogether with 250 labelled job titles results in 625 annotated job titles.

Movement Preferences. Figure 4(a) shows the career movements between career tracks based on observed job transitions covered by the selected 625 job titles. We observe three key preferences: (1) people are more likely to make vertical career movements, (2) the most common destination track of lateral movements is SAM (Sales and Marketing) track, and (3) the most common source track of lateral movements is IFR (Infrastructure) track, i.e., 58%. Diagonal cells from upper left to right bottom indicate vertical movements. We observe significantly higher diagonal transition probabilities, which suggest that people tend to make career moves in the same track. For instance, 0.59 of job transitions from track_x=SAM to track_y=SAM. Off-diagonal cells indicate lateral movements. The higher track movement probabilities at the third column suggest that SAM track is a common destination track to move into, e.g., 0.26 from track_x=PS (0.29 from track_x=IFR) to track_y=SAM. Off-diagonal cells at the second row suggest that IFR track is a common source track of lateral movements in Singapore, e.g., 0.19 (0.29) from track_x=IFR to track_y=PS (track_y=SAM).

Movement Challenges. Figure 4(b) offers a macro view on how challenging to make a career move between career tracks by measuring skill similarities between career tracks. Specifically, we represent each track as a skill vector, where each entry represents the TF-IDF value of a skill. We compute the cosine similarity between each pair of tracks. The skill similarity is symmetric with diagonal cells equal to 1. Figure 4(c) offers a micro view on movement challenges between jobs

by measuring job-level skill similarities. Each job is represented as a skill vector. To verify Figure 4(b), we group jobs in 625×625 job skill similarity matrix by labelled career tracks. As shown, there exist clear block structures around the diagonal, where jobs in a block (track) are more similar in skill sets.

The observations on movement challenges are two folds. First, the skill set is in general fairly dissimilar at off-diagonal cells. This may explain why people tend to make vertical career movements rather than lateral ones because of skill gaps. Second, skill set required by IFR track is very unique from the rest with similarity scores ranging from 0.015 to 0.085. The two tracks with most similar skill set with IFR are PS and SAM tracks. This also coincides with our observations in Figure 4(a), where the two most common destination tracks of lateral movements from IFR track are indeed PS and SAM tracks. In summary, we conclude that skill transferability plays a key role when it comes to lateral career movements.

5 Conclusion and Future Work

We address the problem of career track labelling for jobs by exploiting limited knowledge from existing career track knowledge database. Two research goals are achieved. We propose a data-driven approach to automatically determine career tracks of jobs harvested from LinkedIn, which may reduce the human annotation efforts in maintaining the career track knowledge bases over time across geographical regions. Second, we gain insights on career movements by analyzing upon predicted career track labels. Extensive experiments are conducted to demonstrate the labelling accuracy using LinkedIn dataset and to gain insights on career movements particularly in three aspects: (1) movements between career tracks, (2) track-level movement challenges, and (3) job-level movement challenges measured in skill gap. To the best of our knowledge, this is the first attempt in understanding career movements in a job market. We plan to explore evolving natures of career movements across different job markets. We also plan to explore advanced methods (e.g., deep learning) to improve labelling accuracy on heterogeneous job networks.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. Wang-Chien Lee's work is supported in part by National Science Foundation under Grant No. IIS-1717084.

References

1. Yehuda Baruch: Transforming careers: from linear to multidirectional career paths: Organizational and individual perspectives. Career Development International, Vol. 9 Issue: 1, pp. 58–73 (2004)

2. Thomas W. H. Ng and Kelly L. Sorensen and Lillian T. Eby and Daniel C. Feldman: Determinants of job mobility: A theoretical integration and extension. Journal of Occupational and Organizational Psychology, Vol. 80 Issue: 3, pp.363–386 (2007)
3. Wang, Jian and Zhang, Yi and Posse, Christian and Bhasin, Anmol: Is It Time for a Career Switch?. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1377–1388 (2013)
4. Xu, Huang and Yu, Zhiwen and Xiong, Hui and Guo, Bin and Zhu, Hengshu: Learning Career Mobility and Human Activity Patterns for Job Change Analysis. In: Proceedings of IEEE International Conference on Data Mining, pp. 1057–1062 (2015)
5. Nie, Min and Yang, Lei and Ding, Bin and Xia, Hu and Xu, Huachun and Lian, Defu: Forecasting Career Choice for College Students Based on Campus Big Data. Asia-Pacific Web Conference, pp, 359–370 (2016)
6. Li, Liangyue and Jing, How and Tong, Hanghang and Yang, Jaewon and He, Qi and Chen, Bee-Chung: NEMO: Next Career Move Prediction with Contextual Embedding. In: Proceedings of the 26th International Conference on World Wide Web, pp. 505–513 (2017)
7. Li, Huayu and Ge, Yong and Zhu, Hengshu and Xiong, Hui and Zhao, Hongke: Prospecting the Career Development of Talents: A Survival Analysis Perspective. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 917–925 (2017)
8. Zhu, Xiaojin and Ghahramani, Zoubin and Lafferty, John: Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. In: Proceedings of the 20th International Conference on International Conference on Machine Learning, pp. 912–919. AAAI Press (2003)
9. Y Bengio, O Delalleau, N Le Roux: Label Propagation and Quadratic Criterion. Semi-supervised Learning, 10. (2006)
10. Chapelle, Olivier and Schlkopf, Bernhard and Zien, Alexander: Semi-Supervised Learning. The MIT Press (2010)
11. Sun, Yizhou and Han, Jiawei: Mining Heterogeneous Information Networks: Principles and Methodologies. Morgan & Claypool Publishers (2012)
12. Yamaguchi, Yuto and Faloutsos, Christos and Kitagawa, Hiroyuki: OMNI-prop: Seamless Node Classification on Arbitrary Label Correlation. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 3122–3128. AAAI Press (2015)
13. Yamaguchi, Yuto and Faloutsos, Christos and Kitagawa, Hiroyuki: CAMLP: Confidence-Aware Modulated Label Propagation. In: Proceedings of the SIAM International Conference on Data Mining. (2016)
14. Yamaguchi, Yuto and Hayashi, Kohei: When Does Label Propagation Fail? A View from a Network Generative Model. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3224–3230. (2017)
15. Ding, Chris and Li, Tao and Wang, Dingding: Label Propagation on K-partite Graphs. In: Proceedings of Fourth International Conference on Machine Learning and Applications, pp. 273–278 (2009)
16. Deng, Dingxiong and Bai, Fan and Tang, Yiqi and Zhou, Shuigeng and Shahabi, Cyrus and Zhu, Linhong: Label Propagation on K-partite Graphs with Heterophily. CoRR (2017)
17. Hadiji, Fabian and Mladenov, Martin and Bauckhage, Christian and Kersting, Kristian: Computer Science on the Move: Inferring Migration Regularities from the Web via Compressed Label Propagation. In: Proceedings of the 24th International Conference on Artificial Intelligence, pp. 171–177. AAAI Press (2015)
18. Duran, Alberto Garcia and Niepert, Mathias: Learning Graph Embeddings with Embedding Propagation. NIPS (2017)